PONTIFICIAE ACADEMIAE SCIENTIARVM ACTA 21

Edited by

Werner Arber
Jürgen Mittelstrass
Marcelo Sánchez Sorondo

The Scientific Legacy of the 20th Century

The Proceedings of the Plenary Session
28 October–1 November 2010

Vatican City 2011

The Scientific Legacy of the 20th Century

Pontificiae Academiae Scientiarvm Acta 21

The Proceedings
of the Plenary Session on

The Scientific Legacy of the 20th Century

28 October-1 November 2010

Edited by

Werner Arber
Jürgen Mittelstrass
Marcelo Sánchez Sorondo

EX AEDIBVS ACADEMICIS
IN CIVITATE VATICANA • MMXI

The Pontifical Academy of Sciences
Casina Pio IV, 00120 Vatican City
Tel: +39 0669883195 • Fax: +39 0669885218
Email: pas@pas.va

The opinions expressed with absolute freedom during the presentation of the papers of this meeting, although published by the Academy, represent only the points of view of the participants and not those of the Academy.

ISBN 978-88-7761-101-7

PONTIFICIA ACADEMIA SCIENTIARVM • VATICAN CITY

Certainly the Church acknowledges that "with the help of science and technology…, man has extended his mastery over almost the whole of nature", and thus "he now produces by his own enterprise benefits once looked for from heavenly powers" (*Gaudium et Spes*, 33). At the same time, Christianity does not posit an inevitable conflict between supernatural faith and scientific progress. The very starting-point of Biblical revelation is the affirmation that God created human beings, endowed them with reason, and set them over all the creatures of the earth. In this way, man has become the steward of creation and God's "helper". If we think, for example, of how modern science, by predicting natural phenomena, has contributed to the protection of the environment, the progress of developing nations, the fight against epidemics, and an increase in life expectancy, it becomes clear that there is no conflict between God's providence and human enterprise. Indeed, we could say that the work of predicting, controlling and governing nature, which science today renders more practicable than in the past, is itself a part of the Creator's plan.

Science, however, while giving generously, gives only what it is meant to give. Man cannot place in science and technology so radical and unconditional a trust as to believe that scientific and technological progress can explain everything and completely fulfil all his existential and spiritual needs. Science cannot replace philosophy and revelation by giving an exhaustive answer to man's most radical questions: questions about the meaning of living and dying, about ultimate values, and about the nature of progress itself.

Address of His Holiness Benedict XVI to the Members of the Pontifical Academy of Sciences, 6 November 2006.

Contents

Prologue

The 20th century was an important century for the sciences. In physics, revolutionary theoretical and experimental breakthroughs happened, modern biology began its triumphal march. And what is true for physics and biology is also true for other neighbouring disciplines like astrophysics, chemistry, earth and environmental sciences and the neurosciences. At the same time, more attention was paid to epistemological issues, which were discussed and solved in cooperation with the sciences and philosophy of science. This is reason enough to evaluate the entire development and to picture how this development finds its demanding continuation in modern scientific research.

All PAS members have contributed to the remarkable progress of scientific knowledge in the past century. Personal testimonies from these actors represent valuable documents for the history of science and for future generations. We thus encourage all PAS members to attend the Plenum and to comment on the progress and on still open questions in their special fields of scientific competence. Contributions can represent a general overview of research strategies, of the acquired new scientific knowledge, of its applications and of the expected future impact on the world view and welfare of human societies and on the sustainability of our natural environment. Of course, inclusion of personal recollections, including anecdotes, can render these reviews more lively.

Werner Arber & Jürgen Mittelstrass

Word of Welcome

Dear Participants, good morning and welcome to our 2010 Plenary Session of the Pontifical Academy of Sciences. In the absence of Professor Nicola Cabibbo, our longstanding President, I would like to thank you for coming and for your participation in this very important meeting that touches upon one of the central topics of science. As you well know, we can say that the 20th century marked the greatest progress in science in the history of mankind.

Before this meeting gets under way, I would like to pass the floor to Professor Werner Arber, who has very graciously accepted to say a few words to commemorate our late President. Moreover, I would like to suggest that Professor Arber take on the role of President for this meeting, because of his active role in inspiring and organizing this conference, his scientific prestige and his active membership as a Councillor in the Academy for many years.
Professor Arber, the floor is yours.

Marcelo Sánchez Sorondo

Programme

▶ Thursday, 28 October 2010

Welcome, Commemorations and Self-Presentations

9:00 *Welcome*
H.E. Msgr. Marcelo Sánchez Sorondo, Chancellor of the Pontifical Academy of Sciences

9:15 *Commemorations*
Nicola Cabibbo (W. Arber), Aage Bohr (A. Zichichi), Héctor Croxatto (R. Vicuña), Paul Germain (N. Le Douarin), Crodowaldo Pavan (W. Arber)

10:15 *Self-Presentation of New Members*
Edward M. De Robertis, Gerhard Ertl

11:15 Coffee Break

11:30 Papal Audience

13:30 Lunch at the Casina Pio IV

Introduction

15:00 Prof. Jürgen Mittelstrass
Philosophical Foundations of Science in the 20th Century

Session I: Astrophysics

Chair: J. Mittelstrass

15:30 Prof. Wolf J. Singer
The Explanatory Gap in Neuroscience

16:00 Discussion

16:30 Coffee Break

17:00 Prof. Govind Swarup
Great Discoveries Made by Radio Astronomers During the Last Six Decades and Key Questions Today

17:30 Discussion

18:00 Departure from the Casina Pio IV by bus to attend the concert at Palazzo Boncompagni Ludovisi

18:30 Concert followed by dinner

21:30 Bus leaves Palazzo Boncompagni Ludovisi to take participants back to the Domus Sanctae Marthae

▶ Friday, 29 October 2010

8:30 Visit to the Vatican Museums (departure by bus from Domus Sanctae Marthae)

Chair: H. Tuppy

11:30 Yves Quéré
Science and Language, Twin Sisters

12:00 Discussion

12:15 Prof. Rudolf Muradian
Scaling Laws in Particle Physics and Astrophysics

12:40 Discussion

13:00 Lunch at the Casina Pio IV

Session II: Physics

14:30 Prof. Antonino Zichichi
My Personal Experience on the Scientific Legacy of the 20th Century

15:00 Discussion

15:30 Prof. Walter E. Thirring
Emergence of Order by Chance

16:00 Discussion

16:30 Coffee Break

17:00 Prof. Charles H. Townes
The Laser and How it Happened

17:30 Discussion

18:00 Closed Session for Academicians

19:30 Dinner at the Casina Pio IV

▶ Saturday, 30 October 2010

9:00 *Commemorations*
Stanley L. Jaki (J.-M. Maldamé), Marcos Moshinsky (W.E. Thirring), Marshall W. Nirenberg (M. Sela), George E. Palade (C. De Duve), Robert J. White (A. Szczeklik)

Session III: Earth and Environment Sciences

Chair: A.M. Battro

10:00 Prof. Ignacio Rodríguez-Iturbe
Hydrologic Drivers of Biodiversity: The Impact of Climate Change

10:30 Discussion

11:00 Coffee Break

11:30 Prof. Jean-Michel Maldamé
La place de l'humanité dans le monde des vivants: nouvelles perspectives

11:50 Discussion

12:10 Prof. Enrico Berti
The Discovery of DNA as a Contribution to Understand the Aristotelian Theory of Generation

12:30 Discussion

12:50 Lunch at the Casina Pio IV

Session IV: Cell and Molecular Biology

Chair: N.M. Le Douarin

15:00 Prof. Christian de Duve
The Evolutionary Lottery

15:30 Discussion

16:00 Prof. Michael Sela
Therapeutic Vaccines Against Cancer and Autoimmune Diseases

16:30 Discussion

17:00 Coffee Break

17:30 Prof. Rafael Vicuña
The Evolving Concept of the Gene

18:00 Discussion

18:30 Prof. Werner Arber
Molecular Darwinism and its Relevance for Translational Genetic Research

19:00 Discussion

19:30 Dinner at the Casina Pio IV

▶ Sunday, 31 October 2010

9:00 Departure from Domus Sanctae Marthae to visit the Papal Villa at Castel Gandolfo

10:00 Visit to the Papal Villa

11:00 Presentation of the Pius XI Medal (Prof. Patrick Mehlen)

13:00 Lunch at the Papal Villa

15:00 Departure from Castel Gandolfo and return to the Domus Sanctae Marthae

17:30 Holy Mass at the Casina Pio IV

19:00 Dinner at the Casina Pio IV

▶ Monday, 1 November 2010

Chair: R. Vicuña

9:00 Prof. Edward M. De Robertis
Merging of Evolution and Development

9:30 Discussion

10:00 Prof. Nicole M. Le Douarin
New Developments in Stem Cell Biotechnology

10:30 Discussion

11:00 Coffee Break

11:30 Prof. Takashi Gojobori
Genomic Exploration of the RNA Continent

11:50 Discussion

12:10 Prof. Peter H. Raven
Transgenic Crops and the Future of Agriculture

12:30 Discussion

12:50 Lunch at the Casina Pio IV

15:00 Prof. Ingo Potrykus
Genetic Engineering of Plants

15:30 Discussion

Session V: Neuroscience and Immunology

Chair: M. Sela

16:00 Prof. Andrzej Szczeklik
Discovery of the Defensive System of Endothelium, the Lining of Arterial Wall

16:20 Discussion

16:40 Prof. Aaron J. Ciechanover
Discovery of the System That Destroys Body Proteins

17:00 Discussion

17:20 Coffee Break

End of Conference

17:50 Prof. Werner Arber *Conclusions*

18:30 Dinner at the Casina Pio IV

List of Participants

Prof. Werner Arber
Biozentrum, Department of Microbiology
University of Basel
Basel (Switzerland)

Prof. Antonio M. Battro
Battro and Denham
Buenos Aires (Argentina)

Prof. Enrico Berti
Università degli Studi di Padova
Dipartimento di Filosofia
Padova (Italy)

Prof. Aaron J. Ciechanover
Technion, Israel Institute of Technology
The Rappaport Faculty of Medicine
and Research
Institute, Vascular and Cancer Biology
Research Center
Haifa (Israel)

Prof. Stanislas Dehaene
Inserm-CEA, Cognitive
Neuroimaging Unit
CEA/SAC/DSV/DRM/NeuroSpin
Gif-sur-Yvette (France)

Prof. Edward M. De Robertis
University of California, Los Angeles
Howard Hughes Medical Institute
MacDonald Research Laboratories
Los Angeles, CA (USA)

Prof. Christian de Duve
Christian de Duve Institute
of Cellular Pathology
Brussels (Belgium)
The Rockefeller University
New York, NY (USA)

Prof. Gerhard L. Ertl
Technical University of Hannover
Ludwig Maximilians University
of Munich
(Germany)

Prof. José G. Funes, S.J.
Director of the Vatican Observatory
(Vatican City)

Prof. Takashi Gojobori
Centre for Information Biology
and DNA Bank of Japan
National Institute of Genetics
Mishima (Japan)

Prof. Nicole M. Le Douarin
Collège de France, C.N.R.S.
Institut d'Embriologie Cellulaire
et Moléculaire
Nogent-sur-Marne (France)

Prof. Jean-Michel Maldamé
Institut Catholique de Toulouse
Toulouse (France)

Prof. Jürgen Mittelstrass
University of Constance
Center for Philosophy of Science
Constance (Germany)

Prof. Rudolf Muradian
Universidade Federal da Bahia
Instituto de Física
Salvador Bahia (Brazil)

Prof. Ingo Potrykus
Emeritus Professor
Institute of Plant Sciences, ETH Zürich
Zürich (Switzerland)

Prof. Yves Quéré
Professor and Co-Chair
Académie des Sciences
Paris (France)

Prof. Peter H. Raven
Missouri Botanical Garden
St. Louis, MO (USA)

Prof. Ignacio Rodríguez-Iturbe
Princeton University
Dept. of Civil
and Environmental Engineering
Princeton, NJ (USA)

H.E. Msgr. Marcelo Sánchez Sorondo
Chancellor
The Pontifical Academy of Sciences
(Vatican City)

Prof. Michael Sela
The Weizmann Institute of Science
Department of Immunology
Rehovot (Israel)

Prof. Govind Swarup
National Center for Radio Astrophysics
Tata Insitute of Fundamental Research,
Pune University Campus
Pune (India)

Prof. Andrzej Szczeklik
Jagiellonian University School of Medicine
Department of Medicine
Kraków (Poland)

Prof. Charles H. Townes
University of California, Berkerley
Department of Physics
Berkeley (USA)

Prof. Walter E. Thirring
University of Vienna
Institut for Theoretical Physics
Wien (Austria)

Prof. Hans Tuppy
University of Vienna
Institute of Biochemistry
Vienna (Austria)

Prof. Rafael Vicuña
Pontificia Universidad Católica de Chile
Facultad de Ciencias Biológicas
Santiago (Chile)

Prof. Antonino Zichichi
Università degli Studi di Bologna
Dipartimento di Fisica
Bologna (Italy)

Address to the Holy Father

28 October 2010

Holy Father,

Your Pontifical Academy of Sciences comes before you on the opening day of its Plenary Session, which has as its subject 'The Scientific Legacy of the 20th Century' and on the final day of our satellite meeting on 'Neuroscience and Education'. Last August this Academy suffered the loss of its longstanding President, Professor Nicola Cabibbo, whom we have commemorated with gratitude this morning. We shall keep him – and all of our Academicians who have passed away – in our prayers, especially in the Holy Mass on Sunday.

Holy Father, we are deeply thankful for Your constant solicitude towards our Academy, which is also evident in the appointment of our four new Academicians in these last two years, Francis S. Collins from the USA, Edward M. De Robertis from the USA, Gerard L. Ertl from Germany and Miguel A.L. Nicolelis from USA, whom we are honoured to introduce to you today. In addition, this year Your Holiness will also award the prestigious Pius XI Medal to the young French biologist Patrick Mehlen.

In the course of the development of science in the prodigious 20th century, which our Plenary intends to examine, nothing came to contradict the definition of truth as *adaequatio intellectus et rei*,[1] to which *Fides et ratio* referred. Obviously the truth in this case specially refers to natural realities. Scientists are conscious that this truth of nature is a participation of the Truth, and they do not want science to replace the other manifestations of the truth that derive from philosophy, theology or Revelation. They are also aware that the good of science cannot be disjointed from the good of the human person and from justice in an ever more globalised world. We know about Your concern and project for a new evangelization. Max Planck, one of the major protagonists of the twentieth century that we are going to study, the father of quantum physics and an illustrious member of this Pontifical Academy, believed that incredulity and superstition could be fought by the union of natural science and religion, with his exhortation to 'Get closer to God' (*Hin zu Gott!*).[2] Therefore, Your Holiness, we look forward

[1] Cfr. *Fides et ratio* § 82.

[2] 'Es ist der stetig fortgesetzte, nie erlahmende Kampf gegen Skeptizismus und Dogmatismus, gegen Unglaube und Aberglaube, den Religion und Naturwissenschaft

to listening to Your wise words that will certainly enlighten and orient the course of our Plenary.

We are mindful of the many duties that your high office places upon you, not least the recent and demanding Synod of Bishops for the Middle East and your next trip to Santiago and Barcelona to inaugurate the Church of the Holy Family, and are especially grateful to you for granting us this audience today.

It only remains for me to ask Your Holiness to bless this Academy and all those who will generously share their wisdom with us in the next few days.

H.E. Msgr. Marcelo Sánchez Sorondo

gemeinsam führen. Und das richtungweisende Losungswort in diesem Kampf lautet von jeher und in alle Zukunft: Hin zu Gott!' (*Religion und Naturwissenschaft,* in *Vorträge und Erinnerungen*, Stuttgart 1949, p. 333).

Address of His Holiness Benedict XVI to Participants in the Plenary Session of the Pontifical Academy of Sciences

Clementine Hall • Thursday, 28 October 2010

Your Excellencies,
Distinguished Ladies and Gentlemen,

I am pleased to greet all of you here present as the Pontifical Academy of Sciences gathers for its Plenary Session to reflect on 'The Scientific Legacy of the Twentieth Century'. I greet in particular Bishop Marcelo Sánchez Sorondo, Chancellor of the Academy. I also take this opportunity to recall with affection and gratitude Professor Nicola Cabibbo, your late president. With all of you, I prayerfully commend his noble soul to God the Father of mercies.

The history of science in the twentieth century is one of undoubted achievement and major advances. Unfortunately, the popular image of twentieth-century science is sometimes characterized otherwise, in two extreme ways. On the one hand, science is posited by some as a panacea, proven by its notable achievements in the last century. Its innumerable advances were in fact so encompassing and so rapid that they seemed to confirm the point of view that science might answer all the questions of man's existence, and even of his highest aspirations. On the other hand, there are those who fear science and who distance themselves from it, because of sobering developments such as the construction and terrifying use of nuclear weapons.

Science, of course, is not defined by either of these extremes. Its task was and remains a patient yet passionate search for the truth about the cosmos, about nature and about the constitution of the human being. In this search, there have been many successes and failures, triumphs and setbacks. The developments of science have been both uplifting, as when the complexity of nature and its phenomena were discovered, exceeding our expectations, and humbling, as when some of the theories we thought might have explained those phenomena once and for all proved only partial. Nonetheless, even provisional results constitute a real contribution to unveiling the correspondence between the intellect and natural realities, on which later generations may build further.

The progress made in scientific knowledge in the twentieth century, in all its various disciplines, has led to a greatly improved awareness of the place that man and this planet occupy in the universe. In all sciences, the

common denominator continues to be the notion of experimentation as an organized method for observing nature. In the last century, man certainly made more progress – if not always in his knowledge of himself and of God, then certainly in his knowledge of the macro- and microcosms – than in the entire previous history of humanity. Our meeting here today, dear friends, is a proof of the Church's esteem for ongoing scientific research and of her gratitude for scientific endeavour, which she both encourages and benefits from. In our own day, scientists themselves appreciate more and more the need to be open to philosophy if they are to discover the logical and epistemological foundation for their methodology and their conclusions. For her part, the Church is convinced that scientific activity ultimately benefits from the recognition of man's spiritual dimension and his quest for ultimate answers that allow for the acknowledgement of a world existing independently from us, which we do not fully understand and which we can only comprehend in so far as we grasp its inherent logic. Scientists do not create the world; they learn about it and attempt to imitate it, following the laws and intelligibility that nature manifests to us. The scientist's experience as a human being is therefore that of perceiving a constant, a law, a *logos* that he has not created but that he has instead observed: in fact, it leads us to admit the existence of an all-powerful Reason, which is other than that of man, and which sustains the world. This is the meeting point between the natural sciences and religion. As a result, science becomes a place of dialogue, a meeting between man and nature and, potentially, even between man and his Creator.

As we look to the twenty-first century, I would like to propose two thoughts for further reflection. First, as increasing accomplishments of the sciences deepen our wonder of the complexity of nature, the need for an interdisciplinary approach tied with philosophical reflection leading to a synthesis is more and more perceived. Secondly, scientific achievement in this new century should always be informed by the imperatives of fraternity and peace, helping to solve the great problems of humanity, and directing everyone's efforts towards the true good of man and the integral development of the peoples of the world. The positive outcome of twenty-first century science will surely depend in large measure on the scientist's ability to search for truth and apply discoveries in a way that goes hand in hand with the search for what is just and good.

With these sentiments, I invite you to direct your gaze toward Christ, the uncreated Wisdom, and to recognize in His face, the *Logos* of the Creator of all things. Renewing my good wishes for your work, I willingly impart my Apostolic Blessing.

Commemorations of Deceased Academicians

Nicola Cabibbo († 16.VIII.2010)

Nicola Cabibbo, born in Rome on 10 April 1935, was President of the Pontifical Academy of Sciences for 17 years till his death on 16 August 2010. He was initially appointed as a Pontifical Academician on 9 June 1986 by Pope John Paul II, who also appointed him on 30 March 1993 as President of the Academy. He was, undoubtedly, one of the most important theoretical physicists of our time, and yet he was one of the most humble persons – a perfect gentleman, as well as a highly esteemed President of the Academy.

Nicola Cabibbo was one of the fundamental pioneers in the development of high-energy physics. His contributions were in the field of weak interactions; and he did a large part of this work right here in Rome. It was here that the great Enrico Fermi taught physics at the University of Rome, La Sapienza, which became one of the leading Centres of theoretical physics.

The work of Italian physicists in the development of high-energy physics has been excellent, and a part of it was awarded the Nobel Prize. There was, of course, the great work of Enrico Fermi on the theory of Beta Decay. One of Fermi's colleagues in Rome was Emilio Segrè, who also won a Nobel Prize for the discovery of the anti-proton. Amongst Fermi's collaborators was the young Ettore Majorana, who was highly gifted in mathematics and whose work on neutrino masses, the famous Majorana Equation and the Majorana Neutrino, is so well known.

Several other famous Italian physicists, including Nicola Cabibbo, contributed importantly to this well-established Italian school of physics. The Cabibbo angle appears to be a simple parameter, but it was fundamental to the development of the theory of high-energy physics, particularly of what is called today 'The Standard Model'. We can also recall the experimental work, 60 years ago, which was characterized by what was then referred to as the 'theta-tau' puzzle, raising the question of how the same strange particle could, on one occasion, go into two pions or, otherwise, into three pions. On this basis, Richard Feynman and Murray Gell-Mann proposed in 1958 their famous current x current form of weak interaction theory, which extended Fermi's theory and made it applicable to all forms of weak decays.

In a late recognition of Nicola Cabibbo's important contributions to this field of research, the International Council of Theoretical Physics (ICTP) awarded the 2010 Dirac Medal to Nicola Cabibbo and E.C. George Sudar-

shan on 9 August 2010, just one week before Cabibbo passed away. The underlying theory was a beautiful generalization of the Fermi theory and it also included the axial vector current, in addition to the original vector current of Fermi, in the form of the V-A interaction of Sudarshan and Marshak, thus incorporating the maximal parity violation discovered by T.D. Lee, C.N. Yang and C.S. Wu in 1956.

An important feature of the Fermi/Gell-Mann theory was the idea that weak interactions are universal. Cabibbo succeeded in reformulating the hypothesis of universality in such a way that the discrepancy noted in the comparison of the decay rates of the strange particles with those of the non-strange particles could be explained beautifully. This was done in his seminal paper of 1963.

By 1963, SU(3) symmetry of Gell-Mann and Ne'eman which was at that time called Unitary Symmetry, was already becoming a part of high energy physics. Following Gell-Mann, Cabibbo took the strangeness-conserving and strangeness-violating weak hadronic currents as members of the same octet representation under SU(3). This allowed him a more precise formulation of the weak interaction. He proposed instead his form of universality that came to be known as Cabibbo Universality.

All this was done within the framework of SU(3) and hence the agreement with experimental results helped to establish the usefulness of SU(3) not only in the classification of hadrons, but also in correlating their weak decays.

Cabibbo's paper refers to the Gell-Mann-Levy paper, but we must give substantial credit to Cabibbo as the author of Cabibbo Universality; the Gell-Mann-Levy remark remained as a footnote for almost three years and it was Cabibbo who took it seriously and developed it into a full-fledged theory of the leptonic decays of hadrons with predictions that were verified experimentally. Cabibbo himself in his original paper was rather modest; he wrote, '*I will restrict himself to a weaker form of universality*'. On this scientific basis other scientists have successfully further developed this field of research.

It is Nicola Cabibbo's seminal work that laid the foundation for our modern understanding of the weak interactions among the quarks. Theorists who boldly followed his idea of universality of weak interactions to its logical completion not only successfully predicted the existence of the charmed quark, but also the top and bottom quarks which were necessary for CP violation. Finally, the universality that was formulated by Cabibbo got enshrined in the Standard Model of High-Energy Physics, in the form of the equality of gauge coupling to all the particles. It became a cornerstone of the Standard Model. It can thus be seen that Nicola Cabibbo is one of the several Italian distinguished scientists who laid the building blocks of the particle physics.

Nicola Cabibbo's interests and concerns were not limited to theoretical physics but they extended to all fields of scientific investigations and to the application of scientific knowledge. This is obvious to all members of the Pontifical Academy of Sciences present in this commemorative session and who very highly appreciate and recognize his guidance as our President for the last 17 years.

M.G.K. Menon and Werner Arber

Aage Bohr († 8.IX.2009)

Aage Bohr was born in Copenhagen a few months before his father won the Nobel Prize. His father was Niels Bohr, one of the giants of physics in the early 20th century, who was able to untangle the confusing mysteries of quantum mechanics. Aage Bohr's childhood was one in which a pantheon of great physicists were friends visiting the family home. The remarkable generation of scientists who came to join his father in his work became uncles for him. These uncles were Henrik Kramers from the Netherlands, Oskar Klein from Sweden, Yoshio Nishina from Japan, Werner Karl Heisenberg from Germany and Wolfgang Pauli from Austria. These are all giants of physics, so Aage Bohr is an example of what science means and what political violence means. In fact, three years after he was born, Hitler ordered the deportation of Danish Jews to concentration camps but the Bohrs, along with most of the other Danish Jews, were able to escape to Sweden.

I would like to recall that, despite these great tragedies and political problems, Aage Bohr was able to follow his father's track, contributing to clarify a very important problem in nuclear physics. He was able, together with Ben Roy Mottelson, to explain why the nuclei were not perfectly spherical. I remind you that the volume of the nucleus is 1 millionth of a billion, 10^{-15} smaller than the atom, and when Bohr was young the general feeling about nuclear physics was that the nuclei were perfectly symmetric, perfect spheres, platonically perfect spheres, and here comes the contribution of Aage Bohr with Mottelson, because some experiments were showing that this was probably not true. How can it happen that such a small volume, I repeat, one millionth of a billion times smaller than the atom, cannot be perfectly symmetrical and spherical? Aage Bohr and Mottelson explained that the rotational motion of protons and neutrons inside this extremely small sphere could distort the shape of the nucleus. This had very important developments in nuclear fusion, which generates ten million times more energy than the standard transformation of mass into energy via electromagnetic forces.

What Aage Bohr was able to do was, in fact, the last and most important step in the field of nuclear physics. Before, nuclear physics was shown not to be a fundamental force of nature but a result of the fundamental force of nature, which now we call quantum chromodynamics. It is remarkable that this field of physics, which was unknown in the 1930s, gave rise to an incredible series of unexpected events, the last one being, as I said, understood and explained by Aage Bohr. The first incredible event was the discovery that the particle proposed by Yukawa to be the nuclear glue was not a nuclear glue, it was a particle that had nothing to do with the nuclear forces and which we now call the muon, which is a lepton, not a meson. The decay of the particle considered to be the nuclear glue, namely the pion, was shown to violate two fundamental invariance laws, parity and charge conjugation, and when the first example of real nuclear glue was discovered, namely the pion – this was in 1947 – everybody believed this was the last step in nuclear physics.

However, it was later shown that this so-called elementary particle would result of two quarks, a quark and an antiquark, glued by the fundamental force, which is quantum chromodynamics. In this impressive series of unexpected events, it was the privilege of Aage Bohr to demonstrate his reasons why the last one should be there. So the scientific credit of Aage Bohr will remain in the history of physics as the last step in understanding nuclear physics before a new era started, the one which is now called the physics of quantum chromodynamics, from which nuclear physics derives, but no one knows how to make this apparently elementary passage, namely from quantum chromodynamics to nuclear physics and here comes the latest interest of Aage Bohr. He was very interested to know how the passage takes place, and this reminds me of the first transition from the vacuum to inert matter, a topic which was of great interest to Aage Bohr, namely, does the transition take the simplest elementary case to be or does this transition involve less elementary passages? The simplest way out is not followed by nature in the transition from vacuum to inert matter.

Bohr's interests were not limited to these fundamental problems of physics. He was a member of a scientific committee called Science for Peace, which played an important role during the half century of east-west confrontation. He was also interested in scientific culture, the way in which our understanding of physics comes in, in the sense of involving a great number of people. We should not forget that we live in a world where, as he used to repeat, our culture is as if science had never been discovered. He was very active in promoting the value of Galileo Galilei, who is the father of first-level science. In giving our gratitude to his remarkable work in physics and modern culture

we should not forget that we live in a world were it is our responsibility to let people know the great values of science. It is the most civilized of all possible battles, because it does not produce tragedies but improves our culture. In this battle Aage Bohr was a great leader.

Antonino Zichichi

Héctor Croxatto († 28.IX.2010)

This morning we commemorate one of the most notable men that Latin American science has produced. Loved by all those who had the privilege to know him and held in esteem by both his colleagues and friends, Héctor Croxatto has left in his legacy an inerasable imprint that is becoming of a man of such exceptional qualities. From the beginning, his scientific vocation was strongly influenced by Dr. Eduardo Cruz Coke, who was also a distinguished member of this Pontifical Academy of Sciences from 1948 up until his death in 1974.

Hector Croxatto was incorporated as a Professor of Physiology at the Pontifical Catholic University of Chile in 1934, the institution in which he remained for the duration of his life. These were times when science in Chile was in its infancy. Hundreds of young medics, biologists and biology professors had the honor to pass through his laboratory, being marked by the wisdom and simplicity of their teacher. The scientific production of Prof. Croxatto, notable as much by its abundance as by its high quality, courted numerous and merited recognitions. These gave justice to a life of effort and dedication, and served as inspiration to a generation that could now look upon science in Chile as a genuine option for their future.

The multifaceted personality of Prof. Croxatto drove his participation in diverse initiatives that would have a positive impact on both the scientific community and society as a whole. He was cofounder of the Latin American Academy of Sciences and Third World Academy of Sciences (TWAS). Dr Croxatto was also Director of the Center of Improvement of the Ministry of Education in Chile and founder of the Chilean Society of Hypertension.

However, beyond a life full of achievements, there was without doubt another quality to admire about Prof Croxatto, namely, his profound human values. He permanently demonstrated a special preoccupation for those who worked in his laboratory, ensuring that their passage through his tutorship was an unforgettable period of scientific training and also of personal enrichment. His love for his pupils was always expressed through warmth and friendliness. Art, history, philosophy and science, were touched upon

daily by a man who dazzled us all with his inner greatness, humble attitude and passion for knowledge.

His enthusiasm with scientific discoveries was contagious. On numerous occasions we witnessed the astonishment that these caused him, as they allowed him to teach us the harmony of the forces operating in nature. Without a doubt, his amazement was fed by the special privilege that he possessed to see in nature the hand of God. Prof. Croxatto always gave testimony of a deep religious faith. His particular sensitivity in front of the wonders of creation delivered a transcendent perspective to his task as an investigator. His inner vigor was also fed by his love of art. A dedicated painter during his days of rest, he elaborated a deep comparative analysis of scientific endeavor and artistic creation, both in relation to the way they are generated and in the way they are appreciated by the observer.

God granted to Prof. Croxatto a long life and gave him talents that he cultivated to become a great scientist, humanist and teacher. His life inspires us.

Rafael Vicuña

Paul Germain († 26.II.2009)

Notre confrère et ami Paul Germain, nous a quittés le 26 février 2009 dans sa quatre-vingt neuvième année. Mathématicien de formation, il a consacré son oeuvre scientifique à la Mécanique, science du mouvement. Chercheur très doué, pédagogue hors pair, personnalité exceptionnellement dynamique, Paul Germain a fait bénéficier toutes les entreprises auxquelles il a participé de sa vision lumineuse de la science, qu'il considérait comme une composante essentielle de la culture au service de l'Homme.

Paul Germain était né à Saint-Malo en Bretagne, le 28 août 1920, au lendemain de la grande guerre. Son père, professeur de chimie, avait participé aux batailles. Victime des gaz asphyxiants, il mourut prématurément alors que le jeune Paul, aîné de trois enfants, n'avait que neuf ans. Cette disparition du père ne manqua pas de développer chez Paul Germain le sens des responsabilités et de l'engagement qui le caractérisa toute sa vie.

Il se révéla très tôt être un élève puis un étudiant brillant, ce qui lui valut d'être admis à l'Ecole Normale Supérieure, une des plus sélectives et prestigieuses des Grandes Ecoles françaises. Son but initial était de devenir mathématicien mais, après avoir passé l'agrégation qui le préparait à l'enseignement, sa rencontre avec Joseph Pérès en 1944 changea le cours de son destin.

Joseph Pérès, récemment nommé Professeur à la Sorbonne, venait de créer le laboratoire de Mécanique des fluides dans lequel il avait besoin d'un habile

mathématicien. Paul Germain sera celui-là. Il se passionnait déjà pour les développements de l'aéronautique, dont la guerre qui s'achevait avait montré toute l'importance. Il réussit à mettre au point un modèle mathématique et une méthode numérique applicables aux problèmes posés par la mécanique des fluides en aéronautique bien supérieurs à ceux qui existaient jusque là.

En 1945, alors qu'il séjournait au National Physics Laboratory en Angleterre, sa méthode reçut un excellent accueil, ce qui conforta sa vocation naissante. De retour à Paris, il décida de poursuivre ses recherches dans cette direction. A cette époque, le gouvernement français créait un institut spécialisé dans ce domaine: L'Office National d'Etudes et de Recherches Aéronautiques (ONERA), où Paul Germain fut recruté et où il prépara sa thèse. Celle-ci, qui portait sur «l'Aérodynamique Supersonique», aboutit à la résolution d'équations concernant la mécanique des fluides et fut soutenue en 1948.

Ces travaux permirent à son groupe de calculer et d'optimiser de nombreuses configurations d'ailes, de fuselages et d'empennages, dès le début des années cinquante. Ils seront une base essentielle pour le développement des avions supersoniques qui était alors à l'étude. Parallèlement à son travail de recherche, il commença alors une carrière d'enseignant à l'Université de Poitiers. Il la poursuivit à la Faculté des sciences de Lille (1954-1958) puis à l'Ecole Polytechnique et enfin à l'Université Pierre et Marie Curie à Paris.

Très reconnu dans son domaine d'expertise, Paul Germain effectua, au cours de sa vie, plusieurs séjours aux Etats Unis, à l'Université de Brown (de 1953 à 1954) puis au Californian Institute of Technology à Pasadena (1957) et à Stanford et Berkeley (1969-1970).

Au cours de son année à Brown University, il fit l'expérience du fonctionnement d'un département où l'enseignement et la recherche tout comme les mathématiques appliquées et l'Ingénierie étaient intimement liés. Il en retira une vision synthétique de la Mécanique rompant avec la traditionnelle distinction entre Fluides et Solides désormais regroupés en Milieux Continus. Il n'aura de cesse par la suite de développer dans son enseignement, dans ses recherches et même dans ses actions de structuration de la Mécanique, cette vision unitaire des concepts fondamentaux, allant bien au delà des outils techniques.

En 1962, son groupe de recherche avait suffisamment évolué pour devenir le Laboratoire de Mécanique Théorique dont il rêvait. On lui demanda alors d'être le directeur de l'Office d'Etudes et de Recherches Aéronautiques (ONERA) où il avait, seize ans plus tôt, fait ses débuts dans la recherche. A cette époque, le programme de l'avion supersonique de transport de passagers en coopération avec les Britanniques battait son plein. Enfin, en 1969, le Concorde vole. On peut dire que Paul Germain aura grandement contribué

à ce succès par ses travaux personnels sur les écoulements transsoniques et supersoniques et par son action à la tête de l'ONERA.

De retour à la vie scientifique, Paul Germain s'investit à nouveau dans l'enseignement de sa discipline et de ses concepts fondamentaux qu'il renouvela profondément. Ses travaux sur les puissances virtuelles puis sur la Thermodynamique des Milieux Continus, dont la motivation initiale était de présenter à ses étudiants une vision moderne de la discipline, constituent en fait une œuvre de recherche de haut niveau. Le livre sur la *Mécanique des Milieux Continus* qu'il a publié en 1969 est un classique utilisé à l'époque par tous les étudiants de cette discipline. Il fut suivi par un second ouvrage en 1973 qui connut le même succès.

En 1975, toujours aussi passionné par l'enseignement, Paul Germain acceptait un poste de Professeur de Mécanique à l'Ecole Polytechnique. Il demanda à y enseigner la totalité du programme et délivra un cours, dont sera tiré un livre en deux tomes intitulé *Mécanique,* donnant une vision unifiée, et unique, de tous les aspects de la discipline.

Il avait une très noble idée de son métier d'enseignant et une très haute considération pour son public. Il s'estimait responsable de la formation de la génération d'étudiants qui lui était confiée, ne mettait jamais en cause la capacité de compréhension de son auditoire et s'interrogeait constamment sur la clarté de son cours et sur la pertinence de ses choix.

Il tirait de l'enseignement un très grand plaisir. Je cite ce qu'il a écrit à ce sujet dans le livre testament qu'il nous a laissé, intitulé *Mémoires d'un scientifique chrétien*: *"Eveiller les étudiants à une discipline scientifique, leur en faire découvrir l'intérêt et la beauté, aider déjeunes chercheurs à participer à la grande œuvre de la recherche et être témoin de leur émerveillement dans la fraîcheur des commencements, cela fut et reste pour moi l'une des joies les plus sûres qui m'ait été donnée".*

A sa vie de chercheur et d'enseignant s'ajoute celle d'Académicien. Elu Membre de l'Académie des sciences en 1970, il est choisi par ses confrères en 1975 pour occuper la fonction prestigieuse de Secrétaire perpétuel, qu'il occupera pendant vingt ans. Il fut un des acteurs influents d'une réforme des statuts de l'Académie des sciences qui eut lieu dans les années qui ont suivi. Le besoin de renouveau se faisait sentir dans une institution dont les statuts très anciens remontaient à 1816. La réforme qu'il a menée et qui s'imposait se caractérise notamment par l'augmentation du nombre des membres ainsi que leur rajeunissement, qui l'un et l'autre reflètent le dynamisme de la recherche scientifique au cours de la seconde moitié du vingtième siècle.

Il a de plus développé la mission de conseil de l'Académie à l'adresse du gouvernement français en préparant avec ses confrères de nombreux rapports et notes sur des sujets scientifiques divers.

Paul Germain a été élu Membre de l'Académie Pontificale des Sciences en 1986. Il prenait très à cœur son rôle dans cette Académie, aussi bien en tant que scientifique qu'en tant que chrétien. Il a été pendant de nombreuses années membre de son Conseil.

Il a su concilier une vie professionnelle, intellectuelle et sociale d'une exceptionnelle richesse avec une vie familiale qu'il plaçait très haut dans ses priorités. Ses qualités remarquables ont été saluées par de nombreux honneurs et récompenses. Il était membre de plusieurs autres Académies y compris l'Accademia dei Lincei et titulaire d'un Doctorat Honoris Causa de plusieurs Universités à travers le Monde. Il était Grand Croix dans l'Ordre National du Mérite et Commandeur dans l'Ordre de la Légion d'Honneur.

C'est un homme d'une stature exceptionnelle que nous avons perdu.

Fait à Paris le 20 octobre 2010.

Nicole Le Douarin

Crodowaldo Pavan († 3.IV.2009)

Crodowaldo Pavan was born on the 1st. of December 1919 in Campinas, Brazil. He became member of our Academy in 1978. For the last 31 years of his life he was a very active participant in the activities of the Pontifical Academy of Sciences.

Crodowaldo Pavan had Italian roots. In fact, his grandfather had emigrated from Italy to Brazil. The young Crodowaldo Pavan studied natural sciences and he obtained a PhD in Zoology in 1944. Already a year before he finished his PhD studies in Brazil, he had started to work as a zoologist on the genetics of *Drosophila* flies in collaboration with one of the famous *Drosophila* geneticists, Theodosius Dobzhansky, who was working in the USA. These studies concerned animal evolutionary genetics implying the observation of chromosomes in the optical microscope, which revealed specific banding patterns. This was facilitated by giant chromosomes which had been seen with some, but by far not all kinds of flies.

In the first decades of the 20th century, cytogenetic studies had suggested that the genetic information is carried in the chromosomes. This stimulated more intensive studies. In the search for appropriate animals for such studies, Crodowaldo Pavan isolated on the coastline near Sâo Paulo different kinds of flies. One of these isolates revealed in the optical microscope to have particularly nice giant chromosomes which facilitated Pavan's research.

May I just remind you that in 1944, precisely in the year in which Pavan submitted his PhD thesis, Avery and his collaborators at the Rockefeller Uni-

versity in New York were able to convincingly show that the material basis of genetic information is the nucleic acid DNA, a component of chromosomes.

Giant chromosomes can be found in some flies and in a number of other animals, but they are not very widespread. Giant chromosomes display a remarkable banding pattern along the chromosomal filamentous structures. The intensity of these bands depends on the particular tissue under observation. Particularly intensive bands are called puffings (or shortly "puffs"). One had assumed that these tissue-specific variations of band intensities depend on tissue-specific gene activities. It was Crodowaldo Pavan who found in his working with fly chromosomes good experimental evidence for this assumption. He could show (to his own surprise) that puffs contained a high number of DNA copies which cause the high intensity of the bands. Interestingly, puffs cannot only be seen in chromosomes of adult flies but also on larval developmental tissues. This observation facilitated studies of the embryonic development. Upon the publication of his findings Pavan encountered a lot of disbelief in the scientific community. One could not understand that parts of a chromosome become highly amplified when this particular gene activity is needed. But in the long term it turned out that Crodowaldo Pavan was right: puffs contain many DNA copies of the heavily expressed genes and this in a tissue-specific manner.

As a young scientist, Crodowaldo Pavan spent several stays in the USA, the first one for his collaboration with Dobzhansky. In the 1960's he spent some time at the Oak Ridge National Laboratory, where he studied the effects of radiation as well as the effects of virus infections on the chromosomal morphology, i.e. the banding pattern and gene expression. Around 1970 he worked for some time in Austin, Texas. From there he returned to Brazil. While his early scientific investigations had been done in Sâo Paulo, his later work was carried out in Campinas. Here he used fruit flies to control agricultural pests, applying his scientific knowledge to the benefit of humanity. This shows his wide interests for both basic and applied scientific research at the frontline of genetics.

In his home country Crodowaldo Pavan also collaborated with Johanna Döbereiner, a late member of our Academy. I remember many interesting discussions with them on their experimental results on nitrogen fixation by microorganisms residing in various tissues of agricultural crops such as sugar cane. As a matter of fact, these researchers found out that nitrogen fixation by various kinds of bacteria is much more widespread in plants than one had assumed before. In general, different nitrogen fixing bacteria can be found as symbionts throughout the plant tissues, at least as long as no (or only small) doses of nitrogen fertilizer are applied. This insight is of high relevance for a sustainable agriculture.

Crodowaldo Pavan exerted also activities in science politics in his home country. For some years he was the President of the National Science Council and also of the Brazilian Society for the Advancement of Science. He actively propagated an improvement of the public understanding of science. With his death, our Pontifical Academy of Sciences has lost an active member with an impact on basic scientific knowledge and on the beneficial application of this knowledge, particularly in favor of developing countries and with respect to sustainability of agricultural practices.

Werner Arber

Stanley L. Jaki († 7.IV.2009)

Professor Stanley Jaki was a Catholic priest of the Benedictine order. He was born in Hungary in 1924 and his country's history affected him deeply. He confided in me how traumatized he was by the communists coming to power backed by the Soviet Army. Consequently, his monastic order was a victim of the oppression. After finishing his studies in Rome, he wasn't allowed to return home and emigrated to the USA. That experience strongly influenced his historian work. The passion of his words and work can be divided into three points.

Firstly, Fr Stanley Jaki's work target was to clarify the relations between the sciences of nature and the Catholic Church. He did it on the epistemological level. He promoted Gödel's theorem on philosophical interpretation concerning the incomplete formal systems in order to thwart the rationalist philosophy which set science as an absolute knowledge. He did it in physics and cosmology. His books, *The Relevence of Physics* (1967), and *God and the Cosmologists* (1980) brought him to receive the *Lecomte de Nouy* Prize (1970) and *The Templeton Prize* (1987). And in addition, he was invited to serve as a Gifford Lecturer at the University of Edinburgh.

The second significant feature of Professor Stanley Jaki's works was historical. In his writings he presented his wide perspectives about science since its Greek origins. As a consequence, he was objective enough to put the options of the modern science into perspective. His works were inspired by Pierre Duhem, a famous French historian of sciences he venerated so much. Unfortunately, Pierre Duhem wasn't well-known including in France, so Stanley Jaki's research compensated such a great deficiency.

The third feature of Professor Stanley Jaki's works was theological in a particular way. Actually, he intended to use science to reveal the spiritual dimension of intellectual research. He did it as a Gifford Lecturer, as well as in his last

works titled *The Savior of Science* and *Means to Message: A Treatise on Faith* (1998). He spread his apologetic will both against materialists and scientists engaged in basic research as we can see in the Study of his interpretation of the first Chapter of the Book of Genesis, *Genesis through the Ages* (1997).

This multidisciplinary approach combined history and epistemology. He passionately argued his spiritual position of the scientific research and the immanent rationality to cosmos, which made Professor Stanley Jaki somebody who imparted between different worlds and made him somebody who awoke those who refused to confine themselves in a unique specialization.

Professor Stanley Jaki's generosity and critical mind were inseparable from his spiritual behaviour which became evident at the end of his life, based on his great admiration for John Henry Newman, showing his ultimate aim to be Peace.

Jean-Michel Maldamé

Marcus Moshinsky († 1.IV.2009)

The impressive biographical data of Marcus Moshinsky is very well documented in the Yearbook that we have in front of us, so I think there's no sense in reading it to you. I think I can do justice to his genius better by telling you my personal recollections on four encounters with his work, which show the breadth of his intellectual horizon.

The first was very long ago in 1951. I was in Gottingen and there appeared a paper in *Physics Review* which derived the essential consequences of general relativity just by solving the Schrödinger Equation in an external gravitational potential, and we were very perplexed. It was not talking about the non-Euclidian geometry, and it came from Princeton, where Einstein was still alive and dominating. The paper sort of said that, actually, we don't need Einstein, we can derive all this very simply. We tried to find a mistake, as there were many crazy theories around, but we couldn't find anything so we decided it was correct. Then we didn't know what to do, since he had apparently found an effect which is present, as is the one of general relativity, so perhaps you should add them since both are correct. But if you do that then you get a factor 2 and destroy all the agreement with the experiment, so that's not what you want to do. We saw that this had some philosophical implications, namely you can either look at this situation by saying that space is flat, there's nothing like a curved space, but there's the gravitational potential. However, it changes measuring rods and clocks – in this case it was the hydrogen atom which was pretty good as a measuring

rod and as a clock – thus the gravitational potential changes the size of the measuring rod in such a way that you get the illusion that space is curved. On the other hand, you can also look at it the other way and say that measuring the hydrogen atom has never changed, it's always the unit of length by definition, so we have to take it as the definition of the length. It is just that the space is curved and therefore it can be different at different places. So you can say, as it is now said in general relativity, that the gravitational field is an illusion and, in reality, that space is curved. You see that this already has a philosophical flavour. It shows that what is reality may in some cases be undecidable, because there are two points of view looking at it. Since they are mathematically equivalent, you cannot say one is wrong and the other is correct, and yet they say something different about what is reality. So this was one of his works and you will see that the other works that I encountered always touched upon some general principles.

The next encounter was when I was working with Goldberg and Gellman on the question, to what extent the requirement that nothing propagates faster than light gives some restraints on the scattering amplitude. We found that, strangely enough, it's the mathematically analytic properties of the scattering amplitude that can say whether something propagates faster than light or not. And then we were informed that these analytic properties could all nicely be found in the work of Wigner and Moshinsky, who at this time was working with Eugene Wigner on nuclear physics. So, again, his work had some influence on general principles.

Well, he did many things but the next time we had a complete different approach. It came from the quark model of elementary particles. In the proton you have three quarks, so you have to handle the three-body problem and somehow you have to classify the states of this three-body problem. But it's not just any three-body problem, but they are confined in a potential which doesn't let them go out. The simplest model for a confining potential would be harmonic potential and so it is worthwhile to look at that and, in fact, there you can solve the three-body problem and it was just a sort of heavy task of group theoretic investigations of the properties of the states with respect to the exchange of particles. Then it turned out that this was something that Marcus Moshinsky had done some time ago, and it had implications on what came out of elementary particle physics. In fact, the classification of the states then lead eventually to the notion of colour, that there must be another quantum number hidden in the quark systems.

The fourth system is again something different, it has to do with the Heisenberg commutation relations. They were introduced by Heisenberg but at that time people were mathematically not prepared to make any sense

of it. Heisenberg knew nothing about what it would be, but warned and said, well, maybe they are matrices because for matrices the commutative laws of multiplications did not hold. Then people found it couldn't be matrices because, if you take a trace of a commutator it is always zero, so you have a contradiction. Then people said, perhaps they are just abstract operators. However, this didn't help because the operators can be shown as not to be bounded. Unbounded operators lead to a lot of mathematical complications. The best tool in this game is the coherent states that come from these commutation relations. And this is, again, one of the fields where Marcus Moshinsky was very active.

In summary, I would say he was a master in all theoretical physics of the 20th century and therefore he was just the man that a country like Mexico needed. Although he was not Mexican by origin, he was born in the Ukraine, I think he became one of the dominant figures of South American theoretical physics altogether and I think it's well deserved that he became almost a cult figure in Mexico.

Walter E. Thirring

Marshall Nirenberg († 15.I.2010)

Marshall Nirenberg was born in New York on 10 April 1927 and died this year, on 15 January 2010. As a child he was diagnosed with rheumatic fever and, because of this, the family moved to Florida. In due course he studied for his undergraduate degree at the University of Florida in Gainesville and did his Ph.D. in biochemistry in 1957, at the University of Michigan in Ann Harbor. He joined the National Institutes of Health shortly after and devoted his research to the relationship between DNA, RNA and proteins, and more specifically, to deciphering the genetic code.

I would now like to tell you an anecdote of this period, in which I was directly involved. I was at the NIH at that time, as a professor on sabbatical, and one day Marshall came to my lab and asked me whether by chance I had some poly-L-phenyl-alanine. I answered him that I did not have any, but when he asked me whether I knew in which solvent poly-L-phenylalanine was soluble, I looked up a paper which I had published some years earlier in JACS, on mechanism of polymerization, and found that poly-L-phenylalanine was not soluble, even in such strong solvents as dimethylformamide and dimethylsulfoxide, but was soluble in a saturated solution of hydrogen bromide (HBr) in glacial acetic acid. Such a reagent is used to remove blocking groups such as carbobenzoxy from amino functions, and is

used a lot in peptide synthesis. At that moment, for reasons of my research, I was bubbling in my hood HBr through glacial acetic acid. I gave Marshall some of this reagent, and thanks to it, he found that UUU (uridine-uridine-uridine) dictates the formation of Phe, and thus broke the genetic code. I am actually the only person he thanks in the PNAS paper, for which he received the Nobel Prize in Physiology and Medicine in 1968, which he shared with Robert Holley and Gobind Khorana.

Now, who in his right mind would use this reagent as a solvent? Some years earlier, I had two test tubes in a jar in my office. One contained poly-L-phenylalanine and the other poly-carbobenzoxylysine. My colleague picked what he believed to be blocked polylysine, and after a few minutes he returned and said: 'I cannot understand – it dissolved but did not give the characteristic carbon dioxide bubbles'. I answered: 'Oh gosh, I made a mistake and gave you the wrong test tube', but I also made a note that poly-L-phenylalanine dissolves in HBr in glacial acetic acid. If that mistake had not occurred, I could not have helped Marshall in his discovery.

A few months later, we were all together at the International Biochemical Congress in Moscow in August 1961. Marshall gave a short talk at one of the many workshops which Francis Crick attended. With the force of his personality, he demanded that Marshall should repeat the talk in the big Hall before thousands of listeners and the rest is history.

In 1965 he received the National Medal for Science. In 1974 he was appointed to the Pontifical Academy of Sciences. He was a member of many other learned societies, including the American Philosophical Society, American Academy of Arts and Sciences, National Academy of Sciences USA, and National Institute of Medicine. From 1966 he was Chief of the Laboratory of Biochemical Genetics of the National Heart Institute of NIH, Bethesda. He continued the extension of deciphering the genetic code of all amino acids. At a later stage he entered the field of neurobiology and established many clonal lines of mouse neuroblastoma cells. A neuroblastoma glioma somatic hybrid cell line was generated that expresses abundant opiate receptors, which was used as a model system to explore the mechanism of opiate dependence. These cells were also used as model systems to study many properties of neurons.

Marshall Nirenberg was modest, friendly and a lovable character. He died just a few months ago. Blessed be his memory.

Michael Sela

George Emil Palade († 7.X.2008)

George Emil Palade was born on November 19, 1912, in Jassy, Romania. He studied medicine at the University of Bucharest, graduating in 1940. Already as a student, he became interested in microscopic anatomy and its relation to function and decided early to relinquish clinical medicine for research. After serving in the Romanian army during the Second World War, he moved to the United States in 1946, soon joining the laboratory of Albert Claude at the Rockefeller Institute for Medical Research, where, after Claude's return to Belgium in 1949, he developed an independent laboratory, first in association with Keith Porter and later, after Porter's departure in 1961, on his own. He stayed at what had become the Rockefeller University until 1973, when he moved to Yale University. His later years were spent at the University of California, San Diego, where he acted as Dean of Scientific Affairs. He passed away on 7 October 2008, after suffering major health problems, including macular degeneration leading to total blindness, a particularly painful ordeal for a man who had used his eyes all his life in a particularly creative way. He leaves two children from his first marriage with Irina Malaxa: Georgia Palade Van Duzen and Philip Palade. He married Marilyn G. Farquhar, a cell biologist, in 1971, after the death of his first wife.

Palade's scientific work followed in the wake of Albert Claude's pioneering achievements, using the two new major technical approaches developed by his mentor for the coordinated investigation of cellular structure and function: electron microscopy and cell fractionation. With the help of these tools, to which he provided a number of important improvements, he accomplished some of the major advances made by cell biology after the last war.

From the structural point of view, he described the fine structure of mitochondria, including the *cristae*, to which he gave their name; the dense granules, first called Palade granules and now known as ribosomes, that line the membranes of what his colleague Porter had named the endoplasmic reticulum; as well as detailed features of the Golgi complex, of endothelial cells and many other structures.

In the functional domain, in collaboration with the late Philip Siekevitz and with an international team of first-class coworkers, he unravelled the fundamental pathway whereby secretory proteins are synthesized by membrane-bound ribosomes and simultaneously delivered into the cisternae of the rough endoplasmic reticulum, further processed and channelled, by way of smooth parts of this structure, toward the Golgi complex, where they are packaged into secretion granules, to be finally discharged outside the cells by exocytosis.

Elected to the Pontifical Academy of Sciences on 2 December 1975, George Palade was also a member, among others, of the National Academy

of Sciences, USA, and of the Royal Society. His achievements have been recognized by several important awards, including the Lasker Award (1966), a Gairdner Special Award (1967), the Louisa Gross Horwitz Prize (1970), and the Nobel Prize in Physiology or Medicine (1974).

Christian de Duve

Robert J. White († 16.IX.2010)

Robert Joseph White was born in 1926 in Duluth, Minnesota. There he went to school, received his Bachelor of Science at the University of Minnesota and a medical degree from the Harvard University School of Medicine. He did his surgical residency at Peter Brigham Hospital and a neurosurgical fellowship at the Mayo Clinic, which he subsequently joined as a member of the staff to become eventually the chairman of the Department of Neurosurgery. He was a brilliant surgeon, interested in experimental research. His group was the first to accomplish the total isolation of the brain in the experimental animal. They succeeded in maintaining its viability through the use of extracorporeal systems. He was also the first to successfully transplant and hypothermically store the mammalian brain, with survival for extended periods of time. When I asked him four years ago, here in this place, after his lecture at our Academy, 'Why don't you present a movie in colour of your transplanted brains?', he answered, 'Very few in the audience would stand the view of blood in such an amount, covering not only the operating table but most of the operating room'. It was a demanding skill to do such research, plus, of course, he was an excellent clinical surgeon.

But it is especially the study of cerebral physiology and biochemistry at the very low temperatures that have been among his most important contributions. Just to approach it a little bit, for those of you who don't deal with patients every day, here in Rome and in my country as well, almost every day in intensive care we get at least one patient who underwent resuscitation following sudden cardiac arrest. In some of them, the heartbeat is restored and they breathe spontaneously but their consciousness does not return. They move to the state of existence unknown to man in all his history. They live what we call *vita vegetativa*. The crucial point is the time of starting resuscitation. If it exceeds three to four minutes, the cerebral damage is done. Hypothermia could extend this vital period up to several minutes. The results of Dr White brought about an understanding of why the brain is protected during periods of circulation reduction or arrest under hypothermic conditions. His studies led to the introduction of a number of

new techniques in operating neurosurgery, including the utilization of low temperature states for the treatment of acute spinal cord trauma and protection of the brain during intracranial surgery. Over the last three to four years we have started to use, in intensive care, cool blankets right after resuscitation, and this is also an impact of Dr White's studies.

Many honours and awards were conferred upon Robert White for his outstanding surgical skills and experimental work. He regularly attended Academy meetings, accompanied by his wife. He was a warm, friendly, open person. He died a few weeks ago. The memory of his valuable contribution to medicine and to the growth of our Academy will always be cherished. We shall remember him in our thoughts as our beloved colleague and friend.

Andrzej Szczeklik

Self-presentation of the New Academicians

Edward M. De Robertis

I was born in Boston, Massachusetts, of Argentinian parents. My mother was a poet and my father a neuroscientist doing postdoctoral training at MIT. At that time they were exiled by dictator General Perón, and therefore in 1950 our family moved to Montevideo, Uruguay. Montevideo in the 1950s was a wonderful place to grow up in.

I attended a grammar and high school run by American Methodist missionaries, which provided a good moral education. There, in kindergarten, I met Ana Marazzi, who at age 15 became my sweetheart and later mother to our three beautiful children. My parents divorced when I was five, but my poet mother provided a wonderful home, and made sure I became a confirmed Catholic.

Medical school in Montevideo offered excellent training in the French tradition. I graduated at age 24. We married the day after my final exam, enjoyed a very brief honeymoon, moved to Buenos Aires, and on the third day began Ph.D. studies in Chemistry at the Institute Leloir in Argentina.

Upon completion of my Ph.D., I was accepted into the lab of the eminent embryologist Sir John Gurdon and shortly thereafter we arrived in Cambridge, England. Gurdon was a wonderful mentor, who taught by example. My debt to him is immense.

After three years as a postdoctoral fellow, and the three more as an independent Scientist in Cambridge, I received a call from the Biozentrum of the University of Basel, Switzerland and became Professor of Cell Biology at age 33. The Director of the Biozentrum at that time was Prof. Werner Arber, who is here today. Thank you, Werner.

In Switzerland we had joint group meetings with the great geneticist Walter Gehring. These were very exciting times, for Gehring's group had discovered a gene sequence conserved in several fruit fly genes that regulated anterior-posterior cell differentiations. We collaborated to determine whether similar sequences might be cloned from vertebrate gene libraries. This resulted in the isolation of the first development-controlling gene from a vertebrate in 1984. The study of these genes, now called Hox genes, opened the door for understanding the genetic control of mammalian development.

Twenty-five years ago, I was offered an endowed Chair of Biological Chemistry at the University of California at Los Angeles. There, we carried

out a systematic dissection of the molecules that mediate embryonic induction in frog embryos. We isolated several genes responsible for the induction of cell differentiation. Most were inhibitors of growth factor signaling and one of them, a protein named Chordin, provided the key to the regulation of dorsal-ventral tissue differentiations, not just in vertebrates, but in all bilateral animals.

Thus, our work contributed to the remarkable current realization that embryonic cell differentiations are controlled by regulatory gene networks common to all animals. These discoveries initiated the young discipline of Evolution and Development, called Evo-Devo for short.

I would like to end on a personal note. I join the discussions of this Academy both as a scientist and as a practicing Catholic. I was therefore deeply touched to receive last year's Christmas card from Bishop Sánchez Sorondo. It started: "*In principio erat Verbum*". Above this, the same passage was written in Greek and one could clearly read that *Verbum* translates as *Logos*. As Pope Benedict XVI reminds us, *Logos* in Greek also means Reason. The next line read: "*Et Verbum caro factum est*". This is of course from St. John's gospel, which is read at the end of every traditional Holy Mass. To believe that the "the Word was made flesh", or *Logos*, or Reason, is not easily achieved. Faith needs nurturing surroundings. I was fortunate to have them during my life.

The Pontifical Academy serves to build bridges between Faith and Science – *Pontifex* means the bridge-builder. Biology, which is my field, has been used as an excuse to create false oppositions between Faith and Reason. I therefore welcome this opportunity to help in your task of building bridges between Science and Faith.

Thank you.

Gerhard Ertl

I was born in 1936 in Stuttgart, Germany, the son of a miller. I became interested in the natural sciences already as a boy. It was mainly chemistry, but also physics. It was not clear to me which would be my preferred subject until we got teachers in these subjects. We had a good teacher in physics and a poor teacher in chemistry, so I became a physicist. I got my first degree at the Technical University of Stuttgart but was still interested in chemistry, so I moved for my Ph.D. thesis into the field of physical chemistry, which means the investigation of problems of chemistry with the techniques of physics.

Together with my mentor, Heinz Gerischer, I moved to the Technical University of Munich where I got my Ph.D. in 1965 on a topic that became my lifelong interest, namely reactions at solid surfaces. Twenty-five years ago I moved to Berlin to become one of the directors of the Fritz Haber Institute of the Max Planck Society, succeeding my teacher Heinz Gerischer.

Chemical reactions – that means transformations of molecules into new ones – usually involve the collision of molecules to form new ones. But not every collision is successful, only a small probability exists and this probability that determines the rate of a chemical reaction is determined by an activation energy at its start. This has to do with the fact that chemical transformations always involve breaking of bonds and forming of new bonds, and the energy barrier we have to overcome in this context is the activation barrier. The higher the barrier, the lower the probability. If we offer the chemical reaction an alternative path, we can have a higher rate, a higher probability, and this is done by a catalyst. A catalyst forms intermediate compounds with the molecules involved in the reaction. This catalyst can be in the same phase, these are homogeneous catalysts, in biological systems these are macromolecules or enzymes and in industry practical applications these are mainly solid surfaces, and this is heterogeneous catalysis.

The principle of heterogeneous catalysis comprises the interaction of molecules from the gas phase with a surface of a solid which exposes its topmost atoms with unsaturated bonds, so new bonds can be formed, so-called chemisorption bonds, which can also modify existing bonds, i.e. a molecule may dissociate. These chemisorbed species may diffuse across a surface and form new molecules which are released into the gas phase. The overall sequence of all these steps offers a reaction mechanism with a higher reaction probability. This is the principle of catalysis.

Heterogeneous catalysis is the basis of the chemical industry. About 85% of all products in the chemical industry are made through catalysis. But also solution of problems concerning the environment, energy, or climate change will require the application of catalysis.

One example of heterogeneous catalysis is the car exhaust catalyst where toxic molecules like carbon monoxide or nitric oxides or hydrocarbons are transformed into less harmful ones. In this case, carbon monoxide is oxidised to carbon dioxide and this is through interaction of the molecules from the gas phase with the catalyst's surface. The carbon monoxide is bonded to the surface through the carbon atom, while the oxygen molecule is dissociatively adsorbed, where the oxygen-oxygen bond is breaking and then the adsorbed species diffuse across the surface and form the new CO_2 molecule.

Where do we know this information? This is obviously chemistry restricted to two dimensions, so we need new techniques also to look at the processes involved in these catalytic reactions on the surfaces of the small catalyst particles involved in such a reaction. Small particles because the reactivity, of course, depends on the overall magnitude of the surface area and the more finely divided the particles, the higher the surface area. So the size of these small particles is of the order of one nanometer. Catalysis has been nanotechnology long before this term was invented. And this also shows the problem: to investigate the chemistry of these particles we need tools which are able to analyse the topmost atomic layer on a very, very microscopic level. One way to do this is by just looking at separate crystal planes of these catalytically active species. This can be done on the atomic scale, for example, with the scanning tunnelling microscope.

If a platinum surface interacts with molecules from the gas phase, bonds may be formed, chemisorption bonds, and we have seen one of the essential points is the breaking of bonds, dissociation of a molecule. If we expose such a surface to a diatomic molecule, there will be interaction between the atoms of the molecule with the surface and the bond between the atoms will be weakened and will be eventually dissociated. With our platinum catalyst, in the car exhaust catalyst, at a temperature of about -100°C the oxygen molecule is dissociated, and the atoms formed are separated by about 5 to 8 angstroms because they have to release their energy to the solid and this takes some time, around 300 femtoseconds. At these low temperatures the oxygen atoms stay where they have been formed. If we increase the temperature they can jump from one site to the next site and their residence time becomes shorter. These oxygen atoms move across the surface, they randomly jump across the surface: whenever they come close to each other their lifetime becomes a little bit longer because the oxygen atoms interact with each other, they attract each other weakly. As a consequence, with increasing surface concentration there is no longer a perfect random distribution of the atoms on the surface but they form two new phases: a more condensed phase, like a two-dimensional crystal, and a gaseous phase, like

in solid/gas equilibria. As a consequence of the long-range order which is formed by these adsorbed pieces on the surfaces we can determine the structure of these adsorbed phases by a diffraction technique, electron diffraction in this case. On such a platinum surface, the oxygen atoms form an open mesh, while the CO molecules tend to form a densely packed overlayer, so that the CO molecules inhibit the adsorption of oxygen. On the other hand, if the surface is saturated by oxygen, CO may be adsorbed as a unit cell. And this gives us information about the mechanism of reaction. CO is adsorbed, forms a densely-packed layer, oxygen dissociatively adsorbs into a relatively open layer. If the CO coverage is too high the oxygen cannot be adsorbed, which means CO inhibits the reaction and that's why your car exhaust catalyst doesn't work in the cold, you have to go up with temperature a little bit so that part of the CO molecule can desorb.

On the other hand, CO can adsorb inside the oxygen overlayer and the close neighbourhood between CO and oxygen then enables formation of the CO_2 molecule which is then released into the gas phase. If we start with CO and oxygen in the gas phase, we end up with CO_2. Thereby we gain a substantial amount of energy and most of this energy is already liberated in the first step as heat when these two molecules are chemisorbed on the surface. If we have CO and oxygen adsorbed on the surface they may recombine by overcoming a small activation barrier and CO_2 is formed, which is released into the gas phase. This is the simplest catalytic reaction we can think of. This is probably the *drosophila* of catalysis and many many studies have been made in this way. There are many other much more complicated reactions. For example, one of the most important industrial processes, nitrogen fixation and formation of ammonia from nitrogen and hydrogen in the Haber-Bosch process is much more complicated, but has also been resolved in this way. As I mentioned at the beginning, heterogeneous catalysis will be one of the most demanding fields also for the future solution of the problems of mankind. Thank you very much.

The Pius XI Medal Award

Patrick Mehlen

Brief Account of Scientific Activity

The Dependence Receptor Notion: Apoptosis, from Cell Biology to Targeted Therapy

Since 1998, P. Mehlen's work has been devoted to the development of the dependence receptor notion. P. Mehlen, while working in Dale Bredesen's laboratory in San Diego, proposed that some transmembrane receptors may be active not only in the presence of their ligand as usually believed, but also in their absence. In this latter case, the signaling downstream of these unbound receptors leads to apoptosis. These receptors were consequently named "dependence receptors", as their expression renders the cell's survival dependent on the presence in its environment of its respective ligand (Mehlen *et al.*, 1998, *Nature*). To date, more than 15 dependence receptors have been identified and this functional family includes RET (rearranged during transfection), TrkC, ALK, EPHA4, the netrin-1 receptors DCC (Deleted in Colorectal Cancer) and UNC5H1-4 (Unc-5 homologue 1-4), neogenin, some integrins, and the Sonic Hedgehog receptor Patched (Ptc). P. Mehlen then proposed that the pro-apoptotic activity of these dependence receptors is crucial for the development of the nervous system as a mechanism to "authorize" guidance/migration/localization in settings of ligand presence (Thibert *et al.*, 2003, *Science*; Matsunaga *et al.*, 2004, *Nature Cell Biology*; Tang *et al.*, 2008, *Nature Cell Biology*; Mille *et al.*, 2009, *Nature Cell Biology*). Interestingly, P. Mehlen's group also demonstrated that these dependence receptors represent an important mechanism which limits tumor progression (Mazelin *et al.*, 2004, *Nature*; Mehlen and Puisieux, 2006, *Nature Review Cancer*). The current view is that tumor cells expressing such dependence receptor should undergo apoptosis as soon as primary tumor growth reaches ligand limitation or as soon as tumor cells metastasize in tissues with low ligand content. The demonstration that these dependence receptors were novel types of tumor suppressors was an important discovery in terms of academic research (Prof. Bert Vogelstein interviewed on Mehlen's work by a reporter from the San Francisco Chronicle commented «The results indicate a fascinating and novel mechanism for (tumor) growth control processes»). However, even more interestingly, recent studies conducted by Mehlen's laboratory propose that this notion of dependence receptor may also lead to appealing anti-cancer strategies. Indeed, Mehlen's

group has proposed that in a wide fraction of cancer, the selective advantage that tumors have selected to bypass this dependence for survival on ligand presence is an autocrine secretion of the ligand. Thus, Mehlen's group has shown that in these tumors (e.g., for netrin-1, 47% of lung cancer, 66% of metastatic breast cancer, 40% of neuroblastoma, ect…), disruption of the interaction between the auto-secreted ligand and its dependence receptor reactivates cell death in vitro and is associated with tumor regression in vivo (Fitamant *et al.* 2008, *PNAS*; Delloye-Bourgeois *et al.*, 2009; *JNCI*, Delloye *et al.*, 2009, *JEM*; Bouzas *et al.*, 2010, *JCI*, 6 filed Patents since 2006). This has led to the creation of a spin-off company Netris Pharma in June 2008, dedicated to develop candidate drugs which act as interferents to the ligand/dependence receptors interaction. Regarding the first ligand auto-secreted (netrin-1), a candidate drug has been selected and is in pre-clinical development. Interference to two other autosecreted ligands (NT-3 and SHH) is in early development at Netris Pharma. Thus, if the hypothesis is correct, this discovery may lead to clinical phase I study scheduled for 2012. Thus, from a basic cell biology concept, P. Mehlen and his laboratory may, within the next few years, provide new tools to fight against cancer with a wide societal impact.

Philosophical Foundations of Science in the 20th Century

Jürgen Mittelstrass

The 20th century was an important century in the history of the sciences. It deserves to be called a *scientific century*. It generated entirely novel insights in foundational issues and established a previously unknown intimate connection between science and technology. Whereas physicists at the end of the 19th century had thought of themselves as having reached the end of basic research and had believed the principles of physics to have been discovered in their entirety, in the first third of the 20th century we witness revolutionary changes, comparable to the scientific revolution of the 17th century.

With the development of the Special and the General Theory of Relativity as well as quantum theory, the central theoretical frameworks of modern, non-classical physics were introduced. Theoretical investigations into the statistical interpretation of thermodynamics and infrared radiation lead to the development of quantum mechanics, which in turn prompted modifications of the atomic model and allowed an explanation of the photoelectric effect. The development of the Special Theory of Relativity as a theory of the spatio-temporal relationships between inertial systems moving relative to each other, which yields an explanation of the properties of transformations of the Maxwell-Hertz equations, and of the General Theory of Relativity as theory of the classical (non-quantised) gravitational field, leads to entirely new conceptions of space, time and gravity. Essential steps in the development of quantum mechanics are the development of quantum statistics and of the uncertainty principle, which sets limits on the measurement of atomic processes. In contrast to classical physics, natural laws preclude determinate measurements of the system's state. At the same time, essential clarifications and specifications are made to fundamental concepts of epistemology (or natural philosophy) such as the concepts of space and time in the Theory of Relativity, of causality and locality in quantum theory, of matter and field in the physics of elementary particles.

Besides physics, the discipline of biology, especially molecular biology and biophysics, which, together with biochemistry, conceives of itself as a molecular research programme, as well as evolutionary theory, become a leading science. Within biology, due to the discovery of the chemical structure of the DNA and the deciphering of the genetic code, the 20th century has been

called the century of the gene.[1] Developments in other parts of the natural sciences, such as astrophysics, chemistry, in the earth and environmental sciences as well as in the neurosciences are of comparable significance. In addition, there is an ever-closer connection between science and technology. Scientific research has reached a point where idealisations may be overcome and the controlled laboratory may be left behind. Rather, science is now in the position to do justice to the complexity of the real world.

These developments are accompanied by *epistemological* reflections. On the one hand, these are directly connected to the scientific developments and, as in the case of the concepts of space and time, are part of scientific theory construction; on the other hand, general philosophy of science experiences an increase in importance and influence within that part of philosophy which is close to science. Science does not just yield important discoveries, it also becomes reflexive – in the sense of making its own procedures, theoretical, methodic and empirical, the subject of critical scrutiny. This is especially true concerning the foundations of science.

In what follows, I present a few brief remarks on the topic of philosophical foundations. I want to address three different epistemological approaches: one that is scientific in the narrow sense, emerging out of scientific theorising itself, one that is both scientific and philosophical (mediating, in a sense, between science and philosophy), and one that is of a general philosophical nature (general in the sense of *general philosophy of science*). They are all representative of the connection between science and epistemology, and they all illustrate the high standard of scientific thought in the 20th century. To conclude, a few remarks on developments relating to new forms of organising research and a revised concept of research follow.

1. An approach that is scientific in the narrow sense is connected to epistemological problems which are primarily of scientific importance. Questions raised by quantum mechanics belong to this area. In the so-called Copenhagen Interpretation, a correspondence principle bridges the gulf between classic and quantum-theoretic explanations of the structure of matter. At the same time, the differences between quantum mechanics and classical physics lead to different epistemological interpretations, for instance an *instrumentalist* reading, according to which quantum mechanics is not about the physical reality as such, but about a world as perceived by the epistemological view of the physicist, or a *realist* interpretation, for instance

[1] E.F. Keller, *The Century of the Gene*, Cambridge Mass. and London 2000.

that advocated by Albert Einstein, according to which the physical objects exist independently of each other and the context of measurement.

An instrumentalist approach also implies the view that there are principled epistemological limits to knowledge or human cognition, whereas a realist approach implies the (problematic) view of the incompleteness of quantum mechanics, which might be overcome by assuming hidden parameters. Other examples might be the issue of the conventional nature of simultaneity within Special Relativity and the debate in the foundations of mathematics, in which formalist, Platonist and constructivist conceptions were competing as the bases of mathematics.

2. Connected to epistemological problems of this kind, resulting directly from scientific research, are ones of scientific as well as of philosophical significance. Among these are, for instance, the topics of determinism, emergence, and (again) realism. Everything we know about the world, in science and philosophy, seems to depend on the question whether we live in a *deterministic world*. A well-known example for this is chance in quantum mechanics.[2] Quantum mechanics imposes serious limitations on the predictability of events. The central principle of the theory is 'Schroedinger's equation', which serves to determine the 'state function' or 'wave function' of a quantum system. The state function is generally taken to provide a complete description of quantum systems; no properties can be attributed to such a system beyond the ones expressed in terms of the state function. Schroedinger's equation determines the time development of the state function unambiguously. In this sense, quantum mechanics is a *deterministic* theory.

However, apparently irreducible chance elements enter when it comes to predicting the values of observable quantities. The measurement process in quantum mechanics is described as the coupling of the quantum system to a particular measuring apparatus. Schroedinger's equation yields, then, a range of possible measuring values of the quantity in question, each of these values being labelled with a probability estimate. That is, Schroedinger's equation only provides a probability distribution and does not anticipate particular observable events. Heisenberg's so-called indeterminacy relations

[2] On this and the following point on 'emergence', compare the more extensive treatment in J. Mittelstrass, 'Predictability, Determinism, and Emergence: Epistemological Remarks', in: W. Arber *et al.* (eds.), *Predictability in Science: Accuracy and Limitations (The Proceedings of the Plenary Session 3-6 November 2006)*, Vatican City (The Pontifical Academy of Sciences) 2008 (*Pontificia Academia Scientiarum Acta* 19), pp. 162-172.

are a consequence of Schroedinger's equation, although historically they were formulated independently of this equation and prior to its enunciation. The Heisenberg relations place severe limitations on the simultaneous measurement of what are called 'incompatible' or 'incommensurable' quantities such as position or momentum or spin values in different directions. The more precisely one of the quantities is evaluated, the more room is left for the other one. Like the constraints mentioned before, the limitations set by the Heisenberg relations have nothing to do with practical impediments to increasing measurement accuracy that might be overcome by improved techniques. Rather, the relations express limitations set by the laws of nature themselves. This element of genuine, irreducible chance troubled Albert Einstein very much. It challenges the thesis of a deterministic world.

Concerning the concept of *emergence*, what is at issue is the relationship of properties of wholes to properties of its component parts, equally relevant in science and philosophy. Originally, it made reference to the conceptual contrast, in a biological context, between 'mechanicism' (as a particular variant of materialism) and 'vitalism.' Systematically, it says that it is insufficient to use characteristics of elements and their interrelations to describe characteristics of ensembles or make predictions about them[3] (the whole is more than its parts[4]). According to the *emergence thesis*, the world is a levelled structure of hierarchically organised systems, where the characteristics of higher-level systems are by and large fixed by the characteristics of their respective subsystems, yet at the same time essentially different. Different characteristics and processes occur in the respective levels. Furthermore, weak and strong emergence theses can be distinguished.

The core element of the strong emergence thesis is the non-derivability or non-explainability hypothesis of the system characteristics shaped from the characteristics of the system components. An emergent characteristic is non-derivable; its occurrence is in this sense unexpected and unpredictable. Weak emergence is limited to the difference of the characteristics of systems and system components and is compatible with the theoretical explainability of the system characteristics. Weak emergence is essentially a phenomenon of complexity. Of scientific interest is particularly the *temporal* aspect

[3] For the following see M. Carrier, 'emergent/Emergenz', in: J. Mittelstrass (ed.), *Enzyklopaedie Philosophie und Wissenschaftstheorie*, vol. 2, 2nd ed., Stuttgart and Weimar 2005, pp. 313-314.

[4] See K. Lorenz, 'Teil und Ganzes', in: J. Mittelstrass (Ed.), *Enzyklopaedie Philosophie und Wissenschaftstheorie*, vol. 4, Stuttgart and Weimar 1996, pp. 225-228.

of the emergence thesis, i.e. for ensemble characteristics that occur in developments. Limits of reducibility (of the whole to its parts) figure here as limits of explanation and predictability, which is an important criterion of a justified theory and thus its achievement. This temporal novelty is described by the concept of *creative advance of nature*.

All these epistemological reflections, in science as well as in philosophy, are related to the already-mentioned *realism* debate. In philosophy, one distinguishes between two kinds of realism. *Ontological realism* is the position that the world of objects exists independently of human perception, knowledge and thought; *epistemological realism* – in contrast to idealism, which thinks of the world as being a construction of the self or a representation of the world, respectively – is the position that in the process of discovery, the objects of discovery play an independent role. So epistemological realism assumes essential elements of ontological realism, put simply, the existence of an 'external world'. To the extent that in (philosophical or scientific) theories a realist stand is taken, these are called *empiricist* when they make reference to the relation of the object of discovery and the subject of discovery, or *Platonist* when they make reference to the status of general concepts, so-called *universals*. Accordingly, a distinction may be made between empiricist and Platonist positions on scientific theory formation.

The status of a theory furthermore depends, also from the epistemological point of view, on the interpretation chosen, also concerning determinism and realism. An example would be the interpretation of the electromagnetic field as a state of a mechanical ether in the mechanistic tradition of the 19th century. Departing from this interpretation, Albert Einstein conceived of this field as an independent magnitude. Both are different (possible) interpretations of the same Maxwellian theory of electrodynamics. Furthermore, it is disputable whether a relational theory of space, according to which space represents merely a relation among objects and does not itself exist beside the objects or outside them, is really adequate to the General Theory of Relativity – as Einstein himself believed. Depending on how one translates classical relationalism into the concepts of relativity theory, one receives different answers to the question. At the moment at least, it is impossible definitely to privilege a particular one of these translations. In other words: One and the same theoretical approach can be differently interpreted; interpretations in these scientific cases, too, are not unequivocal. On the contrary, they display characteristic uncertainties that cannot be completely removed even by a rational reconstruction of the basic principles underlying a theory. The interpretation of quantum theory is not essentially different in this regard from an interpretation (say) of Kant's theory of space and time.

In all of these cases we are dealing with questions and areas of research whose results are not clearly attributed to physics or philosophy. This is well illustrated by physicist-philosophers such as Albert Einstein, who first endorsed an operationalist and later a realist epistemology, or Werner Heisenberg, who pursued the project of finding a theory of everything, believing in homogeneous mathematical symmetry, or Stephen Hawking, who writes on quantum cosmology from a general epistemological perspective, endorsing a falsificationist position in the sense of Karl Popper.

3. A properly philosophical status may be attributed to epistemological reflections which in the 20th century gained significance as a discipline entitled *philosophy of science*. These in general deal with problems of structure and development of science, starting from a distinction between *research form* and *theory form* of science. In its research form science is trying to discover what is the case, in its theory form it represents what it has discovered. Science in the research form is an expression of object rationality (including questions regarding the constitution of objects), science in the theory form is an expression of rationality in justification. Epistemology in the domain of science essentially refers to the theory aspect, namely to questions regarding the *structure*, *dynamics* and *explication* of theories. Under the heading 'theory structure' it analyses the structures of the language of science and of scientific explanations and the formation of theories. Under the heading 'theory dynamics' it deals with the developmental structures of scientific theories and with questions concerning the criteria of comparative theory assessment. The heading 'theory explication' applies to questions such as 'is there a physical basis for the direction of time?' or 'does the wave function of quantum mechanics refer to individual particles or an ensemble of particles?' (the Copenhagen versus the statistical interpretation). As examples for such forms of thinking about science the influential approaches of Logical Empiricism (Rudolf Carnap being the main representative) and that of Karl Popper may be mentioned.

Logical Empiricism, which epistemologically may be characterised by its appeal to the conventionalism of Henri Poincaré and its criticism of the thesis of the *synthetic a priori* of Immanuel Kant, conceives of theory development as a continual progress of discovery in which earlier theories are reduced into later ones. Epistemologically speaking, it endorses a two-level view of the conceptual structure of scientific theories, according to which in the structure of science all true propositions are either logically or analytically true propositions, or alternatively empirically or synthetically true propositions.

On this basis, it at the same time pursues the project of the *unity of science*:[5] all states of affairs can be expressed in a physicalist language and by introducing *theoretical concepts*, i.e. concepts which refer to entities not directly observable and which cannot be defined in terms of observational concepts. They are introduced by the postulates of a theory and their function and role is explicated accordingly by the appropriate theoretical context. While theoretical concepts are generally coordinated with observational indicators by correspondence rules, nonetheless, these concepts cannot be translated into such empirical indicators. The reason for their introduction is that they help to order and unify experimental laws successfully. Concepts such as electromagnetic field or the quantum-mechanical wave function, to which empirical characteristics can be assigned only indirectly, partially, and in a manner mediated by theory, are considered legitimate, because with their help the explanatory power of the theories can be increased. Theoretical concepts are thus legitimate explanatory constructs. The conceptional structure of scientific theories according to this position is shaped accordingly.

Karl Popper's approach was very different. Opposing the idea of how the reducibility of theories into each other leads to scientific progress in Logical Empiricism, Popper defends the incompatibility of successive theories. In his methodology of empirical science or *logic of scientific discovery*, entitled 'falsificationist', the term 'corroboration' takes the place of the concept of justification, in particular, empirical justification, as Popper – again, in opposition to Logical Empiricism – appeals to the asymmetry of verification and falsification: general propositions, mostly natural laws, may only be refuted (falsified), but not verified, relative to an empirical basis. Basic propositions, which according to this conception figure as premises of an empirical falsification, are interpreted as corroborating a falsifiable hypothesis. The degree of corroboration of a theory in turn depends on its degree of testability, expressed by the concept of falsifiability. The principle of a critical examination characterising a logic of scientific discovery accordingly requires a pluralism of theories so as to be able to select a 'successful' one, which later (against Popper) was extended by a pluralism of methods by Paul Feyerabend. Progress among theories is due to the ongoing process of critical revision of existing theories from the perspective of truth or at least verisimilitude.

[5] See M. Carrier and J. Mittelstrass, 'The Unity of Science', *International Studies in the Philosophy of Science* 4 (1990), pp. 17–31.

In his later works, Popper tried to describe the formation of theories as an evolutionary process, as the expansion of knowledge in problem-solving contexts, the components of which are creative guesswork and the rational elimination of error. This process is supposed to be based on a 'third world of objective contents of thought', existing alongside the 'first world' of physical objects and the 'second world' of mental states. Opposing this we find *historicist* approaches (Thomas Kuhn), *reconstructivist* approaches (Imre Lakatos), *structuralist* approaches (Joseph Sneed, Wolfgang Stegmüller) and *constructivist* approaches (Paul Lorenzen, Jürgen Mittelstrass), which mostly differ in the degree of emphasis they give to the descriptive or normative perspectives. In all these approaches, the aspect of theory dynamics is dominant.

4. Philosophy, orienting itself on the task of a philosophy of science, stays close to science, and increasingly so even as science is entering in ever closer union with technology and finding new forms of organisation. A new approach towards *technology*, as it emerged in the 20th century, is displayed, for instance, in medicine, microelectronics, and laser technology – science is leaving its academic home and is relating its knowledge to the problems of this world more and more often[6] –, a change towards new *organisational forms* through strengthening the extra-university research in the area of basic as well as in the area of applied research – with big centres of sciences such as CERN, EMBL, the Weizmann Institute and the love of large science groups (centres, clusters, networks, alliances).

With these institutional developments, not only has the organisational structure of science changed, but also the *concept of research*. Originally, this concept was closely linked to the researching subject – researchers and not institutions researched – but now the link between research the verb and research the noun is pulling apart. The community of researchers has become Research with a capital 'R'; the (re)search for truth, central to the idea of science and at the very bottom of any scientist's self-image of what makes him or her a researcher, has become research as a business operation, an organisable and organised process in which individual scientists, thought

[6] See J. Mittelstrass, *Leonardo-Welt: Über Wissenschaft, Forschung und Verantwortung*, Frankfurt am Main (Suhrkamp), pp. 47-73 ('Zukunft Forschung: Perspektiven der Hochschulforschung in einer Leonardo-Welt' [1990]); H. Nowotny and P. Scott and H. Gibbons, *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*, Cambridge etc. (Polity Press) 2001, 2007; P. Weingart and M. Carrier and W. Krohn, *Nachrichten aus der Wissensgesellschaft: Analysen zur Veraenderung der Wissenschaft*, Weilerswist (Velbruck Wissenschaft) 2007.

to be as interchangeable as individuals in the business world, disappear. The mentioned predilection for core areas, centres, clusters, alliances and networks in research is the embodiment of this change. The change is reinforcing the industrialization of science, but is also weakening science's ability to self-reflect. Self-reflection is a distinctive mark of enlightened science. It is characterised by the right ratio of proximity and distance. This is just as true in institutional terms and, when it is achieved, it constitutes the rationality of institutions, in this case scientific institutions. It is also true where scientific self-reflection is paired with social reflection (in the form of advising politics and society), a link in which modern society can find its true 'scientific' character.

There is also a *normative* aspect connected to the idea of self-reflection. Not just epistemological questions, but also aims and objectives are at issue here, and thus questions of orientation, both theoretical and practical. The ethical consequences of an increasing scientification of the world, for instance, belong to these. Philosophical foundations – these are not just epistemological, but also practical and ethically relevant foundations, through which science is normatively reconciling itself with itself and society. The fact that also foundational questions such as these have been addressed in the 20th century, together with the significant theoretical breakthroughs and the epistemological debates accompanying them, characterise it as a truly scientific century. At the same time, this character epitomises demanding requirements which science and philosophy have to satisfy today and in the future.

Scientific Papers

Session I: Astrophysics

The explanatory gap in Neuroscience

Lucia Melloni[1†] & Wolf Singer[1,2,3]

Introduction

We all know what it is like to be conscious, to be aware of something, to be aware of ourselves. However, even though there is this consensus and even though most would agree that consciousness is intimately related to brain functions, a universally accepted definition is still lacking. It is one of the goals of cognitive neuroscience to arrive at a better characterization of consciousness through a better understanding of the underlying neuronal mechanisms – and there have been remarkable advances in the study of the *Neuronal Correlates of Consciousness* (NCC). Since the seminal paper of Crick and Koch (1990), several theoretical proposals (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; Lamme, 2006; Singer, 2000) as well as important empirical contributions regarding NCC have been made. However, important questions remain unresolved. In the following we propose that the search for a neuronal *correlate* of consciousness is equivalent with the search for neuronal *mechanisms* that have to account for several constitutive properties of consciousness. Therefore, we shall first summarize these characteristics and then explore putative mechanisms. Based on theoretical considerations and empirical evidence we propose that neuronal synchrony between distant neuronal assemblies might be a key mechanism for perceptual awareness. Finally, some pressing questions in the field of consciousness research will be discussed.

Some basic definitions

Rather than aiming at a comprehensive definition of consciousness or awareness we prefer an operational definition of these terms. We consider a

[1] Department of Neurophysiology, Max Planck Institute for Brain Research, Deutschordenstraße 46, 60528 Frankfurt am Main, Germany

[2] Frankfurt Institute for Advanced Studies, Goethe University, Ruth-Moufang-Str. 1, 60438 Frankfurt am Main, Germany

[3] Ernst Strüngmann Institute for Brain Research

[†]Correspondence should be addressed to: Wolf Singer, Max Planck Institute for Brain Research, Department of Neurophysiology, Deutschordenstraße 46, 60528 Frankfurt am Main (Germany) wolf.singer@brain.mpg.de.

cognitive process as conscious, if the subject is aware of it and can report about it. If a subject can report the presence (or absence) of a stimulus (detection) or its identity (discrimination), we assume that the subject was conscious of the respective cognitive content. Failure to detect or to identify will be taken as a sign of unawareness. Even though this is not an exhaustive definition, it allows for an objective evaluation of subjective data, a scientific program called 'heterophenomenology' (Dennett, 1992). Because of the brain's complexity it is difficult to induce specific and selective modifications of neuronal activation patterns. This often precludes establishment of causal relations between neuronal and cognitive processes and restricts research to the collection of correlative evidence. Therefore, consciousness research has focused on the search for neuronal correlates that are defined as the minimal set of neuronal events which are necessary and/or sufficient for perceptual awareness (Chalmers, 2000). This definition, however, confounds the search for a mechanism with the identification of mere correlates. For example, if a response in a certain brain region is associated with a consciously perceived stimulus, activation of this area cannot per se be taken as an NCC, because the observed activity could be either the prerequisite for or the consequence of conscious processing. What is needed instead is a model that causally relates certain neuronal mechanisms to consciousness. This in turn requires induction of predicted changes of consciousness by interfering with the putative mechanisms. In the following we define several requirements that need to be met by a mechanism that supports consciousness. Subsequently we shall discuss whether neuronal synchronization fulfills these premises.

Properties of consciousness constraining neuronal implementations

One central property of perceptual awareness is its unified character. However, the architecture of the brain is distributed, supporting multiple, parallel operations in numerous specialized but highly interconnected processing areas. This raises the question of how the unity of conscious experience can arise from the distributed processing of the brain.

Another characteristic of conscious processing is that the contents of our experience constantly change over time but, at any given moment, are discrete and distinct. Thus, the neuronal implementation of consciousness has to meet the requirement to support a seamless flow of ever changing contents that are experienced as coherent in space and time.

A particularly striking feature of consciousness is its limited capacity. At any moment in time, only a small fraction of the ongoing processes in the brain gains access to consciousness. Thus, being conscious of something al-

ways implies prevention of other contents from reaching consciousness at the same time, suggesting the action of a powerful selection process that gates access to consciousness. This in turn raises the question as to what distinguishes processes that enter consciousness from those that do not. As most results of sensory and executive processing have in principle access to consciousness – with the exception of certain vegetative signals – the distinguishing signature cannot be content specific but should be assignable in a dynamic way to all activity patterns that can, in principle, give rise to conscious experience.

Subjectively it appears to us that our actions and decisions depend mainly on those processes that we are conscious of. Experimental evidence indicates, however, that stimuli and processes which have no access to consciousness also have an impact on behavior. They can either trigger or inhibit a particular behavior. These unconscious determinants of behavior are not restricted to low level sensory processes but include the results of deep semantic decoding and the programming of complex motor responses (Dehaene *et al.*, 1998; van Gaal, Ridderinkhof, Fahrenfort, Scholte, & Lamme, 2008). Thus, theories about the neuronal correlates of consciousness have to address the question of how signals are selected for access to awareness and which signatures distinguish conscious from unconscious information processing.

Neuronal synchrony as a key correlate of perceptual awareness

We shall argue that neuronal synchrony possesses most of the features that we have defined above as constitutive for consciousness or for processes leading to conscious awareness.

In order to create a unified conscious experience of multimodal contents, the modular architecture of the brain has to be overcome. One solution would be convergence of all information in a singular center. This option is considered theoretically implausible (Dennett, 1992), and so far no such center has been identified. Furthermore, behavioral and brain imaging studies have shown that unconscious processing engages very much the same cerebral areas as conscious processing, including frontal and prefrontal cortex (Lau & Passingham, 2007; van Gaal *et al.*, 2008). Thus, there is no compelling evidence for specific areas supporting conscious processing. An alternative to convergence is that the distributed processes are bound together dynamically by rendering the signals coherent in time (Hebb, 1949), e.g. by synchronization (Singer & Gray, 1995). In this framework, the presence or absence of a particular feature is signaled through the activity of feature-selective neurons, while the grouping of those elementary

features into coherent representations of cognitive contents is achieved via synchronization of the respective neurons, forming a distributed representation of a perceptual object. Neuronal synchronization is a self-organizing process that allows rapid formation and dissolution of neuronal assemblies defined by coherence. Such dynamic binding is an economical and flexible strategy to cope with the representation of the virtually unlimited variety of feature constellation characterizing perceptual objects. Taking the unified nature of conscious experience and the diversity of possible contents into account, coherence (synchrony) offers itself as a mechanism allowing representation of ever changing constellations of content in a unifying format. In this way the assemblies that represent the different but unified contents of the continuous flow of consciously processed items could be bound together in ever changing constellations into a coherent, but permanently changing whole (meta-assembly). The updating of this meta-assembly would then be achieved by continuous cooption and exclusion of subassemblies. In this framework the rate-limiting factor for the formation of a new meta-assembly corresponds to the time needed to establish stable coherence. In case coherence is expressed by synchrony, this would be the time needed to establish stable phase relations. Stable states would then be reached once the relative phase between local oscillations ceases to change (Tognoli & Kelso, 2009).

Synchronization is also ideally suited to contribute to the selection of contents for access to consciousness. Synchronization enhances the saliency of signals and thereby facilitates their propagation in sparsely connected networks such as the cerebral cortex (Abeles, 1991; Fries, 2009; Jensen, Kaiser, & Lachaux, 2007). Gamma band synchronization, in particular, assures coincidence among distributed inputs with millisecond precision. Furthermore, when neuronal responses engage in synchronized oscillations, frequency and phase adjustments can be exploited for the selective routing of activity and the dynamic gating of interactions between interconnected neurons. At the level of individual neurons, oscillations are associated with periodic alternations of phases with high and low excitability, the latter resulting from the barrage of synchronized IPSPs that have both a shunting and a hyperpolarizing effect (Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). Excitatory inputs that arrive at the depolarizing slope of an oscillation cycle generate large responses, whereas inputs arriving at the falling slope and trough are shunted and ineffective. Hence, neuronal oscillations define temporal windows for effective communication between neurons, providing a mechanism to selectively and flexibly bias the communication between neuronal groups (Fries, 2009). When two groups of neurons open their windows of susceptibility

(their excitatory phases) at the same time, they are more likely to interact with each other, to increase their synchrony and, as a consequence, to also have enhanced influence on other groups of neurons. By adjusting oscillation frequency and phase, groups of neurons can either be linked together into tightly coupled assemblies or be segregated into functionally isolated subgroups. This mechanism can act both within and across areas, and can, in principle, account for selective and flexible routing of information within networks with fixed anatomical architectures. Taken together, the oscillatory patterning of activity and the option to adjust frequency and phase of the oscillations could serve three complementary functions: gain control, selective and flexible routing of information between neuronal groups, and formation of coherent representations. Furthermore, if depth of processing is determined by the extent of propagation of information in the brain, this can also account for the observation that conscious perception is associated with deeper processing than unconscious perception.

As previously mentioned, any neuronal correlate of consciousness should exhibit signatures that differ between conscious and unconscious forms of information processing. Regarding neuronal synchrony, this prerequisite is fulfilled with respect to its spatial scale: Following the distinction between local and global scale integration by Varela *et al.* (2001), processing carried out unconsciously (which is automatic and modular) should be based mainly on local integration in divergent-convergent feed-forward architectures, whereas conscious processing should involve large-scale integration via the extended networks of re-entry loops that couple neurons both within and across the various levels of the cortical processing hierarchies (for a similar proposal see Dehaene *et al.*, 2006). Anatomical and physiological studies suggest that neuronal, in particular cortical, architectures share features of 'small world networks' (reviewed in Bassett & Bullmore, 2006). These allow for the coexistence of both local modular and global, distributed processes. Important properties of this architecture are minimization of path length between any pairs of nodes, optimization of the number of connections and the possibility of coexistence of multiple local processes and globally ordered states. Moreover, such networks can operate in critical states, allowing for fast reconfigurations of network dynamics (Bassett & Bullmore, 2006; Sporns & Zwi, 2004; Yu, Huang, Singer, & Nikolic, 2008). Cortico-cortical connections can be subdivided in two major subgroups. Local, intra-cortical connections that run tangentially to the layers and link neurons that share similar response properties and are separated by only a few hundred micrometers, and long-distance connections that often but not always run through the underlying white matter and link neurons in different cortical areas. The latter serve exchange of

information between distinct cortical sites and can establish globally coordinated activation patterns (Varela *et al.*, 2001).

We started this chapter describing four conditions that any theory of NCC should satisfy. We then discussed how neuronal synchronization could fulfill each of them, namely that neuronal synchronization could account for (i) the unity of experience as well as for (ii) its temporal dynamics, (iii) for the selection of signals entering conscious processing and for (iv) the distinction between conscious and unconscious processing related to the spatial scale of the synchronized activity. We shall now review evidence relating to neuronal synchronization to consciousness.

Evidence relating to long-range synchronization and consciousness

Masking offers an interesting possibility to distinguish between conscious and subconscious processing, since the same physical stimuli can be either perceived or not perceived depending on the temporal and spatial sequence of stimuli that surround them. In our studies, we capitalized on this phenomenon and presented words that could be perceived in some trials and not in others (by adjusting the luminance of the mask) and simultaneously performed electroencephalographic (EEG) recordings (Melloni *et al.*, 2007). Several measures were analyzed: time-resolved power changes of local signals, the precision of phase synchronization across recording sites over a wide frequency range, and event-related potentials (ERPs). A brief burst of long-distance synchronization in the gamma frequency range between occipital, parietal and frontal sensors was the first event that distinguished seen from unseen words, while local synchronization was similar between conditions. Interestingly, after this transient period of synchronization, several other measures differed between seen and unseen words: we observed an increase in amplitude of the P300 ERP for visible words which most likely corresponds to the transfer of information to working memory. In addition, during the interval period in which visible words had to be maintained in memory, we observed increases in frontal theta oscillations. Theta oscillations have been related to maintenance of items in short-term memory (Jensen & Tesche, 2002).

To test whether the increase in long-distance synchronization relates to awareness or depth of processing, we further manipulated the depth of processing of invisible words. It has previously been shown that invisible words can be processed up to the semantic and motor level (Dehaene *et al.*, 1998). In a subliminal semantic priming experiment we briefly presented words (invisible) that could either be semantically related or not related to a second visible word on which subjects had to carry out a semantic classification

task. Invisible words were processed up to semantic levels as revealed by modulation of the reaction times depending on the congruency between invisible and visible words: congruent pairs exhibited shorter reaction times than incongruent ones. We observed increases in power in the gamma frequency range for unseen but processed words. For visible words we additionally observed increases in long-distance synchronization in the gamma frequency range (Melloni & Rodriguez, 2007). Thus, local processing of stimuli is reflected in increases in gamma power, whereas long-distance synchronization seems to be related to awareness of the stimuli. This suggests that conscious processing requires a particular dynamical state of the cortical network that is characterized by the dynamic configuration of widely distributed networks through transient synchronization. The large-scale synchronization that we observed in our study could reflect the transfer of contents into awareness and/or their maintenance. We favor the first possibility given the transient nature of the effect and argue that the subsequent theta oscillations might support maintenance. It is conceivable that short periods of long-distance synchronization in the gamma band reflect the update of new contents, while the slower pace of theta oscillations might relate to the sustained integration and maintenance of local results. The interplay between these two frequency bands might underlie the phenomenon of continuous but ever changing conscious experience (see below).

Recently, Gaillard *et al.* (2009) revisited the question of processing of visible and invisible words. In intracranial recordings in epileptic patients they observed that invisible words elicited activity in multiple cortical areas which quickly vanished after 300 ms. In contrast, visible words elicited sustained voltage changes, increases in power in the gamma band, as well as long-distance synchronization in the beta band that showed bidirectional Granger causality. In contrast to our study, Gaillard *et al.* observed a rather late (300-500 ms) rise of long-distance synchronization. However, it is important to note that in the study of Gaillard *et al.*, phase-synchrony was analyzed mostly over electrodes within a given cortical area or at most between hemispheres, and thus this analysis raises an important methodological issue. Earlier synchronization events could have passed undetected because of lack of electrode coverage. Since with intracranial recordings the electrode placement is based on medical and not experimental considerations, analyses are restricted to the available electrodes and locations. Despite of these restrictions, this study provides one of the most compelling pieces of evidence for a relation between long-distance synchronization and consciousness.

Another commonly used paradigm in studies of consciousness is binocular rivalry. When two images that cannot be fused are presented simultaneously

to each of the two eyes, perception fluctuates between the two images. Thus, despite constant stimulation conditions, perception alternates. This suggests a selection mechanism that gates access to consciousness. Several studies using different stimulus materials as well as recording methods (single cell analysis and local field potential recordings in animals and electroencephalographic and magnetoencephalographic registration in human subjects) have shown increased synchronization and phase locking of oscillatory responses to the stimulus that was consciously perceived and controlled the pursuit eye movements (Cosmelli *et al.*, 2004; Fries, Roelfsema, Engel, König, & Singer, 1997; Srinivasan, Russell, Edelman, & Tononi, 1999). Cosmelli *et al.* (2004) extended the findings obtained in human subjects by performing source reconstruction and analyzing phase-synchrony in source space. These authors observed that perceptual dominance was accompanied by coactivation of occipital and frontal regions, including anterior cingulate and medial frontal areas. Recently, Doesburg *et al.* (2009) provided evidence for a relation between perceptual switches in binocular rivalry and theta and gamma band synchronization. Perceptual switches were related to increments in long-distance synchronization in the gamma band between several cortical areas (frontal and parietal) that repeated at the rate of theta oscillations. The authors suggested that transient gamma-band synchronization supports discrete moments of perceptual experience while theta oscillations structure their succession in time, pacing the formation and dissolution of distributed neuronal assemblies. Thus, long-range gamma synchronization locked to ongoing theta oscillations could serve to structure the flow of conscious experience allowing for changes in content every few hundred millisecond. Further research is required to clarify the exact relation between the two frequency bands and their respective role on the generation of percepts and the pacing of changes in perception.

Another paradigm in consciousness research exploits the attentional blink phenomenon. When two stimuli are presented at short intervals among a set of distractors, subjects usually detect the first (S1) but miss the second (S2) when the two stimuli are separated by 200-500 ms. Increases in long-range neuronal synchrony in the beta and gamma frequency ranges have been observed when S2 is successfully detected (Gross *et al.*, 2004; Nakatani, Ito, Nikolaev, Gong, & van Leeuwen, 2005). Furthermore, Gross *et al.* (2004) observed that successful detection of both S1 and S2 was related to increased long-distance synchronization in the beta range to both stimuli, and this enhanced synchrony was accompanied by higher desynchronization in the inter-stimulus-interval. Thus, desynchronization might have facilitated the segregation of the two targets, allowing for identification of the second stimulus (also see Rodriguez *et al.*, 1999). Source analysis revealed, as in the

case of binocular rivalry, dynamical coordination between frontal, parietal, and temporal regions for detected targets (Gross *et al.*, 2004).

In summary, studies of masking, binocular rivalry, and the attentional blink support the involvement of long-range synchronization in conscious perception. Recent investigations have suggested further that a nesting of different frequencies, in particular of theta and gamma oscillations, could play a role in pacing the flow of consciousness. Furthermore, the study of Gross *et al.* (2004), suggests that desynchronization could serve to segregate representations when stimuli follow at short intervals. These results are encouraging and should motivate further search for relations between oscillatory activity in different frequency bands and consciousness, whereby attention should be focused not only on the formation of dynamically configured networks but also on their dissolution.

Pressing question in the field of consciousness research

One influential view on the function of consciousness posits that consciousness allows for exchange of contents between several independent processing modules, which in turn makes this information available to a variety of processes including perceptual categorization, unification in short- and long-term memory, evaluation, etc. (Baars, 1997). If consciousness is a prerequisite of these cognitive processes, how can the requirements for and the consequences of consciousness be distinguished? If conscious and unconscious processing did not only differ with respect to their respective qualitative properties, but also with respect to their consequences, such a distinction might indeed be impossible: If consciousness, as proposed, enhances the depth of processing then a method that contrasts differences in perceptual states of physically identical stimuli will reveal not only activity related to awareness per se but also activity related to its consequences, i.e. deeper processing, episodic memory formation, etc. (for a similar argument see Lau, 2008). This might explain why different research groups find signatures of consciousness at different moments in time. Neurons in the medial temporal lobe (MTL) show an all or none response profile depending on whether stimuli are processed consciously. When briefly presented images are made unperceivable through backward masking, neurons in MTL do not respond. In contrast, when the same images are recognized, MTL neurons exhibit clear responses with a latency of about 300 ms (Quiroga, Mukamel, Isham, Malach, & Fried, 2008). Following the logic of the contrastive approach, this result can be taken as a correlate of consciousness. The question is: is this a correlate of consciousness or a correlate of the consequences of consciousness? Given the intimate

relation between MTL and hippocampus and hence with networks responsible for the management of episodic memory, it is also conceivable that the rather late MTL responses have to do with memory formation following conscious perception rather than with the access to consciousness per se. This view is supported by the fact that complete resection of the hippocampus and adjacent temporal cortex does not lead to deficits in conscious perception per se (Postle, 2009). However, this does also not imply that MTL does not contribute to conscious perception. It is important to distinguish between neuronal signatures which correlate with the content of our moment-to-moment conscious awareness, and those that contribute to the feeling of a stream of consciousness (continuous present). Thus, MTL activity might not contribute to the former, but be essential for the latter. In order to experience a continuous present, memories from events that have just occurred should be linked to those occurring at the present time. Thus, the delayed MTL activity could serve to assure the continuity of the flow of consciousness.

The conundrum of cause and effect

How else could one solve the conundrum to distinguish between cause and effect? There is no panacea. Obviously, the most straightforward approach is to obtain comprehensive data on the sequence of effects distinguishing conscious from unconscious processing, assuming that causes precede effects. This could be complemented with attempts to interfere with these sequential processes using transcranial magnetic stimulation. For instance, if blocking the late activity in MTL does not abolish conscious perception but its continuity we would be one step further. Following this strategy those processes could be discarded one by one that figure as consequences of consciousness, leaving us with those more closely related to consciousness per se. Secondly, one could evaluate introspective reports as a direct measure of the quality of the experience. In this case one would not contrast conscious with unconscious processing, but arrive at a parametric estimate of 'consciousness'. For instance, in visual experiments, we could ask subjects about how clear their content of awareness is, and then correlate increases in subjective clarity with measures of neuronal activity.

Conclusions

It is evident that the study of consciousness has greatly profited from the search for neuronal correlates. However, simply showing that the brain makes a difference between conscious and unconscious processing is not sufficient.

We propose here to go one step further and to develop mechanistic explanations that establish plausible and, hopefully, at later stages even causal relations between brain processes and consciousness. Oscillatory synchrony is one candidate mechanism, and it has the advantage that it can be measured relatively directly in humans who are able to give detailed descriptions about their conscious experience. However, oscillations and synchrony seem to be mechanisms that are as intimately and inseparably related to neuronal processing as the modulation of neuronal discharge rates. Thus, without further specification these phenomena cannot be addressed as NCC apart from the triviality that consciousness does not exist without them. We and others (Varela *et al.*, 2001) propose that the spatial scale and perhaps also the precision and stability of neuronal synchrony might be taken as more specific indicators of whether the communication of information in the brain is accompanied by conscious experience or not. In this framework, conscious experience arises only if information that is widely distributed within or across subsystems is not only processed and passed on to executive structures but in addition bound together into a coherent, all-encompassing, non-local but distributed meta-representation. This interpretation is compatible with views considering consciousness as the result of the dynamic interplay of brain subsystems; but it poses the challenging question related to the 'hard problem of consciousness' research: is there something on top of this distributed meta-assembly that makes us experience? From a truly dynamicist point of view, the answer is 'probably not'. However, this should not discourage us from trying to separate the contribution of each subsystem and then go on to characterize how they interact. This would be a real alternative to the ever-lurking homunculus.

Acknowledgements

This work was supported by the Max Planck Society. We are indebted to Caspar M. Schwiedrzik for insightful comments on this manuscript.

References

Abeles, M. (1991). *Corticotronics: Neural circuits of the cerebral cortex*. Cambridge, MA: Cambridge University Press.

Baars, B.J. (1997). *In the theatre of consciousness: the workspace of the mind*. New York, NY: Oxford University Press.

Bassett, D.S., & Bullmore, E. (2006). Small-world brain networks. *Neuroscientist, 12*(6), 512-523.

Chalmers, D.J. (2000). What is a neural correlate of consciousness? In T. Metzinger (ed.), *Neural correlates of consciousness: em-*

pirical and conceptual questions (pp. 17-40). Cambridge, MA: MIT Press.

Cosmelli, D., David, O., Lachaux, J.P., Martinerie, J., Garnero, L., Renault, B., *et al.* (2004). Waves of consciousness: ongoing cortical patterns during binocular rivalry. *Neuroimage, 23*(1), 128-140.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Semin Neurosci, 2*, 263-275.

Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn Sci, 10*(5), 204-211.

Dehaene, S., Naccache, L., Le Clec, H.G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., *et al.* (1998). Imaging unconscious semantic priming. *Nature, 395*(6702), 597-600.

Dennett, D.C. (1992). *Consciousness explained*. London: Penguin.

Doesburg, S.M., Green, J.J., McDonald, J.J., & Ward, L.M. (2009). Rhythms of consciousness: binocular rivalry reveals large-scale oscillatory network dynamics mediating visual perception. *PLoS One, 4*(7), e6142.

Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annu Rev Neurosci, 32*, 209-224.

Fries, P., Roelfsema, P.R., Engel, A.K., König, P., & Singer, W. (1997). Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proc Natl Acad Sci USA, 94*(23), 12699-12704.

Gaillard, R., Dehaene, S., Adam, C., Clemenceau, S., Hasboun, D., Baulac, M., *et al.* (2009). Converging intracranial markers of conscious access. *PLoS Biol, 7*(3), e61.

Gross, J., Schmitz, F., Schnitzler, I., Kessler, K., Shapiro, K., Hommel, B., *et al.* (2004). Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. *Proc Natl Acad Sci USA, 101*(35), 13050-13055.

Hebb, D.O. (1949). *The organization of behavior*. New York, NY: Wiley.

Jensen, O., Kaiser, J., & Lachaux, J.P. (2007). Human gamma-frequency oscillations associated with attention and memory. *Trends Neurosci, 30*(7), 317-324.

Jensen, O., & Tesche, C.D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *Eur J Neurosci, 15*(8), 1395-1399.

Lamme, V.A. (2006). Towards a true neural stance on consciousness. *Trends Cogn Sci, 10*(11), 494-501.

Lau, H.C. (2008). Are we studying consciousness yet? In L. Weiskrantz & M. Davies (eds), *Frontiers of Consciousness. The Chichele Lectures* (pp. 245-258). Oxford: Oxford University Press.

Lau, H.C., & Passingham, R.E. (2007). Unconscious activation of the cognitive control system in the human prefrontal cortex. *J Neurosci, 27*(21), 5805-5811.

Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., & Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. *J Neurosci, 27*(11), 2858-2865.

Melloni, L., & Rodriguez, E. (2007). Non-perceived stimuli elicit local but not large-scale neural synchrony. *Perception, 36* (ECVP Abstract Supplement).

Nakatani, C., Ito, J., Nikolaev, A.R., Gong, P., & van Leeuwen, C. (2005). Phase synchronization analysis of EEG during attentional blink. *J Cogn Neurosci, 17*(12), 1969-1979.

Postle, B.R. (2009). The hippocampus, memory, and consciousness. In S. Laureys & G. Tononi (eds), *The Neurology of consciousness: Cognitive Neuroscience and Neuropathology* (pp. 326-338). London: Academic Press.

Quiroga, R.Q., Mukamel, R., Isham, E.A., Malach, R., & Fried, I. (2008). Human single-neuron responses at the threshold

of conscious recognition. *Proc Natl Acad Sci USA, 105*(9), 3599-3604.

Rodriguez, E., George, N., Lachaux, J.P., Martinerie, J., Renault, B., & Varela, F.J. (1999). Perception's shadow: long-distance synchronization of human brain activity. *Nature, 397*(6718), 430-433.

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci, 12*(3), 106-113.

Singer, W. (2000). Phenomenal awareness and consciousness from a neurobiological perspective. In T. Metzinger (ed.), *Neural correlates of consciousness: empirical and conceptual questions* (pp. 121-137). Cambridge, MA: MIT Press.

Singer, W., & Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu Rev Neurosci, 18*, 555-586.

Sporns, O., & Zwi, J.D. (2004). The small world of the cerebral cortex. *Neuroinformatics, 2*(2), 145-162.

Srinivasan, R., Russell, D.P., Edelman, G.M., & Tononi, G. (1999). Increased synchronization of neuromagnetic responses during conscious perception. *J Neurosci, 19*(13), 5435-5448.

Tognoli, E., & Kelso, J.A. (2009). Brain coordination dynamics: true and false faces of phase synchrony and metastability. *Prog Neurobiol, 87*(1), 31-40.

van Gaal, S., Ridderinkhof, K.R., Fahrenfort, J.J., Scholte, H.S., & Lamme, V.A. (2008). Frontal cortex mediates unconsciously triggered inhibitory control. *J Neurosci, 28*(32), 8053-8062.

Varela, F., Lachaux, J.P., Rodriguez, E., & Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. *Nat Rev Neurosci, 2*(4), 229-239.

Yu, S., Huang, D., Singer, W., & Nikolic, D. (2008). A small world of neuronal synchrony. *Cereb Cortex, 18*(12), 2891-2901.

Great Discoveries Made by Radio Astronomers During the Last Six Decades and Key Questions Today

Govind Swarup

1. Introduction

An important window to the Universe was opened in 1933 when Karl Jansky discovered serendipitously at the Bell Telephone Laboratories that radio waves were being emitted towards the direction of our Galaxy [1]. Jansky could not pursue investigations concerning this discovery, as the Laboratory was devoted to work primarily in the field of communications. This discovery was also not followed by any astronomical institute, although a few astronomers did make proposals. However, a young electronics engineer, Grote Reber, after reading Jansky's papers, decided to build an innovative parabolic dish of 30 ft. diameter in his backyard in 1935 and made the first radio map of the Galaxy in 1940 [2].

The rapid developments of radars during World War II led to the discovery of radio waves from the Sun by Hey in 1942 at metre wavelengths in UK and independently by Southworth in 1942 at cm wavelengths in USA. Due to the secrecy of the radar equipment during the War, those results were published by Southworth only in 1945 [3] and by Hey in 1946 [4]. Reber reported detection of radio waves from the Sun in 1944 [5]. These results were noted by several groups soon after the War and led to intensive developments in the new field of radio astronomy.

In Section 2 are summarized radio observations of the Sun and of the massive coronal mass ejections that disrupt satellites and terrestrial power grids. In Section 3 are described discoveries of the powerful radio galaxies and quasars that indicate the presence of supermassive Black Holes of millions of solar mass at the centre of galaxies. In Section 4 is described the great controversy that arose between the Steady State theory and the Big Bang Model in 1961, after Martin Ryle and colleagues noted excess counts of weaker radio sources in the catalogue made by them using radio interferometers. I then describe observations of angular size of a large number of weak radio sources made with the Ooty Radio Telescope using the method of lunar occultation; their statistics indicated the evolution of the radio sources with cosmic epoch, consistent with the Big Bang Model. In Section 5 are described the important discovery of the Cosmic Microwave

Background radiation (CMB) by Penzias and Wilson in 1965 and later its detailed observations with higher angular resolution by Mather *et al.* in 1990 with the COBE satellite and by Bennett *et al.* in 2003 with the WMAP satellite; these observations have given a firm support to the Big Bang Model, indicating that the Universe is dominated by 74% dark energy, 22% dark matter, and 4% ordinary matter. Observations of the HI emission from the spiral galaxies and attempts to measure the epoch of re-ionization are summarized in Section 6. The serendipitous discovery of the Pulsating Radio Sources (Pulsars) is described in Section 7. Observations of more than a hundred molecules in the interstellar medium and megamasers are summarized in Section 8. Developments of earth's rotation synthesis radio telescopes for high-resolution observations of celestial radio sources are described in Section 9. In Section 10 are discussed some of the Key Questions today concerning the Universe. Conclusions are given in Section 11.

In this brief review, pioneering observations and discoveries are described at first, followed by descriptions of the current status. The references are not exhaustive and only indicative.

2. Radio Studies of the Sun and Solar Wind

Soon after the end of the War in 1945, a few groups, particularly in Australia and the UK, started detailed observations of radio emission from the Sun, using existing radar equipment to begin with and later with interferometers. In 1946 and 1947, Pawsey and colleagues found that: (a) solar corona has a temperature of about one million degrees, (b) solar radio emission has a slowly varying component related to sunspot area and (c) there occur intense radio bursts associated with the flare activity [6]. Ryle and colleagues also measured angular sizes of solar emission associated with sunspots and also its circular polarization confirming the predictions by Martyn and by Appleton and Hey. These discoveries led to the development of two major facilities in Australia for investigating characteristics of the solar radio emission. Wild and colleagues [7] developed a swept frequency solar radio spectrograph that led to major classifications of solar radio bursts: (a) Type I, as noise storms, (b) Type II caused by outward ejections of matter with velocities of hundreds of km that cause plasma oscillations at successive higher levels of the solar corona and (c) Type III, caused by ejections of matter of ~1/3rd of the velocity of light. Type IV was later identified by French workers, Type V by Wild and colleagues and Type U by Maxwell and Swarup. In 1953, Christiansen and Warburton constructed an innovative grating array in order to make two-dimensional maps of the radio emission from the Quiet Sun [8]. During the last 60 years, these

pioneering observations have been pursued in great detail by scores of workers and have provided very valuable information about the solar activity [9]. Of particular importance are the massive coronal mass ejections (CMEs) that derive their tremendous energy from the stressed magnetic fields by the sunspot activity on the Sun causing large disturbances on the earth. CMEs have also been associated with the coronal holes. Observations of the interplanetary scintillations of ~1000 compact components of distant radio galaxies and quasars are being done on a daily basis over a large part of the sky around the Sun by Manoharan and colleagues using the Ooty Radio Telescope in India [10]. These observations provide information about variations of the solar wind and also acceleration of the coronal mass ejections affecting the earth. During the last 15 years, X-ray and coronagraphic observations of the Sun by the SOHO satellite of NASA have provided valuable data about the quiet and active Sun. NASA's STEREO has revealed the 3D structure of the CMEs. Japanese and Russian agencies have also built solar observatories in Space.

3. Radio Galaxies, Quasars, and Black Holes

3.1. Radio Galaxies

I describe firstly the remarkable story of the discovery of Cygnus A and its optical identification with a distant galaxy. In 1945 Hey, Parson and Phillips in the UK noted fluctuations in the intensity of cosmic radio noise towards the direction of the Cygnus constellation [11]. Their antenna had a very broad beam. In 1947 Bolton and Stanley determined its source size as ~8 arc-minute using a 'sea interferometer', consisting of an antenna placed on a hill at Dover Heights towards the Pacific Ocean in Australia that produced interference fringes as the source rose from the horizon [12]. In 1951, Graham-Smith measured its position to an accuracy of ~1 arc-minute using a radio interferometer [13]. Thereafter, Baade and Minkowski made observations in that direction with the 200 inch (5 m) Mt. Palomar telescope and identified Cygnus A with a perturbed galaxy having a recession velocity of 17000 km s^{-1}, corresponding to a redshift of 0.06 implying a distance of ~1000 million light years, much further than any other known optical galaxy at that time [14]. In 1953 using an intensity interferometer of about one arc-minute resolution, Jenison and Das-Gupta found that Cygnus A is a double radio source [15]. Since Cygnus A has a very high flux density, it became clear that it should be possible to detect much weaker radio sources up to large distances using sensitive radio telescopes and thus distinguish between various cosmological models, as discussed in the next Section.

At that time there was great controversy about the physical processes giving rise to the very powerful radio emission. Brehmstrahlung radiation by hot bodies was totally inadequate. It was concluded in 1954 that radio emission is caused by 'synchrotron radiation' when electrons with relativistic velocities spiral in the presence of magnetic fields resulting in radiation of extremely high power. Observations of the predicted polarization gave support to the theory.

By 1950, using rather modest equipment, Australian and UK radio astronomers had catalogued ~50 discrete radio sources. A few were associated with known galaxies such as Virgo A and Centaurus A. Later, Martin Ryle and his group constructed radio interferometers using parabolic cylinders with large collecting area at Cambridge in UK and catalogued more than 250 radio sources across the northern sky by 1960. By then, Bernie Mill and colleagues in Australia also catalogued a few hundred radio sources, mostly across the southern sky, using the Mills-Cross using dipole arrays. At first, there was great controversy concerning the overlapping portions of the two catalogues but it was resolved soon with better measurements by the Cambridge group, resulting in the well-known 3C catalogue.

Since the wavelength of radio waves is quite large, radio interferometers with spacing of many kilometers are required for making detailed radio images of celestial sources with arcsec resolution, as is now possible using synthesis radio telescopes that are described in Section 9. Today, thousands of radio galaxies have also been mapped with sub-arc second resolution. Very long baseline interferometers (VLBI) have provided even milli-arcsec resolution. About 13 years ago, Japanese astronomers placed a 10m diameter parabolic dish orbiting in space and combined it with ground radio telescopes on Earth in order to study a few compact radio sources with 0.0001 arcsec resolution. To date, millions of extragalactic radio sources have been catalogued by various workers. A major challenge has been to make optical identification, although it has become easier after the usage of CCDs on optical telescopes. Yet, a large number of radio sources have remained unidentified with galaxies, as most of the weaker radio sources are likely to have much higher redshifts, requiring large optical telescopes to observe fainter galaxies.

We next summarize the nature of radio galaxies. As described earlier, radio galaxies are millions of time more energetic than normal galaxies. A radio galaxy is generally a double or triple source, with two outer radio lobes and a central component associated with a supermassive Black Hole at the centre of the galaxy. The central active galactic nuclei (AGN) give rise to jets of relativistic electrons and positrons, and also slower protons, in two opposite direction that result in radio lobes at the two opposite extremities (see Figure 1, p. 357).

3.2. Quasars (QSO)

3C273 is a compact quasi-stellar radio source (quasar). In 1963, Martin Schmidt concluded from the known spectral lines of 3C273, which hitherto were found to be very puzzling as their occurrence could not be explained by any stellar process, that the spectra consisted of Balmer lines of hydrogen and were Doppler shifted corresponding to a redshift of 0.158, the highest known redshift at that time [17]. This conclusion indicated immediately the existence of a new class of celestial objects in the Universe. 3C273 is an optically bright galaxy with a magnitude of 13. It is a compact radio source having not only a radio jet but also optical and X-ray jets. Subsequently, a large number of quasars have been discovered, the brighter ones mostly by Australian radio astronomers using the Parkes Radio Telescope at 5GHz. Radio and optical surveys have indicated that quasars are associated with galaxies that have active galactic nuclei (AGN). Many AGNs are found only at optical wavelengths and are called Quasi Stellar Objects (QSO). A large number of QSOs have also been catalogued by optical surveys; a few have been identified even at redshifts > 6. Many QSOs are found to be radio loud but a large number are radio quiet. Many QSOs are also strong X-ray sources, with the X-ray emission arising close to the Black Holes located at the centre of AGNs. X-ray observations have provided important information about Black Holes, such as their spin and properties of the surrounding Accretion disks that feed matter into the Black Holes from the associated galaxies.

Unified models of active galaxies indicate that jets of energetic particles emanate from the AGNs moving outwards at relativistic velocities. If the jet is beamed towards the observer, the radio emission from the central component, that generally has a flat spectrum, gets relativistically beamed and thus AGN is observed as a quasar with flat spectrum. If one of two jets is at a larger angle to the observer, radio emission only from that jet is seen as shown in Figure 1 (see. page 357). The unified model also explains observed optical spectra of the AGNs, which depend on the orientation of the axis of tori around the central Black Holes.

3.3. Black Holes in the Universe

Although the presence of massive Black Holes at the centre of active galaxies was firstly established by radio astronomy observations, their existence has been firmly established by extensive radio, optical and X-ray observations. It has now been concluded that almost all galaxies have supermassive Black Holes of several millions, some even billions, of solar mass. Our Galaxy has a Black Hole of about one million solar mass. The

matter from the outer parts of a galaxy spirals into the Black Hole, forming a torus and an accretion disk around the Black Hole. Although it is not yet clear as to how jets are created, it has been suggested that a large magnetic field near the centre of AGN gives rise to the jets of energetic particles. However, not all galaxies are powerful radio sources. Stellar-size Black Holes have also been discovered in our Galaxy.

OJ287 is the largest Black Hole observed to date. It is considered to be an object of ~100 million Suns in a binary orbit around a Black Hole of ~17 Billion Suns. OJ287 was first detected in 1970s in the Ohio Radio Sky Survey. From its radio and optical observations, it is classified as a BL Lac object (with featureless continuum emission). It is located 3.5 billion light years away. It has produced many quasi-periodic optical outbursts going back approximately 120 years, as apparent on photographic plates from 1891 (Wikipedia).

4. Radio Astronomy and Cosmology

In 1929 Hubble made a remarkable discovery that *the farther away is a galaxy from us, the faster it is moving away*. His observations indicated that the Universe is expanding and provided support to Father Lemaitre's 1927 paper that was based on the framework of the General Theory of relativity. These results gave rise to the Big Bang Model, according to which the Universe was extremely energetic and tiny in the beginning and has continued to expand thereafter. As an alternative to the Big Bang theory, in 1948 Hoyle, Bondi and Gold proposed the Steady State Theory, according to which new matter is created continuously as the Universe expands [18].

By 1960, Martin Ryle from Cambridge had catalogued ~250 radio galaxies. He concluded from the 'radio source counts' that weaker radio sources are much more numerous supporting the Big Bang model [19], (see Figure 2). However, Hoyle questioned the implicit assumption by Ryle that the weaker sources are located far away. The above arguments immediately led to a great controversy between the Big Bang model and the Steady State Theory.

In order to distinguish between the Big Bang and Steady State models, I decided in 1963 to measure angular sizes of a large number of weak radio sources, as radio sources located far away were expected to have smaller angular sizes. At that time, angular sizes of extra-galactic radio sources had been determined with sufficient accuracy for a few dozen sources only. Further, no suitable radio interferometers with sufficiently large sensitivity and arcsec resolution were in the offing. Hence, I decided to use the lunar occultation

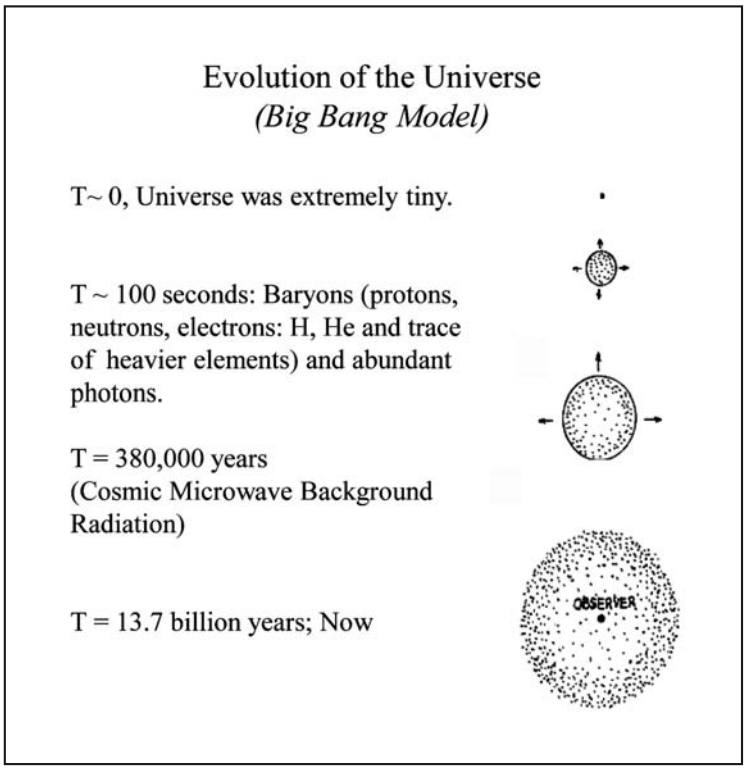


Figure 2. The sketch indicates evolution of the Universe according to the Big Bang Model. The Model predicts that weaker radio sources located far away would be much more numerous compared to those nearby (assuming no evolution of luminosities with cosmic epoch).

method for the purpose. I proposed and directed construction of a cylindrical radio telescope of a large size, 530m long and 30m wide that was located on a hill at Ootacamund (Ooty), with its long axis parallel to that of the earth, taking advantage of India's location close to earth's equator. Using the Ooty Radio Telescope (ORT), we measured angular sizes of ~1000 weak radio sources with arcsec resolution using the method of lunar occultation. We found that the weaker sources of smaller flux densities had smaller angular size, compared to angular size of stronger 3C sources with higher flux densities. Also, our results indicated cosmic evolution of radio sources, thus providing independent evidence of the Big Bang Model [20, 21].

5. Big Bang Model and the Cosmic Microwave Background (CMB)

5.1. Big Bang Model

According to Big Bang Model, the Universe originated from an extremely energetic state at a very early epoch. Based on the measured values of the fundamental constants, we can derive the Planck time of 5×10^{-44} seconds as a possible era of the Universe soon after its creation. It is considered that Space and Time originated at that epoch. Active areas of quantum mechanics, quantum gravity and string theories are attempting to answer the question as to what was before the Big Bang but this area of research is still at an early stage. In 1980 Guth proposed that soon after the Big Bang the Universe went to a very rapid inflation by a factor of 10^{120}. The inflationary model explained certain dilemmas such as the homogeneity of the Universe and the horizon problem. The inflationary model also predicted occurrence of small-scale irregularities with a flat power spectrum that have been measured by the Cosmic Microwave Background radiation, as described below. Certain predictions of the above cosmological model are constituents of the Grand Unification Theory of Particle Physics and are subject to experimental verification that is being done by the Large Hadron Collider and other large accelerators.

In the Big Bang Model, matter consisting of protons, electrons and neutrons formed in the Universe when its temperature decreased from ~10^{11} to ~10^{9} K at epochs of ~0.1 s to 100 s. The Model predicts abundance of ~0.76% of hydrogen and ~0.24% of helium-4, and a much smaller fraction of deuterium, helium-3 and lithium-7 with respect to the baryon density of the Universe. Many observations of abundance of these elements made over the last few decades are found to be consistent with the predictions of the Big Bang Model. The Model also predicts occurrence of a very large density of photons due to annihilation of electrons, positrons, etc. at the early cosmic epoch of tens of seconds. The Cosmic Microwave Background (CMB) Radiation originated after the photons stopped interacting with matter as the temperature of the Universe decreased to ~3000 K when electrons and protons combined. Detailed observations of the CMB have provided a strong foundation to the above Model as described below.

5.2. Cosmic Microwave Background (CMB)

In 1965, Penzias and Wilson made a serendipitous discovery that there exists an all sky Microwave Background Radiation (CMB), corresponding to blackbody temperature of ~3 K [22]. The Nobel Prize was given to them in 1978. This radiation had been predicted to occur in the Big Bang Model

and therefore gave strong support to the Model. It could not be explained in the Steady State Theory.

During the 1990s the COBE satellite designed by John Mather and George Smoot showed that the CMB radiation corresponded to a perfect blackbody radiation of 2.7 K (Figure 2), with minute anisotropies as expected in the Big Bang Model (Figure 3) [23]. The Nobel Prize was given to them jointly in 2006.

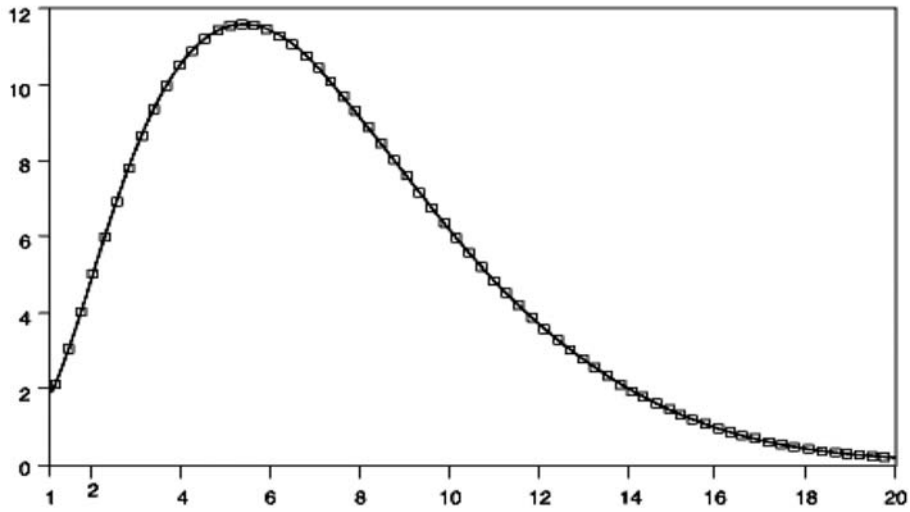


Figure 3. Vertical axis gives brightness (10^{-4} erg/sec/cm^2/steradian/cm^{-1}). Horizontal axis gives Frequency (cycles/cm) [23].

Soon after observations of the CMB made by Penzias and Wilson, a question arose as to what are the seeds of structure in the CMB that gave rise to the formation of galaxies later by gravitational collapse. A major support to the Big Bang Model came after the detections of predicted fluctuations in the CMB of about 1 part in 100,000 by the COBE data, as described by Smoot *et al.* [24]. The WMAP observations described by Bennett *et al.* have a much higher resolution of ~0.3 degrees that show detailed structure up to sub-horizon scale as shown in Figure 4 [25] (see p. 358).

5.3. Standard Cosmological Model

Evidence for the occurrence of dark matter in the Universe was first postulated by Zwicky in the 1930s by measuring velocities of galaxies in galaxy clusters. Its occurrence became established from the Flat Rotation curves that were observed optically by Rubin and Ford in 1970 [26], and by HI observations made by Robert and Whitehurst in 1975 [27] and Bosma in 1978 [28]. The presence of dark matter in the Universe is supported independently from observations of gravitational lensing of distant sources by the intervening clusters. Further, detailed observations of the velocities of various galaxies in clusters have provided an estimate of the dark matter density.

Based on observations of distant supernovae, a surprising conclusion was made about a decade ago, by Riess *et al*. [28] and Perlmutter *et al*. [29] that the Universe is accelerating, indicating the presence of dark energy (with negative pressure). From a detailed analysis of the WMAP observations, also using some of the earlier measurements, Spergel *et al*. [31] concluded that the Universe is dominated by 74% dark energy, 22% dark matter, and 4% ordinary matter as observed in stars and galaxies. The WMAP data provides accurate estimates to $\sim \pm 10\%$ of several fundamental parameters, such as H_0, Ω_0 etc. of the cosmological model. Thus, it has been suggested that we have now entered an era of 'Precision Cosmology'. The recently launched Planck satellite will provide more precise measurements of the CMB concerning the cosmological model. Polarization measurements of CMB may test predictions of inflation.

6. The 21 cm Emission Line of Neutral Hydrogen (HI)

After reading the 1940 paper by Reber giving the radio map of the Galaxy, Professor Jan Oort wondered whether there were any spectral lines in the radio window (he got a copy of the paper even though the Netherlands was occupied during the World War II). His student, van de Hulst, made a path-breaking calculation in 1944 that the two closely-spaced hyperfine energy levels in the ground state of the neutral hydrogen atom (HI) emit radiation at a frequency of 1420.4058 MHz. Emission of the predicted HI line was observed in 1951 towards the Galaxy by Ewen and Purcell [32], and was soon confirmed by Australian and Dutch astronomers. Subsequently, its high-resolution observations delineated the spiral structure of the Galaxy for the first time. It may be noted that optical observations of the spiral structure are limited to close distances only, due to the presence of dust in the Galaxy. HI observations have now been carried out for a large

number of spiral galaxies and dwarf galaxies. HI observations also provide information about the formation and evolution of galaxies, as hydrogen is the basic ingredient of stars and galaxies that are formed by its gravitational collapse. All the heavier elements are formed at the centre of stars by fusion. Their subsequent collapse leads to supernova and also interstellar clouds where molecules are found.

Observations of HI from very distant galaxies provide important clues about the evolution of galaxies. The most distant galaxy detected so far has a redshift of 0.18. By co-adding HI emission of a large number of galaxies in a cluster with measured redshifts, average content of HI in distant galaxies up to z ~0.8 has been estimated. With the Square Kilometer Array (SKA), it would be possible to measure HI of individual galaxies up to z ~3. HI absorption studies towards distant quasars have also been very fruitful. An important area of research is to search for the epoch of reionization of HI that is expected to occur prior to the redshift of about 6, soon after formation of first stars and galaxies. LOFAR, MWA and SKA described in Section 9 may determine this epoch.

7. Pulsars

In 1967 Antony Hewish and his student Jocelyn Bell were observing interplanetary scintillations of compact components of radio sources at Cambridge in the UK using a recently constructed array for that purpose. They discovered serendipitously that pulsed radio emission with a highly accurate periodicity occurred in the direction 1919+21 (right ascension and declination) [33]. Soon a few other pulsed radio sources were found and were called Pulsating Radio sources (Pulsars). The Nobel Prize was given to Anthony Hewish for the discovery of Pulsars in 1974; also to Martin Ryle for developing innovating radio techniques and for studies of radio galaxies, as described in Sections 3 and 9.

Due to their highly accurate periodicities, Gold [34] suggested that pulsars are associated with neutron Stars, being the end product of stars when their nuclear fuel runs out (giving rise to a supernova remnant and a neutron Star). Due to the collapse of the parent stars and conservation of their angular momentum, neutron stars start spinning at a fraction of second. Their magnetic field becomes tens of billions of Gauss, resulting in coherent beamed radiation in the direction of their magnetic poles. If the rotation and magnetic axes are misaligned, periodic pulsed emission is observed on each rotation, analogous to that observed from a lighthouse. About 1700 pulsars have been catalogued so far. Many pulsars with millisec (ms) peri-

odicities have also been discovered. Their precise periods correspond to highly accurate clocks, matching or exceeding atomic clocks, and therefore provide important tests of the General Theory of Relativity. A set of millisecond pulsars may allow detection of the primordial gravitational radiation, which is predicted by the inflationary model. An emission mechanism of Pulsars was suggested by Goldreich and Julian in 1969 soon after their discovery [35], but it has not been able to explain many observations. Many attempts are being made to find a satisfactory emission mechanism.

The first Binary Pulsar, 1913+16, was discovered by Hulse and Taylor in 1975. According to the General Relativity theory, a binary star system should emit gravitational waves. The loss of orbital energy results in shrinkage of its orbit. From accurate timing of the pulse period of the binary pulsar 1913+16, Weisberg and Taylor concluded in 1983 that the orbit of the pulsar is shrinking due to the gravitational radiation as predicted by the Einstein's General Theory of Relativity [36]. Recent observations have confirmed that its orbit continues to shrink as expected. The Nobel Prize was given to Hulse and Taylor in 1993 for this discovery.

8. Molecules and Megamasers

In 1963 Weinreb *et al.* discovered absorption lines of the interstellar molecule OH at 1665 and 1667 MHz [37]. The NH_3 molecule was found in the interstellar medium by Townes and collaborators in 1968. Over 140 molecules have been discovered till now in interstellar and circumstellar clouds in the Galaxy, including some with 13 atoms. The molecule CO has also been detected in nearby galaxies and in faraway galaxies including the distant quasar at $z = 6.42$. A recent finding by Iglesias-Groth *et al.* (MNRAS in Press) suggests that some of the key components in terrestrial prebiotic chemistry are present in the interstellar matter.

The OH maser in emission was discovered towards HI regions in the Galaxy in 1965. The first OH megamaser was observed in 1982 towards the nearest ultra-luminous infrared galaxy, Arp 220 [38]. Radiation occurs close to galactic nuclei. Their measurements have provided mass of the central Black Hole. Due to their powerful emission, OH megamasers have now been observed in many faraway galaxies. Megamasers of water (H_2O), formaldehyde (H_2CO) and methane (CH) have also been observed. Recently, Kanekar, Chengalur and Ghosh have made accurate measurements of the OH transition lines in galaxies at higher redshifts to investigate whether Fundamental Constants, such as the fine structure constants, change with cosmic epoch. Their results are suggestive but require further investigations.

9. Radio Telescopes

I list here only some of the major radio telescopes in order to indicate the type of instruments that have allowed frontline research in radio astronomy.

9.1. Synthesis Radio Telescopes

Since the wavelength of radio waves is quite large, it becomes necessary to use radio interferometers with large spacing in order to map radio sources with adequate resolution. In the 1940s Australian radio astronomers led by Pawsey used firstly a sea interferometer and later spaced interferometers. It was recognized that a pair of interferometer measures one Fourier component of a radio source and by using many spacings its map can be obtained by inverse Fourier transform. In the 1950 Christiansen built a grating interferometer for mapping the Sun. Mills and Little built a cross-type antenna, known as Mills Cross. Ryle and colleagues also started to build radio interferometers soon after 1946. During the late 1950s the Cambridge group, led by Ryle, built several sensitive interferometers and later employed for the first time the principle of earth's rotation synthesis radio interferometer for carrying out radio surveys [39] (Nobel Prize 1974). This technique has been exploited by a number of radio telescopes built in the world over the last 6 decades. Currently, the prominent radio telescopes using this principle are: (1) Westerbork Synthesis Radio Telescope in the Netherlands in 1971; (2) Very Large Array in USA in 1980; (3) Australia Telescope Array in 1990 and (4) Giant Metrewave Radio Telescope in India in 1999.

Several major synthesis-type radio telescopes with large sensitivity, high resolution and wide frequency coverage are being built now using new technologies, particularly, ALMA in Chile, LOFAR in Europe, LWA in USA, MeerKat in South Africa and ASKAP in Australia. MWA being developed by MIT, RRI and others is also likely to be located in Australia. Most ambitious is the international SKA radio telescope likely to be built by 2020 with contributions by more than 17 countries. SKA will provide extraordinary capability, with a collecting area up to 1 million sq. km, baselines of several hundred km and a wide frequency range.

9.2. Large aperture single radio telescopes

Single radio telescopes with a large collecting area allow special observations, such as spectral line observations, pulsar research etc. A major breakthrough was the construction of a 76 m diameter parabolic dish in 1957 conceived by Bernard Lovell. A 64 m diameter parabolic dish became operational at Parkes in 1961. A 100 m diameter dish was built at Bonn in 1972. A 100 m diameter dish with 2000 servo controlled panels to allow

operation at mm wavelengths was built at Green Bank in the USA in 2002. The Arecibo Telescope built in 1971 has the largest collecting area amongst single aperture radio telescopes. It consists of part of a sphere of 300 m diameter fixed on the ground, with a steerable dish of about 25 m diameter placed near its centre allowing pointing in different directions over ~ ± 20 degrees of the zenith. A similar radio telescope of 500 m diameter, called FAST, is under construction in China.

10. Some Key Questions Today

I list below five key questions that are amongst the major objectives of the proposed SKA. The document 'Science with the Square Kilometer Array' gives extensive details of a wide number of astronomical objectives of the SKA [40].

Q.1: What is ultra-strong field limit of relativistic gravity?

Increased sensitivity of radio telescopes will discover many more pulsars and may find a pulsar in orbit around a Black Hole and near the galactic centre. Accurate timing of a large number of milli-second pulsars may detect primordial gravitational radiation as has been predicted to occur during very rapid inflation of the Universe in the Big Bang Cosmology.

Q.2: What is origin and evolution of the Cosmic Magnetism?

Radio Astronomy is uniquely placed to determine the evolution of a magnetic field from early times to now, through studies of Faraday rotation of polarization of synchrotron radiation in distant radio galaxies, and also observations of Zeeman splitting. As an example, the radio map of the nearby galaxy M51 at a wavelength of 6 cm show distribution of a large-scale magnetic field along the spiral arms. Observations of Faraday rotation of synchrotron emission of radio halos in distant clusters at high redshifts could become possible with SKA and thus may give a clue about the origin and evolution of cosmic magnetism.

Q.3: What are the processes about the formation and evolution of galaxies in the Universe?

The measurements of the unique 21 cm (1420 MHz) radiation of neutral hydrogen (HI) from a large number of galaxies up to large distances would provide important information about the formation and evolution of galaxies and of cosmology.

Q.4: When did the first stars form and neutral hydrogen get reionized?

Theoretical predictions, computer simulations and WMAP measure-

ments indicate that the first stars and galaxies collapsed gravitationally from the primordial neutral hydrogen (HI) at redshifts of about 30. Later, neutral HI got ionized by UV by about redshift of ~6. Details of the epoch of Reionization are of great importance for studies of structure formation in the Universe, requiring measurements of emission and absorption of HI in the frequency range of about 50 to 200 MHz.

Q.5: Is there intelligent life search elsewhere in our Galaxy?

The proposed Square Kilometer Array (SKA) will allow studies of radio emission from extra-solar planets and Search for Extraterrestrial Intelligence (SETI) towards millions of stars.

10. Conclusion

Radio astronomy observations have revealed occurrence of truly violent phenomena in the Universe such as those occurring in radio galaxies and quasars, indicating the presence of supermassive Black Holes in the nuclei of the galaxies. One of the greatest discoveries of the last century was that of Cosmic Microwave Background Radiation which provided support to Big Bang cosmology. Its detailed observations combined with other astronomical data have indicated that visible matter is only 0.04%, dark matter 0.26% and dark energy 0.7% of matter density in the Universe. Discovery of line emission of neutral hydrogen (HI) allows investigations of the formation and evolution of galaxies and their dynamics. Observations of a binary Pulsar by Hulse and Taylor have provided evidence of gravitational radiation predicted by the General Theory of Relativity. Over 140 molecules have been discovered in the interstellar medium of our Galaxy. These molecules are ingredients of life on earth and raise the question of whether life exists elsewhere. There are many major questions today for which the ambitious SKA project that is likely to materialize by 2020 may provide a clue.

Over the last few decades, additionally to the radio window, sensitive observations made at X-Rays, UV, optical and infrared parts of the electromagnetic spectrum have provided important information about the physical processes in stars and galaxies. This multi-wavelength astronomical research may give us further insight into the mysteries of the Universe.

References

1. Jansky K., *Proc. Inst. Rad. Eng.*, 21, 1387, 1933.
2. Reber, G., *Astrophys. J.*, 91, 621, 1940.
3. Hey, J.S., *Nature*, 157, 47, 1946.
4. Southworth, G.C., *J. Franklin Inst.*, 239, 285, 1945.
5. Reber, G., *Astrophys. J.*, 100, 279, 1944.
6. Pawsey, J.L., Payne-Scott, R., and McCready, L.L., *Nature*, 157, 158, 1946.
7. Wild, P., *Australian J. Sci. Res.*, A4, 36, 1951.
8. Christiansen, W.N., and Warburton, J.A., *Aust. J. Phys.*, 5, 262, 1953.
9. Bastian, T.S., Benz, A.O., Gary, D.E., *Annual Review of Astronomy and Astrophysics*, 36: pp. 131-188.
10. Manoharan, P.K., *Solar Physics*, 265, 137, 2010.
11. Hey, J.S., Parsons, S.J. and Phillips, J.W., *Nature*, 158, 234, 1946.
12. Bolton, J., and Stanley, G.J., *Nature*, 162, 312, 1948.
13. Smith, F.G., *Nature*, 168, 962, 1951.
14. Baade, W. and Minkowski, R., *Astrophys. J.*, 119, 206, 1954.
15. Jennison, R.C., and M.K. Das Gupta, *Nature*, 172, 996, 1953.
16. Bridle *et al.*, *Astron. J.*, 108, 766, 1994.
17. Schmidt, M., *Nature*, 197: 1040, 1963.
18. Bondi, H. and Gold, T., *Mon. Not. R. Astron. Soc.,* 108, 252, 1948; Hoyle, F., *Mon. Not. R. Astron. Soc.*, 108, 372, 1948.
19. Ryle, M., and Clarke, R.W., 122, 349, 1961.
20. Swarup, G., *Mon. Not. R. Astron. Soc.* 172, 501, 1975.
21. Kapahi, V.K., *Mon. Not. R. Astron. Soc.*, 172, 513, 1975.
22. Penzias, A.A., and R.W. Wilson, *Astrophys. J.*, 142, 419, 1965.
23. Mather, J.C., *et al.*, *Astrophys. J. Lett.* 354, L37, 1990; *Astrophys. J.*, 420, 439, 1994.
24. Smoot, G.F., *et al.*, *Astrophys. J. Lett.*, 396, L1, 1992.
25. Bennet, C.L., *et al.*, *Astrophys. J. Suppl.*, 148, 1, 2003.
26. Rubin, Vera C., and Ford, W.K., *Astrophys. J.*, 159, 379, 1970.
27. Roberts, M.S., and Whitehurst, R.N., *Astrophys. J.*, 201, 327, 1975.
28. Bosma, A., *The distribution and kinematics of neutral hydrogen in spiral galaxies of various types*, Ph.D. thesis, University of Groningen, 1978.
29. Riess, A.G., *et al.* 1998, *Astron. J.*, 116, 1009, 1998.
30. Perlmutter, S., *et al.*, *Astrophys. J.*, 517, 565, 1999.
31. Spergel, D.N., *et al.*, *Astrophys. J. Suppl.*, 148, 175, 2003.
32. Ewen, H.I. and Purcell, E.M., *Nature,* 168, 356, 1951.
33. Hewish, A., *et al.*, *Nature*, 217, 709, 1968.
34. Gold, T., *Nature*, 218, 731, 1968.
35. Goldreich, P. and Julian, W.H., *Astrophys. J.*, 157, 869, 1969.
36. Taylor, R.A. and Weisberg, J.M., *Astrophys. J.*, 345, 434, 1989.
37. Weinreb, S., *et al.*, *Nature*, 200, 829, 1963 [38].
38. Baan, W.A., Wood, P.A.D., Haschick, A.D., *Astrophys. J.*, 260, L49, 1982.
39. Ryle, M., and Neville, A.C., *Mon. Not. R. Astron. Soc.*, 125, 39, 1962.
40. Carrili, C., and Rawlings, S., *New Astronomy Reviews*, 48, pp. 979-1563, Dec. 2004.

Session II: Physics

My Personal Experience on the Scientific Legacy Of The 20th Century

Antonino Zichichi

1. Introductory Remarks Concerning the 'Convictions Spread by Modern Culture'

Let me, first of all, express my gratitude to our Chancellor, Monsignor Marcelo Sánchez Sorondo, Professors Werner Arber and Jürgen Mittelstrass, for having organised this extremely interesting and 'up-to-date' series of plenary sessions of our Academy, dedicated to the *Scientific Legacy of the 20th Century*.

1st point. The Scientific Legacy of the 20th Century cannot be independent from and must be coupled with the Culture of our Time [1].

2nd point. This Culture is defined as being 'modern' but in fact it is pre-Aristotelic [2]. Proof: neither Rigorous Logic nor Science are part of the Culture of our Time.

Let me recall a statement by H.H. Benedict XVI, concerning the Culture of our Time. The Pope has pointed out that it is necessary to speak about the *elements that challenge the convictions spread by Modern Culture*. The most important of these 'convictions' is the link between Science and Faith. Here comes my second point: namely the fact that, in the Culture of our Time, Rigorous Logic and Science are absent. It is generally believed that the reason why people have Faith is because the great public knows little, very little, about Rigorous Logic and Science.

Modern Culture maintains that if people knew more about Mathematics and Physics which, according to Enrico Fermi, is the fulcrum of all sciences, people would realise that Faith has nothing to do with either Logic or Science and that Faith is in contradiction with the great achievements of Mathematics and Physics. A widespread conviction of Modern Culture is that Atheism is the result of the great achievements in mathematical rigour and in Physics. If our so-called Modern Culture were consistent in its reasoning, it would have to recognise the fact that a rigorous analysis of what Atheism is all about shows that Atheism is an act of Faith about nothing (see Appendix 1).

Here comes my 'personal experience', based on what I have done in Physics. The result is that what I have done is perfectly consistent with all other achievements in the fundamental search for the existence of the 'Logic of Nature'. This is what we have been doing since Galileo Galilei, the father of the 1st Level of Science (the three levels of Science are discussed in Ap-

pendix 2). The results obtained in 1st Level Science show that these results were obtained in a totally unexpected way, i.e.: no one had been able to predict these discoveries. The list of these discoveries is impressive as I have already reported in previous lectures ([3], see also Appendix 5.3). We call these achievements UEECs, which stands for **U**nexpected **E**vents with **E**normous **C**onsequences. What I have done further confirms the existence of UEEC phenomena, which started to be discovered by the father of the 1st Level of Science, Galileo Galilei. Let me show a synthesis of achievements in Physics from Galilei to the first half of the 20th Century (Figures 1 and 2).

"UEEC" TOTALLY UNEXPECTED DISCOVERIES FROM GALILEI TO FERMI-DIRAC, THE "STRANGE" PARTICLES AND THE YUKAWA GOLDMINE	
I	Galileo Galilei: $F = mg$.
II	Newton: $F = G\frac{m_1 \cdot m_2}{R_{12}^2}$
III	Maxwell: the unification of electricity, magnetism and optical phenomena, which allows to conclude that light is a vibration of the EM field.
IV	Becquerell: radioactivity.
V	Planck: $h \neq 0$. The quantum nature of the World.
VI	Lorentz: Space and Time cannot both be real.
VII	Einstein: the existence of time-like and space-like worlds. Only in the time-like world, simultaneity does not change, with changing observer.
VIII	Einstein: the photon.
IX	Weyl: Gauge Invariance.
X	Bohr: the structure of the atom.
XI	de Broglie: wave nature of particles.
XII	Schrödinger: wave function, and its probabilistic interpretation (Born).
XIII	Rutherford: the nucleus.
XIV	Hess: cosmic rays.
XV	Einstein: the Space-Time curvature.
XVI	Von Neumann: the proof that Quantum Mechanics is self consistent (no contradictions).
XVII	Pauli: the Exclusion Principle.
XVIII	Heisenberg: the Uncertainty Principle.

Figure 1.

XIX	Dirac discovers his equation, which opens new horizons, including the existence of the antiworld.
XX	Chadwick: the neutron.
XXI	Wigner: Time Reversal Invariance (T).
XXII	Majorana: relativistic invariance allows not only spin ½, as it is the case for the electron, but any spin value.
XXIII	Majorana: uncharged particles with spin ½ identical to their antiparticles are allowed by relativistic invariance. These particles are now called "Majorana fermions".
XXIV	Fermi–Dirac and Bose–Einstein discover two completely different statistical laws.
XXV	Other Invariance Laws: Charge conjugation (Weyl and Dirac); Parity (Wigner); CPT (Pauli).
XXVI	The neutrino (Pauli, Fermi).
XXVII	Fermi: weak forces.
XXVIII	The Stars are "nuclear-fusion" candles (Fermi, Bethe).
XXIX	Von Neumann: electronic computing.
XXX	The sequence of unexpected Fermi discoveries: Fermi-coupling, Fermi-gas, Fermi-momentum, Fermi-temperature, Fermi-surface, Fermi-transition, Fermi-length (plus the other three quoted above: XXIV, XXVI, XXVII).
XXXI	The "strange particles" are discovered in the Blackett Lab.
XXXII	The Yukawa goldmine. Let me devote some attention to the discussion of UEEC events in nuclear physics (i.e., The Yukawa Goldmine). *Nuclear Physics and UEEC events*. It is considered standard wisdom that nuclear physics is based on perfectly sound theoretical predictions. People forget the impressive series of UEEC events discovered in what I have decided to call the "Yukawa goldmine" [4]. Let me quote just three of them: 1 The first experimental evidence for a cosmic ray particle believed to be the Yukawa meson was a lepton: the muon. 2 The decay-chain: $\pi \rightarrow \mu \rightarrow e$ was found to break the symmetry laws of Parity and Charge Conjugation. 3 The intrinsic structure of the Yukawa particle was found to be governed by a new fundamental force of Nature, Quantum ChromoDynamics: QCD. As you know 2007 was the centenary of the birth of Hideki Yukawa, the father of theoretical nuclear physics [4]. In 1935 the existence of a particle, with mass intermediate (this is the origin of "mesotron" now "meson") between the light electron, m_e, and the heavy nucleon (proton or neutron), m_N, was proposed by Yukawa [5]. This intermediate mass value was deduced by Yukawa from the range of the nuclear forces. Contrary to the general wisdom of the time, Yukawa was convinced that the particles known (electrons, protons, neutrons and photons), could not explain how protons and neutrons are bound into the extremely small dimensions of a nucleus.
XXXIII	The "Majorana fermions" give rise to a sequence of unexpected discoveries not only in the grand unification of all fundamental forces but also in the physics of condensed matter, such as: Majorana spin-flip and ultra-low T physics, topological insulators, Majorana liquids and fermion fractionalization, Majorana fermions in tunable semiconductors, Majorana fermions and topological phase transitions.

Figure 2.

I have included the invention of electronic computers by Von Neumann (XXIX), which no one could have imagined at the beginning of the 20th Century. Point no. XXX refers to the impressive list of Fermi discoveries: once again, all totally unexpected.

THE SECOND HALF OF THE 20TH CENTURY	
XXXIV	The Subnuclear World.
XXXV	The Standard Model and Beyond.
XXXVI	The Superworld.

Figure 3.

The UEECs of the second half of the 20th Century (Figure 3) are grouped into 3 classes:

- one is the 'Subnuclear World'
- the second is the 'Standard Model and Beyond'
- the third is the 'Superworld'.

The existence of the Subnuclear World and the Standard Model are strictly correlated. The third is the frontier of our knowledge which exists as a fascinating mathematical structure, but lacks Galilean experimental proof (Appendix 3).

The reason why no one is able to predict what is discovered in fundamental scientific research is inherent in the fact that the Author of the Logic which governs the world, from its most elementary structures to the frontier of the cosmos, is smarter than us all: philosophers, thinkers, mathematicians, physicists, artistic leaders, musicians, no one excluded.

The Author of the Logic of Nature being smarter than us all, the only way to learn more about the Fundamental Logic is to perform experiments. The most advanced experiment in the frontier of our Physics is, today, the Quark-Gluon-Coloured-World (QGCW) [6] project whose purpose is to understand how the world was one-tenth of a nanosecond (10^{-10} sec.) after the Big Bang. No philosopher, no mathematician, no physicist can tell us if, at that moment, the world was as we think it could have been, i.e.: obeying the Supersymmetry Law which establishes that Fermions and Bosons must be exactly equivalent, i.e.:

$$F \equiv B .$$

This supersymmetry law generates the Superworld. Details about the reasons why the Superworld is needed are in Appendix 3.4. From the Superworld, after 20 billion years, here we are with the world in 4 dimensions (3 for space, one for time, Figure 15a), while the Superworld has 43 dimensions (Figure 15b). These two Figures are on page 21. Where the ashes of the Superworld might be is in Appendix 3.5. The point I want to emphasize is that no one can tell us what will be discovered at CERN with the LHC, the Large Hadron Collider, the world's most powerful collider, which will recreate the conditions the world was in at $\Delta t = 10^{-10}$ sec. after the Big Bang. No one can tell us if the Superworld was there at that time. Only the experimental results will allow us to know if the reasons why the Superworld is needed are correct and the corresponding mathematics do belong to the Logic of Nature that we are trying to decipher.

After these long introductory remarks, I will now devote the last part of this lecture to my activity, which is my contribution to the confirmation that UEEC phenomena exist and represent the proof that the Author of the Logic of Nature is smarter than us all. Here is my personal experience.

2. My Scientific Testimony

A few examples I have been involved in are reported in Figure 4.

① **The 3rd lepton, HL** (now called τ) with its own neutrino, ν_{HL} (now called ν_τ),
despite *the abundance of neutrinos:* ν_e *and* ν_μ.

② **Antimatter**
despite *S-matrix and C, P, CP, T breakings.*

③ **Nucleon Time-like** EM structure
despite *S-matrix*

④ **No quarks** in violent (pp) collisions
despite *scaling.*

⑤ **Meson mixings**
$\theta_V \neq \theta_{PS}$ **:** $(51^o) \neq (10^o) \neq 0$ ***despite*** $SU(3)_{uds}$.

⑥ **Effective energy:** the Gribov QCD-light
despite *QCD-confinement.*

⑦ **The running** of $\alpha_1\ \alpha_2\ \alpha_3$ versus **energy**:
the EGM effect, the GAP between E_{GUT} *and* E_{SU}*, and the absence of the Platonic straight line convergence.*

Figure 4.

I will only discuss four points: 1, 2, 6 and 7.

Point 1

The Third Lepton, and the other unexpected events in Electroweak Interactions are illustrated in Figure 5.

Note that for the Electroweak force, Nature has not chosen the simplest way out SU(2), but unexpectedly SU(2)×U(1).

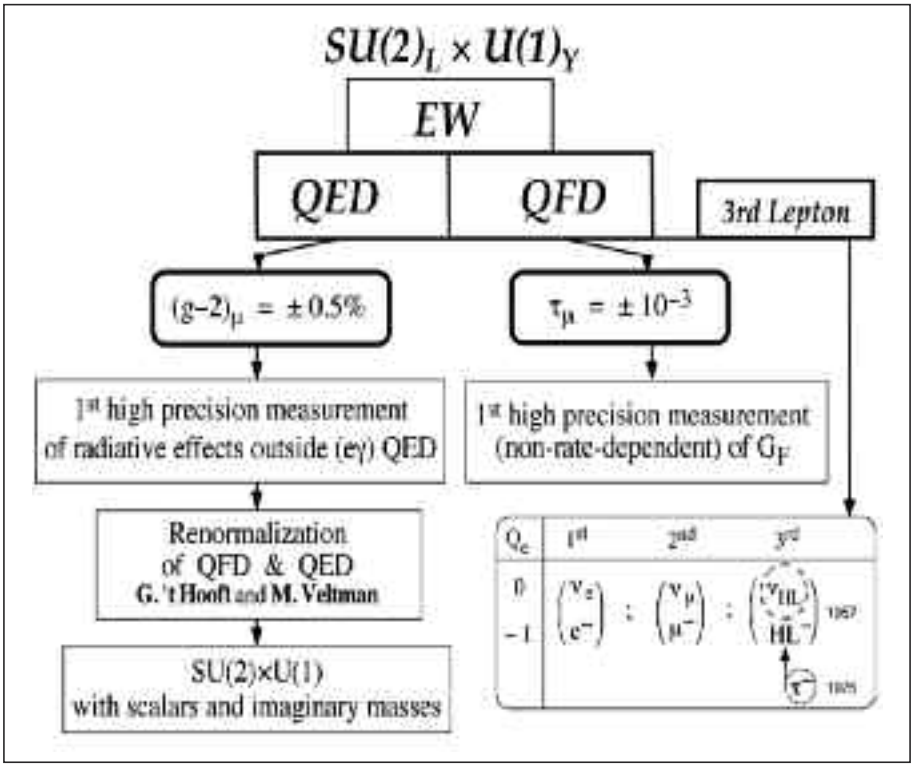


Figure 5.

Point 2

The problem of understanding the difference between mass and matter is illustrated in Figure 6. The incredible series of events which originated with the problem of understanding the stability of matter is shown in Figure 7, together with the unexpected violation of the Symmetry Operators (C, P, T, CP) and the discovery of Matter-Antimatter Symmetry.

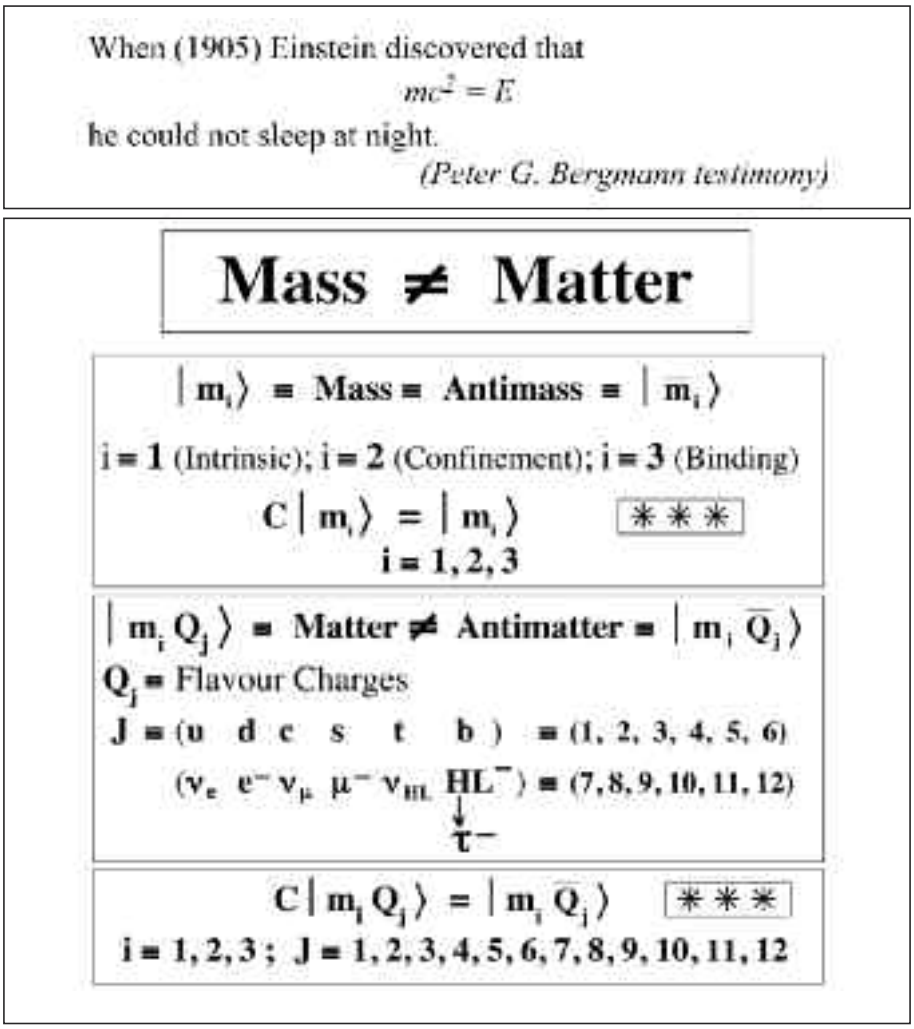


Figure 6.

Figure 7 shows seven decades of developments, which started from the antielectron and C-invariance and brought us to the discovery of nuclear antimatter and to the unification of all gauge forces with a series of unexpected discoveries.

THE INCREDIBLE STORY
TO UNDERSTAND THE ORIGIN OF THE STABILITY OF MATTER
SEVEN DECADES FROM THE ANTIELECTRON TO ANTIMATTER
AND THE UNIFICATION OF ALL GAUGE FORCES

- **The validity of C invariance from 1927 to 1957.**

After the discovery by Thomson in 1897 of the first example of an elementary particle, the Electron, it took the genius of Dirac to theoretically discover the Antielectron thirty years after Thomson.

1927 → Dirac equation [7]; the existence of the antielectron is, soon after, theoretically predicted. Only a few years were needed, after Dirac's theoretical discovery, to experimentally confirm (Anderson, Blackett and Occhialini [8]) the existence of the Dirac antielectron.

1930-1957 → **Discovery of the C operator** [(charge conjugation) H. Weyl and P.A.M. Dirac [9]]; discovery of the P Symmetry Operator [E.P. Wigner, G.C. Wick and A.S. Wightman [10, 11]]; discovery of the T operator (time reversal) [E.P. Wigner, J. Schwinger and J.S. Bell [12, 13, 14, 15]]; discovery of the CPT Symmetry Operator from RQFT (1955-57) [16].

1927-1957 → Validity of C invariance: e^+ [8]; [17]; [18]; $K^0_L \to 3\pi$ [19] but see LOY [20].

- **The new era starts: C ≠ ; P ≠ ; CP ≠** (*) .

1956 → Lee & Yang P ≠ ; C ≠ [21].

1957 → Before the experimental discovery of P ≠ & C ≠, Lee, Oehme, Yang (LOY) [20] point out that the existence of the second neutral K-meson, $K^0_L \to 3\pi$, is proof neither of C invariance nor of CP invariance. Flavour antiflavour mixing does not imply CP invariance.

1957 → C.S. Wu et al. P ≠ ; C ≠ [22]; CP ok [23].

1964 → $K^0_L \to 2\pi \equiv K_L$: CP ≠ [24].

1947-1967 → QED divergences & Landau poles.

1950-1970 → The crisis of RQFT & the triumph of S-matrix theory (i.e. the negation of RQFT).

1965 → Nuclear antimatter is (experimentally) discovered [25]. See also [26].

1968 → The discovery [27] at SLAC of Scaling (free quarks inside a nucleon at very high q^2) but in violent (pp) collisions no free quarks at the ISR are experimentally found [28]. Theorists consider Scaling as being evidence for RQFT not to be able to describe the Physics of Strong Interactions. The only exception is G. 't Hooft who discovered in 1971 that the β-function has negative sign for non-Abelian theories [29].

1971-1973 → β = – ; 't Hooft; Politzer; Gross & Wilczek. The discovery of **non-Abelian** gauge theories. Asymptotic freedom in the interaction between quarks and gluons [29].

1974 → All gauge couplings $\alpha_1\ \alpha_2\ \alpha_3$ run with q^2 but they do not converge towards a unique point.

1979 → A.P. & A.Z. point out that the new degree of freedom due to SUSY allows the three couplings $\alpha_1\ \alpha_2\ \alpha_3$, **to converge towards a unique point** [30].

1980 → QCD has a "hidden" side: the multitude of final states for each pair of interacting particles: (e^+e^- ; p ; πp; Kp; νp; pp; etc.)
The introduction of the Effective Energy allows to discover the Universality properties [31] in the multihadronic final states.

1992 → All gauge couplings converge towards a unique point at the gauge unification energy: $E_{GU} \cong 10^{16}$ GeV with $\alpha_{GU} \cong 1/24$ [32, 33] .

1994 → The Gap [34] between E_{GU} & the String Unification Energy: $E_{SU} \cong E_{Planck}$.

1995 → **CPT loses its foundations at the Planck scale (T.D. Lee)** [35].

1995-1999 → **No CPT theorem from M-theory (B. Greene)** [36].

1995-2000 → A.Z. points out the need for new experiments to establish if matter-antimatter symmetry or asymmetry are at work.

(*) The symbol ≠ stands for "Symmetry Breakdown".

Figure 7.

Point 6

The non-Abelian nature of the Interaction describing quarks, gluons and the Effective Energy with the set of unexpected discoveries is illustrated in Figure 8.

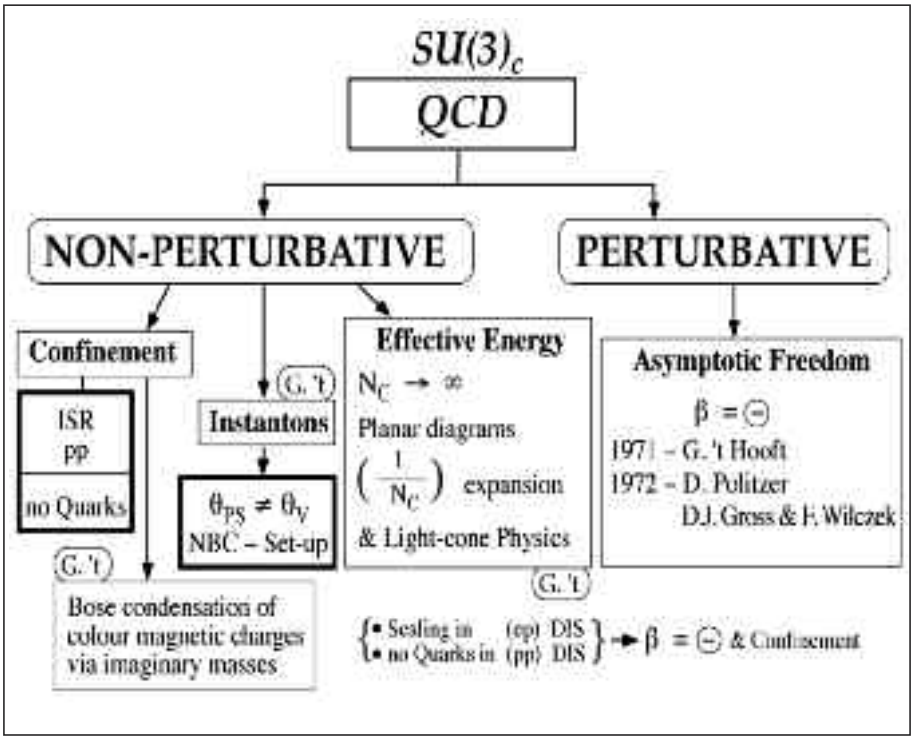


Figure 8.

Point 7

The Unification of all Forces and the Supersymmetry threshold with its problems are reported in Figures 9 and 10 (see pp. 359-360) respectively.

Figure 10 illustrates the EGM effect which lowers by a factor 700 the threshold for the production of the lightest superparticle.

The mathematical formalism used to obtain the results shown in Figures 9 and 10 is a system of three differential non-linear equations (shown in Figure 11) describing how the gauge couplings

$$\alpha_i, \alpha_j \text{ (with } i = 1, 2, 3; \text{ and } J = 1, 2, 3 \text{ but } i \neq j),$$

vary with 'μ', the basic parameter which depends on the energy of a given elementary process.

THE UNIFICATION OF ALL FUNDAMENTAL FORCES

The lines in Figure 6 result from calculations executed with a supercomputer using the following system of equations:

$$\mu \frac{d\alpha_i}{d\mu} = \frac{b_i}{2\pi} \alpha_i^2 + \sum_j \frac{b_{ij}}{8\pi^2} \alpha_i \alpha_j$$

This is a system of coupled non-linear differential equations where the existence of the Superworld is taken for granted. This system describes how the gauge couplings (α_1, α_2, α_3) vary with "μ", the basic parameter which depends on the energy of the elementary process, from the maximum level of Energy (Planck Scale) to the energy level of our world.

Figure 11.

During more than ten years (from 1979 to 1991), no one had realized that the energy threshold for the existence of the Superworld was strongly dependent on the 'running' of the masses.

This is now called: the EGM effect (from the initials of Evolution of Gaugino Masses). To compute the energy threshold using only the 'running' of the gauge couplings ($\alpha_1, \alpha_2, \alpha_3$) corresponds to neglecting nearly three orders of magnitude in the energy threshold for the discovery of the first particle (the lightest) of the Superworld [33], as illustrated in Figure 10.

A different way to describe how the gauge couplings $\alpha_1, \alpha_2, \alpha_3$ vary with energy is reported in Figure 12 (see p. 361). The simplest way to get GUT (the point where all fundamental forces are together: Grand Unification Theory) would be the straight line. But the real world does not follow this 'platonic' straight line. The sequence of points (the big red points), in steps of 100 GeV, is very different from the Platonic line (dotted blue points). The way nature goes is reported by the sequence of the big red points which are the result of the mathematics reported in Figure 11.

3. Where we are in Understanding the Logic of Nature

My scientific testimony, synthetically discussed in the previous paragraphs, is a contribution to where we are now in understanding the Logic of Nature. This is illustrated in Figures 13-17 and 18 (see p. 362).

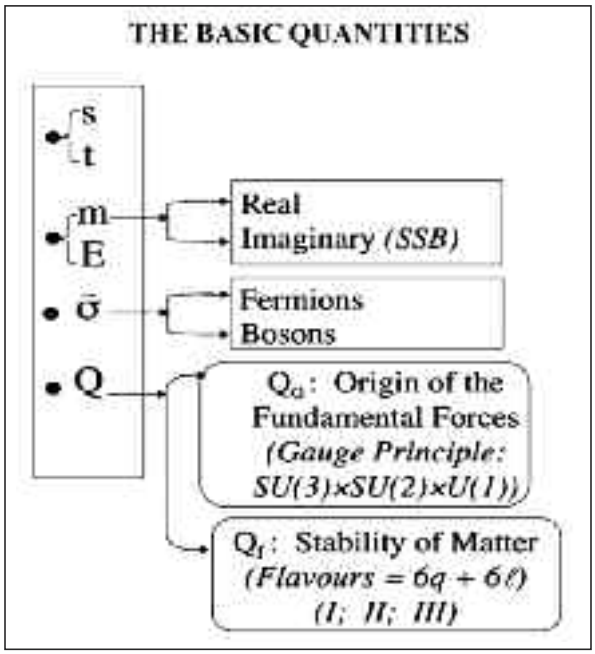


Figure 13.

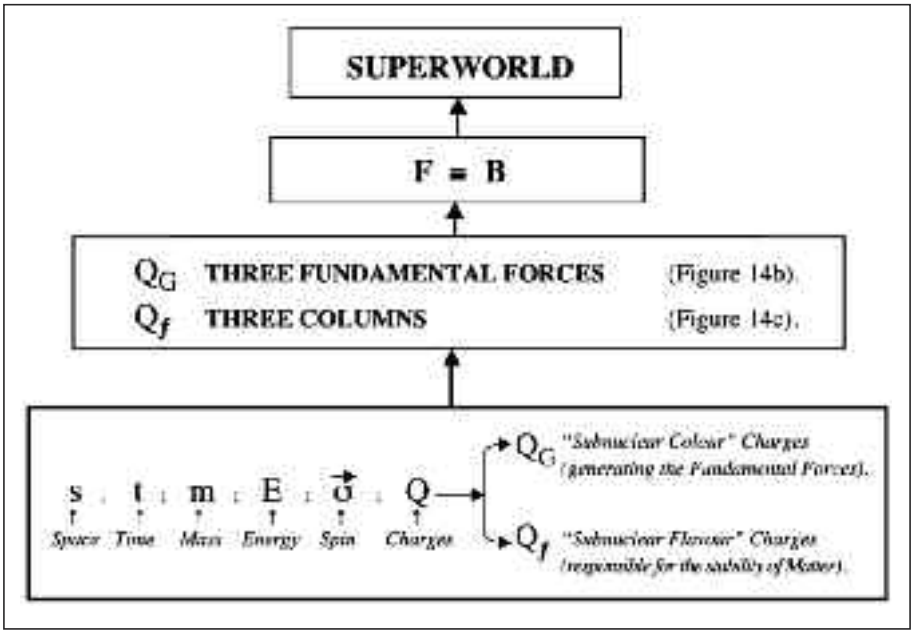


Figure 14a.

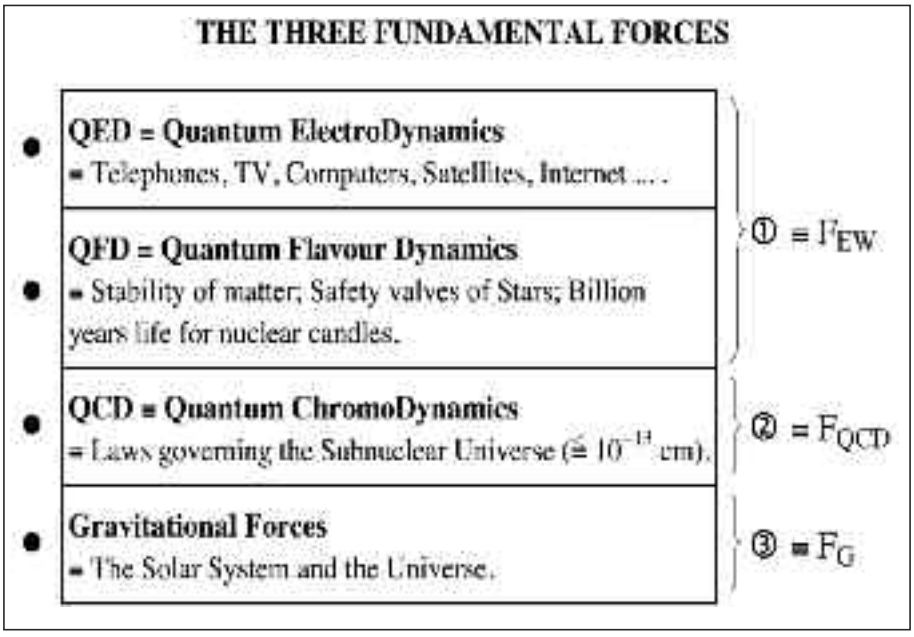


Figure 14b.

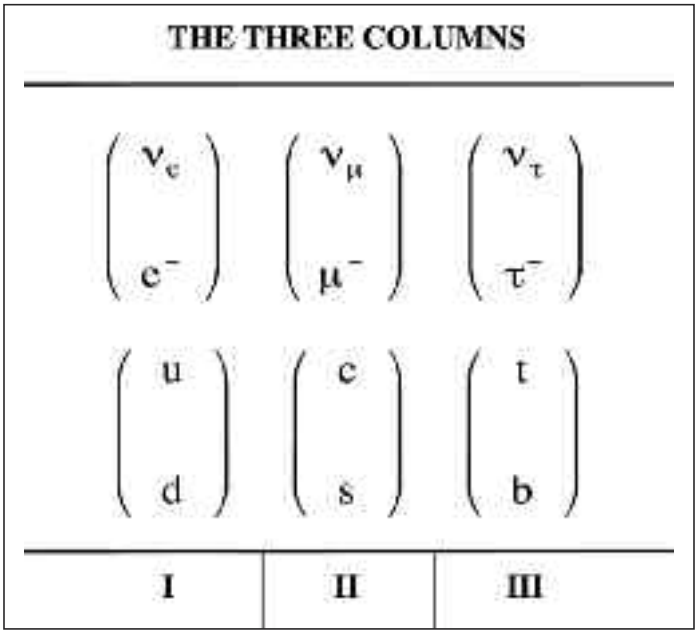


Figure 14c.

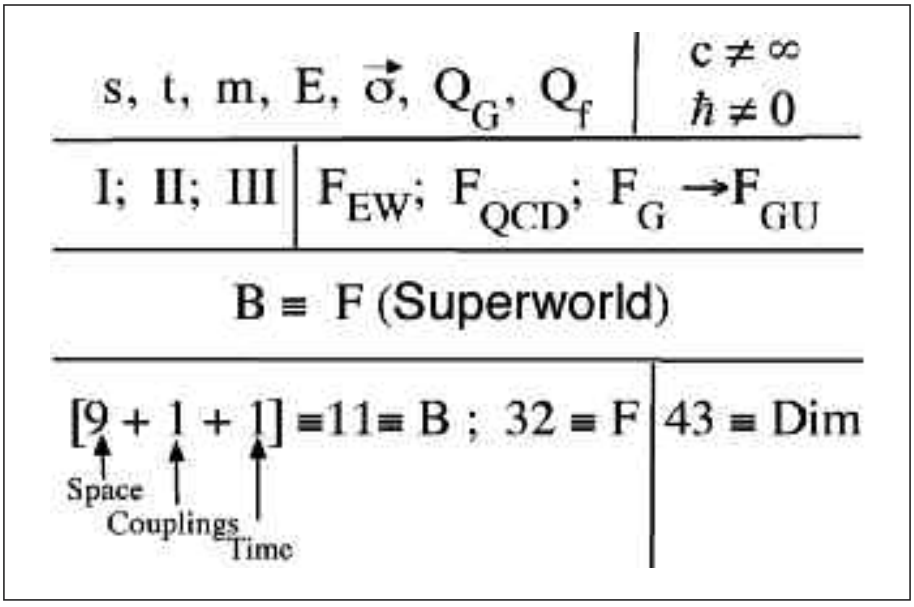


Figure 14d.

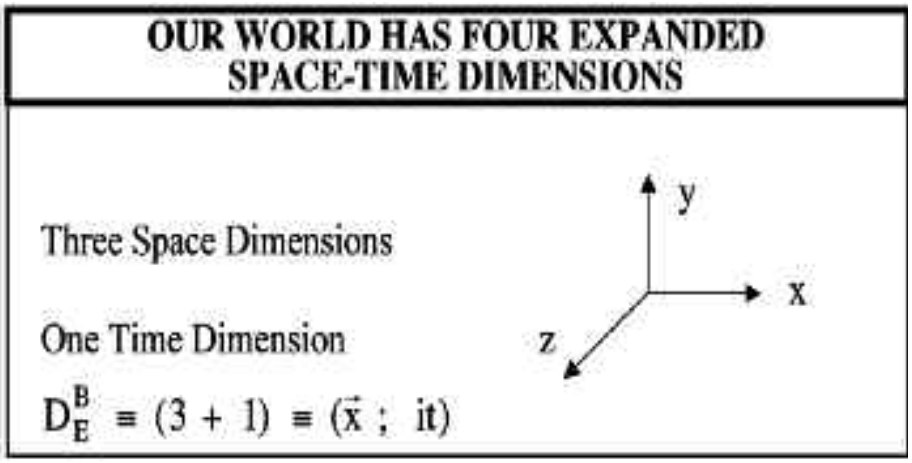


Figure 15a.

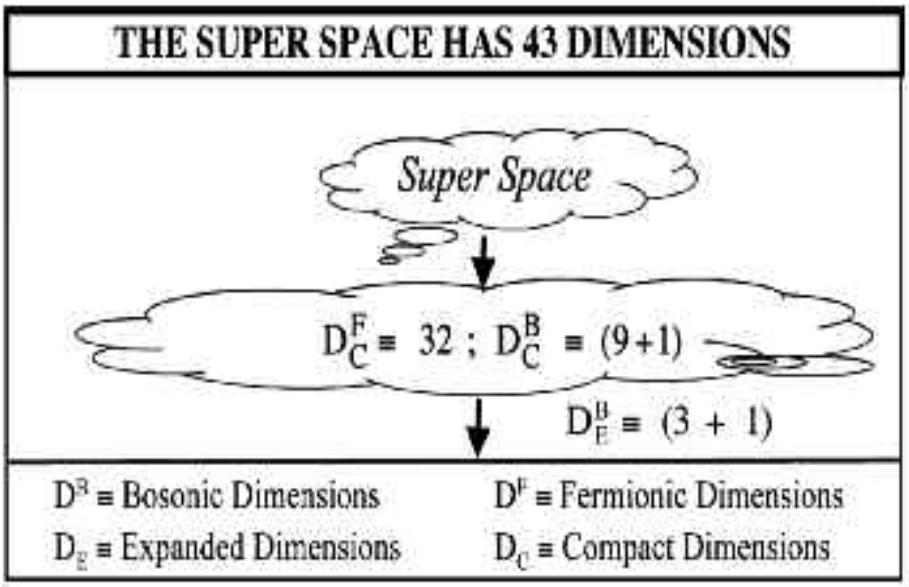
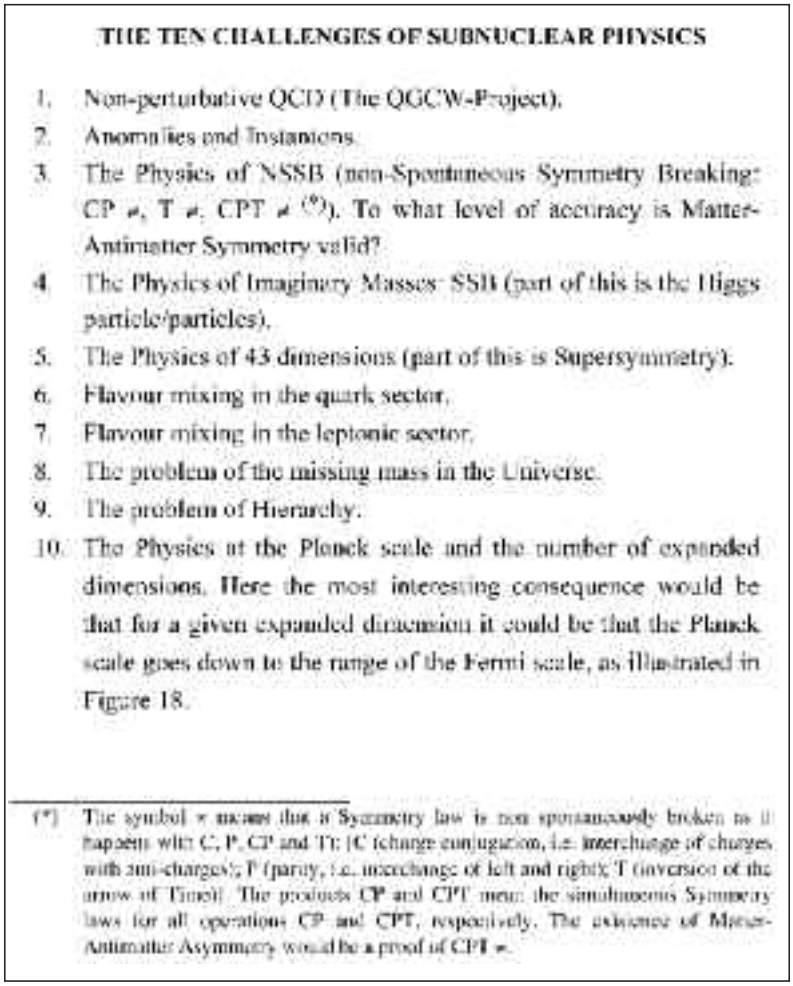


Figure 15b.


SM&B
THE STANDARD MODEL AND BEYOND
① RGEs: $(\alpha_i\ (i = 1, 2, 3);\ m_j\ (j = q, l, G, H));\ j(k^2)$.
• GUT ($\alpha_{GUT} \approx 1/24$) & GAP ($10^{16} - 10^{18}$) GeV.
• SUSY (to stabilize $m_H/m_P \approx 10^{-17}$).
• RQST (to quantize Gravity).
② Gauge Principle (hidden and expanded dimensions).
— How a Fundamental Force is generated: SU(3); SU(2); U(1) and Gravity.
③ The Physics of Imaginary Masses: SSB.
— The Imaginary Mass in SU(2)×U(1) produces masses ($m_{W^\pm}$; m_{Z^0}; m_G; m_f), including $m_\gamma = 0$.
— The Imaginary Mass in SU(5)⇒SU(3)×SU(2)×U(1) or in any higher (not containing U(1)) Symmetry Group ⇒ SU(3)×SU(2)×U(1) produces Monopoles.
— The Imaginary Mass in $SU(3)_c$ generates Confinement.
④ Flavour Mixings & CP ≠ , T ≠ (direct ≠ , not via SSB).
— No need for it but it is there.
⑤ Anomalies & Instantons:
— Basic Features of all Non-Abelian Forces.
Note: q = quark and squark; l = lepton and slepton; G = Gauge boson and Gaugino; H = Higgs and Shiggs; RGEs = Renormalization Group Equations; GUT = Grand Unified Theory; SUSY = Supersymmetry; RQST = Relativistic Quantum String Theory; SSB = Spontaneous Symmetry Breaking.
m_F = Fermi mass scale; m_P = Planck mass scale; k^2 = quadrimomentum; C = Charge Conjugation; P = Parity; T = Time Reversal; ≠ = Breakdown of Symmetry Operators.
The five basic steps in our understanding of nature. ① The renormalization group equations (RGEs) imply that the gauge couplings (α_i) and the masses (m_j) all run with k^2. It is this running which allows GUT, suggests SUSY and produces the need for a non point-like description (RQST) of physics processes, thus opening the way to quantize gravity. ② All forces originate in the same way: the gauge principle. ③ Imaginary masses play a central role in describing nature. ④ The mass-eigenstates are mixed when the Fermi forces come in. ⑤ The Abelian force QED has lost its role of being the guide for all fundamental forces. The non-Abelian gauge forces dominate and have features which are not present in QED.

Figure 16.



THE TEN CHALLENGES OF SUBNUCLEAR PHYSICS

1. Non-perturbative QCD (The QGCW-Project).
2. Anomalies and Instantons.
3. The Physics of NSSB (non-Spontaneous Symmetry Breaking: CP ≠, T ≠, CPT ≠ (*)). To what level of accuracy is Matter-Antimatter Symmetry valid?
4. The Physics of Imaginary Masses: SSB (part of this is the Higgs particle/particles).
5. The Physics of 43 dimensions (part of this is Supersymmetry).
6. Flavour mixing in the quark sector.
7. Flavour mixing in the leptonic sector.
8. The problem of the missing mass in the Universe.
9. The problem of Hierarchy.
10. The Physics at the Planck scale and the number of expanded dimensions. Here the most interesting consequence would be that for a given expanded dimension it could be that the Planck scale goes down to the range of the Fermi scale, as illustrated in Figure 18.

(*) The symbol ≠ means that a Symmetry law is non spontaneously broken as it happens with C, P, CP and T: C (charge conjugation, i.e. interchange of charges with anti-charges); P (parity, i.e. interchange of left and right); T (inversion of the arrow of Time). The products CP and CPT mean the simultaneous Symmetry laws for all operations CP and CPT, respectively. The existence of Matter-Antimatter Asymmetry would be a proof of CPT ≠.

Figure 17.

4. Conclusion: The Scientific Legacy of the 20th Century

Here is the Legacy: *Our father is not chaos. We are the children of a formidable, Rigorous Logic which is valid from the smallest structures of the Subnuclear World to the borders of the Universe*.

The whole of our knowledge is proof of it, as shown in Figure 19. In fact, if we were the children of chaos, the contents of this Figure would not exist. If a fellow could deduce the content of Figure 19 from chaos, the Legacy quoted above would be in trouble. This fellow does not exist.

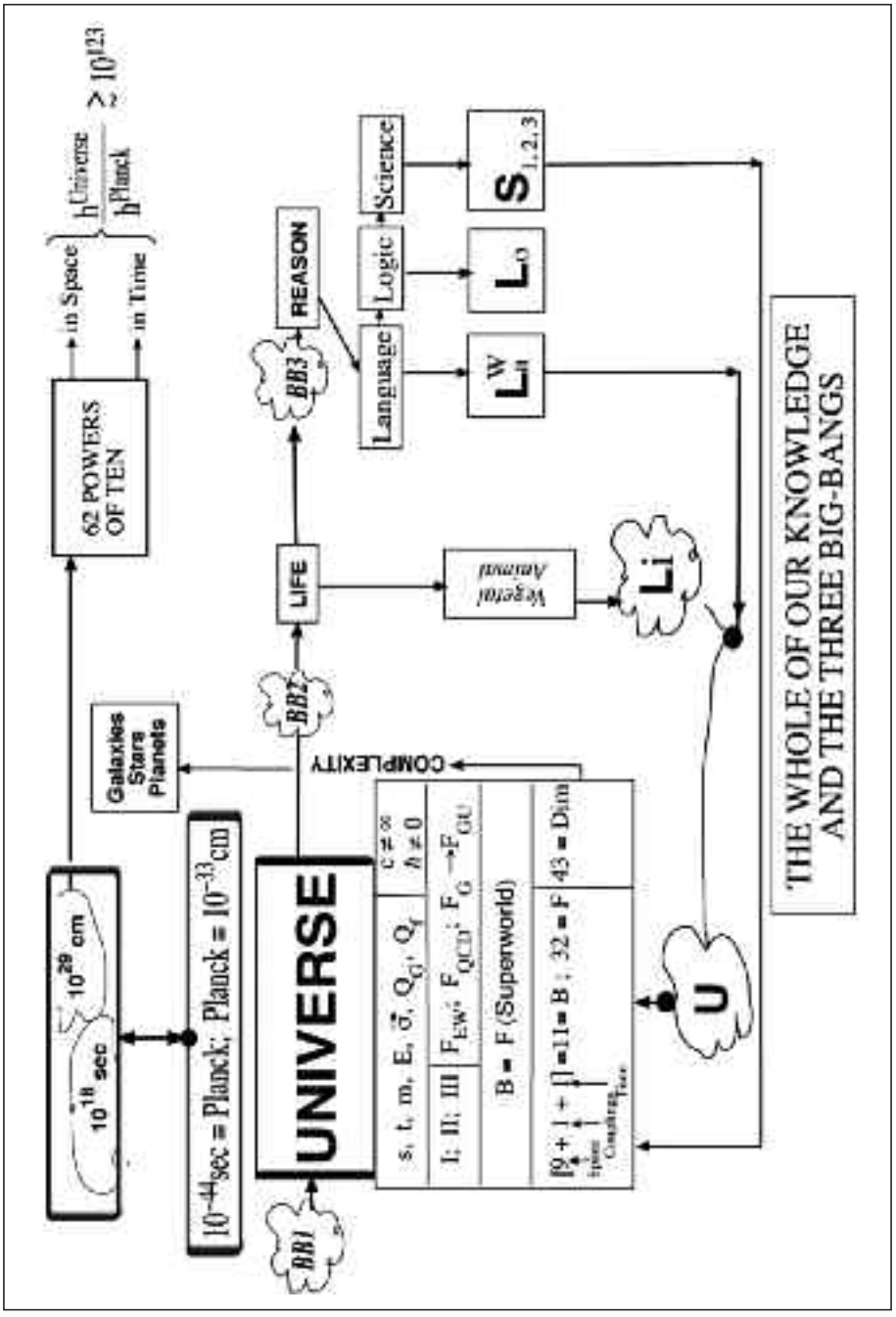

$\geq 10^{123}$
$\frac{h^{Universe}}{h^{Planck}}$
in Space
in Time
62 POWERS OF TEN
10^{29} cm
10^{18} sec
10^{-44}sec = Planck; Planck = 10^{-33}cm
BB1
UNIVERSE
BB2
BB3
REASON
LIFE
Galaxies Stars Planets
COMPLEXITY
Language
Logic
Science
L_i^W
L_O
$S_{1,2,3}$
Vegetal Animal
Li
U
s, t, m, E, σ, Q_G, Q_f
$c \neq \infty$
$h \neq 0$
I; II; III
F_{EW}; F_{QCD}; $F_G \rightarrow F_{GU}$
B = F (Superworld)
43 = Dim
THE WHOLE OF OUR KNOWLEDGE AND THE THREE BIG-BANGS

Figure 19.

APPENDIX 1
Atheism is an Act of Faith about Nothing

1.1. Reason according to Atheists

For Atheistic Culture, Reason is the outcome of the Biological Evolution of the Human Species. The Biological Evolution of the Human Species (BEHS), however, lies below the third level of scientific credibility (see Appendix 2). This can be clearly understood by comparison with the Cosmic Evolution.

BEHS lacks rigorous mathematical formulation and is not based on reproducible experiments at the first level. If BEHS were Science at the first level, then a BEHS equation would exist, leading to the outcome of Reason. And that is not all. There are innumerable forms of living matter. None of these, however, has been able to discover Science, or rigorous Logic, or Collective Memory. BEHS is unable to explain how it is that we are the only form of living matter that has the great privilege of being endowed with Reason.

1.2. Atheism is self-contradictory

Atheism is a contradictory logical construction. In fact, it denies the existence of the Transcendent.

Since the greatest conquests of Reason in the Immanent are Language, Logic and Science, Mathematics (rigorous theoretical Logic) should be able to demonstrate that God does not exist, and Science (rigorous experimental Logic) should be able to discover that God does not exist.

Mathematics has not demonstrated the Theorem of the Denial of God and Science has not discovered the scientific proof of the non-existence of God.

If everything finds expression within the Immanent alone, how is it possible that there is no Theorem of the Denial of God, nor the scientific discovery of the non-existence of God? Here is the contradictory nature of the logical construction of Atheism.

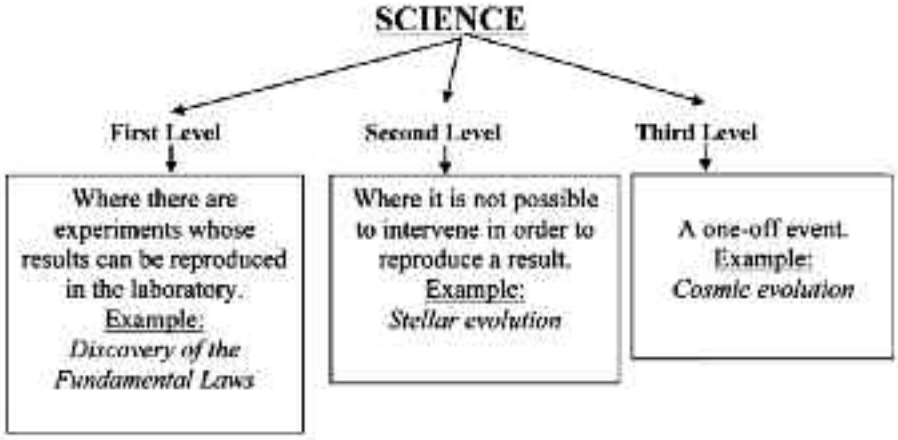
1.3. The Transcendent solves the contradiction of Atheism

In the Logical Structure of the Believer, there exists the Transcendental Sphere, and Reason is a gift of God.

God has given us this unique privilege that has allowed us to make the Three Great Conquests. Logical Mathematics is not able to demonstrate the Theorem of the Existence of God in that, if it could, God would be Mathematics alone. God instead is everything. The same is true for Science. If Science were to manage to discover God, then God would have to be just

Science. But instead, God is everything. It is the task of philosophical thought (see Appendix 6.4) to demonstrate that God exists through the Transcendental Sphere of our existence and its connections with the Immanent Sphere of everyday life.

APPENDIX 2
A Note on the Three Levels of Science



In order to be 'scientific', an activity needs the existence of the first level: i.e., experiments with reproducible results in a laboratory. The results must be expressed in mathematical terms with the correspondent uncertainty quoted.

If the experiment is reproduced in another laboratory and gives results which are in contradiction with previous knowledge it is necessary to establish which one of the experiments is wrong.

In the given activity, it must be possible to put different experiments in a mathematical formalism which allows 'predictions' to be made (see Appendix 5.2). The best example of such an activity is the series of experiments in electricity, magnetisms and optics that after two centuries allowed Maxwell to find four equations from which all results could be derived. The four Maxwell equations gave rise to the most powerful understanding of the effects generated by the electromagnetic forces which allow *predictions* to be made with very high precision. This understanding is known as Quantum ElectroDynamics (QED).

Many activities can become 'scientific' if they follow the example of QED. Otherwise, the existence of the second and third level must be continued until the first level is discovered in the given activity.

When this happens to be the case all three levels must be formulated in a rigorous way, and there should be no contradiction among them. An example of the link between the three levels of Science: Cosmic Evolution formulated in a rigorously mathematical way, and based on the discoveries of the Fundamental Laws made at the first level.

No phenomena known in the Galilean sense (i.e., rigorously reproducible) exist that cannot be explained as a consequence of first level Science. This represents the greatest conquest of Reason in the Immanent.

This study, undertaken by Galilei just four centuries ago, leads us to conceive of the existence of a reality even more exciting than the one we are used to – a reality of extraordinary symmetry which has been called Superworld (see Appendix 3.4).

APPENDIX 3
Language (Permanent Collective Memory), Rigorous Logic and Science (From the Stones to the Superworld)

3.1. The greatest conquests of Reason are Language (with Permanent Collective Memory) Logic and Science

If Language were sufficient to discover Science, it would have been discovered at the dawn of civilisation. If rigorous Logic were sufficient to discover Science, it would have been discovered by the Greeks.

To discover Science, it is not sufficient to think and reflect (Language), or to resort to rigorous reasoning (Mathematical Logic). To discover Science (Logic of Nature), there is one single route: to be able to find rigorously formulated questions. This requires an act of humility: the recognition that the Author of the Logic of Nature is more intelligent than any of us – philosophers, thinkers, mathematicians, logisticians, scientists. It is necessary to surrender before the intellectual Majesty of He who made the world.

It was Galilei who understood this. It was he who said that the footprints of the Creator were to be found in the stones (just as in the Stars). Galilei brought the Logic of the Stars into common matter (stones, string, wood), through an act of Faith on the existence of a fundamental Logic which governs the real world (see Appendix 5.1).

In pre-Galilean thinking, *for Atheists and believers alike, matter could not be a depository of fundamental truth*. The Fathers of the Church were the first to say that Nature is a Book written by God. Galilei had the privilege of understanding that the characters of that Book had to be mathematical, and that it was not enough to reflect on the heavens and Stars.

All previous cultures attributed to the heavens properties that lay above those of the stones. Galilei brought the Logic of Nature into stones and common matter, saying that our intellect has a power below that of the Author of the Logic of Nature. And thus it is necessary to bow before His intellectual Majesty and ask humbly how He has made the world. In other words, what rigorous Logic – of all possible logics – did He follow to make the world as it appears to our eyes and our intellect. The significance of a rigorous and reproducible experiment is precisely what Galilei intended and experienced: to humbly ask a question to the Author of the Logic.

3.2. Ten thousand years compared with four centuries

This is how, in just four centuries, we have managed to decipher a good part of the Logic of Nature. And we have managed to understand just how right was Galilei's humility. In fact, from the dawn of civilisation right up to Galilei – in other words, for a good ten thousand years – all that man thought he had discovered about how the world was made, without ever carrying out an experiment, turned out to be wrong. Still today, Galilean teaching rules the logic of all the scientific laboratories in which the Fundamental Laws of Nature are studied.

Here is a last example of enormous interest today. No one can tell us whether the Superworld exists or not. And yet this theoretical reality is based on rigorous mathematical foundations. It is on these foundations that we believe we have understood so many properties of the world in which we live. But even so, the Galilean proof to be certain of the existence of the Superworld is lacking.

Logical rigour is not sufficient; we need Galilean proof. To know more about the Logic of Nature it is necessary to be able to formulate the right questions to the Author of the Logic who made the world. This is how, in just four centuries, we have reached the threshold of the Superworld.

3.3. From Galilei to the Superworld via Fundamental and Universal Laws

Galilei studied stones in order to discover the Logic of Nature. He could have discovered chaos instead. Had Galilei not existed, we would know nothing about the existence of the Fundamental Laws of Nature. So two questions arise:

- what did Galilei know about the fact that the Fundamental Laws of Nature had to exist?
- and on what foundations was he able to conceive that these Laws had to be Universal and Immutable?

Imagining the existence of Universal and Immutable Fundamental Laws does not involve acts of Reason and nothing else, but of Faith in the existence of a Logic of Nature which governs the world in all its structures.

Were it not for Galilean Science, we would not be able to say that Fundamental Laws of Nature, Universal and Immutable, exist; nor that these Laws lead to the unification of all the phenomena studied in the visible Universe, which appears to us with just four dimensions.

The Grand Unification brings with it the need for a Superworld, a scientific reality with forty-three dimensions: eleven of the 'boson' type and thirty-two of a 'fermion' nature.

3.4. Why we need the Superworld

Here are the problems that make the Superworld a necessity.

1) The two energy scales must be kept separate: 10^{19} GeV (Planck) and 10^{2} GeV (Fermi).
2) The gravitational attraction of light must be prevented from being infinite. Otherwise we would see neither the light of the Stars nor the light of our Sun. The 'gravitino' (Supergravity) allows the gravitational attraction of light to be finite.
3) Gravitational attraction is powerful but it cannot be infinite. We would be stuck to the Sun. Space would not exist between Stars and Galaxies. Cosmic expansion would not exist. In order to have a finite gravitational attraction, theories are needed in which the Euclidean concept of point is abandoned. The point is replaced by a string. No more Pointlike Theories but Superstring Theories. These theories must be supersymmetric: the Supersymmetry Law ($F \equiv B$) must be valid. Otherwise 'tachions' would appear.
4) Aiming at the Unification of all fundamental phenomena – the synthesis of which is provided by three 'gauge couplings', α_1 α_2 α_3, running with the energy – the Supersymmetry Law ($F \equiv B$) must necessarily be introduced.
5) Supersymmetry does not show up at our energy scale. Hence the problem arises to compute the energy above which the ($F \equiv B$) Law starts to act. Thanks to the EGM effect, this energy level is 700 times more accessible than thought so far.
6) An interesting detail: the theoretical model called no Scale-Supergravity is the Infrared solution of Superstring Theory. This model might allow us to understand the extremely small value of the Cosmological Constant.
7) Finally: why Three Columns and Three Forces? The answer to this question should come from the 43-dimensions of the Superspace.

3.5. Where the ashes of the Superworld could be

The ashes of the Superworld (the so-called neutralinos) could explain the compactness of our Galaxy.

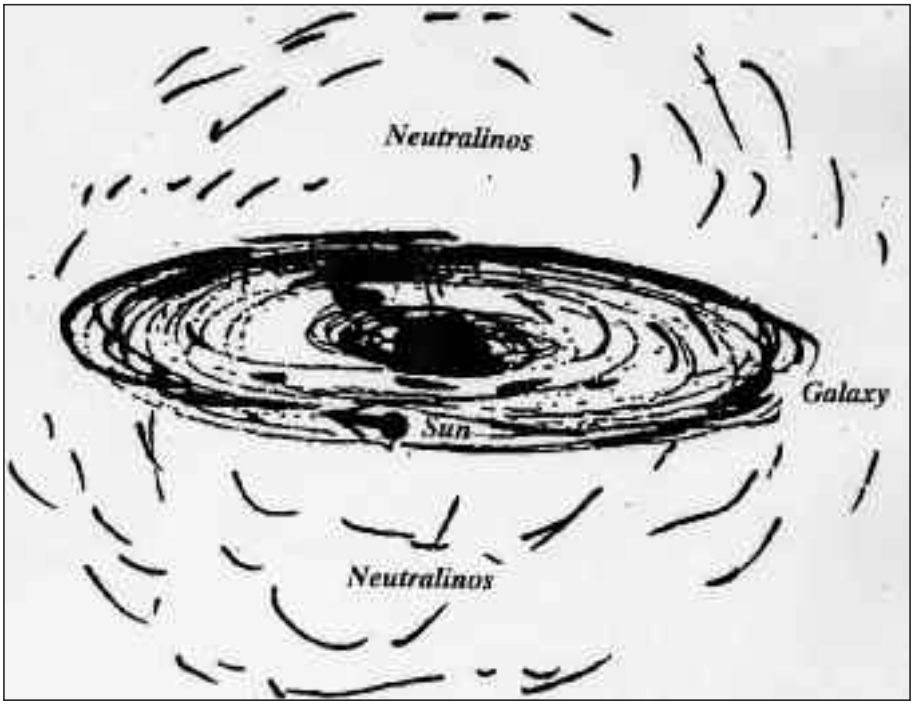


Figure 20.

Neutralinos cannot aggregate into Stars since, being neutral, they lose little energy. This would allow neutralinos to remain in a sphere concentric with our Galactic centre. Even though they aggregated into Stars, neutralinos could not emit light, like ordinary Stars do. Fire needs the plasma of protons and electrons. This is why super Stars cannot emit light.

3.6. Our World and the Planck World

It is interesting to compare the density of our body and the density of the Planck Universe. The scales of length, mass and time of the world we are familiar with, and the scales of the Planck world are shown in Figure 21.

OUR WORLD	THE PLANCK WORLD
Human Body Density $1 gr/cm^3$	Planck Density 10^{37} Universes/cm^3
OUR SCALE of Length, Mass and Time	THE PLANCK SCALE
length cm	length 1.6×10^{-33} cm
mass gr	mass 2.2×10^{-5} gr
time sec	time 5.4×10^{-44} sec
Everyday Reality	The Reality we come from

Figure 21.

APPENDIX 4
The Values of Science and Faith are Closely Linked

We will now see that Science is a source of values, and that these values are in perfect harmony with the values of Faith, not in antithesis. Below is a short summary of the values that Science has in common with Faith. The description of each value follows.

1 *Revolution*

2 *Racism*

3 *Universality*

4 *Elevation of the individual*

5 *Intellectual stimulus*

6 *Humility*

7 *Truth*

8 *Reflection on facts*

9 *Goodness and tolerance*

10 *Fight against preconceptions*

11 *Generosity*

12 *Freedom of thought*

4.1. Revolution

Let's begin with the concept of revolution. When a scientific discovery is made, the dominant Culture loves to point out that a real revolution has taken place.

Scientific revolutions have never produced deaths or injuries. The concept of 'revolution' derives from the discovery that it was the Earth and the other satellites of the Sun that move, going around in their orbits. It was the 'revolution of the orbits' that gave life to Galilean Science. The term 'revolution' intended to emphasise the impact of the 'revolution of the orbits' of the planets on the history of the world. With the passage of time, cultural mystification went to work to change the scientific term 'revolution of the orbits' into the meaning of 'socio-political revolution', like the October Revolution that led to the first example of a Republic with Atheism as State religion, causing many millions of victims.

Instead, following a scientific revolution, everyone is richer than before. It would be more correct to speak of construction, rather than revolution. In Sci-

ence, there is never denial of the past: it is improved on, taken on board and built on. It is as if, when climbing an immense mountain, what we took to be the summit opens up a panorama never before observed – and, as if this were not enough, with it comes the discovery that there is another, even higher, peak.

The term scientific *revolution* does not in any way justify social revolution. But this is what the dominant Atheistic Culture indeed did, in order to persuade that, after all, scientific rigour had necessarily to go down the road of *revolution*, understood in the commonly accepted sense of revolt, with attendant massacres and horrors of every type.

4.2. Racism

A scientist cannot say: 'I am unable to believe in this new scientific discovery because it was made by a man whose skin is a different colour from mine'. Science is an intellectual activity that rejects racism outright.

4.3. Universality

Man has always been in search of universal values. Science shows that Universal Laws exist. The Weak Forces that produce measurable phenomena in our laboratories are the same as those that make the Sun work. The light produced by a match is analogous to that produced by the Stars. Gravitational Force, which makes a stone fall downwards and holds us to the Earth is the same Force that oversees the formation of our Solar System and of the Galaxies.

4.4. Elevation of the individual

Science exalts the individual and his work. The value of a scientist is not established by the power of an army tank, but by his intellect and research efforts.

And here the entire sum of contributions must be recognised. Albert Einstein is inconceivable without Max Planck, James Maxwell, Isaac Newton and Galileo Galilei. All scientists, giants of Science: all believers.

4.5. Intellectual stimulus

Science spurs man on to reach out for further conquests. There is no rest in our endeavour to extend and improve our knowledge. Instead, an ideology is put forward as if it were the final goal of an intellectual conquest. And this holds man back, century after century, on frontiers created from abstract speculations, which in no time at all become dogma.

Science accepts the dogma of the Transcendent. But it rejects the dogma of the Immanent.

4.6. Humility

The scientist in his daily work faces problems he is unable to resolve. Galilei took more than a decade to understand friction and thereby arrive at the formulation of the first law of motion. Einstein devoted eleven years, from 1905 to 1916, to get to the bottom of the significance of Galilei's experiments on the fall of material bodies. Eleven years to succeed in writing one equation. Science is made up of unresolved problems. Something happens, and we move on to the next thing. And there our difficulties begin again. Einstein worked for the last thirty years of his life in the attempt to unify all the Forces of Nature. It was his great, *unfinished* opus. How can a man who is unable to reply to a question be arrogant? Science, as we have said before, is made up of unresolved questions. This is why it is based on a pillar of intellectual humility. Arrogance is born of ignorance.

4.7. Truth

Should a scientist tell a lie, he would be excluded from the scientific context. For Science, something that is true has to be reproducible. The scientist, when he comes to understand something or make a discovery, has to explain in full detail how he has arrived at that result. Whoever, no matter the colour of his skin, has to be able to reproduce that scientific truth wherever, and at any given moment. Mystification and falsehood lie outside scientific activity.

4.8. Reflection on facts

Science teaches us to reflect, not to rush to conclusions without checking every consequence of a discovery in the known sectors of the fundamental structures of Creation. Science trains us for objective, not emotive, judgement. It relies on facts, experimental proof that is reproducible, the baptism of Galilean scientific legitimacy. It does not rely on words and abstract formulae. Nor does it make sense to say that a theory is mathematically beautiful or ugly. It can be only true or false, although it also happens, almost always, that when a piece of research reaches its conclusion, when everything has finally been understood in a specific field, then the mathematical formulation turns out to be more elegant than anticipated.

4.9. Goodness and tolerance

Science teaches intellectual goodness and tolerance. Extremes have to be understood, not defeated. Things that appear to be poles apart can both turn out to be necessary for a description of the fundamental phenomena of Nature. Just one example should suffice: the wave and particle property. Light, for a long time, was considered to be a particle phenomenon. Then wave-

like. And the two descriptions seemed to be mutually exclusive. Instead, light is at one and the same time both wave and particle. Many centuries have been necessary to come to this understanding. The wave-particle *duality* is valid not only for light, but for all particles. This duality is one of the most significant conquests in the history of scientific thought.

4.10. Fight against preconceptions

Science fights an unceasing battle against preconceptions: even if centuries are needed to dismantle them. The great difference between Classical Physics and Modern Physics lies in the fact that a tiny quantity (the so-called *Planck's Constant*) was considered to be exactly zero. Another enormous quantity (the speed of light) was considered infinite. Three hundred years to break down two preconceptions.

4.11. Generosity

Science also has important facets of generosity. Explaining to others the results of a discovery is something that enriches both scientist and listener. Science teaches that there exists an absolutely perfect form of generosity and love for our neighbour. He who gives up a piece of bread does a good deed, but clearly suffers if he has little bread. He who gives away what he knows, loses nothing, even if he ends up giving away everything he has.

4.12. Freedom of thought

Freedom of thought is of vital importance for Science. This includes respect for that form of living matter known as man, and therefore respect for his dignity. Of all the forms of living matter, we in fact are the only one which has been granted the privilege of understanding the Logic He followed in creating the reality in which we live and of which we are made. This unique privilege is the source of the highest dignity to which one can aspire: that of being made in the image and likeness of the Creator of all things visible and invisible. To read the Book of Nature, written by the Creator, one needs to be free of any prejudice, the only guide being the replies given by He who has made the world when we put forward a question. The intellectual freedom to put a question to He who has made the world has to be absolute.

APPENDIX 5
Chaos or Logic?

5.1. If there is Chaos there are no Fundamental Laws. If there is a Logic there must be the Author

Science aims at understanding what God has written, using the rigour of Mathematics. Galilei said and thought that the Fundamental Laws of Nature are in fact expressed as precise mathematical equations. The father of Science did not know that his studies of oscillating pendulums or stones rolling down an inclined plane would have allowed him to deduce rigorous laws. Chaos, randomness, whim might just as possibly have appeared instead: one day like this, a year later quite different. One law for Pisa, another for the Moon.

Galilei instead was thinking in terms of fundamental and universal laws, expressible in rigorously mathematical form. Together, these laws were to represent, and *de facto* do represent, the Logic of Nature.

'In that stone there is the hand of the Lord. By studying *common objects* I will discover the Laws of He who has made the world'. This was the Faith that inspired Galilei to challenge the dominant Culture of his time. He simply wanted to read the Book of Nature, written by the Creator in mathematical characters.

The Book of Nature reveals to us how the world has been made: the work of Creation. This opus could have been written in no other way but rigorously, in mathematical characters. It is the scientist, in the first person, who has to strive in order for everyone to know how to read that astonishing and fascinating Book.

In it is written how the world is made. Since it is dealing with a construction, its language has to be rigorous. Knowing how to read it means making available for the benefit of man the laws that rule the Cosmos, in communion, not in antithesis, with the word of God, that is, the Bible. The Bible is written in a simple way, so that everyone can understand it; its purpose is not to explain how the Immanent part of our existence is made. Instead, it has the goal of tracing out for man the path that leads to the Lord. Science gives us the certainty of not being the children of Chaos, but of a rigorous Logic. Who is the Author of this Logic? Atheism replies: no one. This is why Science, born in the Immanent, brings man towards the Transcendent, because it is absurd that such Rigorous Logic does not have an Author.

5.2. If there is Chaos there are no predictions

Let us see how predictions at the fundamental level of scientific knowledge can exist. The experimental evidences for the existence of predictions are the very many results of scientifically reproducible experiments.

For example the measurement of the anomalous magnetic moment, in symbols (g–2), of the electron (e):

$$(g–2)_e$$

which is theoretically computed at an extraordinary level of precision (few parts in ten billion parts) and is experimentally verified to be correct.

Could the

$$(g–2)_e$$

be predicted before the discovery of the Maxwell equations and the existence of Quantum ElectroDynamics (QED)?

Predictions at the fundamental level of scientific knowledge depend on UEEC events (discussed in the next Chapter 5.3).

For example: it is the discovery of the laws governing electric, magnetic and optical phenomena (all totally unpredicted) which produced the mathematical structure called QED.

Mathematical structure was not invented before the innumerable series of UEEC events in electricity, magnetism and optics which allowed Maxwell to express 200 years of experimental discoveries in a set of 4 equations.

Mathematical formalism comes after a totally unexpected discovery: an UEEC event which no one was able to predict.

In the whole of our knowledge predictions exist only in Science.

These predictions are the analytic continuation of what is already known. The greatest steps in the progress of Science came and will come from totally unpredicted discoveries.

This is the reason why we need to perform experiments, as Galileo Galilei realized, 400 years ago.

Today we have all mathematics needed to describe the Superworld, but in order to know if the Superworld exists we need the experimentally reproducible proof of its existence (as discussed in Appendix 3).

5.3. If there is Chaos there are no UEEC events. UEEC are the proof that the Author of the Logic is smarter than us all, no one excluded

5.3.1. Unexpected Discoveries in Physics

Let me show a synthesis of achievements in Physics from Galilei to the first half of the 20th Century (Figures 1 and 2, pp. 94-95).

I have included the invention of electronic computers by Von Neumann (XXIX), which no one could have imagined at the beginning of the 20th Century. Point no. XXX refers to the impressive list of Fermi discoveries:

once again, all totally unexpected. The UEECs of the second half of the 20th Century (Figure 3, p. 96) are grouped into 3 classes:

- one is the 'Subnuclear World'
- the second is the 'Standard Model and Beyond'
- the third is the 'Superworld'.

The existence of the Subnuclear World and the Standard Model are strictly correlated. The third is the frontier of our knowledge which exists as a fascinating mathematical structure, but lacks Galilean experimental proof (as discussed in Appendix 3).

The greatest synthesis of all times in the study of fundamental phenomena (Figures 13 and 14, pp. 103-105) has been reached through a series of totally unexpected discoveries reported in Figures 16 (p. 107) and 22 (p. 363).

5.3.2. The Standard Model and Beyond

The superb synthesis called the 'Standard Model' is a part of a more general structure, where many problems are open. We call this structure 'The Standard Model and Beyond', 'SM&B' (Figure 16, p. 107).

This Structure brings to the unification of all Fundamental Forces of Nature, suggests the existence of the Superworld and produces the need for a non-point-like description of Physics processes (the so-called Relativistic Quantum String Theory: RQST), thus puving the way to quantizing gravity.

5.3.3. Conclusions about UEEC from Galilei to Subnuclear Physics and other fields

In the field of Subnuclear Physics, totally unexpected discoveries date back to the beginning of Galilean Science.

Question. *What about other fields?* One which is very intensive in number of discoveries is the field of condensed matter.

Let me quote Tony Leggett (University of Illinois, Urbana - Champaign, USA), Nobel Prize 2003 for 'Superfluidity': 'It is relatively rare in Condensed-Matter Physics to predict discoveries; it is a field where you fall over them by accident'.

APPENDIX 6
If Our Culture were Modern, the Cultural Mistifications which are in the 'Present Convictions of a Modern Culture' would not Exist

6.1. If we were to live in the Era of Science everybody would know that Science and Faith share the same values

If we lived in the era of Science, the values of Science would form an integral part of the so-called Modern Culture. In fact, they are truths that render Science an intellectual activity that is in perfect communion with religious thought. We are dealing with two essential components that make up our existence: one that operates within the Immanent, Science; the other that operates within the Transcendent, Faith.

And this is the conclusion one comes to. Science, by studying the Immanent in the most rigorous way that human intellect has ever been able to conceive, discovers a series of truths, whose values (see Appendix 4) are in perfect harmony with those that the same form of living matter, called *man*, learns from Revealed Truth.

Four centuries after the time of Galilei, that which the father of Science was able to see with a pure act of Faith and Love towards Creation becomes visible in dazzling clarity: Nature and the Bible are both works by the same Author.

The Bible – said Galilei – is the word of God. Nature instead is His writing. If we lived in the era of Science, these truths would be the cultural heritage of everyone.

6.2. A few examples of cultural mystifications in 'Scientific' popularisation

Scientific Culture has the duty to correct the cultural mystifications of the popularisation of science, mystifications that might at first sight seem mistakes committed in good faith. But the fact that they are all bound to a common cultural substrate confirms that they are not. In fact, the mystification that Faith and Science are in antithesis is not the only instance where falsehood is elevated to truth by popularisation of science. There are many more. Here are a few examples.

Popularisation of science has:

- confused Science with Technology.
- never explained that the three great conquests of Reason are: Language, Logic and Science (Appendix 3).
- always kept silent regarding the Galilean distinction of the three levels of scientific credibility (Appendix 2).
- attributed to Science the responsibilities of the Planetary Emergencies; responsibilities that belong instead to political violence (planet packed

with chemical, bacteriological and nuclear bombs) and economic violence (irresponsible industrialisation and related effects).
- elected itself spokesman of ideas (for example: scientific materialism) that are in total contradiction with the conquests of scientific thought.
- endorsed as frontiers of true and great Science research activities that still lie below the third level of scientific credibility (for example: BEHS, biological evolution of the human species).

Our epoch will go down in History as that in which cultural mystification has raged: falsehood becomes truth. The main author of this mystification has been the dominant Atheistic Culture.

In this way, Science and Technology have been deliberately confused. And the blame continues to be laid at the feet of Science, a blame that instead belongs to political violence. Violence which, in the 20th Century, had examples of terrifying power in Hitler and Stalin; they both exploited the use of Science (Technology) for political ends, not for progress or civilisation.

6.3. If everything is Science, nothing is Science. It is necessary to distinguish Science from the other conquests of Reason. There is only one Science

'Scientific Culture' is the only form of defence against cultural pollution, maintained Dirac, Kapitza and Fermi. If everything is Science then nothing is Science. And it is impossible to explain that scientific Marxism is the exact opposite of Science. It is thus necessary to distinguish Science from the other conquests of Reason – i.e., from Mathematical Logic and Language.

The umbrella of Language covers Poetry, Art, Philosophy and all intellectual activities that are not concerned with reading the Book of Nature in order to decipher the Logic followed by He who has made the world. Using Language, in all its forms, everything can be said and its contrary. Language – as Borges says – has the supreme aspiration of 'magnificent' structures such as a Poem can have, leaving aside Logic and Science, which is the Logic of the Nature.

Scientific knowledge is engaged full time in studying – in a Galilean reproducible way – this Logic. The key to distinguishing this activity from all others lies in intellectual humility, without which scientific knowledge would never have been born nor able to grow. This intellectual humility, which is vital for scientific knowledge, is not always present – in fact, often quite the reverse – in intellectual activities that contribute to the growth of non-scientific knowledge. This is why there is only one Science, while there are many forms of Art, Literature and Philosophy and other intellectual activities, often in contradiction one with another.

6.4. Humanistic Culture is not in contrast with Scientific Culture. The role of philosophical thought

This has been the case in the past and will continue to be so in the future. Even so, it is philosophical thought that produces fundamental contributions in the study of the Transcendental Sphere of our existence.

The contradiction intrinsic in Language's very structure is surmounted when Philosophy comes into play: its roots allow an understanding of how and why this contradiction does not have to extend beyond the conquests of Language.

In other words, the fact that there are various forms of Poetry, Art, Music cannot be taken as a basis on which to build a *Humanistic Culture* in contrast with *Scientific Culture*. The contradiction lies in the Creativity of Language itself, from which arise various expressions of our way of hearing and seeing the world. It is right that it is so. It is required by Language's very structure. It is here that the links with the Transcendental Sphere of our existence come into being, links that extend to Logic and Science through the creative processes of these great conquests of Reason in the Immanent. Creativity in Language finds its maximum structure in philosophical thought, without which it would not be possible to reflect on the Transcendental Sphere of our life. It is at this frontier that Philosophy expresses the highest creative power.

Creativity in Science has to coincide with the Logic chosen by He who has made the world to create the reality we are made of and in which we live. We scientists are not able to invent the existence of the Third Lepton (see Chapter 2). We can imagine its existence on the basis of experimental results, which can suggest new avenues for us to follow.

But whether the third lepton exists is known to the Creator, before any scientist in the world. It is He who has decided to include this 'third column' in the structure of Creation. We have been granted the privilege of discovering that it does indeed exist. The same is true for the existence of Antimatter and all other discoveries in which I have been directly involved, as reported in Chapter 2.

6.5. Creativity in Mathematics

With Mathematical Logic, the significance of Creativity is different. It is a legitimate act of the intellect to invent a new mathematical structure: with its rules and theorems. This structure does not necessarily have its correspondence in the Logic of Creation.

In order for this mathematical-logical structure to exist, the only condition is the principle of non-contradiction. But the principle of non-contradiction arises in philosophical thought, an integral part of Language. Logic

formulates this principle rigorously, and uses it to underpin any of its structures. A structure – completely invented by the intellect – must not lead to a theorem and the negation of the theorem itself.

Having said this, the problem of the role of Mathematics in the Logic of the Creation remains open: this topic has impassioned the very best mathematicians of all time. There is no doubt that a formidable logical-mathematical structure can exist (and therefore be non-contradictory), without there being any correspondence with the reality of the world in which we live and of which we are made.

This in no way diminishes the fascination of the Creativity in the two conquests of Reason (Language and Logic), which, since they are distinct from Science, do not fall under Galilean-type experimental confirmation.

However, it is of fundamental importance to distinguish Science from the other two conquests of the Reason of the Immanent, in that, if everything is Science, then nothing is Science, with all the devastating cultural consequences, some of which are referred to in this Section.

6.6. Cultural pollution

Kapitza said: 'Cultural pollution is the most difficult Planetary Emergency to overcome'. Here is an example. In the USSR, very few knew of the ecological disasters caused by the triumphs of the 'five-year plans' made known everywhere through propaganda campaigns, even in the western world, where they were taken as models of unprecedented development. In Italy, Communist Party members made great reference to them. No one, however, spoke of the ecological disasters of *Semipalatinsk* (100 times worse than Chernobyl), the *Aral Sea* (50% of its waters destroyed), the *City of Sulphur* (an area as large as half of Piedmont, contaminated to the point where the population had to go around wearing gas masks). These were the times of the cold war and no one dared to hope for the collapse of the USSR. But even so, the hero of Science, Pëtr Kapitza, considered it necessary to start immediately to fight cultural pollution in countries that were free; in those dominated by the USSR it was unthinkable. Dirac said: 'It is easy to declare ourselves as free men where there is democracy and freedom. Try to do this where political violence rages. Kapitza suffered the consequences during years and years of his life'.

Cultural pollution has its roots in political and economic violence, which, by dominating the media (TV, radio, press and other channels), has enabled so many flagrant cultural mystifications to become 'truth'.

A terribly effective weapon of cultural pollution is pseudo-scientific confusion, an essential component of popularisation. To cite meaningless data as if they were Galilean proofs of scientific truth; to introduce apparently valid ar-

guments with bibliographic references that add nothing to the inexistent proof of the point in question: this is the technique of cultural pollution that destroys valuable energies from the struggle for the triumph of Scientific Culture.

6.7. An example of cultural confusion: Science, Art and Mysticism

According to a number of scholars, the pillars supporting our existence are: 'Science' (rational approach), 'Art' (aesthetic approach) and 'Mysticism' (religious approach). These theories have nothing new to say about the conquests of Reason. Rather, they go backwards in time because they ignore Galilean teaching. In fact, they confuse the Transcendental Sphere of our existence (to which Mysticism belongs) with the Immanent Sphere (to which Science belongs). Furthermore, they include in the so-called 'rational approach' both Science and Mathematics, confusing Science with Logic. Galilei teaches that, to discover Science, the rigour of Mathematical Logic (thus, the rational approach) is not sufficient.

If it were so, the Logic of Creation would have been discovered by the Greeks, two thousand years before Galilei. If mathematical rigour sufficed, we could say that the Superworld exists. The Galilean thesis is based on 'Language', 'Logic' and 'Science' and it could not be more rigorous in distinguishing the three conquests of Reason. Art in fact belongs to Language.

APPENDIX 7
A Great Alliance is Needed Between Science and Faith

In the 1980s this alliance strove to make a real contribution to overcoming the risk of a Nuclear Holocaust. Then, with the fall of the Berlin Wall came the need to avoid the danger of an Environmental Holocaust created by the political and economic violence that triggered the undeclared War between the planet's North (the rich) and South (the poor). Once again, Scientific Culture in communion with Faith acted to avoid the latent danger of an Environmental Holocaust, by implementing pilot projects related to the Planetary Emergencies, thanks to volunteer work carried out by its scientific community.

We have discussed how the dominant Atheistic Culture, using as its weapon the public dissemination of what is passed off as Science, has instead wanted everyone to believe that Science and Faith are enemies. It has always confused Science with Technology, has never explained that the three towering conquests of Reason are: Language, Logic and Science, never mentioned the Galilean distinction between the three levels of scientific credibility, and has laid at Science's feet the responsibility for the Planetary

Emergencies – responsibility that instead belongs to political violence (planet packed with chemical, bacteriological and nuclear bombs) and economic intemperance (unaccountable industrialisation). Atheistic Culture too has acted as a spokesperson of ideas, such as scientific materialism, that are in utter contradiction with the conquests of scientific thought, and has endorsed as frontiers of real and true Science, research activities that still lie below the third level of scientific credibility (for example: biological evolution of the human species: BEHS).

Had Atheistic Culture itself discovered Science, then the *Great Alliance* could never have been conceived. This Alliance represents the cultural guide for the third millennium. The birth of a Scientific Culture in communion, not in antithesis, with Faith has enabled the danger of a Nuclear Holocaust to be overthrown (Erice Statement), and allowed the creation of scientific and technological foundations from which to confront issues of the Environmental Holocaust (pilot projects for the Planetary Emergencies).

As said before, the 20th Century will take its place in History for having seen the fall of the Berlin Wall and the start of an undeclared War between North (the rich) and South (the poor). The third millennium needs the Great Alliance between the two most important conquests of Reason, which are Science, in the Immanent of our existence, and the God-given gift connected with Reason in the Transcendent of our being, Faith. We would do well to recall that St. Paul and all our theological tradition define Faith as a gift from God. A gift linked to Reason, as described by St. Thomas Aquinas: 'Naturalis ratio per creaturas in Dei cognitionem ascendit, fidei vero cognitio a Deo in nos e converso divina revelatione descendit'(*) (*ScG* IV 1, 3349). While emphasising the rational aspect of Faith, the entire Christian biblical tradition attributes it to the inner touch by the Spirit of God (*instinctus Dei invitantis*: St. Thomas Aquinas) that awakens the dynamism of free will. Faith is thus considered by Christian theology as a gift from God within man's Reason, which under the impulse of this same free will, and aided by the Holy Spirit, accepts the gift.

We are the only form of living matter that has been granted the privilege of the gift of Reason and free will. Let us seek to use it well. The third millennium must open up man's heart to hope through a Scientific Culture in synergy with Faith, not in antithesis. This is why – Benedict XVI teaches – Science must do everything in its power to ensure the triumph of the values of Galilean Scientific Culture.

(*) 'Natural reason ascends to a knowledge of God through creatures and, conversely, the knowledge of faith descends from God to us by divine revelation'.

References

1. 'Scientific Culture and the Ten Statements of John Paul II', A. Zichichi, Plenary Session on *The Cultural Values of Science*, 8-11 November 2002, Vatican City, The Pontifical Academy of Sciences – in Proceedings of the Plenary Sessions, *Scripta Varia* 105, pp. 288-313, PAS, Vatican City (2003).
2. 'Rigorous Logic in the Theory of Evolution', A. Zichichi, Plenary Session on *Scientific Insights into the Evolution of the Universe and of Life*, 31 October-4 November 2008, Vatican City, The Pontifical Academy of Sciences – in Proceedings of the Plenary Sessions, *Acta* 20, pages 101-178, Vatican City (2009); see also 'Elements of Scientific Rigour in the Theory of Evolution', A. Zichichi, Addendum in Plenary Session on *The Cultural Values of Science*, 8-11 November 2002, Vatican City, The Pontifical Academy of Sciences – in Proceedings of the Plenary Sessions, *Scripta Varia* 105, pp. 314-330, PAS, Vatican City (2003).
3. 'Totally Unexpected Discoveries: A Personal Experience', A. Zichichi, Plenary Session on *Paths of Discovery*, 5-8 November 2004, Vatican City, The Pontifical Academy of Sciences – in Proceedings of the Plenary Sessions, *Acta* 18, pp. 130-153, PAS, Vatican City (2006).
4. 'From the Yukawa Particle to the QGCW', A. Zichichi, in Proceedings of the *Symposium for the Centennial Celebration of Hideki Yukawa*, 23rd International Nuclear Physics Conference, Tokyo, Japan, June 3-8, (2007), *Nuclear Physics A*, Vol. 805, Issues 1-4 (eds S. Nagamiya, T. Motobayashi, M. Oka, R.S. Hayano and T. Nagae), pp. 36-53 (2008); and 'Yukawa's Gold Mine', A. Zichichi, *AAPPS Bulletin,* Vol. 18, n. 3 (ISSN 0218-2203), pp. 50-54, June (2008); see also: *CERN Courier*, Vol. 47, n. 7, pp. 43-46, September (2007).
5. 'Interaction of Elementary Particles', H. Yukawa, Part I, Proc. *Physico-Math. Soc. Japan* 17, 48 (1935); 'Models and Methods in the Meson Theory', H. Yukawa, *Reviews of Modern Physics* 21, 474 (1949).
6. *The QGCW Project*, A. Zichichi *et al.*, CERN-LAA Preprint, October 2006; see also 'Logical Reasoning in Experimental Physics: Past and Future', A. Zichichi, in *Gerardus 't Hooft Liber Amicorum to celebrate his 60th anniversary* (2006).
7. P.A.M. Dirac, 'The Quantum Theory of the Electron', *Proc. Roy. Soc.* (London) A117, 610 (1928); 'The Quantum Theory of the Electron, Part II', *Proc. Roy. Soc.* (London) A118, 351 (1928).
8. 'The Positive Electron', C.D. Anderson, *Phys. Rev.* 43, 491 (1933); 'Some Photographs of the Tracks of Penetrating Radiation', P.M.S. Blackett and G.P.S. Occhialini, *Proc. Roy. Soc.* A139, 699 (1933).
9. H. Weyl, *Gruppentheorie und Quantenmechanik*, 2nd ed., 234 (1931).
10. E.P. Wigner, 'Unitary Representations of the Inhomogeneous Lorentz Group', *Ann. Math.*, 40, 149 (1939).
11. G.C. Wick, E.P. Wigner, and A.S. Wightman, 'Intrinsic Parity of Elementary Particles', *Phys. Rev.* 88, 101 (1952).
12. E.P. Wigner, 'Über die Operation der Zeitumkehr in der Quanten-mechanik', *Gött. Nach.* 546-559 (1931). Here for the first time an anti-unitary symmetry appears.
13. E.P. Wigner, *Ann. Math.* 40, 149 (1939).
14. J. Schwinger, *Phys. Rev.* 82, 914 (1951).
15. J.S. Bell, 'Time Reversal in Field Theory', *Proc. Roy. Soc.* (London) A231, 479-495 (1955).
16. To the best of my knowledge, the CPT theorem was first proved by W. Pauli in his article 'Exclusion Principle, Lorentz Group and Reflection of Space-Time and Charge', in *Niels Bohr and the Development of Physics* [Pergamon Press, London, p. 30 (1955)], which in turn is

an extension of the work of J. Schwinger [*Phys. Rev.* 82, 914 (1951); 'The Theory of Quantized Fields. II.', *Phys. Rev.* 91, 713 (1953); 'The Theory of Quantized Fields. III.', *Phys. Rev.* 91, 728 (1953); '*The Theory of Quantized Fields. VI.*', *Phys. Rev.* 94, 1362 (1954)] and G. Lüders, 'On the Equivalence of Invariance under Time Reversal and under Particle-Anti-particle Conjugation for Relativistic Field Theories' [*Dansk. Mat. Fys. Medd.* 28, 5 (1954)], which referred to an unpublished remark by B. Zumino. The final contribution to the CPT theorem was given by R. Jost, in 'Eine Bemerkung zum CPT Theorem' [*Helv. Phys. Acta* 30, 409 (1957)], who showed that a weaker condition, called 'weak local commutativity' was sufficient for the validity of the CPT theorem.

17. 'Observation of Antiprotons', O. Chamberlain, E. Segrè, C. Wiegand, and T. Ypsilantis, *Physical Review* 100, 947 (1955).
18. 'Anti-Neutrons Produced from Anti-Protons in Charge Exchange Collisions', B. Cork, G.R. Lambertson, O. Piccioni, W.A. Wenzel, *Physical Review* 104, 1193 (1957).
19. 'Observation of Long-Lived Neutral V Particles', K. Lande, E.T. Booth, J. Impeduglia, L.M. Lederman, and W. Chinowski, *Physical Review* 103, 1901 (1956).
20. 'Remarks on Possible Noninvariance under Time Reversal and Charge Conjugation', T.D. Lee, R. Oehme, and C.N. Yang, *Physical Review* 106, 340 (1957).
21. 'Question of Parity Conservation in Weak Interactions', T.D. Lee and C.N. Yang, *Phys. Rev.* 104, 254 (1956).
22. 'Experimental Test of Parity Conservation in Beta Decay', C.S. Wu, E. Ambler, R.W. Hayward, D.D. Hoppes, *Phys. Rev.* 105, 1413 (1957); 'Observation of the Failure of Conservation of Parity and Charge Conjugation in Meson Decays: The Magnetic Moment of the Free Muon', R. Garwin, L. Lederman, and M. Weinrich, *Phys. Rev.* 105, 1415 (1957); 'Nuclear Emulsion Evidence for Parity Non-Conservation in the Decay Chain $\pi^+\mu^+e^+$', J.J. Friedman and V.L. Telegdi, *Phys. Rev.* 105, 1681 (1957).
23. 'On the Conservation Laws for Weak Interactions', L.D. Landau, *Zh. Éksp. Teor. Fiz.* 32, 405 (1957).
24. 'Evidence for the 2π Decay of the K_2^0 Meson', J. Christenson, J.W. Cronin, V.L. Fitch, and R. Turlay, *Physical Review Letters* 113, 138 (1964).
25. 'Experimental Observation of Anti-deuteron Production', T. Massam, Th. Muller, B. Righini, M. Schneegans, and A. Zichichi, *Nuovo Cimento* 39, 10 (1965).
26. *The Discovery of Nuclear Antimatter*, L. Maiani and R.A. Ricci (eds), Conference Proceedings 53, Italian Physical Society, Bologna, Italy (1995); see also A. Zichichi in *Subnuclear Physics – The first fifty years*, O. Barnabei, P. Pupillo and F. Roversi Monaco (eds), a joint publication by University and Academy of Sciences of Bologna, Italy (1998); World Scientific Series in 20th Century Physics, Vol. 24 (2000); see also 'Why antihydrogen and antimatter are different', A. Zichichi, *CERN Courier*, Vol. 49, n. 4, pp. 15-17, May (2009).
27. The first report on 'scaling' was presented by J.I. Friedman at the 14th International Conference on *High Energy Physics* in Vienna, 28 August-5 September 1968. The report was presented as paper n. 563 but not published in the Conference Proceedings. It was published as a SLAC preprint. The SLAC data on scaling were included in the Panofsky general report to the Conference where he says '... the apparent success of the parametrization of the cross-sections in the variable ν/q^2 in addition to the large cross-section itself is at least indicative that point-like interactions are becoming involved'. 'Low q^2 Electrodynamics, Elastic and Inelastic Electron (and Muon) Scattering', W.K.H. Panofsky in Proceedings of 14th Inter-

national Conference on *High Energy Physics* in Vienna 1968, J. Prentki and J. Steinberger (eds), page 23, published by CERN (1968). The following physicists participated in the inelastic electron scattering experiments: W.B. Atwood, E. Bloom, A. Bodek, M. Breidenbach, G. Buschhorn, R. Cottrell, D. Coward, H. DeStaebler, R. Ditzler, J. Drees, J. Elias, G. Hartmann, C. Jordan, M. Mestayer, G. Miller, L. Mo, H. Piel, J. Poucher, C. Prescott, M. Riordan, L. Rochester, D. Sherden, M. Sogard, S. Stein, D. Trines, and R. Verdier. For additional acknowledgements see J.I. Friedman, H.W. Kendall and R.E. Taylor, 'Deep Inelastic Scattering: Acknowledgements', *Les Prix Nobel 1990*, (Almqvist and Wiksell, Stockholm/Uppsala 1991), also *Rev. Mod. Phys.* 63, 629 (1991). For a detailed reconstruction of the events see J.I. Friedman, 'Deep Inelastic Scattering Evidence for the Reality of Quarks' in *History of Original Ideas and Basic Discoveries in Particle Physics*, H.B. Newman and T. Ypsilantis (eds), Plenum Press, New York and London, 725 (1994).

28. *Quark Search at the ISR*, T. Massam and A. Zichichi, *CERN preprint*, June 1968; 'Search for Fractionally Charged Particles Produced in Proton-Proton Collisions at the Highest ISR Energy', M. Basile, G. Cara Romeo, L. Cifarelli, P. Giusti, T. Massam, F. Palmonari, G. Valenti and A. Zichichi, *Nuovo Cimento* 40A, 41 (1977); and *A Search for quarks in the CERN SPS Neutrino Beam,* M. Basile, G. Cara Romeo, L. Cifarelli, A. Contin, G. D'Alì, P. Giusti, T. Massam, F. Palmonari, G. Sartorelli, G. Valenti and A. Zichichi, *Nuovo Cimento* 45A, 281 (1978).
29. A. Zichichi in *Subnuclear Physics – The first fifty years*, O. Barnabei, P. Pupillo and F. Roversi Monaco (eds), a joint publication by University and Academy of Sciences of Bologna, Italy (1998); World Scientific Series in 20th Century Physics, Vol. 24 (2000).
30. 'New Developments in Elementary Particle Physics', A. Zichichi, *Rivista del Nuovo Cimento* 2, n. 14, 1 (1979). The statement on page 2 of this paper, 'Unification of all forces needs first a Supersymmetry. This can be broken later, thus generating the sequence of the various forces of nature as we observe them', was based on a work by A. Petermann and A. Zichichi in which the renormalization group running of the couplings using supersymmetry was studied with the result that the convergence of the three couplings improved. This work was not published, but perhaps known to a few. The statement quoted is the first instance in which it was pointed out that supersymmetry might play an important role in the convergence of the gauge couplings. In fact, the convergence of three straight lines ($\alpha_1^{-1}\,\alpha_2^{-1}\,\alpha_3^{-1}$) with a change in slope is guaranteed by the Euclidean geometry, as long as the point where the slope changes is tuned appropriately. What is incorrect about the convergence of the couplings is that, with the initial conditions given by the LEP results, the change in slope needs to be at $M_{SUSY} \sim 1$ TeV as claimed by some authors not aware in 1991 of what was known in 1979 to A. Petermann and A. Zichichi.
31. V.N. Gribov, G. 't Hooft, G. Veneziano and V.F. Weisskopf, *The Creation of Quantum ChromoDynamics and the Effective Energy*, L.N. Lipatov (ed.), a joint publication by the University and the Academy of Sciences of Bologna, Italy (1998); World Scientific Series in 20th Century Physics, Vol. 25 (2000).
32. 'The Effective Experimental Constraints on M_{SUSY} and M_{GUT}', F. Anselmo, L. Cifarelli, A. Petermann and A. Zichichi, *Nuovo Cimento* 104A, 1817 (1991).
33. 'The Simultaneous Evolution of Masses and Couplings: Consequences on Supersymmetry Spectra and Thresholds',

F. Anselmo, L. Cifarelli, A. Petermann and A. Zichichi, *Nuovo Cimento* 105A, 1179 (1992).
34. 'A Study of the Various Approaches to M_{GUT} and α_{GUT}', F. Anselmo, L. Cifarelli and A. Zichichi, *Nuovo Cimento* 105A, 1335 (1992).
35. 'Are Matter and Antimatter Symmetric?', T.D. Lee, in Proceedings of the *Symposium to celebrate the 30th anniversary of the Discovery of Nuclear Antimatter*, L. Maiani and R.A. Ricci (eds), Conference Proceedings 53, p. 1, Italian Physical Society, Bologna, Italy (1995).
36. 'String Theory: the Basic Ideas', B. Greene, Erice Lectures – Discussion 1999 in *Basics and Highlights in Fundamental Physics*, A. Zichichi (ed.), World Scientific (2001).
37. 'Search for Supersymmetric Particles using Acoplanar Charged Particle Pairs from Z^0 decays', ALEPH Collab., D. Decamp *et al.*, *Phys. Lett.* B236, 86 (1990).
38. 'Search for Neutral Higgs Bosons from Supersymmetry in Z decays', ALEPH Collab., D. Decamp *et al.*, *Phys. Lett.* B237, 291 (1990).
39. 'Search for Neutralino Production in Z decays', ALEPH Collab., D. Decamp *et al.*, *Phys. Lett.* B244, 541 (1990).
40. 'Search for the Neutral Higgs Bosons of the MSSM and other two Doublet Models', ALEPH Collab., D. Decamp *et al.*, *Phys. Lett.* B265, 475 (1991).
41. 'Search for Heavy Charged Scalars in Z^0 decays', DELPHI Collab., P. Abreu *et al.*, *Phys. Lett.* B241, 449 (1990).
42. 'Search for Pair Production of Neutral Higgs Bosons in Z^0 decays', DELPHI Collab., P. Abreu *et al.*, *Phys. Lett.* B245, 276 (1990).
43. 'Search for Scalar Quarks in Z^0 decays', DELPHI Collab., P. Abreu *et al.*, *Phys. Lett.* B247, 148 (1990).
44. 'A Search for Sleptons and Gauginos in Z^0 Decays', DELPHI Collab., P. Abreu *et al.*, *Phys. Lett.* B247, 157 (1990).
45. 'Mass Limits for Scalar Muons, Scalar Electrons and Winos from e^+e^- Collisions near S**(1/2)=91–GeV', L3 Collab., B. Adeva *et al.*, *Phys. Lett.* B233, 530 (1989).
46. 'Search for the Neutral Higgs Bosons of the Minimal Supersymmetric Standard Model from Z^0 Decays', L3 Collab., B. Adeva *et al.*, *Phys. Lett.* B251, 311 (1990).
47. 'Search for the Charged Higgs Boson in Z^0 decay', L3 Collab., B. Adeva *et al.*, *Phys. Lett.* B252, 511 (1990).
48. 'A Search for Acoplanar Pairs of Leptons or Jets in Z^0 decays: Mass Limits on Supersymmetric Particles', OPAL Collab., M.Z. Akrawy *et al.*, *Phys. Lett.* B240, 261 (1990).
49. 'A Search for Technipions and Charged Higgs Bosons at LEP', OPAL Collab., M.Z. Akrawy *et al.*, *Phys. Lett.* B242, 299 (1990).
50. 'A Direct Search for Neutralino Production at LEP', OPAL Collab., M.Z. Akrawy *et al.*, *Phys. Lett.* B248, 211 (1990); P.D. Acton *et al.*, preprint CERN-PPE/91-115, 22 July 1991.
51. 'Searches for Supersymmetric Particles Produced in Z Boson decay', MARK II Collab., T. Barklow *et al.*, *Phys. Rev. Lett.* 64, 2984 (1990).
52. *Searches for New Particles at LEP*, M. Davier, LP-HEP 91 Conference, Geneva, CH, Preprint LAL 91-48, December 1991.
53. 'The Evolution of Gaugino Masses and the SUSY Threshold', F. Anselmo, L. Cifarelli, A. Peterman and A. Zichichi, *Nuovo Cimento* 105A, 581 (1992).
54. 'A Detailed Comparison of LEP Data with the Predictions of the Minimal Supersymmetric SU(5) GUT', J.R. Ellis, S. Kelley, D.V. Nanopoulos, preprint CERN-TH/6140-91, *Nucl. Phys.* B373, 55 (1992).

The Emergence of Order

Walter Thirring

Explanation

We study the mutation-selection dynamics. The dynamical variables are the p_i. They are positive and their sum is normalized to unity. The α_{ik} are parameters which determin the evolution. The p_i can be thought of probabilities of populations. Their randomness is given by the entropy S. The α reflects the accidental situation in which the system is embedded. The main question is whether they will lead to order (low S) or chaos (high S)

$$\frac{dp_i}{dt} = \sum_k \alpha_{ik} p_k - p_i \sum_{\substack{j,k \\ i,j,k = 1,2,\ldots,d}} \alpha_{jk} p_k$$

$$\sum_i p_i = 1$$

$$S = \sum_k p_k (1 - p_k)$$

It turns out that the p_i generically tend to a limit which is independent of their initial values and are determined by the α_{ik}. As first orientation we do not restrain the α and let them be random numbers within certain bounds. This leads to a distribution of the final entropies which depends only on the dimension d of the p_i space, ($i=1\ldots d$).

The next figure shows this distribution and we see that for each d the entropy clusters around the maximal value. There is no creation of order from disorder. Next we consider a hierarchical structure of the p_i so that the α tends to a triangular matrix. In this case order is created out of disorder.

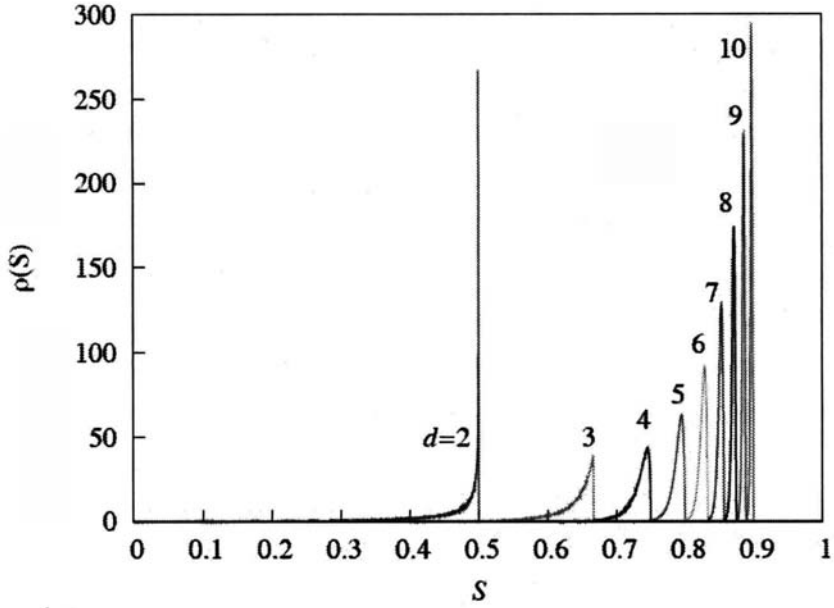
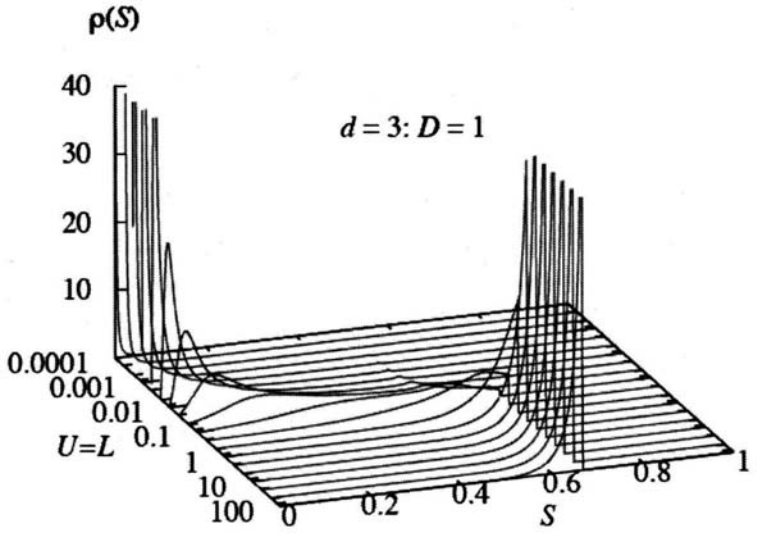


Figure 1.



Die Entropie als Funktion vom Verhältnis der Nichtdiogonalelemente U, L zu den Diagonalelementen D. Die Schranken für erstere wurden gleich gewählt.

Figure 2.

Such a behaviour can be easily understood by interpreting the p_i financially of the assets of d players. The dynamically equation solves the payments of player i to player k and the order of the indices i reflect the richness of player i. Triangularity of the α means that the payments mainly go from the poor to the rich. Therefore eventually the latter end up having all the money. It is interesting to see that this situation changes already if the non-triangularity reaches a few percent.

The Laser and How it Happened

Charles H. Townes

I'm going to discuss the history of the laser and my own personal participation in it. It will be a very personal story. On the other hand, I want to use it as an illustration of how science develops, how new ideas occur, and so on. I think there are some important issues there that we need to recognize clearly.

How do new discoveries really happen? Well, some of them completely by accident. For example, I was at Bell Telephone Laboratories when the transistor was discovered and how? Walter Brattain was making measurements of copper oxide on copper, making electrical measurements, and he got some puzzling things he didn't understand, so he went to John Bardeen, a theorist, and said, 'What in the world is going on here?' John Bardeen studied it a little bit and said, 'Hey, you've got amplification, wow!'. Well, their boss was Bill Shockley, and Bill Shockley immediately jumped into the business and added a little bit. They published separate papers but got the Nobel Prize together for discovering the transistor by accident.

Another accidental discovery of importance was of a former student of mine, Arno Penzias. I'd assigned him the job of looking for hydrogen in outer space using radio frequencies. I'd been doing radio spectroscopy and I thought, well, there's a chance of maybe finding hydrogen out there with radio waves, so he looked. He didn't find hydrogen but he did a good job, got his PhD, went on to Bell Telephone Laboratories and there joined up with Bob Wilson and they continued to look. They were using a maser amplifier, which was the most sensitive amplifier available – and it still is. They didn't find hydrogen but found a low intensity continuous radiation coming in from all directions, radio radiation coming in from all directions in the microwave region. What was that? Well, they looked at it and talked to other people and published it, and people recognized that it was a residual of the original Big Bang of the Universe, the first real demonstration that yes, there was a Big Bang. The Universe did have a beginning and this was the discovery of the beginning of the Universe by accident.

Many discoveries happened that way but some are the result of a steady, directed effort and the latter is the case of the laser. In both cases it is necessary to use great care, thoroughness and intensity. You see, Walter Brattain was doing very careful work and so was Arno Penzias, and they made great accidental discoveries. In the case of the laser, it came about as a very systematic effort on my part. Why did I want to do this? Well, in 1939 I got my PhD and I

wanted to do physics research at the university, but there were no interesting university jobs at that time, so I took a job at Bell Telephone Laboratories. They initially let me do some physics research but pretty soon the war was coming on so they said, 'We want you to start building a radar system'. Oh dear, I had to become an engineer and design a radar system! They wanted a radar system with shorter wavelengths than anything they had, 1¼ cm – about half an inch – wavelength. Well, OK, so I had to do that and learned a lot of engineering, which has become very valuable to me. However, when we almost finished building it, we discovered that wavelength is absorbed by water vapor in the atmosphere. Oh dear, the waves wouldn't get through the atmosphere, so we had to discard the whole thing. As a result of that I decided maybe I would try to study and check out this water vapor, so in the laboratory I made measurements of water vapor absorption at this wavelength. I recognized then a new kind of spectroscopy in the radio region, very precise. We had very narrow bandwidths, with very precise frequencies, and so I started studying other molecules, including ammonia and so on. Bell Laboratories let me do that and particularly after the war I could stop engineering and do physics. I studied molecules and got great precision not only about molecular structure, but also the nature of the nuclei in the molecules, their spins and shapes. I found I could measure how the nuclei differed from sphericity, for example, looking at the spectra. I published and that became important physics, important enough that I was offered a job at Columbia University to continue to do such work. OK, well, that's great. So I got to Columbia University and continued to work. I recognized, though, that I really wanted to get on down to shorter wavelengths. Now, electronics could, at that time, produce fairly shorter wavelengths, down to about 2 or 3 mm but not much shorter. I wanted to get on down to still shorter wavelengths, maybe down to the infrared – that is below 1 mm as we define infrared. I had my students try various electronic things, but they didn't work. Now, the Navy knew that I was interested and the Navy wanted to get to shorter wavelengths too, so they asked me to form a national committee to find some way to get to shorter wavelengths. I formed a national committee and got a lot of important scientists and engineers together. We travelled all over the country visiting various laboratories and talking with people. After a year's time we hadn't found any answers about how to produce shorter wavelengths. At our last meeting in Washington D.C. we had to write a report saying sorry, we hadn't found anything. I woke up early in the morning worrying about it. I went out and sat on a park bench on a lovely sunny morning and I thought, 'Why haven't we been getting any ideas? Now, what can possibly do this?' I thought, well, of course molecules and atoms can produce short waves but, of course, thermo-

dynamics says they can't produce more than a certain intensity. Intensity depends on the temperature at which you have the molecules and atoms, so you can't get much intensity. Wait a minute! Wait a minute! They don't have to obey thermodynamics. We can get molecules in states that don't follow thermodynamics. Thermodynamics says you have to have more molecules in a lower state than in an upper state, so there's more absorption than emission and that limits the total emission that they can produce. The upper states drop down and give energy, the lower states, of course, absorb energy. Hey, wait a minute, we can pick out molecules mostly in the upper state. If we get enough of them in the upper state they can amplify, because they will all drop down and nothing absorbs, so they'll all emit radiation.

Now, at Columbia, Professor I.I. Rabi had, for some time, been doing molecular and atomic beam work where he separated states of molecules and atoms by deflecting them in a beam. The beam was deflected by electromagnetic fields so you can pick out various states, and I recognized that that was one way I could do it. I persuaded Jim Gordon, a student at Columbia, to do this for his thesis. We worked on ammonia molecules because I thought I should do it in the microwave region first – I wanted to get into the infrared, but I thought I'd do it in the microwave region first because I had a lot of microwave equipment and that was the simplest thing to do. So Gordon and I worked on it. We were building equipment to try to send the ammonia molecules in a beam, deflect them so that the high energy ones could be focused into a cavity and the low energy ones could be thrown away. Well, Gordon was working on this for a couple of years. But then Professor Rabi, who was the former chairman of the department, and Professor Kusch, who was then chairman – and they were both excellent physicists, they got Nobel Prizes – came into my laboratory and said, 'Look Charlie, you've got to stop! That's not going to work, we know it's not going to work, you know it's not going to work, you're wasting the department's money, you've got to stop!' Fortunately, in a university, a professor can't be fired just because he's stupid. He can be fired if he does something morally wrong but not simply if he's stupid and has done something scientifically wrong, so I knew they couldn't fire me and I said, 'No, I think it has a reasonable chance of working, I'm going to continue'. So they marched out of my lab, angrily. Well, we kept going and about two months later Jim Gordon came to my classroom and said, 'Hey, it's working!' Well, all my students and I went to the laboratory to see this thing that was working. We were sending enough molecules into the cavity, they emitted some radiation, the radiation bounced back and forth and stimulated the other molecules to give up more energy and so it produced an oscillation. This

oscillation was a very, very pure frequency and wow, that was exciting and a lot of people got interested. I was due to have a sabbatical leave, so I took a sabbatical leave and went to Paris. In Europe I knew Aage Bohr, who had been at Columbia with me (he died recently, his father was Niels Bohr, a very famous physicist and both of them got Nobel Prizes). So I went to visit Aage Bohr and I was walking along the street with Niels Bohr and he asked me what I was doing. I told him we had this oscillator, giving very pure frequency from molecules. 'Oh', he said, 'No, no that's not possible'. I said, 'Well, we've got it'. And he said, 'No, you've misunderstood something, no, no'. He just wouldn't talk to me about it, 'No, no, you're wrong'. Why was that? I suspect he was thinking of the uncertainty principle. You send a molecule through a cavity, and if you try to measure its frequency, the uncertainty principle says that the frequency can be measured only with an accuracy of one over the time that it passes through the cavity, and that's pretty short. That's the uncertainty principle. Bohr was sure it wouldn't work and that it didn't give such pure frequencies. He didn't recognize I was using a big collection of molecules and I had feedback and so on. Any engineer recognizes that feedback amplifiers, or feedback oscillators, can give very pure frequencies. Any engineer knows that but Bohr didn't recognize this and he just shut me up. He wouldn't listen.

There was also John von Neumann, a very famous mathematical physicist. I ran into him at a cocktail party and he asked me what I was doing. I told him we had this very pure frequency and he said, 'No, that's not possible, you're doing something wrong, you haven't measured it right, you misunderstand'. 'No, I've got it!' 'No, no, no'. Well, he went off to get another cocktail. About 15 minutes later he came back and said, 'Hey, you're right, tell me more about it!' Somehow he had suddenly woken up to the idea. He was a little bit more of an engineer than Bohr was. Well, you see, getting engineering and physics together was important. As I said, the field became very exciting with a lot of people working on it and it grew and grew. I had gotten the first one going in 1954 and after about a year and a half I said, 'Well, I really want to get on down to the shorter wavelengths. Let me see how I'm going to do that, I want to sit down and see just how far down we can get in wavelength'. I sat down and wrote some notes and equations and hey, it looked like I could get right on down to light waves. Wow, right on down to light waves, oh boy! Well, the field was such a hot field then I knew I shouldn't say anything about it because a lot of people would immediately compete with me. When I was building the maser, absolutely nobody competed. People would come by and say, 'Oh that's an interesting idea', but nobody competed. The only other people doing this were the

Russians, Nikolay Basov and Alexander Prokhorov, who had an independent idea and I didn't know they were working on it and they didn't know I was working on it. They didn't actually get one going first but they got the Nobel Prize with me for generally thinking of the idea. So nobody else was interested until it got going, then everybody was interested and it was very competitive. So I decided, well, let me see what I can do first and publish something, rather than saying anything to anybody and have them immediately compete with me.

I was a consultant at Bell Telephone Laboratories and Arthur Schawlow – who had been a post doc with me and married my kid sister, which I was very pleased about – was working at Bell Telephone Laboratories. I went and talked with him, told him about it, and he said, 'Well, you know, I've been wondering about that, can I work on this with you?' I said, 'Well, sure, OK', so he worked on it with me and he added an idea which was important. I was going to send the molecules into the cavity and then the light would bounce around the cavity in all directions. But with two parallel mirrors as suggested by Art Schawlow, light would bounce back and forth only in one direction and produce a beam, a nice beam. He had that idea and he added it so we decided, 'Well, this ought to be patented but I guess we probably ought to give the patent to Bell Labs. Let's take this idea to Bell Laboratories' lawyers and have them patent it and then we'll publish a paper about it'. So he went to Bell Laboratories' lawyers, but he called me back a couple of days later and said, 'Well, the lawyers told me they're not interested in patenting that, because they say that light has never been used for communication and it wouldn't be interesting for Bell Labs, so if we want to patent it they will just give us the patent'. I said, 'Well, it just shows they don't understand. Of course it can be used for communication, they don't understand, you go back and tell them. We shouldn't take the patent away from them just because their lawyers don't understand, so you go back and tell them, yes, it can be used for communication'. The lawyers then responded, 'Well, if you can show us how we can use it for communication then OK, we will patent it for Bell Labs'. So we did that and wrote a patent, entitled *Optical Masers and Communication*. Now, we had named this original thing the maser. My students and I sat down and said, what shall we name it, and we *named* it maser for Microwave Amplification by Stimulated Emission of Radiation, MASER. That was the maser. It was an original thing and the maser became a very popular name, so Art and I wrote the patent and we called it an *Optical Maser and Communication* for the patent. The lawyers went ahead and patented it then, because we showed them how it could be used for communication, which was obvious to us.

Well, now we were going to publish, I knew if we started trying to do the experiment and make one then we would have a lot of competition and somebody else might beat us to it anyhow, so we'd better publish the theory showing how it could be done. So Art Schawlow and I wrote a paper saying how it could be done, calling it *Optical Maser*, and we published. And then everybody got interested and wow, there was a lot of competition. Everybody jumped into the field. Yes, and at that time also I had been asked to go down to Washington to head a big scientific group to advise the government. I felt, well, that's a kind of a public duty, I should probably do it, so I went down and I was vice president of this group of scientists to advise the government. I agreed to go down for two years, hence couldn't work very well on trying to build the first laser. My students were working on it and I was hoping they would get along pretty fast, but the actual first one was built by Theodore Maiman at Hughes. Now, Ted had read our paper, with a lot of other people, and everyone jumped in the field trying to build one and he built the first system. The name was pretty quickly changed from optical maser to laser, for *Light Amplification by Stimulated Emission of Radiation*. We just called it originally an optical maser but that was too long a term, so laser became the name for it and now the term is used for any wavelength shorter than 1 mm. Anything longer than 1 mm is still a maser, so we have masers and lasers. They are basically the same thing, just different wavelengths. Well, so Ted Maiman built the first one and then Javan, one of my former students who was at Bell Telephone Laboratories, built the next one with some other people working with him, William Bennet and Donald Herriot. The next one was built by another student of mine at General Electric, Mirek Stevenson, who was working with a guy named Peter Sorokin. All the first lasers were built in industry, not by universities. Why? Because they could concentrate, work hard, and furthermore, because of the maser, industry had gotten interested in the field and had hired a lot of students from many different universities working in that field, so they had young people who were interested and knew the field. Thus all the first lasers were built in industry. As they were built, there was much growth in the field and a lot of people contributed different things. We had originally thought of lasers being produced in a gaseous discharge. But now they are produced by solids and in all kinds of ways, some of them are very small and some are very large. And by now there are lots of scientific things that have been done with lasers and masers. There have been thirteen Nobel Prizes in addition to the original prize for the invention – thirteen Nobel Prizes based on the use of masers or lasers as a scientific tool, so it has produced an enormous amount of science for which

I'm very pleased. They provide very high precision frequency as atomic clocks, they can measure distances very, very precisely. Even long distances can be measured precisely; we have measured the distance to the moon to about 1 cm accuracy (half an inch accuracy in the distance to the moon).

Laser beams also produce great directivity. So there's a lot of science and I'm just delighted to see all the good science that has been produced, as well as a lot of industry. Now, you see, the field I was working in, microwave spectroscopy, was not of interest to industry. They said, 'That's just physics and won't pay off', but it did pay off and that's typical of science, of course. New ideas come along and every once in a while there are lots of commercial applications. Well, the laser is now producing industry of some tens of billions of dollars a year at least. There is lots of industry. Lasers come big and small and do all kinds of things in industry. There are a lot of medical applications. Light now is a fantastic communicator. It has such a wide bandwidth, you can get about a billion channels of communication on one beam of light. It's just changed communication enormously. It's used in the military. It's not powerful enough to produce a weapon exactly, but it can direct weapons and it's used for pointers. It's also used for cutting and welding and manufacturing, all kinds of things, both very large and very small.

The biggest laser now is what's called the National Ignition Facility. It's built in order to ignite uranium to make uranium energy, to produce nuclear energy by shining intense light on these nuclei and allowing them to burn. This light intensity produces the highest temperature anyone has ever achieved. The NIF, National Ignition Facility, built by a laboratory of the University of California, has 192 laser beams. It's about 30 m high, 192 laser beams all focused together into a diameter as small as about 1.5 microns, just a couple of light wavelengths in size, and the total energy going in is 600 thousand billion watts. Just think of the temperature that represents, higher than any temperature anybody's ever achieved before. Not only do lasers provide very high temperatures, they also produce the lowest possible temperatures ever achieved. And these low temperatures produced one of the Nobel Prizes, for somebody who achieved extremely low temperatures with lasers. Thus lasers have an enormous variety of applications scientifically and commercially and, again, as I say, I'm delighted. Just think about the Bell Laboratories' lawyers, how they didn't think it would have commercial applications, certainly not communications applications.

Everybody was initially surprised about the possibility of masers producing lightwaves. However, once we published the paper on the possibility of doing this, then a lot of people jumped into the field, including industry, and recognized some of the applications, but not all of them. I recognized a lot

of the applications but one I missed completely: I didn't realize the medical applications, but those have been very widespread. Now, let's look at this. It is fairly typical in the development of science – it is now a very big science, it's very big industrially, it was completely ignored by industry initially – not a field of interest – but it's become very important. Well, we must be open to new ideas. Also, note that somebody tried to stop me, even important physicists tried to stop me. Industry wasn't interested at all, important physicists told me it wasn't going to work, and even after I had it going some important physicists told me *no*, *that's impossible*, *that's crazy*, *you don't understand*, *you've done something wrong*, *you don't know what you're doing*. New ideas are new, we've got to be open to new ideas and encourage people to explore new things, even the things that we're not very sure are going to work, or we think won't work, but it's good to explore. Another thing to remember is unpredictability. We frequently can't predict new things. And so we must again allow people to stick to new possibilities, explore new things, because we don't know where we are and what we're missing. In fact, all the scientific information needed for lasers was recognized as early as about 1920. We knew all the physics involved by as early as 1920. But the only possible suggestion of this, before we made things work, was about 1922. Richard Tolman, a theoretical physicist, was writing about quantum electronics and the excitation of atoms and said, well, of course, if you had more atoms in the upper state than the lower state then they would increase the energy of the waves a bit, but he went on to say that it would probably be very small, and this was just in the middle of a paragraph where he was discussing the theory of it all, so nobody paid any attention. He didn't pay any attention to it and nobody said anything more about it from 1922 until we had the idea and got it going, and so the first laser was built in 1960, about 40 years later. Humans wasted 40 years because we didn't use the theory that was there. The basic theory was all understood but nobody applied it, nobody explored it, nor recognized it.

So what are we missing now? Let's think about it, let's explore, be open-minded, encourage young people to look at new things. That, as I say, is kind of a history of how science develops. Many, many people have contributed to this idea, many people have done things which I didn't imagine initially, and they've added on things. That is what allows science to grow. We all work together, and now I want to emphasize interaction among scientists.

When I went to Paris on a sabbatical leave – I mentioned Niels Bohr, and so on – I ran into one of my former students who was there and was working on electronic spins. He found that, if he put some of them up in energy in a magnetic field, they would stay up for some time. I said, 'Hey,

wait a minute, maybe we can make a maser using electron spins in the upper state. We can amplify and we can tune it, and thus get a tunable amplifier with electron spins'. So we published a little paper about that possibility. Then Nicolaas Bloembergen, who was at Harvard, read this and had a still better idea. Because he had been working with electron spins, with two of them joined together so they could have three energy states – an upper level, a middle level and a lower level – he recognized that you could pump from the lower level to the upper level and then fall down to the middle level and get amplification. So that produced the first really good maser amplifier. Then I went to Japan after that and I continued my sabbatical leave. I ran into a biologist that I had known at Columbia, Francis Ryan, and I said, 'What are you doing?' And he said, 'You know, I've been trying to work out a theory of the fluctuations in the populations of microorganisms. Microorganisms can die or they can double and multiply and I'm trying to figure out the equation of how they vary in numbers'. And I said, 'Wait a minute, that's just what I want!' To get the fluctuations in a maser amplifier I had to add one term, namely spontaneous emission – stimulated emission is like birth of a new microorganism, not just splitting, but you can automatically produce the birth of a new photon by just emission – and so I added one term to the equation and we worked out the answers to that equation. That allowed me to provide a theory of the fluctuations of the maser amplifier and oscillator, as well as fluctuations in the number of microorganisms. That is another example of interaction between science and scientists, you see. These are just examples of the importance of interactions of different fields and different scientists. We must talk together, we must let fields interact. The engineering experience I had was enormously important in producing a new field and now there are more and more people who perhaps know the whole thing about lasers but what is it we are missing now. We missed lasers for forty years, what are we missing now? We must be open minded, encourage new ideas, encourage exploration and hopefully produce a lot of new interesting results, scientifically and economically. Thank you.

Session III: Earth and Environment Sciences

Quantifying the potential impacts of climate change on vegetation diversity at large spatial scales

Megan Konar, Ignacio Rodríguez-Iturbe

1. Introduction

Climate change is likely to be the most significant threat to biodiversity worldwide after 2050 (Strengers *et al.*, 2004). For this reason, quantification of the potential impacts of climate change on biodiversity is urgently needed (Sala *et al.*, 2000; Clark *et al.*, 2001; Botkin *et al.*, 2007). The various features associated with climate change (e.g. temperature, precipitation patterns, CO_2 concentrations, sea level rise, etc.) will likely impact different species in unique and unpredictable ways, making it particularly challenging to model.

It is important to consider biodiversity at the appropriate spatial scale when studying the impact of climate change, since projections of environmental variables under climate change are typically provided as large spatial scales (*Intergovernmental Panel on Climate Change*, 2007). Biodiversity is scale-dependent. In fact, one of the oldest and most well documented patterns in community ecology is the species-area curve, which describes the observed increase in species richness as area increases (Preston, 1962; Rosenzweig, 1995). This relationship has long fascinated ecologists, leading to an extensive literature devoted to the scale dependence of diversity patterns (Currie, 1991; Crawley and Harral, 2001; Hui, 2009). While the increase in the number of species with area is a widely recognized empirical phenomenon, the mechanisms driving this observed relationship are still widely debated in the literature. Since biodiversity is scale-dependent, the spatial scale must be appropriate when coupling biodiversity and climate change models. For this reason, we focus on quantifying the impact of climate change on biodiversity at large spatial scales in this paper.

In this paper, we highlight some recent efforts to quantify the potential impacts of climate change on biodiversity, with a particular emphasis on vegetation driven by hydrologic variables. We focus on the diversity of vegetation in two very different ecosystems. The first is the Mississippi-Missouri River System (MMRS), the largest watershed in North America, comprising 2,980,000 km^2, approximately 40% of the surface area of the continental United States. The second is the Everglades National Park (ENP), encompassing nearly 5,700 km^2, which is comprised of a mosaic of different vege-

tation communities. Hydrology has long been recognized as a driving feature in wetland systems and numerous studies have demonstrated a relationship between hydro-patterns and vegetation communities in the Everglades (Ross *et al.*, 2003; Armentano *et al.*, 2006; Zweig and Kitchens, 2008, 2009). However, the recognition of hydrology as a key driver of vegetation diversity in the MMRS has only recently been shown (Konar *et al.*, 2010).

2. Modeling biodiversity patterns

Many modeling efforts are currently underway to understand and predict the loss of biodiversity. In this paper, we utilize two different, yet complementary, approaches to model vegetation diversity at large spatial scales. For the ENP, we develop a community-distribution model, in which vegetation communities are correlated with hydrological regimes (Todd *et al.*, 2010). Projections of hydrologic variables in the ENP, as given by global climate models, are then used to obtain projections of vegetation communities, assuming that the relationship between vegetation communities and their hydrological niche remains constant in the future (Todd *et al.*, 2011). In the MMRS, we utilize a neutral meta-community model, based on population dynamics, with precipitation as a key driver. Precipitation values are obtained for future scenarios from global climate models, and the impacts on tree diversity patterns are quantified (Konar *et al.*, 2010).

2.1. Vegetation communities in the Everglades National Park

The Everglades National Park (ENP) (shown in Fig. 1, p. 364) encompasses nearly 5,700 km^2 and is a mosaic of different vegetation communities (Gunderson and Loftus, 1993). In total, the park has at least 830 vegetation taxa and includes all of the major habitats found within the larger Everglades ecosystem (Avery and Loope, 1983). Prior to the 1900s, the Everglades was a broad, slowly flowing wetland, originating in Lake Okeechobee and flowing south to the Gulf of Mexico. Flow velocities are often less than 1 cm s^{-1} due to the low slope (3 cm km^{-1}) and vegetative interference. Today, the Everglades is a hydrologically altered landscape due to human action and drainage, with flow controlled through an extensive system of levees, pumps, and canals. Even the ENP, designated as a national park, is impacted by human modification to the hydrology. In this section, we briefly describe the community-distribution model of vegetation in the ENP. The interested reader is referred to Todd *et al.* (2010) for additional details.

The Everglades Depth Estimation Network (EDEN) was used to obtain information on hydrological characteristics across the ENP. Namely, this data

set provides daily water level information for the entire freshwater portion of the Everglades. EDEN data is provided at the scale of 400 m × 400 m, based on over 250 monitoring wells, and covers the entire ENP and beyond. We used this information to calculate the number of hydroperiods in a year, the conditional mean depth of each hydroperiod, the mean duration of a hydroperiod, and the percentage of time inundated. For this analysis, we define a hydroperiod as an individual inundation episode. Our calculations are based on the EDEN data from 2000-2007.

Vegetation data was taken from the Center for Remote Sensing and Mapping Science at the University of Georgia and the South Florida Natural Resources Center (Welch and Madden, 1999). In this study, a 20 m × 20 m grid was laid over the ENP study area, for which the dominant vegetation type was extracted, producing over 5 million vegetation pixels. Since the vegetation and hydrology data are provided as difference scales, a hydrology pixel encompasses 400 vegetation pixels. There are 52 plant communities in the ENP provided by the vegetation database, though 13 vegetation communities comprise greater than 1% of the landscape.

The relationship between a vegetation community and the four hydrological variables was evaluated by extracting all pixels with the same dominant vegetation type and then creating histograms of the hydrologic measures. This allows us to differentiate the vegetation communities based upon their hydrological niches. Plotting the distribution of a vegetation community for a particular hydrologic measure allows us to determine where that community is disproportionately represented. From Fig. 2a (p. 364) it is clear that Muhly Grass was predominantly found in drier locations with a mean depth less than 14 cm that were inundated less than 54% of the time. Bay-Hardwood Scrub, on the other hand, tended to be found in wetter locations, with a clear preference for locations that were most constantly inundated (refer to Fig. 2b, p. 364), while Sawgrass, which is the most abundant vegetation type in the ENP by an order of magnitude, demonstrated indifference to the amount of time that a site was inundated, but tended to be found less frequently at sites with a mean depth between 50 and 80 cm (refer to Fig. 2c, p. 364). Our finding that sawgrass is relatively tolerant to the percent time inundated, but more sensitive to the depth of inundation, is supported by previous studies (Gunderson, 1994).

We believe that this study provides a good representation of the linkages between vegetation and hydrological processes because of the large sample size (>5 million vegetation pixels), the use of mean hydrologic conditions over a long period of record (8 years), and the mapping of dominant vegetation type, rather than every community present, thereby limiting the

chance of a change throughout short periods of time. Fig. 2 (p. 364) supports the contention that many vegetation communities within the ENP are structured on hydrological gradients. While multiple factors are undoubtedly important in determining the presence of a particular vegetation type at a given location in a landscape as diverse and dynamics as the ENP, our results decidedly show that hydrological processes are indeed a major influence structuring vegetation communities. In particular, we found that the percent time inundated and the mean depth of inundation are the major discriminatory variables, supporting the findings of Gunderson (1994).

2.2. Tree species in the Mississippi-Missouri River System

The ecologist Richard Levins (1970) was the first to use the term 'metapopulation' to indicate a set of local populations within a larger system. Several models have applied this concept to the study of extinction processes (Hanski and Gaggiotti, 2004). Recently, metapopulation models, using neutral ecological dynamic, have been shown to accurately characterize large-scale biodiversity characteristics of both fish (Muneepeerakul *et al.*, 2008; Bertuzzo *et al.*, 2009) and trees (Konar *et al.*, 2010). In this section, we briefly describe the model used to characterize tree diversity in the Mississippi-Missouri River System (MMRS), shown in Fig. 3 (p. 365). For further detail, the interested reader is referred to Konar *et al.* (2010).

We implemented a neutral metacommunity model of tree diversity in the MMRS. The 824 DTAs of the MMRS were chosen to represent the local communities of the system. Occurrence data for 231 tree species was compiled for each DTA of the MMRS from the U.S. Forest Service Forest Inventory and Analysis Database. These data were then analyzed for two key biodiversity signatures. First, we consider the distribution of local species richness (LSR). LSR is simply the number of species found in a DTA. The spatial distribution of LSR in the MMRS is shown in Fig. 3 (p. 365), and its corresponding histogram is shown in Fig. 4 (p. 365). The frequency distribution of LSR is bimodal due to the environmental heterogeneity of the MMRS, where species-rich DTAs in the east contribute to the peak around 40-50 species, while those DTAs in the west make up the species-poor peak in the histogram. Second, we consider the species rank-occupancy, the number of DTAs in which a particular species is found as a function of its rank.

To model this system, each local community is assigned a tree habitat capacity (H), defined as the number of 'tree units' that are able to occupy each DTA. A tree unit can be thought of as a subpopulation of trees of the same species. A habitat capacity value is assigned to each DTA that is proportional to the forest cover of that DTA. This is because forest cover

is assumed to be the best determinant of the number of trees that are able to exist within a local community.

The model is based on key population dynamics: birth, death, dispersal, colonization, and diversification. Since the model is neutral, all processes implemented in the model are equivalent for all species. At each time step a randomly selected tree unit dies. Another tree unit is selected to occupy the newly available resources. With probability ν, the immigration rate, the empty spot will be occupied by a tree species that does not currently exist within the system; while, with probability $1-\nu$, the empty spot will be colonized by a species that already exists within the system.

The dispersal process determines how individuals move and how the empty spot will be colonized. Since neutral dynamics operate in the model, the probability that an empty spot is colonized by a certain species is dependent only on the relative abundance of the offspring of that species present at the empty location following the dispersal process.

Tree offspring move through the system based on the dispersal kernel, a mathematical representation of how individuals move. Here, two kernels are used to represent the movement of trees in the MMRS: one for colonization within the system (denoted by the subscript C) and a second for immigration into the system from outside (denoted by the subscript I). The colonization kernel is assumed to take the exponential form and uses the two-dimensional landscape structure: $K_{ij} = C_C \exp(-D_{ij}/\alpha_C)$, where K_{ij} is the fraction of tree offspring produced at DTA j that arrive at DTA i after dispersal; C_C is the normalization constant ($\Sigma_i K_{ij} = 1$); D_{ij} is the shortest distance between DTA i and j measured in 2D space; and α_C is the characteristic dispersal length of colonizing individuals. The immigration kernel allows trees to move across the system boundaries as they would in real life. Immigration across the MMRS boundaries is incorporated into the model by making ν_i, the immigration rate at DTA i, a function of distance to the system boundary and the habitat capacity of the associated boundary DTA, since it is reasonable that immigration would occur more frequently through hospitable environments. The immigration rate is thus calculated as: $\nu_i = C_I H_{bi} exp(-D_{bi}/\alpha_I)$, where H_{bi} and D_{bi} are the habitat capacity of the boundary DTA closest to DTA i and the distance between them, respectively; C_I is the normalization constant ($\Sigma_i \nu_i = \psi$), where ψ is the average number of immigrant species in one generation (defined as the period over which each tree unit dies once on average); and α_I is the characteristic distance travelled by immigrants.

As illustrated in Fig. 4 (p. 365), the model provides an excellent fit to the empirical patterns of tree diversity in the MMRS as well as its sub-regions.

Of key importance, this modeling approach allows for the direct linkage of large-scale biodiversity patterns to environmental forcings (i.e. precipitation). A common point of confusion in the use of neutral models is that they ignore environmental variation. However, we would like to stress that neutral models are able to capture the impact of changing environmental drivers. Individuals in neutral models respond to environmental changes; however, they do so in an equivalent manner.

3. Impacts of climate change

In the previous section, we showed that hydrology structures both vegetation communities and diversity patterns at the ecosystem scale in two very different environments, namely, the Everglades National Park and the Mississippi-Missouri River System. In this section, we briefly describe the potential impacts of climate change on vegetation in both systems. The interested reader is referred to Todd *et al.* (2011) and Konar *et al.* (2010) for additional description and results.

In the ENP, vegetation communities were shown to associate with different hydrological niches. By comparing a vegetation community's relative abundance at given depths and percent time inundated, relative to its system-wide abundance, we have shown that vegetation communities react differently to hydrologic conditions. For example, a community like Sawgrass is able to persist in a variety of hydrologic conditions, while the distribution of a community like Bay-Hardwood Scrub is more narrowly controlled by hydrologic environments. In order to determine the impact of climate change on these vegetation communities, we assume the relationship between the vegetation communities and hydrologic niche remains constant, and project these same hydrologic variables under climate change.

Using our computed changes in hydrologic class frequency and the developed vegetation-hydrology relationship, we predicted the percent cover of individual vegetation communities across the entire ENP. Here, we focus on the changes observed between present conditions and the high emissions climate change scenario, since all emissions scenarios showed a similar impact on vegetation community change. Community changes under the high emissions scenario showed the most extreme departures, so they are presented here for the 'worst-case' scenario.

Recall that there were 13 vegetation communities that individually comprise >1% of the ENP landscape under the current climate scenario. Under the high emissions scenario, this drops to 11 vegetation communities (refer to Table 1). Five communities that had percent coverage greater than 1%

Vegetation Type	Present	High	% Change
Sawgrass	60.68	55.21	-9.0
Mixed Gramminoids	6.55	8.82	34.7
Tall Sawgrass	5.80	2.24	-61.4
Muhly Grass	4.07	10.25	152.0
Spike Rush	2.98	1.38	-53.5
Red Mangrove Scrub	2.16	0.92	-57.4
Bayhead	1.72	0.83	-51.7
Pine Savanna	1.59	5.17	224.3
Willow Shrublands	1.47	1.36	-7.9
Dwarf Cypress	1.45	0.69	-52.1
Bay-Hardwood Scrub	1.44	0.49	-66.1
Brazilian Pepper	1.22	2.50	104.4
Cattail Marsh	1.09	0.29	-73.5
Slash Pine with Hardwoods	0.88	2.96	237.2
Hardwood Scrub	0.71	1.57	121.9
Subtropical Hardwood Forest	0.75	1.43	91.1

Table 1. Percent coverage of dominant vegetation types within Everglades National Park under the present and high emissions scenarios. The percent change of dominant vegetation types between the present and high emissions scenarios are also provided. Only those vegetation types constituting more than one percent of the total landscape are listed. Taken from Todd *et al.* (2011).

under present conditions fell below the 1% threshold (i.e. Red Mangrove Scrub, Bayhead, Dwarf Cypress, Bay-Hardwood Scrub, and Cattail Marsh), while three communities that represented less than 1% of the landscape under present conditions increased above this threshold under climate change (i.e. Slash Pine with Hardwoods, Hardwood Scrub, and Subtropical Hardwood Forest). Under climate change, Sawgrass remained the most dominant vegetation community, though its relative abundance decreased from 60.7% to 55.2%. Other communities showed large decreases in percent cover, such as Cattail Marsh, Bay-Hardwood Scrub, and Tall Sawgrass. In contrast, Slash Pine with Hardwoods, Pine Savanna, Muhly Grass, Hardwood Scrub, and Brazilian Pepper all showed large increases in abundance under climate change.

Thus, changes in the hydrologic landscape under the most extreme emissions scenario led to profound changes in the frequency and distribution of vegetation communities in the ENP. There was a net loss of two vegetation communities under climate change. Some vegetation communities declined under climate change, while some demonstrated a positive reaction to climate change. Specifically, communities that tend to prefer xeric con-

ditions became more numerous, whereas communities that prefer more hydric conditions became more scarce. One surprising finding was that the forecasted drier conditions may allow other vegetation communities to competivitely displace Sawgrass.

For the MMRS system, we showed that a neutral metacommunity model effectively reproduces several characteristic patterns of tree diversity simultaneously when coupled with an appropriate indicator of habitat capacity and dispersal kernel. It is important to highlight that a single climatic variable (i.e. mean annual precipitation, MAP) was used to represent the habitat capacity of trees. Establishing a functional relationship between forest cover and mean annual precipitation allows us to force the model with new values of habitat capacity under climate change and quantify changes in the tree diversity patterns. This is an important step in quantifying the potential impacts of climate change on biodiversity patterns.

Projections of MAP were used to obtain new values of habitat capacity for the 824 DTAs in the MMRS. Specifically, the mean annual precipitation from 2049-2099 was determined for 15 statistically downscaled climate projections from the Coupled Model Intercomparison Project 3 (CMIP3) for the A2 emissions path CMIP3 (2009). The A2 emissions path is the most extreme pathway given by the Intergovernmental Panel on Climate Change (2007). However, recent carbon dioxide emissions are above those in the A2 scenario, indicating that this scenario may be more conservative than initially though, though future emissions remain uncertain (Karl *et al.,* 2009).

A schematic of how new values of habitat capacity were calculated from projections of MAP is provided in Fig. 5 (p. 366). Potential forest cover under the current climate scenario is depicted by points 'A'. To obtain P values under the climate change scenarios, the projected MAP for DTA *i* is located on the graph and the new corresponding potential forest cover is noted. These new values of P are represented on Fig. 5 by points 'B'. This new value of potential forest cover was then used in the equation $H_i = C_H P_i I_i$ to calculate the habitat capacity of DTA *i* under climate change. Both I and C_H are assumed to remain constant under climate change. This ensures that any differences between model realization are due only to climate change.

With these resulting new habitat capacities, we determine how various climate change scenarios are projected to affect tree diversity patterns in the MMRS. Each of the 15 climate change scenarios given by CMIP3 was implemented in the model. Here, the results that pertain to the most dramatic lower (i.e. 'species-poor') and upper (i.e. 'species-rich') bounds in the biodiversity patterns are reported in Fig. 6 (p. 366). Note that the probability of any particular outcome in macrobiodiversity patterns is heavily reliant

on the probabilities associated with the projected precipitation patterns provided by the global climate models. For this reason, the patterns reported here should be interpreted as envelopes of plausible biodiversity scenarios, rather than as predictions of biodiversity outcome.

With the tree diversity patterns under the current climate as a benchmark (i.e. the black line in Fig. 6, p. 366), there is a decrease in the frequency of high diversity local communities and an increase in the frequency of low diversity local communities across all systems in the species-poor scenarios. Additionally, the peaks of the LSR histograms associated with the MMRS and all sub-regions shift leftward, i.e., in the species-poor direction. Of importance, the tail of the rank-occupancy curve exhibits the largest contraction, which is where rare species in the system are represented. In other words, rare species are likely to be disproportionately impacted under climate change, a finding shared with niche-based model Morin and Thuiller (2009).

Tree diversity patterns are impacted more under the species-poor scenarios than under the species-rich scenarios, with the exceptions of the North and Northwest sub-regions, where impacts are of comparable magnitudes under both scenarios. This is due to the changes in the habitat capacities of these regions under both scenarios, as DTAs in these regions are located on the increasing portion of the function (i.e. the blue points in Fig. 5, p. 366), such that increases to MAP translate to increased values of habitat capacity. This is not the case in the the South sub-regions, for example, where increases to MAP do not lead to increased values of habitat capacity, since the function saturates in this region (i.e. note the red points in Fig. 5, p. 366).

Although changes to MAP do not solely determine how the tree diversity patterns will be impacted, it is an important component. The species-poor and species-rich scenarios tend to correspond to those scenarios in which the MAP was among the lowest or the highest, respectively, for a given system. However, there are situations in which this is not the case, such as in the South sub-region, where CNRM-CM3 is classified as the species-poor scenario, even though the average MAP is lowest under the GFDL-CM2.0 model (refer to Table 2).

A map of projected changes to mean local species richness under the species-poor scenario is provided in Fig. 7 (p. 367). Note the decreasing trend in the percentage of species lost from West to East. However, DTAs west of 97.5°W are low-diversity, while those east of 97.5°W are species-rich (similar to the case of fish explored in the previous section). Thus, there is an increasing trend in the absolute number of species lost from West to East. The largest decrease in region-averaged LSR occurs in the South sub-region, where 6.3 species are projected to be lost on average.

Scenario	MMRS	North	Southwest	East	South	Northwest
Current	790.08	831.62	571.38	1177.27	1237.87	432.70
BCCR-BCM2.0	840.73	898.64	515.04	1360.49	1301.49	460.07
CGCM3.1 (T47)	901.58	963.37	584.28	1356.10	1370.95	527.48
CNRM-CM3	778.80	882.67	436.89	1295.90	**1176.41**	440.99
CSIRO-Mk3.0	853.54	918.88	581.37	1292.85	1279.78	489.99
GFDL-CM2.0	711.61	784.22	**382.52**	1203.40	*1062.83*	411.35
GISS-ER	899.59	1032.24	508.99	**1479.87**	1418.20	451.23
INM-CM3.0	709.41	777.85	467.90	*1069.25*	1058.09	412.90
IPSL-CM4	715.84	784.20	477.33	1089.84	1001.62	428.19
MIROC3.2	**654.41**	**684.62**	412.58	1040.21	994.41	**355.29**
ECHO-G	868.32	918.59	646.06	1283.78	1337.44	485.51
ECHAM5/MPI-OM	850.37	911.29	554.26	1339.91	1303.14	463.10
MRI-CGCM2.3.2	868.41	*965.67*	571.59	1316.05	1350.87	480.27
CCSM3	*890.18*	930.47	614.50	1396.40	1358.08	496.96
PCM	887.12	898.73	*644.60*	1335.32	1321.70	*526.24*
UKMO-HadCM3	795.37	824.48	488.05	1297.19	1231.82	437.27

Table 2. Mean annual precipitation (MAP) of the systems considered in this study for the current climate scenario and fifteen climate change scenarios. All values are in mm. Nomenclature of the climate change scenarios follows that of CMIP3. Numbers highlighted in bold indicate the species-poor climate change scenario for a given system, those in italics indicate the species-rich climate change scenario.

Thus, we have quantified the potential impacts of climate change, with hydrologic variables acting as the conduit, on vegetation diversity, both at the community and at the species level. Both models that we implemented are appropriate for use at large spatial scales, an important consideration for climate change impact analysis. One advantage of the neutral model is that it does not assume that the relationship between species and environmental variables remains constant in the future. However, a drawback to the neutral model, is that we are not able to directly map between species in the real world and those in the model, to determine how climate change will impact a particular species, as we are in the community distribution approach. Thus, these modeling approaches are complementary in nature to one another. Both approaches suggest that climate change may dramatically alter key diversity patterns at large spatial scales. These complementary analyses allow us to quantify the potential impacts of climate change on biodiversity, with far reaching implications for conservation biology, restoration efforts, and resource management.

References

Armentano, T., J. Sah, M. Ross, D. Jones, H. Cooley, and C. Smith (2006), Rapid responses of vegetation to hydrological changes in Taylor Slough, Everglades National Park, Florida, USA, *Hydrobiologia*, 569, 293-309, doi:10.1007/s10750-006-0138-8.

Avery, G., and L. Loope (1983), *Plants of the Everglades National Park: A preliminary checklist of vascular plants*, 2nd ed., U.S. Department of the Interior.

Bertuzzo, E.R., R. Muneepeerakul, H.J. Lynch, W.F. Fagan, I. Rodríguez-Iturbe, and A. Rinaldo (2009), On the geographic range of freshwater fish in river basins, *Water Resources Research*, 45.

Botkin, D.B., H. Saxe, M.B. Araujo, R. Betts, R.H.W. Bradshaw, T. Cedhagen, P. Chesson, T.P. Dawson, J.R. Etterson, D. P. Faith, S. Ferrier, A. Guisan, A.S. Hansen, D.W. Hilbert, C. Loehle, C. Margules, M. New, M.J. Sobel, and D.R. B. Stockwell (2007), Forecasting the effects of global warming on biodiversity, *BioScience*, 57 (3), 227-236.

Clark, J.S., S.R. Carpenter, M. Barber, S. Collins, A. Dobson, J.A. Foley, D.M. Lodge, M. Pascual, R.P. Jr., W. Pizer, C. Pringle, W.V. Reid, K.A. Rose, O. Sala, W.H. Schlesinger, D.H. Wall, and D. Wear (2001), Ecological forecasts: An emerging imperative, *Science*, 293, 657-660.

CMIP3 (2009), Statistically Downscaled WCRP CMIP3 Climate Projections, http://gdo-dcp.ucllnl.org/downscaled_cmip3_projections/dcpInterface.html.

Crawley, M.J., and J.E. Harral (2001), Scale dependence in plant biodiversity, *Science*, 291, 864-868.

Currie, D.J. (1991), Energy and large-scale patterns of animal and plant species richness, *The American Naturalist*, 137 (1), 27-49.

Gunderson, L. (1994), Vegetation of the Everglades: determinants of community composition. In: Davis S.M. and Ogden J.C., editors. *Everglades: the ecosystem and its restoration*, 323-340 pp., St. Lucie Press.

Gunderson, L., and W. Loftus (1993), The Everglades. In: Martin W.H., Boyce S.G., Echternacht A.C., editors. *Biodiversity of the Southeastern United States: Lowland terrestrial communities*, 199-255 pp., John Wiley & Sons, Inc.

Hanski, I., and O. Gaggiotti (2004), *Ecology, genetics, and evolution of metapopulations*, Elsevier Academic Press.

Hui, C. (2009), On the scaling patterns of species spatial distribution and association, *Journal of Theoretical Biology*, 261, 481-487.

Intergovernmental Panel on Climate Change (2007), *Climate Change 2007: The Physical Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press.

Karl, T.R., J.M. Melillo, and T.C. Peterson (2009), *Global climate change impacts in the United States*, Cambridge University Press.

Konar, M., R. Muneepeerakul, S. Azaele, E. Bertuzzo, A. Rinaldo, and I. Rodríguez-Iturbe (2010), Potential impacts of precipitation change on large-scale patterns of tree diversity, *Water Resour. Res.*, 46, W11,515.

Morin, X., and W. Thuiller (2009), Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change, *Ecology*, 90 (5), 1301-1313.

Muneepeerakul, R., E. Bertuzzo, H.J. Lynch, W.F. Fagan, A. Rinaldo, and I. Rodríguez-Iturbe (2008), Neutral metacommunity models predict fish diversity patterns in Mississippi-Missouri basin, *Nature*, 453, 220-222, doi:10.1038/nature06813.

Preston, F. (1962), The canonical distribution of commonness and rarity, *Ecology*, 43 (185-215), 410-432.

Rosenzweig, M.L. (1995), *Species diversity in space and time*, Cambridge University Press.

Ross, M., D. Reed, J. Sah, P. Ruiz, and M. Lewin (2003), Vegetation: environment relationships and water management in Shark Slough, Everglades National Park, *Wetlands Ecol. Manag.*, 11, 291-303.

Sala, O.E., F.S.C. III, J.J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sandwald, L.F. Huenneke, R.B. Jackson, A. Kinzig, R. Leemans, D.M. Lodge, H.A. Mooney, M. Oesterheld, N. L. Poff, M.T. Sykes, B.H. Walker, M. Walker, and D.H. Wall (2000), Global biodiversity scenarios for the year 2100, *Science*, 287, 1770-1774.

Strengers, B., R. Leemans, B. Eickhout, B. de Vries, and A. Bouwman (2004), The land-use projections and resulting emissions in the IPCC SRES scenarios as simulated by the image 2.2 model, *GeoJournal*, 61, 381-393.

Todd, M.J., R. Muneepeerakul, D. Pumo, S. Azaele, F. Miralles-Wilhelm, A. Rinaldo, and I. Rodríguez-Iturbe (2010), Hydrological drivers of wetland vegetation community distribution within Everglades National Park, Florida, *Advances in Water Resources*, 33, 1279-1289.

Todd, M.J., R. Muneepeerakul, F. Miralles-Wilhelm, A. Rinaldo, and I. Rodríguez-Iturbe (2011), Possible climate change impacts on the hydrological and vegetative character of Everglades National Park, Florida, *Ecohydrology*.

Welch, R., and M. Madden (1999), Vegetation map and digital database of South Florida's National Park Lands, final report to the U.S. Department of the Interior, National Park Service, Cooperative Agreement Number 5280-4-9006, Tech. rep., Center for Remote Sensing and Mapping Science, University of Georgia, Athens, GA.

Zweig, C., and W. Kitchens (2008), Effects of landscape gradients on wetland vegetation communities: information for large-scale restoration, *Wetlands*, 28, 1086-96.

Zweig, C., and W. Kitchens (2009), Multi-state succession in wetlands: a novel use of state and transition models, *Ecology*, 90, 1900-9.

The place of Man in nature
Epistemological notes

Jean-Michel Maldamé

An image haunts western conscience, which dates back to Plato and neoplatonic philosophy; it has been taken over by the Christian tradition and is present in all scientific works, from Aristotle to modern times. That image is that of the 'Great Chain of Beings'.[1]

The metaphysical conviction which presides over that image is linked to the notion of creation as conveyed by the neoplatonic scheme of emanation. Such a vision of creation is based on the conviction that the Creator is supremely kind, and that it is in the nature of kindness to give itself away and to show its richness in a total way. Theologians have made a principle of it: *bonum diffusivum sui*: the good shines forth its own goodness. So, the world is made distinctive by its plenitude and, for that reason, all manners of beings are found in nature. Each creature is the realisation of a well-defined perfection: a limited perfection, but which is part of a whole, of a perfection which brings together all perfections and harmonizes them.

This conviction has given rise to an image, the image of the great chain of beings. According to that image, all beings are hierarchized into a scale: at the lower end, one finds the *materia prima*, or shapeless matter, followed by the material beings which have a form, organised into a hierarchy from the simplest to the more complex. Then come plants, then animals, again organised into a hierarchy according to their complexity and richness. Then comes mankind, followed by angels, or spiritual beings; the culminating point is reached with the perfect being, a purely spiritual entity. It may be important to underline the fact that, in this scheme, the steps are contiguous,[2] and that there are intermediate beings,

[1] See Arthur O. Lovejoy, *The great Chain of Being*, Cambridge, Harvard University Press, 1936 & 1964.

[2] In the *Sum against Gentiles*, Thomas Aquinas wrote: 'if you look attentively, you will observe that there is a gradation in the diversity of beings. Above inanimate objects, are the plants. Above the latter, the animals, deprived of reason. Above them, intelligent substances. And at each stage, a diversity which makes some more perfect than others, so that the first of the beings at the bottom of the hierarchy is close to those at the top, and *vice versa*' (*Contra Gentiles*, III, 97).

which belong to both levels,[3] just as man participates of the riches of both matter and spirit.

This image falls within the province of a vision which modern science has challenged, because it is too simplistic. But the general scheme of thought has remained, by transforming the meaning of the image concerning the place of mankind in nature, since according to the image of the chain of beings, Man is both matter and spirit, at the top level of material beings, and at the bottom level of spiritual beings. In my paper, I shall use this image to discuss how today's science has reassessed the place of mankind in nature, chiefly because of the theory of evolution and neurosciences.

1. The Emergence of Mankind

From the end of the eighteenth century onward, the observation of a great multiplicity of different beings has prevented us from placing the species along one same line. It has been necessary to give up the idea of tracing a linear scale, where man would sit at the top, and propose another image which would assert the unity and the diversity of the living.

1.1. The tree of the living

For the inventory of the species getting ceaselessly richer and richer with the discovery of new species – in particular those coming from countries discovered and explored by Europeans – a new classification was needed.

Buffon had anticipated it, but it was Lamarck who realised it. Lamarck abandoned the linear series system. The wealth of observations was so great, that he had to imagine branchings. When he realised that there were huge differences between the vertebrate and the invertebrate, he separated them into two branches. He was then compelled to multiply the classes for the invertebrate. Starting with the two classes defined by Linnaeus (the insects and the worms), he came up with five classes in 1794 (mollusca, insects, worms,

[3] G.W. Leibniz: 'Since the law of continuity requires that when the essential attributes of one being approximate those of another all the properties of the one must likewise gradually approximate those of the other, it is necessary that all the orders of natural beings form but a single chain, in which the various classes, like so many rings, are so closely linked one to another that it is impossible for the senses or the imagination to determine precisely the point at which one ends and the next begins – all the species which, so to say, lie near to or upon the borderlands being equivocal, and endowed with characters which might equally be assigned to either of the neighbouring species', quoted by Arthur O. Lovejoy, *The great Chain of Being*, p. 145.

echinodermata, and polyps); then, in 1809, he was able to produce 10 classes. In 1815, in the introduction to his *Natural History of Invertebrate Animals,* Lamarck deepened his approach to the march of nature. He proposed a general distribution of animals according to their nervous systems, and at the same time a phylogenetic presentation where branchings and gaps are numerous. A linear image was quite out of the question. But there remained, of the idea of the great chain of beings, a gradual vision of nature, according to which the complexity of the organisation is on the increase, but opening into different branchings, according to criteria which are specific to each branch of the reconstituted arborescence. Thus, an entirely new vision of nature appeared, excluding two elements of the old vision: the hierarchisation, and the eminent place occupied by the human being in the world of the living.

For a long time it had been habitual to place mankind at the top of the modified tree. But things changed in the 20th century. This top position was challenged. As a matter of fact, the general outline of the tree makes it necessary to define a criterion for all the serialized elements. The classification may be done according to different factors: a greater aptitude for survival, a better ability to adapt, fecundity, sociability, numerous offspring, longevity... According to whether such or such a criterion has been chosen, the outline will be different. The resulting hierarchy has a different profile. What seemed to be first comes second. It follows that it is relativized. What applies to the usual forms of taxonomy becomes a prevailing factor where genes and their expression are concerned. In this new classification, the place of man varies in accordance with the chosen criterion. Scientific thought was faced with a new anthropological requirement, where the decisions concerning the place of mankind became the touchstone by which all the options chosen in the course of the research were revealed.

So, modern anthropology was born in the shade of palaeontology, concerned with noting the constituent elements of mankind. The quest for what is the essence of man remains the major challenge for thought, and on this point, an important change has occurred in these last few years. Studies in human palaeontology have brought a brilliant confirmation of Charles Darwin's conclusions in *The Descent of Man*, and confirmed the legitimacy of an approach to man within the framework of evolutionary thought. The multiplication of discoveries, however, has been marked by a situation which must be mentioned: it is paradoxical, because it leads the way for both certitudes and interrogations.

1.2. A problematical arborescence

In the best part of the 20th century, a theory progressively emerged. A compelling certainty was reached concerning the emergence of mankind.

There have been numerous discoveries of fossils in Western Africa. Their classification has obtained a certain amount of consensus among the scientific community, who agreed to recognize its value in tracing the prehistorical past of the modern man. It became acceptable to draw a genealogical tree starting in the African Rift; in this diverging development, the scientific books placed the separation between the human world, and the world of monkeys. History offered a certain coherence, when it spoke of *homo habilis* and *homo erectus*, following a large family of Australopitheci.

In the light of such views, anthropology has continued classification according to the criteria established by Linnaeus.[4] Works on human palaeontology marked out populations from specimens, joining to the word 'homo' adjectives often related to the places where the remains had been found: *homo sapiens*, *homo neanderthalis*, *homo heidelbergensis*, *homo antecessor*, *homo ergaster*, *homo habilis, homo rudolfensis*…

However, the outline of a continued arborescence remained uncertain. The most prudent scholars contented themselves with tracing lines in the forms of segments, in a tentative approach to a tree, without pronouncing themselves on the branchings. A sign of such a tension between two elements appears in the use of the words used for the classification of historical stages. One speaks of hominoids, hominids and the terms don't have the same meaning with different authors. One still stumbles on difficulties, when fossils are discovered which we don't know how to integrate into a lineage: such is the case of *Kenyanthropus platyops*, which is a singularity. It is then safer to keep to a classification which only speaks of *homo habilis*, *homo erectus*, *homo neanderthalis* and *homo sapiens*, in a presentation which accepts to be minimal.[5]

It is the same with the ancestors of the *homo* kind, in the classification of Australopitheci. When speaking of the australopithecus, we use qualifying adjuncts borrowed from the fossil world or from their morphological characteristics: *australopithecus habilis*, *australopithecus garhi*, *austropithecus rudolfensis, australopithecus bahreilghazali, australopithecus anamensis, australopithecus afarensis.* The dividing line is blurred, because in such an enumeration, one same qualifier (*habilis*, *rudolfensis*) is attributed to the *homo* genre and to the *australopithecus* genre. Same thing earlier in time when one finds 'ancestors' to

[4] It is a general principle. See Guillaume Lecointre et Hervé Le Guyader, *Classification phylogénétique du vivivant*, Paris, Berlin, 2001.

[5] See Éric Crubezy, José Braga, Georges Larrouy, *Anthropobiologie: Évolution humaine*, Paris, Elsevier-Masson, 2008.

australopitheci *ardipithecus ramidus* and also *paranthropus*. Here again, one stumbles on diverging interpretations.

This multiplication of viewpoints does not allow one to draw a continued genealogy with any amount of certainty. When it is done, the arborescence can take on several forms. It has now become habitual to draw segments which do not intersect. The use of parallel segments allows a view of contemporaneous populations, without marking the diverging points of the arborescence.

And yet, in scientific works, one still keeps trying to find a lineage of some sort: the debate aroused by the famous fossil named 'Toumai' is exemplary. The controversy about this ancestor which shows diverging characteristics proves that faithfulness to the founding principles of biology (*natura non facit saltum*) invites one to look for an essential point for the origins, in a science whose precariousness must be acknowledged.

1.3. A creative tension

It is important, in order to clarify this discussion, to note that whatever divergences exist, they result from the methods of analysis. First comes the morphological approach, resting on the structure of the bones which have been discovered and the anatomical characteristics which they enable one to infer. But fossils are disparate, incomplete, and raise problems of dating and interpretation. This is why another method soon imposed itself. It rests on molecular biology, since the genome in its linear sequence of nucleotides composing the DNA gives access to the totality of information on our biological heritage. Molecular data give access to a genealogical structure based on DNAmt (mitochondrial DNA). The genealogical study is then more precise and allows one to assert the unity of the *homo* genre with certainty for the closer periods in history (200,000 years backwards). But it is impossible to go further back. The genealogical tree of the modern man (*homo sapiens* or *homo sapiens sapiens*) is thus very difficult to outline.

From the presentation of this debate, there remains the fact that this way of placing mankind on the great tree of the living shows a conflict between two elements which are the key to the problem of deciding what the essence of man is.[6] Namely, on the one hand, to underline the insertion of

[6] The philosophical aspect of the question has been addressed by Jean-Marie Schaeffer, in *La Fin de l'exception humaine*, *nrf*-essais, Paris, Gallimard, 2007, and by Jean-Michel Maldamé, En quête du propre de l'homme, *Revue Thomiste*, Toulouse, 2009, n. 2, pp. 253-307.

mankind into the world of the living and, on the other, to show its irreducible originality towards the other forms of superior animals. It is quite clear, in fact, that mankind forms a specific ensemble. If it has roots in the animal world, and a real parenthood with the animals which are closer to it, it is also related to the history of life.

Not to stop at this abrupt observation, while remaining within the framework of the theory of evolution, it seems useful to pay attention to a phenomenon which Charles Darwin had already paid attention to. In the *Descent of Man*, Darwin remarked that the movement of evolution was not the same among humans, and among the animals that were the closer to them. As a matter of fact, human evolution implies a way of assuming the evolutive constraints which results in their displacement. Of course, mankind does not suppress them. But it can displace them as a counterbalancing effect. This is possible because of the plasticity of the laws of biology, and therefore funds the notion of culture.

2. Neurosciences

Darwin's remark rested on the observation of comportment. It has been given a scientific basis through another route, that of the investigation of the brain, which represents another great adventure of the scientific mind throughout the 20th century. From this point of view, the contribution of the neurosciences is not only medical or biological, it also plays an essential part in anthropology and allows a better understanding of the place of mankind in the world of the living.

2.1. The neurosciences

What we know about the brain is the result of a considerable progress in our ability to explore cerebral activities. In the first place came the discovery and the observation of the neuron, that singular cell of the human body. Then came the time of the exploration of functions, made possible through a better knowledge of networks, with the understanding of the connexions and interactions first between cells, then between networks. It was the foundation of the neurosciences strictly speaking, as a science that unified various disciplines.

In the early days, neurosciences have taken into account the study of comportment. Then, neuro imagery has given access to the observation of the life of the brain, doing away with the simplistic side of the early conclusions. It allows us, today, to follow the activity of the observed subject, and thus to link his cerebral condition to his activity. That was a considerable

progress. As a matter of fact, medical studies, in the beginning, were conducted starting from the examination of lesions and perturbations in movements, language and the expression of thought. The means of observation allow the scrutiny of subjects in their normal activity, and permit to establish series where observations make sense towards the establishing of a general anthropology.

2.2. The evolution of the brain

Neurosciences make it possible to understand that the processes of evolution elsewhere observed result in the apparition of a circular structure. In fact, evolution takes place through the commitment of the living to the world where they are supposed to act for survival. Such a commitment is a safeguard for patrimony, and at the same time a redeployment of aptitudes. Selection is then, in a certain way, canalised and even orientated. What is valid for life in general is also valid for mankind, whose plasticity makes it possible to develop the fundamental elements of the relation to the world, to others, to oneself.

The knowledge of the brain shows how its development progresses according to a particular mode of evolution. Such a particularity lies in the importance of the action which the subject does on himself. The concept of reflexion can here be used in its primary sense, that of a mirror. By acting upon himself, the human being becomes himself. He carries latent aptitudes from the sphere of the possible into that of reality.[7]

The judgment passed on this point is, more than for the other elements of the history of life, a retrospective one. It is in the light of the present aptitudes of mankind that mankind judges the history which has led to its present state. One then considers that the field of the possible, and therefore the contingency of situations and the random nature of events, have been actualized, and fall within the scope of a certain continuity. In this retrospective perspective, the development of the brain and its internal structuring make it possible for the human being to be in the world in an original manner; a manner which is reinforced in places which concern the proper nature of man. The concept of neoteny or juvenilisation makes it possible to locate it. This scientific concept acknowledges the fact that, at his birth, the little human child lacks the necessary aptitudes to survival, and that this is linked to a certain immaturity, by comparison with the animals that are

[7] See John L. Bradshaw, *Human Evolution. A Neuropsychological Perspective*, Taylor Francis Group, 1997.

closer to him, like the chimpanzee, who is immediately capable of acting for his survival. Such a deficiency is in fact an advantage, as it allows a development which is the fruit of a prolonged education, where the associative cortex is being structured.

Evolutive neurosciences thus show how the human brain has specified itself, diversified and complexified itself.[8] The studies made on language and gestural communication prove it. The language has its roots in specific zones of the brain. It implies a strict hierarchisation. Linked to the possibility of language is a possibility of conscience linked to intelligence. In the scientific approach, intelligence is the ability to solve new problems. This very general definition makes it possible to incorporate different definitions. It bridges the gap between ethology and anthropology.

2.3. A science of mind?

A side effect of the neurosciences has been to introduce into the field of science elements which traditionally belong to philosophy, anthropology or psychology: emotions, imagination, consciousness and the unconscious... Successful research has made it possible for us to speak of the elaboration of a spiritual science, giving to the word 'spiritual' the adversative meaning which it has in ordinary language to name observable behaviours. But to speak of 'spiritual science' has a strange ring to ears that have been used to acknowledge the transcendence of the human in relation to the animal world. The debate has entered the world of philosophy.[9]

What is at stake, then, is to know whether the expression 'science of mind or spirit' is accurate, or not. The question is an epistemological one, because the question is to know what the nature is, of the 'reduction' produced by the inscription of the study of human activity through the scientific method. It is indispensable, as a matter of fact, to introduce a distinction between the reductionism of the scientific method and systematic reductionism. The former contents itself with presiding over the scientific activity properly speaking, whereas the latter is a metaphysical option facing the human specificity.

One must, however, introduce at this stage a critical remark on the process. In fact, the study of the nervous system may be at fault and makes the mistake inherent to any specialized research: namely, to take into con-

[8] See François Clarac & Jean-Pierre Ternaux, *Encyclopédie historique des neurosciences*. Du neurone à l'émergence de la pensée, Bruxelles, de Boeck, 2008.

[9] See *Philosophie de l'esprit*, t. I: *Psychologie du sens commun et sciences de l'esprit*, t. II: *Problèmes et perspectives*, Textes clés de philosophie de l'esprit, Paris, Vrin, 2003.

sideration the part under scrutiny, and ignore the totality of the living being. Such a reproach has often been made – alas with reason – where medical treatments are concerned and bioethics reminds us that it is a person who must be taken care of, not a case, a function or an organ. Our concern with the brain, which is necessary for the improvement of science, must therefore be envisaged in a strictly philosophical perspective.

3. The anthropological question

Recent scientific discoveries tend to prove that mankind is perfectly integrated into the world of the living. We therefore think that the first lesson which must be drawn from what science has brought us, is to do away with the dualist vision of the human being.

3.1. Doing away with dualism

The research which has been done in the neurosciences shows the difficulties and the inadequacies of the spiritualist tradition which, emblematically, ever since Descartes, follows a dualist way.[10] It draws a line between the body and the soul, which are pronounced of a different nature: the former is purely material and governed by mechanical laws, according to the metaphor of the mechanical animal; the latter is immaterial, purely spiritual. Such a vision of the dualist tradition, thus characterized, is probably exaggerated; but this presentation enables one to understand why it does not account for what we know of life. This obviously means a liberation for the mind, which can concern itself with the living and proceed with an approach which is not situated at the only level of analysis...

But this rupture with the philosophical tradition, which is more subtle and profound than its detractors claim, should not be an excuse for going to the opposite extreme: reductionism and monism. Reductionism is easy to denounce: it explains away the whole by its constituent parts. The human being is considered from the sole point of view of chemistry. It is important to be aware of the unity of the living.

3.2. The dynamic unity of the human being

The present state of knowledge concerning the neurosciences enables us to give a more accurate representation of them. I shall use the word 'integra-

[10] See Antonio Damasio, *Descartes' Error: Emotion, Reason and the Human Brain*, Putnam books, 1994; *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*, Harcourt, 2003.

tion' to describe it. The word signifies that elements which can exist elsewhere are captured and integrated into a system. Because of the unity of such a system, the elements are situated at the level of organisation; their own development is regulated for the welfare of the whole. Integration makes it possible for one element to realise what it is. But since it is associated to others, it produces a greater effect than it would if isolated, or placed in a different context.

The unity of the system comes first and foremost. The perception is not that of analysis, but of a systemic vision. Resorting to it means expressing a philosophical option, the option which favours unity rather than analysis. Such an option respects the development of scientific knowledge, as is shown by the emergence of the so-called *evo-devo* theory to account for the theory of evolution.

Going back to the original image of the arborescence and in accordance with the spirit of the phylogenetic classification of the living, it appears that evolution is not static, but dynamic. The central principle is that any living being is animated by a dynamic force which drives him forward towards accessing its perfection – *conatus essendi* as the philosophers say.

As far as our question concerning 'the place of man in nature' is concerned, it may help to come back to the theme mentioned at the beginning with the image of the scale of beings, which has become the tree of beings. In such a presentation, there is a continuity between the degrees of living beings. The present vision shows that such a continuity is still considered as fundamental. It is the foundation of scientific work, even if the results invite us to change our way at looking at this continuity.

In the traditional vision, continuity was a sign of perfection, a scale where very specific natures were inscribed. Now the evolutionist paradigm which presides over the scientific thought invites a consideration which is not at the level of essences, but at the level of an element which is essential to life. Life is indeed characterized by functions (food, growth, reproduction...). But also by a tension, which is an invitation to fully realise one's potentialities. What is potential tends to realisation. On the great tree of the living, divergence is the result of this urge to develop potential riches. This is true for human beings. The notion of *exaptation*, first introduced by Stephen Gould and often used since, accounts for it.[11] When a decisive mutation is about to take place, potentialities already inscribed in the genome can operate. Arborescence is thus understood as an internal dynamism of the liv-

[11] See Stephen Jay Gould, *The Structure of Evolutionary Theory*, Cambridge, Harvard University Press, 2002.

ing. Any living being is tensely concentrated on realising the rich potentialities which are in him. He tends to realise what in him is 'potential' (a word which is preferable to 'virtual'). To grow is the distinctive characteristic of the living. But the notion of growth is not only valid for individuals: it also concerns the tree of the living where the rich forms of life multiply.

Such a vision of life enlightens the understanding of the place of man in nature. Its unity is part of the force which has been at work ever since the beginning. This unity is fundamental, but it is not the only element for anthropology.

3.3. The recognition of alterity

Mankind is regarded by scholars as a new issue, because a human being is capable of language, in the full sense of the word. The human language generates access to alterity, in other words to the recognition of the other. The human being stands face to face with nature. Such a situation comes up with language, in the widest sense of the word. The ability to designate is the proper condition of man; ethologists have observed that a little child points his finger, which animals cannot do – except if they have been trained to do so. Alterity is made even more obvious when the language is articulate, and when words are associated with the real world, not only singular objects, but classes of objects, acts that link, gestures that relate them to human activity. A human being is then in a situation of irreducible alterity. The psychological process of recognising others in the human world is inscribed in such a perspective of the discovery of nature. In the realm of mankind, the relation is different; empathy does not suffice; a new stage is necessary, where the dynamics of life assert themselves: which means the establishment of a social connexion on a basis of transcendence. This dimension is traditional and is part of universal culture. It appears clearly in historical times; it is also patently obvious where art displays itself in the forms that we know; it can also be observed in the fabrication of a tool which is not limited to its immediate use, since a tool looks forward to the future, the anticipation of the future taking place in similar circumstances.

Conclusion

The development of science is not limited to a few scholarly pieces of information locked up in a specialized field. It is an invitation to found a new anthropology. It is an invitation to revise a certain number of founding principles of all cultures, since what is concerned here is mankind and the quest of the essence of man.

The philosopher who reads contemporary scientific journals discovers with glee that the problems posed by the expansion of science are quite traditional. Aristotle's ideas, like the ideas of the stoic philosophers on the human body, are quite in favour at the present time. What is under discussion here, is the stature of man, his manual capacity, his ability to imagine, to have representations and to take part in a social organisation. The place of man in nature nonetheless takes a new outlook. It has a twofold aspect.

The former can be described as the notion of precariousness. The history of life shows that there is no absolute necessity or determinism in its development. Phenomena occur at random; results are never ensured definitively. The greatness of the human being results from his richness, which makes him vulnerable.

Secondly, the greatness of the human being does not lie with the notion of strength, but with the ability to recognize the existence of others. Such an ability can be observed in four areas. First, the relation to nature and the biological and physical environment; then, the relation to other human beings in society, and in interpersonal relationships; moreover, in the relation to oneself, what is at work here is the reflexion on action and the meaning of action. To these three elements one must add a type of relation which reaches beyond immediate perceptions: a relation to a being whose action accounts for the origin and the end of human existence. Such a transcendence operates through various channels: the channel of science should be put aside on this subject, but it leads to a new philosophical approach, where the unity and the dynamism of the human being are recognized.

The Discovery of DNA as a Contribution to Understand the Aristotelian Theory of Generation

Enrico Berti

My contribution to this conference is quite unusual, because, unlike almost all the other members of the illustrious Academy gathered here, I'm not a scientist, but a philosopher, or rather a historian of philosophy. Consequently, the work I have done in over fifty years of research has not contributed at all to the development of twentieth century science, while the opposite is true, i.e. the scientific progress of the twentieth century has helped me solve some problems of a philosophical nature that I had come across in the course of my historical research. The main object of my research is the thought of Aristotle, which I have studied in its historical context and philosophical value. I have also pieced together its fate over the centuries and tried to highlight its present-day importance. At first I concentrated on the metaphysics of Aristotle, which I believed to be the most valuable aspect of his thought, because of its historically proven ability to provide a solid philosophical basis for a religious – not only Christian – conception of reality. Then, in the wake of the revival of practical philosophy that occurred in the second half of the twentieth century by philosophers such as Gadamer and Ritter in Europe, MacIntyre and Nussbaum in the U.S., and many others, I discovered the value of Aristotle's ethics and politics, which still seem feasible today, even independently of a religious vision of reality. More recently, that is in the last ten years, I have turned my attention to the study of Aristotle's biology, particularly his theory of the generation of animals, and this is where I have encountered some major problems.

As it is well known, in the *De generatione animalium*, the treaty expressly devoted to the breeding of animals, Aristotle explains this phenomenon by means of his theory of the four types of cause: material, formal, efficient and final. He asserts that, in animals that reproduce by mating, the female provides the matter, while the male provides the 'principle of movement and generation', i.e. the moving or efficient cause, which, as we shall see, is also form. Indeed, for Aristotle, generating means giving form to matter. More precisely, the 'principle of generation', provided by the male, according to Aristotle is the 'semen', while the matter, provided by the female, is the menses. Obviously Aristotle did not have a microscope, which would have allowed him to see

the eggs, but simply noted that menstruation ceases in the pregnant female, deducing that menstrual blood was used to form the 'first mixture of male and female', which he calls 'the fruit of conception'. The male seed does not become part of the fruit of conception, that is, it does not in any way constitute its matter, which is provided entirely by the female, but it transmits impulses or movements to it, which give it form.

At this point we must remember that, for Aristotle, the form of the living beings is the soul. Indeed, for Aristotle, the soul is no longer a demon, as it was in the Orphic-Pythagorean tradition, i.e. an intermediate being between man and god, which pre-exists the body, embodies in the latter at birth, and exits it at the time of his death, to transmigrate into another body. Nor is it a substance similar to ideas, temporarily united to a body and destined to survive it, as it was for Plato. According to the famous definition of *De anima*, the soul for Aristotle is 'the form of a natural body that has life in potency', namely the form not of an artificial body but, precisely, of a natural one, which, thanks to it, is capable of living. It is therefore the principle of life, the very capability to live.

However, given that by life we mean many things, first of all self-nutrition and growth, which are proper to plants, then movement and feeling, which are proper to animals, and finally thought and will, which are peculiar to human beings, there will be three kinds of souls: the vegetative soul, the ability to absorb nutrient and grow, which belongs to plants, the sensitive soul, the ability to move and feel, which belongs to animals, and the intellective soul, the ability to think and want, proper to human beings. All living beings, therefore, have souls: plants, animals and humans. But note the following: each genus of living beings has only one kind of soul, plants only the vegetative soul, animals only the sensitive soul and humans only the intellective soul. Indeed, as Aristotle explains, the types of soul are like geometrical figures, where the following contains in potency the previous one, for example the quadrangle contains in potency the triangle. Thus the sensitive soul contains in potency the vegetative soul and the intellective soul contains in potency both the vegetative and the sensitive soul. This means that the ability to perform superior activities, such as thinking and wanting, contains in itself the ability to perform lesser activities, such as eating, growing, moving and perceiving. The human soul, therefore, is the effective presence of all these capabilities in a single body, i.e. the ability to live in the fullest sense. The soul is generally the formal cause, that is, the cause of being, i.e. of living, because, for living things, being is living, and it is also the final cause, that is, the goal, because, according to Aristotle, the goal of living things is experiencing life in all its dimensions, namely carrying out all the functions of

which living things are capable, from the lowest, such as absorbing nutrients, to the highest, such as thinking.

A first problem of this theory arises when Aristotle observes that the various parts of the new body, i.e. heart, lung, liver, eye, do not form together at the same time, but are formed consecutively, 'in the same way as the knitting of a net'. It is well known that Aristotle observed the gradual formation of a chick by examining the development of the embryo contained in the hen's egg. So he can say that, through observation, at a given moment some parts of the embryo are already there and others not yet, and not because they remain hidden because of their smallness: indeed, the lung, which is larger than the heart, appears later. Thus the parts of the embryo are formed one after the other, not because one generates the other, but because the essential form of each part is contained in potency in the part that exists already in agency. According to Aristotle, the body part that is formed first is the heart, because it is the cause of nourishment and thus of the growth of all the other parts. After the heart comes the brain, then the internal organs and finally the external ones. The generation of the different parts is described as a series of consecutive actions, which automatically follow one another according to a sort of programme that is contained in the semen, or in the fruit of conception.

The form is called also *logos*, i.e. the ratio between the various components of each tissue or of each organ, and which causes a tissue or an organ to be what it is. But *logos* also means 'notion' or 'discourse', that is, 'formula'; therefore form is a notion, or a speech, or a formula: today we would call it an 'information'. While the matter of the tissues and organs that are formed in generation comes from heat and cold, that is, from the properties of matter from which they derive, provided by the female, their form derives from the form contained in actuality in the male parent and transmitted through the motion produced by the sperm. How is such a process possible? This was the first problem that Aristotle's theory of generation posed to me.

Another problem that arises further is what kind of soul is transmitted in the generation of animals. Indeed from certain passages of *De generatione animalium* it seems that the vegetative soul is present first of all in the animal embryo, by which it lives the life of a plant, because the first tasks that the embryo carries out consist essentially in its growth, which is consequent to nourishment; then the formation of the sensitive soul in it, through which the embryo lives the life of an animal; and finally, in the case of a human embryo, it seems that the intellective soul is formed in it, after the entry of the intellect in it, which apparently comes 'from outside'. This interpretation was enormously successful both in late antiquity, and in the Middle Ages, that is,

in ages dominated by a creationist vision, whereby the human soul did not come from the parents, but appeared to be created directly by God. This however seems incompatible with the doctrine contained in *De anima*, according to which the animal only has one soul, the sensitive soul, and consequently it must be assumed that man too possesses a single soul, the intellective one. The late ancient and medieval interpreters therefore had to assume that real substantive changes were produced in the development of the embryo, i.e. that the embryo was initially a plant, equipped with the vegetative soul only, and then turned into an animal, equipped with the sensitive soul only, and finally became a man, equipped with the intellective soul only. But there is no trace of these substantial mutations in the Aristotelian doctrine of generation, rather generation appears as an ongoing process, driven by a single form, which remains the same.

The solution of both problems came to me after reading of an article by a biologist about the discovery of DNA. As you know, DNA was discovered during the 1950s by James Watson and Francis Crick who, also on the basis of the research carried out by other scientists, were able to describe the structure of deoxyribonucleic acid, i.e. of one of the two acids which form the nucleus of cells. Watson and Crick found that DNA molecules consist of two chains of nucleotides in the shape of helixes intertwined with each other. At the time of cell division the two helixes separate and on each of them another is built, in order to reconstitute the original structure. Thus, DNA can reproduce without changing its structure, except for occasional errors or mutations. For this discovery, Watson and Crick obtained the 1962 Nobel Prize for Medicine.

The philosophical significance of this discovery for the interpretation of Aristotle was brought to light some years later by the physicist and biologist Max Delbrück (1906-1981), who in turn won the Nobel Prize for Medicine in 1969 for his research on bacteriophage viruses, in an article dedicated to Aristotle with the ironic title, *Aristotle-totle-totle*.[1] In it Delbrück argued that, if it were possible to give a Nobel Prize in memory of someone, it should be given to Aristotle for the discovery of the principle implied in DNA. He then quoted the passages from the biological works, where Aristotle argues that the male parent contributes to generation by providing the principle of motion through his semen, giving rise to form, and the female parent provides matter, with her menses, translating phrases like 'principle of motion' with

[1] M. Delbrück, Aristotle-totle-totle, in J. Monod and E. Borek, *Of Microbes and Life*, New York 1971, pp. 50-55.

'plan of the development', and 'form and essence' with 'program of development'. He wrote:

> Put into modern language, what all of these quotations say is this: The form principle is the information which is stored in the semen. After fertilization it is read out in a preprogrammed way; the readout alters the matter upon which it acts, but it does not alter the stored information, which is not, properly speaking, part of the finished product. In other words, if that committee in Stockholm, which has the unenviable task each year of pointing out the most creative scientists, had the liberty of giving awards posthumously, I think they should consider Aristotle for the discovery of the principle implied in DNA.

Meanwhile, a similar conclusion had been reached by Marjorie Grene, a scholar of Aristotle and biology expert, who argued that the Aristotelian notion of 'form' operates in many ways like the concept of organization (or information) in modern biology, which is an example of the DNA sequence.[2] But Delbrück's thesis was authoritatively confirmed by the great biologist and historian of biology, Ernst Mayr, who wrote:

> Some of today's authors have had the courage to use modern terms in exposing Aristotelian thought: the words that Aristotle would probably have used had he lived today. I refer to the use of the term 'genetic program' by Delbrück to clarify the intentions with which Aristotle used *eidos* in the description of the development of the individual'. And 'it has been said, not without justification, that the Aristotelian separation of a formative principle (*eidos*) from the matter on which it acts, does not deviate much from the modern concept according to which the genetic program controls the modelling of the phenotype (Delbrück, 1971)'.[3]

More recently, as a partial criticism of Delbrück and Mayr, it has been stated that, according to modern genetics, the function of DNA is limited by the environment of the cell with which it interacts, so that rather than a 'genetic program' guiding development, we should speak of an 'epigenetic program'; however, it was recognized at the same time that this corresponds in a way to what Aristotle said in his concept of 'potential form', which interacts with matter, therefore the Aristotelian theory must be interpreted

[2] M. Grene, *Aristotle and Modern Biology*, Journal of the History of Ideas, 33, 1972, pp. 395-424.

[3] E. Mayr, *The Growth of Biological Thought*, Cambridge Mass. 1982 (It. transl. *Storia del pensiero biologico*, Torino 1990, p. 13).

not as a 'genetic vitality' based only on the notion of 'entelechy', but implies a mechanism, as shown by Aristotle's example of the automated puppets.[4]

If we now return to the problem of human development, we must recognize that, according to today's genetics, what distinguishes the human genome from that of other living species, although in a minimal (but important) percentage, is the 'sequence' of the various components that make up genes, i.e. the DNA segments of which the chromosomes contained in the cell nucleus are made. Well, the components of DNA, which are equivalent to what Aristotle called 'matter', are the same for all living beings, while the 'sequence', i.e. the order in which they are arranged, is different. However this order is equivalent to what Aristotle called 'form' and all the characteristics that develop in the living being depend on this order, just as for Aristotle all the characteristics of plants and animals depend on their form, that is, on their 'soul'.

In conclusion, the discovery of DNA permits to understand the Aristotelian theory of generation in a new way, following which it emerges that, unlike the traditional interpretation, in *De generatione animalium* just as in *De Anima* Aristotle admits one soul, which in the case of animals is the sensitive soul, containing in potency the vegetative one, in the sense that it implements first of all the functions proper to plants and then those proper to animals, and in the case of human beings it is the intellective soul, which implements first the functions of plants, then those of animals and finally those that are proper to human beings. The sentence according to which the intellect comes 'from outside' does not express Aristotle's thought, but expresses what, according to Aristotle, should have been the opinion of the Platonists, that is, of the supporters of the soul pre-existing the body, had they been able to take into account the way in which generation actually takes place. The only part of the soul that, from the point of view of pre-existence, could pre-exist the body, appears to be the intellect, thanks to the immateriality of its functions. But for Aristotle the intellective soul, thus including the intellect, does not pre-exist the body, but is generated in the embryo through the motive action exerted by the father through the sperm, an action which transmits to the embryo the faculties proper to the form possessed by the father, which is a distinctly human form. Therefore the soul generated in the embryo, if it is generated by human sperm, already contains in potency even the intellect, because it is a specifically human soul.

[4] T. Vinci and J.S. Robert, *Aristotle and Modern Genetics*, Journal of the History of Ideas, 66, 2005, pp. 201-221.

Session IV: Cell and Molecular Biology

The Evolutionary Lottery

Christian de Duve

Introduction

It is now established that all living beings, including humans, descend by evolution from a single ancestral form and that this process was largely driven by natural selection, the fundamental mechanism, first discovered by Charles Darwin and independently perceived by Alfred Russell Wallace, whereby forms of life best fit to survive and produce progeny under prevailing conditions obligatorily emerge when several variants compete for the same limited resources. A striking feature of this process is the dominant part played in it by chance, which does so in two distinct ways: first by the mutations that are offered for selection and, next, by the environmental circumstances that condition the selection process.

These facts imply that the extraordinary diversity of living forms on Earth is the outcome of a vast planetary lottery, or, rather, a long string of lotteries, played over almost four billion years and leading, from primitive forms of life, first to bacteria, or prokaryotes, next to unicellular eukaryotes, or protists, and finally to multicellular plants, fungi, and animals of increasing complexity. Humans appear at the very end of the animal line as the products of a lengthy succession of chance events. In the eyes of many of the thinkers who have reflected on the topic, the logical inference from this observation is that the chances of appearance of our species were virtually nil at the start, with as necessary implication the extreme improbability and consequent meaninglessness of the human condition.

The purpose of the present paper is to re-evaluate the validity of this attitude in the light of a closer examination of the data on which it rests.

1. The Rules of the Game

1.1. Mutations

Natural selection depends on the faithful transmission of hereditary traits, to ensure the genetic continuity of selected lineages, and, in a crucially important way, on occasional alterations of this process, or *mutations*, to generate the variants on which selection acts. Such changes may be caused by a number of different factors, including inaccuracies in DNA replication (very rare – one wrongly inserted base in one billion – but nevertheless

significant because of the large size of genomes), rearrangements of DNA sequences by recombination, deletion, insertion, transposition, or other phenomena, chemical alterations of DNA by physical agents, such as UV light, X rays, or radioactivity, by chemical substances (mutagens), or by biological agents such as viruses. In addition, changes affecting other features of DNA chemistry, such as the methylation of certain bases, or the manner in which DNA is associated with proteins in chromosomes may also be involved. These features are covered in the contemporary literature by the term 'epigenetic', which was used previously for non-hereditary changes acquired after birth, in particular in the brain, and is still used in this sense by some developmental biologists and neurobiologists.

A feature common to all mutations is that they are *accidental*. They have specific causes, as just seen, but these causes happen fortuitously and, especially, bear no relation to any foresight of the consequences they may entail.

This notion is important with respect to the theory of 'intelligent design' (ID), which claims that certain critical evolutionary events, ranging from the assembly of cilia and flagella to the formation of eyes and the conversion of reptiles into birds, could not have occurred naturally, but required the operation of some supernatural entity that predefined the outcome and engineered the appropriate genetic changes accordingly. This view differs from strict creationism in that it has no biblical roots and does not negate evolution, but it shares with creationism its call on a supernatural agency. ID is sometimes called 'creationism in disguise' for that reason. It goes back to finalism, or teleology, which is itself closely related to vitalism, the theory, defended by many earlier biologists, according to which life is 'animated' by some kind of 'vital spirit'. Finalism is fuzzier on this issue, claiming simply that life is a goal-directed process, without specifying who or what does the directing. The term 'teleonomy' is sometimes used to express the fact that life has the *appearance* of being goal-directed, but due only to its internal organization and not to any directing agency.

A detailed discussion of ID does not belong in the present paper. Just two comments are in order. First, from the purely scientific point of view, it is readily shown that many of the claims made by ID advocates rest on oversimplified views of the evolutionary process, which ignore factors such as the immense times taken, the circuitous pathways followed, and the large numbers of individuals and generations involved, as well as much of the recent information provided by molecular phylogenies. In fact, plausible explanations have already been offered for several of the allegedly unexplainable evolutionary processes, the formation of eyes, for example. Next, and more importantly, ID is simply *not a scientific theory*. It disqualifies

itself as such by its assertion of unexplainability. Scientific research is based on the postulate that events are naturally explainable. Whether this is true or not is immaterial. There can be no research without this basic assumption. Do away with it, and you can close your laboratory.

The term 'postulate' is important in the above consideration. Science is not entitled to affirm, as is done by some scientists, that everything is naturally explainable. Until everything is explained, such a statement is unwarranted. Subject to this caveat, imposed by scientific objectivity, it must be recognized that spectacular successes have been achieved under the aegis of the naturalist postulate and continue to be achieved at an ever increasing pace. These successes certainly strengthen the postulate enormously and encourage further research under its guidance.

By definition, natural selection can act only on the variants that are offered to it. Better solutions to an environmental challenge may be possible. If they are not provided, they will not be realized. Selection is limited by the kinds of variants that are offered to it by chance

This obvious fact raises the question as to how many of all the possible variants are included in the set provided by chance. At one end of the spectrum, if the set is complete, selection will bring out the best in reproducible fashion; the final outcome will be *optimization* with respect to the environmental challenge faced. At the other end, if only a very small subset of the possible variants is provided, whatever happens in reality will depend on the composition of this subset; the process will be ruled by *contingency*.

For a long time, the second possibility was the ruling opinion, though rarely expressed in quantitative terms. It was simply taken as self-evident that, because of the involvement of chance in the course of evolution and of the vast number of possibilities open to it, this course must by necessity have been dominated by contingency. This view was eloquently defended and propagated by many evolutionists of the past, including George Gaylord Simpson, Ernst Mayr, Jacques Monod, François Jacob, and Stephen Jay Gould, to mention only a few. Coinciding with the rise of existentialism, especially in France, this message from science was interpreted as affording strong support to the philosophy of the absurd then in vogue.

Little attention was paid to the fact that chance always operates within a set of *limits*. Whether at heads-or-tails, roulette, or the lottery, the number of possibilities is finite and given occurrences become increasingly probable as more trials are made. Thus, even a seven-digit lottery number has a 99.9% probability of coming out if 69 million drawings are made. Admittedly, lotteries for gain don't function that way. But the evolutionary lottery is different. Because of the enormous times and large number of individuals

involved, also because of the intrinsic constraints of genomes, many specific mutations have a greater probability of occurring than intuition would lead one to predict.

Several facts support this contention. Take *mimesis*, for example, the property whereby some animals closely resemble their surroundings and thereby evade predators better than those not similarly protected. Acquisition of this property in one shot is clearly impossible. An insect cannot suddenly become almost indistinguishable from the leaf or branch on which it sits; a fish cannot suddenly resemble the sand or pebbles on which it rests. The process, if it occurred naturally, as must be supposed, must necessarily have gone through a large number of stages, at each of which the animals became a little more similar to their environment, sufficiently so to enjoy some selective advantage. It is evident that these stages could not have occurred if the necessary mutations had not been provided each time.

Another impressive fact is the frequency of evolutionary *convergence*, the independent acquisition of the same adaptations to given environmental challenges. Hundreds of examples of this remarkable phenomenon, from saber-toothed tigers to anteaters, have now been recorded, prompting members of the younger school of evolutionists, such as Simon Conway Morris and Richard Dawkins, to defend the view of a largely obligatory and reproducible evolutionary history, in direct opposition to their predecessors.

Note, however, that the view remains conditional: same circumstances, same result. But what if the circumstances change? Here, contingency comes back to the forefront, by linking the history of life to the vagaries of environmental changes. This is the second chance-dependent factor in natural selection.

1.2. The Environment

Natural selection is critically dependent on the prevailing environment. The features that are selected are those that are conducive to, or, at least, compatible with, the survival and proliferation of the individuals and populations involved *under the conditions* to which they are exposed. Change those conditions and the selective response will be different.

An obvious implication of this fact is that evolution must have been molded by the environmental history of the Earth, which makes it unique, whatever the number of life-bearing planets in the universe, as no planet can have exactly the same history. True enough. But how different can one expect the two to be?

Here, a basic distinction must be made between two ways in which the environment plays a role. In one, which may be called *instructive* (with no connotation of design), the environment defines the selected property. Thus,

adaptation to certain external conditions, such as dryness or cold, is clearly influenced by the environment, desert or polar ice field, to which the organisms are exposed. Mimesis is another obvious example of environment-dependent evolutionary change. Without green leaves, no insect would become leaf-like. Most of the innumerable details that define biodiversity fall in this category, reflecting the enormous variety of environmental conditions that have affected natural selection. In this respect, life on our planet is undeniably unique.

The other way in which the environment may affect natural selection may be termed *facilitating*: the elicited phenomenon is intrinsically mandated by the stage reached by evolution, with the environment acting simply to provide the trigger for this potential to materialize. A typical example of such a happening is the rise of the mammals after some global catastrophe, presumably caused by the fall of a large meteorite on the Yucatan Peninsula in Mexico about 65 million years ago, cleared the way for them by wiping out the dinosaurs and many other forms of life. One is clearly not dealing here with an adaptation to a specific environmental situation, but rather with the actualization of an existing propensity by an environmental accident. Indeed, it is most likely that the dinosaurs were fated to disappear in any case, together with the luxurious vegetations from which they drew their subsistence, and that, if not the fall of a meteorite, some other accident would have precipitated their extinction.

Hominization, launched 6-8 million years ago by an upheaval believed by some anthropologists to be the separation of the savannah from the forest by the Great African Rift, which provided selective value to bipedalism and the associated brain expansion, could be another example of environmental facilitation of a latent evolutionary step. The process, once initiated, developed so rapidly – a quadrupling of brain size in only a few million years – as to suggest that the step involved was long present in potential form, awaiting only an environmental trigger to be precipitated. Had the Rift not split the African continent, assuming it played a role, some other accident could have propelled some chimpanzee-like primate on the way to becoming human.

It is possible that many decisive events in evolution belong to this category, imposed by the inner constraints of the evolutionary process and merely triggered into happening by environmental factors. Precise information on this topic is lacking, but the possibility it evokes must be kept in mind as it implies that the history of life on Earth, although subject to the vagaries of environmental conditions, may in its main lines, have followed a course largely imposed by properties, potentialities, and constraints inherent to the living process.

2. A Fresh Look at Evolution

2.1. The Evolutionary Lottery

Our view of evolution as a huge planetary lottery has not changed. What has changed is our appreciation of the probability of a lucky number coming out. Chance, we have learned, does not exclude necessity.

Two factors have to be reconsidered. First, mutations, although governed by chance, are not as 'chancy' as was believed. Because of the immense number of opportunities that are provided on the evolutionary scale, the mutations due to be most effective under the circumstances are often almost guaranteed to occur at some stage, thereby introducing optimizing necessity into the process.

As to the part played by the environment in the lottery, it depends on the nature of the affected event. The role of environmental contingencies is clearly decisive in the myriad instances of adaptation to specific geological, geographical, climatic, ecological, or other adventitious circumstances. Environmental conditions tend to be less decisive and more often merely facilitating when it comes to major transitions. In this new perspective, evolution appears as intrinsic to the living process, with every major step somehow mandated by the stage that preceded it, all the way from the earliest living forms up to humankind.

As to the earliest living forms themselves, I have argued elsewhere that, because of the deterministic nature of chemical events and of the frequency of optimizing selection, the processes that initiated life on Earth must have been imposed by the physical-chemical conditions that prevailed on the prebiotic Earth. Given those conditions, life as we know it – including ATP, RNA, DNA, base pairing, the genetic code, protein enzymes, and lipid membranes – was virtually bound to appear.

The view that emerges from those considerations is of life and mind as *cosmic imperatives*, rather than improbable products of random chance. The reason supporting this statement does not lie in any finalistic or 'anthropic' view of the universe, seen as having been created for the *purpose* of giving rise to life and mind, but rests simply on a *factual* assessment of the events that have governed evolution, including the appearance of humankind. The universe just happens to be such as to necessarily give rise to life and mind. Some observers may derive a theistic view from this realization. Others, however, may content themselves with seeing it as a manifestation of *ultimate reality*.

2.2. The Tree of Life

Evolution is often pictured by a tree rooted in the early chemical phenomena that have given rise to the first living cells, almost four billion years

ago. Like all trees, the tree of life has grown in two directions: vertically and horizontally. The vertical direction, delineated by the trunk and master branches, has given rise to increasing *complexity*. The horizontal direction, traced by the countless lateral ramifications that have sprung at each level of complexity, has led to increasing *diversity*.

The main conclusion to be derived from our new appreciation of evolution is that contingency has affected mostly the horizontal ramifications of the tree of life. On the other hand, the vertical extensions of the tree appear as strongly driven by the inner pressures and resulting constraints created by the evolutionary stage reached, waiting only for some environmental trigger to be set in motion.

2.3. Extraterrestrial Life

A corollary of the above considerations is that, if another Earth-like planet should display conditions conducive to the development of forms of life similar to those that started life on Earth, the resulting tree would most likely differ greatly from the Earth tree in the details of its canopy, but could show a similar vertical structure. Given enough time, the appearance of human-like intelligent beings could even be contemplated.

These points are relevant to the great interest accorded in recent years to the search for life-bearing extrasolar planets and for signs of extraterrestrial intelligence. Such searches are justified by what is known of evolution and by the very large number of sun-like stars believed to exist in the universe (on the order of $3x10^{21}$). We are not likely to be unique with so many opportunities provided for intelligent life to arise. The problem is that most of those countless planets are totally out of reach of present technologies. Even those that have been identified in our nearest neighborhood could not reveal telling signs of life to existing instruments, except, possibly, for the presence of molecular oxygen (not found so far), which, on Earth, is a product of life. What the future will bring can obviously not be anticipated.

2.4. The Future

A major question raised by the above considerations is: Will the tree of life continue growing as it has done before, losing branches and extending new ones in the horizontal direction to create more diversity, and, especially, progressing vertically towards increasing complexity? A priori, there seems to be no valid reason for excluding such an eventuality. There is plenty of time for it. According to astronomers, the Earth should remain physically able to support life for at least 1.5 billion years, perhaps as long as five billion years, when the sun is expected, its energy resources exhausted, to convert

into a red giant, abolishing all possibilities of life on surrounding planets. As to the plausibility of such an event, only human hubris could cause us to rule it out. In all objectivity, there is plenty of room for improvement in human nature. We have no valid reason for considering our advent as the crowning event in evolution. Our recent past is landmarked by the appearance and extinction of hominid species of increasing cranial capacity and, presumably, greater mental power. The remarkable tendency of the human brain to grow bigger and more powerful is presumably still extant, awaiting only the anatomical and developmental changes needed to make it possible for it to manifest itself.

Present circumstances are, however, very different from those that have allowed the appearance of our species and the extinction of our forebears. Instead of small bands subsisting precariously, often completely separated from each other and capable of evolving each in isolation, humanity has invaded the entire surface of our planet, filling it with more than six billion individuals connected by a dense network of communications. Our extinction and replacement by some sort of 'übermensch' would require a massive planetary disaster too horrible for even our imagination to picture. The rise of a better fit species on such ruins would have nothing in common with the displacement of the Neanderthals by our species.

There is an even more fundamental difference. This dire fate is not ineluctable. For the first time in the history of life on Earth, a species has appeared that is not slavishly subject to natural selection. Thanks to their superior brains, humans have acquired the ability to do what natural selection is incapable of: look beyond the immediate present, foresee the outcome of possible future events, elaborate plans as a function of those predictions and responsibly act accordingly, even if it means sacrificing immediate benefits for a greater, later good. The future of life and, with it, of humanity itself, thus depends on the wisdom with which coming generations will make use of this ability.

Summary and Conclusion

- There is less chance, and more necessity, in evolution than has commonly been believed, not because of the intervention of some purposeful influence in the process, but because of the frequency of selective optimization and of the intrinsic constraints of the living process.
- The horizontal growth of the tree of life in the direction of increasing diversity has been largely contingent on environmental peculiarities not expected to be repeated on another planet. Its vertical growth toward

increasing complexity, however, seems to be more obligatory and commanded by the attained evolutionary stage.

– With the advent of humankind, natural selection has ceased to be the only driving force of evolution. Human foresight and ability to purposefully act against natural selection have changed the rules of the game. Henceforth, the future of life and that of humanity itself will depend, at least partly, on human responsibility and wisdom.

Therapeutic Vaccines Against Cancer and Autoimmune Diseases

Michael Sela

In the spirit of the invitation to this Plenary Session, on the topic of 'The Scientific Legacy of the 20th Century', I shall try to mention here some of the highlights of my research in the last sixty years at the Weizmann Institute of Science.

Elucidation of antigenicity and immunogenicity

My laboratory pioneered the design of amino acid oligomers to define the minimal and precise chemical characteristics of antigens – molecules that could be bound by antibodies (1). Using these tools, we determined the functional size of the antibody binding pocket, and characterized the effects of charge, hydrophobicity, and side chain interactions in the antibody-antigen binding complex (1). These studies defined the chemistry of antigen binding by antibodies and laid the foundation for subsequent structural investigations based on x-ray analysis of crystallized preparations and NMR studies. In the course of these studies, I called attention to the essential difference between antigenicity (the capacity of a molecule to bind antibodies) and immunogenicity (a term coined to designate the capacity of a molecule to induce an active immune response). This distinction has become a guiding principle in immunology (1).

Discovery of a chemical basis for the action of immune response genes

In the light of my characterization of antigenicity, I went on to apply amino acid oligomer chemistry to the question of immune response genes. We synthesized amino acid oligomers with defined, minimal chemical differences and, together with Hugh McDevitt, discovered and analyzed the role of MHC genes in mediating genetic control of the immune response (2,3). Work with amino acid oligomers, which proceeded independently of that research, established a solid, synthetic chemical foundation for subsequent biologic studies. The chemical research was seminal in providing the mindset for subsequent biological studies and for the x-ray crystallography that definitely solved the structure of the antigen-binding site of the antibody.

Based on the high water-solubility of poly-DL-alanine, I could open all the disulfide bridges of an immunoglobulin without the product dropping out of solution. Upon reoxidation, all the immunological properties, whether as antigen (4) or as antibody (5), have returned, thus proving the correctness of the selection theory of antibody formation (6).

Vaccines are prophylactic in the sense that they are administered to healthy individuals to prevent a disease. Nevertheless, there is a growing trend to use vaccines to alleviate the suffering of those already with a disease. Great effort is being devoted to develop vaccines against tumors, AIDS, hepatitis, tuberculosis, and possibly against the bacteria that cause gastric ulcers. Copolymer 1 (Copaxone, glatiramer acetate) used today as a vaccine against multiple sclerosis (MS), is a good example of a beneficial treatment for this autoimmune disease, based on its similarity to the myelin basic protein (MBP), one of the putative causes of MS. This finding could lead to the therapeutic vaccines against other autoimmune diseases such as myasthenia gravis, systemic lupus erythematosus (SLE) and rheumatoid arthritis. Furthermore, antibodies prepared against prions raise hopes for a vaccine against bovine spongiform encephalitis and Creutzfeldt-Jacob disease and antibodies to a peptide derived from amyloid plaques could degrade plaques and be used as a therapeutic vaccine against Alzheimer's disease.

By its definition, a preventive vaccine is sufficiently similar in its chemistry to the etiological agent that provokes the disease so that the immune response directed against it can act against the causative agent. This situation is analogous in the case of therapeutic vaccines.

Development of an effective therapy for multiple sclerosis

Therapeutic vaccines become more and more important, especially as life expectancy increases. Efforts to develop vaccines against such diseases as cancer, AIDS, hepatitis, tuberculosis, Alzheimer's disease, and mad cow disease have not yet reached the stage where they can be successfully used on a daily basis. However, significant progress has been made in the realm of autoimmune diseases, resulting, (at least in one case) in an immunomodulatory vaccine against multiple sclerosis that was developed in my laboratory, and that is in daily use by more than 200,000 patients in 50 countries. The drug or therapeutic vaccine against exacerbating-remitting type of multiple sclerosis is a copolymer of four amino acid residues, denoted Copaxone, which is related to myelin basic protein (7-9).

The story began when we started synthesizing a series of amino acid copolymers composed of four amino acids to create an artificial immunogen

that would mimic myelin basic protein (MBP) and might induce the experimental autoimmune disease EAE, a model of MS. This bold step failed; none of the copolymers were encephalitogenic. But we countered this failed idea with an even bolder idea: the copolymer might not induce EAE, but it might, by mimicking MBP, induce the immune system to resist the disease. This turned out to be the case, and for the next two decades we realized the clinical application of Copaxone to human MS. Today, Copaxone is the most widely used treatment for MS. It is remarkably low in undesirable side effects, yet it significantly reduces the attack rates in relapsing-remitting MS and it prolongs considerably the ability of MS patients to maintain a relatively tolerable quality of life.

Speaking historically, the injection of several positively charged amino acid copolymers in aqueous solution into mice, rabbits and guinea pigs, resulted in efficient suppression of the onset of the disease EAE. The Cop 1 primarily used, now called GA or Copaxone, is composed of a small amount of glutamic acid, a much larger amount of lysine, some tyrosine, and a major share of alanine. Thus, its overall charge is positive. There is significant immunologic cross-reaction (both at the antibody and cell levels) between Cop 1 and the MBP. Interestingly, when an analog of Cop 1 made from D-amino acids was tested, it had no suppressing capacity, nor did it cross-react immunologically with the basic protein. Cop 1 is neither generally immuno-suppressive nor toxic. Actually, it is not helpful in any other autoimmune disease except MS and its animal model, experimental allergic encephalomyelitis (EAE). GA (glatiramer acetate, Copaxone) was demonstrated to suppress EAE induced by MBP in a variety of species: guinea, pigs, rabbits, mice and two species of monkeys (rhesus monkeys and baboons). In contrast to rodents, in which GA inhibits the onset of the disease, in primates it was used as treatment of the ongoing disease. After a couple of early clinical trials, it was clear that GA showed efficacy in treating patients with the relapsing-remitting disease. In three randomized double-blind trials, GA, at a dose of 20 mg once daily, administered s.c. in patients, was significantly more effective than placebo for the respective primary endpoint of each trial (proportion of relapse-free patients, relapse rate, and number of enhancing lesions on MRI scans) (10, 11).

Progression to sustained disability, as measured by the Kurtzke expanded disability status scale, was secondary endpoint in the two long-term trials. Patients with relapsing-remitting MS treated with GA in the pivotal US trial were significantly more likely to experience reduced disability, and placebo recipients were more likely to experience increased disability.

Three different clinical trials investigated humoral and cellular immune responses in MS patients treated with Copaxone 1 (12). All patients devel-

oped Cop 1-reactive antibodies, which peaked at 3 months after initiation of treatment, decreased at 6 months, and then remained low. The proliferative response of peripheral blood mononuclear cells to Cop 1 was high initially and gradually decreased during treatment. Several studies showed that MS patients mainly produce the Th2 type of GA-specific T cells after receiving GA (13,14). Cross-reactivity between GA and MBP is seen at several levels: antibodies, T cells, and cross-triggering of cytokines.

Disseminated demyelination is the primary morphological hallmark characterizing multiple sclerosis (MS) and its animal model, experimental autoimmune encephalomyelitis (EAE), leading to axonal loss and neurological impairments. It is, therefore, important to evaluate MS treatments for their neuroprotective capability to prevent demyelination and/or enhance remyelination. The interplay between pathological demyelination and the corresponding repair mechanism remyelination involves, on the one hand, the inflammatory immune cells that mediate the damage and on the other hand, the myelin-producing cells, the oligodendrocytes. The latter are terminally differentiated cells with a limited capacity to respond to injury that are destroyed in the actively demyelinating lesions. Accordingly, remyelination requires the recruitment of oligodendrocyte precursor cells (OPCs) by their proliferation and migration into the demyelinating area and their further differentiation into mature myelinating oligodendrocytes through distinct stages characterized by morphological transformation, and sequential expression of developmental markers.

The interplay between demyelination and remyelination is critical in the progress of MS and its animal model EAE. In a recent study (15), we explored the capacity of glatiramer acetate (GA, Copaxone) to affect the demyelination process and/or lead to remyelination in mice inflicted by chronic EAE, using both scanning electron microscopy and immunohistological methods. Spinal cords of untreated EAE mice revealed substantial demyelination accompanied by tissue destruction and axonal loss. In contrast, in spinal cords of GA-treated mice, in which treatment started concomitantly with disease induction (prevention), no pathology was observed. Moreover, when treatment was initiated after the appearance of clinical symptoms (suppression) or even in the chronic disease phase (delayed suppression) when substantial demyelination was already manifested, it resulted in a significant decrease in the pathological damage.

Presently, Copaxone (GA, Cop 1) is the most used drug against multiple sclerosis. It has already crossed one million years of use without significant side effects.

Reformation of the native structure of a protein.

It was a very early stage (1956) that I spent a most exciting period of my research in the laboratory at the NIH of my close friend and mentor, Christian Anfinsen, the deceased member of the Pontifical Academy of Sciences. By reducing the four disulfide bridges in bovine pancreatic ribonuclease and letting it stay overnight in solution, the enzymatic activity of ribonuclease was largely restored, and this essentially proved that there is no need for additional genetic information to tell the open polypeptidic chain how to refold into the unique protein architecture (16, 17) .

Synergistic effects in immunotherapy of cancer

After synthesizing a peptide corresponding to the amino-terminus of the carcinoembryonic antigen (CEA) we could show that antibodies to the peptide could recognize CEA in the blood of patients. Later on, we used to link by a weak covalent bond a small chemotherapeutic drug to an anticancer antibody (still polyclonal as monoclonal antibodies were not yet discovered). As a spacer between the drug and the antibody we used either dextran or polyglutamic acid. Despite interesting results, we concentrated later on the quality of the monoclonal antibody per se, and thus we found out an important synergistic effect between a small drug and the monoclonal antibody against ErbB1 (referred also as EGFR-epidermal growth factor receptor (18)). As a result of this discovery, the drug Erbitux is used only with a small chemotherapeutic drug, covered by our patent. Later on, we found a strong synergistic effect between two antibodies against the same receptor, provided they were against epitopes sufficiently removed (19, 20). In one case, it was against ErbB1 (19), in the other case against ErbB2 (20). Thus monoclonal antibodies prolong survival of cancer patients. However, the effectiveness of such therapeutic antibodies is low and patients evolve resistance. Thus, there is place for improvement. We found that pairs comprising an antibody reactive with the dimerization site of ErbB-2 and an antibody recognizing another distinct epitope better inhibit ErbB-2-overexpressing tumors than other pairs or the respective individual mAbs. Because the superiority of antibody combinations extends to tumor cell cultures, we assume that nonimmunological mechanisms contribute to mAb synergy. One potential mechanism, namely, the ability of mAb combinations to instigate ErbB-2 endocytosis, is demonstrated. Translation of these lessons to clinical applications may enhance patient response and delay acquisition of resistance.

Conclusion

The common denominator of the studies described is the use of a molecular approach to medical problems, starting with developing the tools of amino acid polymer chemistry, applying them to elucidate fundamental questions in immunology, culminating in a copolymer treatment for a tragic human disease, and in improving cancer treatment by synergy.

References

1. Antigenicity: Some molecular aspects, M. Sela, *Science* 166, 1365 (1969).
2. Genetic control of the antibody response. 1. Demonstration of determinant-specific differences in response to synthetic polypeptide antigens in two strains of inbred mice, H.O. McDevitt and M. Sela, *J. Exp. Med.* 122, 517 (1965).
3. Genetic control of the antibody response. II. Further analysis of the specificity of determinant-specific control, and genetic analysis of the response to (H,G)-A-L in CBA and C57 mice, H.O. McDevitt and M. Sela, *J. Exp. Med.* 126, 969 (1967).
4. Recovery of antigenic activity upon reoxidation of completely reduced polyalanyl rabbit immunoglobulin G., M.H. Freedman and M. Sela, *J. Biol. Chem.* 241, 2383 (1966).
5. Recovery of specific activity upon reoxidation of completely reduced polyalanyl rabbit antibody, M.H. Freedman and M. Sela, *J. Biol. Chem.* 241, 5225 (1966).
6. E.D. Day, *Advanced Immunochemistry*, William and Wilkins Co., Baltimore, 1972, p. 127.
7. Glatiramer acetate (copolymer 1) in the treatment of multiple sclerosis, M. Sela and D. Teitelbaum, *Expert Opinion Pharmacotherapy* 2, 1149 (2001).
8. D. Simpson, S. Noble, C. Perry. Glatiramer acetate: a review of its use in relapsing-remitting multiple sclerosis, *CNS Drugs* 16, 826 (2002).
9. K.P. Johnson, *The Remarkable Story of Copaxone*, Dia Medica Publishing, 2010.
10. A pilot trial of Cop 1 in exacerbating-remitting multiple sclerosis, M.B. Bornstein, A. Miller, S. Slagle, M. Weitzman, H. Crystal, E. Drexler, M. Keilson, A. Merriam, S. Wasserthheil-Smoller, V. Spada, W. Weiss, R. Arnon, I. Jacobsohn, D. Teitelbaum and M. Sela, *The New England Journal of Medicine*, 317, 408 (1987).
11. K.P. Johnson, B.R. Brooks, J.A. Cohen, C.C. Ford, J. Goldstein, B.P. Lisak, L.W. Myers, H.S. Panitch, J.W. Rose, R.B. Seiffer, T. Vollmer, L.P. Weiner and J.S. Wolinski, Copolymer 1 Multiple Sclerosis Study Group, *Neurology* 1, 65 (1995).
12. Humoral and cellular immune responses to Copolymer 1 in multiple sclerosis patients treated with Copaxone, T. Brenner, R. Arnon, M. Sela, O. Abramsky, Z. Meiner, R. Riven-Kreitman, N. Tarcik and D. Teitelbaum, *J. Neuroimmunology*, 115, 152 (2001).
13. Glatiramer acetate (Copaxone) induces degenerate, Th-2-polarized immune response in patients with multiple sclerosis, P.W. Duda, M.C. Schmied, S. Cook, J.I. Krieger, and D.A. Hafler, *J. Clin. Invest.* 105, 967, (2000).
14. Multiple sclerosis: comparison of copolymer 1 – reactive T cell lines from treated and untreated subjects reveals cytokine shift from T helper 1 to T helper 2 cells, O. Neuhaus, C. Farina, A. Yassouridis, H. Wienl, F. Then Bergh, T. Dose, H. Wek-

erle, and R. Hohlfeld, *Proc. Natl. Acad. Sci. USA* 97, 7452 (2000).
15. Demyelination arrest and remyelination induced by glatiramer acetate treatment of experimental autoimmune encephalomyelitis, R. Aharoni, A. Herschkovitz, R. Eilam, M. Blumberg-Hazan, M. Sela, W. Bruck R, and R. Arnon, *Proc. Natl. Acad. Sci. USA* 105, 11358 (2008).
16. Reductive cleavage of disulfide bridges in ribonuclease, M. Sela, F.H. White, Jr, and C.B. Anfinsen, *Science*, 125, 691 (1957).
17. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, C.B. Anfinsen, E. Haber, M. Sela and F.H. White, Jr., *Proc. Natl. Acad. Sci. USA* 47, 1309 (1961).
18. Efficacy of antibodies to epidermal growth factor receptor against KB carcinoma in vitro and in nude mice, E. Aboud-Pirak, E. Hurwitz, M.E. Pirak, F. Bellot, J. Schlessinger and M. Sela, *Journal Nat. Cancer Inst.* 80, 1605 (1988).
19. Synergistic down-regulation of receptor tyrosine kinases by combinations of monoclonal antibodies: implications for cancer immunotherapy, L.M. Friedman, A. Rinon, B. Schechter, L Lyass, S. Lavi, S.S. Bacus, M. Sela and Y. Yarden, *Proc. Natl. Acad. Sci. USA* 102, 1915 (2005).
20. Persistent elimination of ErbB-2/HER-2-overexpresssing tumors using combinations of monoclonal antibodies: relevance of receptor endocytosis, T. Ben-Kasus, B. Schechter, S. Lavi, Y. Yarden and M. Sela, *Proc. Natl. Acad. Sci. USA* 106, 3294 (2009).

The Evolving Concept of the Gene

Rafael Vicuña

> *... where the meaning of most four-letter words is all too clear, that of gene is not. The more expert scientists become in molecular genetics, the less easy it is to be sure about what, if anything, a gene actually is.*[1]

One of the disciplines of biology that showed more progress than any other during the 20th century was genetics, with advances in the area of molecular genetics being perhaps the most outstanding. Statements such as: *The gene concept has certainly been one of the landmarks in the history of science in the 20th century*;[2] *There can be little doubt that the idea of the gene has been the central organizing theme of twentieth century biology*;[3] *During the twentieth century, the gene has emerged as the major driving force of biology*,[4] reflect the relevance of the gene as a focal subject of study during the last century. One might think that an accurate concept of the gene was pivotal for this outcome. However, paradoxically, this is not the case, since defining the gene has always proven to be a difficult task, especially at the present time. In spite of the latter, geneticists have been able to thoroughly study how traits are passed down to progeny and how gene variation constitutes the basis of evolution.

This essay does not pretend to summarize the history of genetics, nor of the gene itself. Rather, its purpose is to highlight the landmarks of the evolving concept of the gene and to depict some recent findings that are making understanding the gene even more difficult. For a comprehensive collection of essays dealing with historical and epistemological perspectives of the concept of the gene, a recent book published by Cambridge University Press is highly recommended.[5]

[1] Pearson, H. What is a gene? *Nature* 441, 399-401, 2006.

[2] El-Hani, C.B. Between the cross and the sword: The crisis of the gene concept. *Genet. Mol. Biol.* 30, 297-307, 2007.

[3] Moss, L. *What genes can't do*. Cambridge, The MIT Press, 2003.

[4] Rédei, G.P., Koncz, C. & Phillips, J.D. Changing images of the gene. *Adv. Genetics* 56, 53-100, 2006.

[5] *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, 2000.

The birth of the gene concept

Heredity began to be studied as a scientific discipline only in the 20th century. In the early days, breeders were more concerned with the statistical analysis of inherited traits than with the material causation of them. The Austrian monk Gregor Mendel, after his discovery of the laws of segregation and of independent assortment published in 1865, was the first to suggest that reproductive cells were endowed with elements, characters or factors that had the potential for a trait. However, he did not make any distinction between whatever was transmitted in the seeds and the traits themselves. Three years later, most probably unaware of Mendel's work but certainly inspired by the teachings of Hippocrates, Charles Darwin presented what he dubbed a 'provisional hypothesis of pangenesis', according to which body cells produce minute particles called gemmules that travel to the reproductive cells, where they congregate prior to fertilization.[6] Although this hypothesis was not supported by observation, it allowed Darwin to explain phenomena such as the intermediate nature of hybrids and the heredity of acquired characters. Interestingly, Darwin's half cousin Francis Galton committed himself to prove the validity of this hypothesis by transfusing blood between dissimilar breeds of rabbits. After examining the features of their offspring, he concluded that there was no evidence of transmission of gemmules by this body fluid.

In 1889, reflecting his disagreement with heredity of acquired characters, Hugo de Vries proposed the theory of 'intracellular pangenesis', according to which animals and plants possess independent characters that are correspondingly associated with distinct particles called pangenes. These particles are located in the nucleus, where they replicate when the cell divides. Daughter cells each receive a complete set of pangenes, including the precursors of reproductive cells during their differentiation. Thus, pangenes do not leave the cells and their travel along the body never takes place. A few years earlier, the German zoologist August Weismann had advanced the similar theory of 'The continuity of the germplasm'. In his own words: *…heredity is brought about by the transference from one generation to another of a substance with a definite chemical, and above all, molecular constitution. I have called this substance 'germ-plasm' and have assumed that it possesses a complex structure, conferring upon it the power of developing into a complex organism*.[7] Weismann

[6] Darwin, Ch. Provisional hypothesis of pangenesis. In: *Animals and plants under domestication*, vol. 2. Orange Judd, New York, pp. 428-483, 1868.

[7] Weismann, A. The continuity of the germ-plasm as the foundation of the theory of heredity. In: Poulton E.B., Schonland S., Shipley A.I.E. (eds) *Essays upon heredity and kindred biological problems by Dr. August Weismann*. Clarendon Press, Oxford, 1885.

thought that the germplasm was only present in the cell lineage that leads to the reproductive cells, whereas somatic cell lineages had received from their progenitors only the material required for the development of the respective organs and tissues. In contrast, studies conducted in plants had rightly convinced de Vries that each nucleus of the body contains the same complete set of pangenes.

By the dawn of the new century, geneticists were mostly using the term *unit-character* for the entities responsible for specific traits that behaved as indivisible units of Mendelian inheritance. But the Danish botanist Wilhelm Johanssen liked de Vries's theory of pangenes and in 1909 coined the term gene for the *special conditions, foundations and determiners which are present in unique, separate and thereby independent ways (by which) many characteristics of the organisms are specified*.[8] The new designation came to replace that of unit-character. Although by inference, genes had to be present in the germ cells, their physical constitution was unknown and the concept proved useful to account for the transmission of traits from one generation to the next. In other words, it was a concept mainly based on a function that could be identified by genetic methods. Johanssen also introduced the terms genotype and phenotype, thus clearly distinguishing between the potential for a trait and the trait itself.[9] From 1910 to 1915, studying segregation of mutations in the fruit fly *Drosophila melanogaster*, Thomas Morgan and his group showed that genes reside in chromosomes and that they occupy specific locations in them, as beads on a string. He figured that the ability of genes to recombine was proportional to their distance in the chromosome. Studies on X-linked inheritance in the same organism allowed him to assign genes to the X chromosome. All together, contemporary work gave rise to the perception of the gene as a unit of function (one trait), a unit of mutation and a unit of recombination, a vision that prevailed until the early 1940s.

However, genes were still considered mainly as entities having the potential for a trait and whose effects could be inferred from them. In other words, efforts were focused more in traits as manifestations of genes rather than in their material counterparts. Morgan himself made this clear during his Nobel Prize Lecture in 1933: *Now that we locate [the genes] in the chromosomes are we justified in regarding them as material units; as chemical bodies of a*

[8] Johannsen, W. *Elemente der Exakten Erblichkeitslehre*. Gustav Fisher, Jena, 1909. Cited by Hall, B.K. The gene is not dead, merely orphaned and seeking a home. *Evol. Develop.* 3(4), 225-228, 2001.

[9] Falk, R. What is a gene. *Stud. Hist. Phil. Sci.* 17, 133-173, 1986.

higher order than molecules? Frankly, these are questions with which the working geneticist has not much concerned himself, except now and then to speculate as to the nature of postulated elements. There is not consensus of opinion amongst geneticists as to what genes are – whether they are real or purely fictitious – because at the level at which the genetic experiments lie, it does not make the slightest difference whether the gene is a hypothetical unit, or whether the gene is a material particle.[10] This reductionist approach, still not constrained to a specific material counterpart, led Raphael Falk to coin the term *instrumental gene*, to imply a hypothetical construct that was accepted as if it was a real entity.[11]

But there were also manifestations of a more material conceptualization of the gene. The very fact that a gene could be mutated[12] or recombined was certainly a consequence of its physical identity. Perhaps this evidence may have influenced Herman J. Muller, a member of Morgan's group, to support the notion that genes are 'ultra microscopic particles' found in the chromosomes rather than a 'pure idealistic concept divorced from real things'.[13] Another inclination towards a material nature of the gene was the genomere model proposed by Eyster to interpret gene instability expressed in variegated traits in fruit flies and spotting in corn kernels. This model stated that genes are composed of different particles that are unequally distributed during mitotic divisions.[14] Investigators such as Correns, Anderson and Demerec favored the genomere hypothesis, until it was disproven few years later by Muller.[15] And there were also the results obtained in 1928 by Griffith, showing that some substance originally present in killed virulent *Pneumococcus* cells was able to transform a non-virulent live strain into a virulent one.[16]

In the early 40s, George W. Beadle and Edward L. Tatum were studying metabolism in *Neurospora* and showed that certain mutations in genes caused errors in specific steps in metabolic pathways. This observation gave rise to

[10] Thomas H. Morgan, The relation of genetics to physiology and medicine. Nobel Lecture, Stockholm, June 1933; cited by R. Falk in: What is a gene? *Stud. Hist. Phil. Sci.* 17, 133-173, 1986.

[11] Falk, R. The gene in search of an identity. *Hum. Genet.* 68, 195-204, 1984.

[12] Muller, H.J. Artificial transmutation of the gene. *Science* 46, 84-87, 1927.

[13] Falk, R. What is a gene? *Stud. Hist. Philos. Sci.* 17, 133-173, 1986.

[14] Eyster, W.H. A genetic analysis of variegation. *Genetics* 9, 372-404, 1924.

[15] Muller, H.J. The problem of genic modification. Proceedings of the Fifth International Congress of Genetics, Berlin, 1927. *Z Induktive Abstammungs Vererbungslehre* [Suppl 1]: 234-260.

[16] Griffith, F. The significance of pneumococcal types. *J. Hyg.* (London) 27, 113-159, 1928.

the 'one gene-one enzyme' hypothesis, supporting the view that genes carried information related to the metabolic processes taking place inside the cells and more specifically, that each individual gene is responsible for the synthesis of a single enzyme.[17] Chemistry also had a role to play. Vernon Ingram showed that changes in two abnormal hemoglobins due to mutations were in each case confined to a single amino acid residue of the globin polypeptide. Since there could be no doubt that genes determine the amino acid residues of polypeptide chains, the expression 'one gene-one enzyme' was modified to 'one gene-one polypeptide'.[18]

The nature of the genetic material became even more tangible when Oswald Avery[19] and collaborators showed that the substance causing transformation in experiments that followed the protocols of Griffith's was DNA. Unambiguous confirmation of the DNA theory of inheritance was obtained few years later by Alfred D. Hershey and Martha Chase.[20] The structure of DNA proposed by Watson and Crick in 1953 gave the definite stroke to the instrumentalist view of the gene in favor of the realistic one, initiating the *classical molecular gene concept*. This states that a gene is a stretch of DNA that encodes a functional product, a single polypeptide chain or RNA molecule. Implicit in it is the idea that this genome unit performs one single function. At last, then, structure and function were blended in the same concept.

The newly revealed structure of DNA also encouraged speculation about the still prevailing idea of the gene as a unity of function, mutation and recombination. Prior to 1955, several investigators had already obtained the first hints that the unit of function might not be indivisible, since not only more than one mutation could be mapped to the same gene but also intragenic recombination had been detected in *D. melanogaster* and the fungus *Aspergillus nidulans*[21] (see also references therein). Who most clearly confirmed this was Seymour Benzer. The so-called *cis-trans* complementation test led him to coin the word *cistron* to imply the unit of genetic function.

[17] Beadle, G.W. and Tatum, E.L. Genetic control of biochemical reactions in Neurospora. *Proc. Natl. Acad. Sci.* 27**:** 499-506, 1941.

[18] Dunn, L.C. Old and New in Genetics. *Bull. New York Acad. Med.* 40(5): 325-333, 329, 1964.

[19] Avery, O.T., MacLeod, C.M., and McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* 79**:** 137-158, 1944.

[20] Hershey, A.D., and Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36, 39-56, 1952.

[21] Portin, P. The origin, development and present status of the concept of the gene: A short historical account of the discoveries. *Current Genomics* 1, 29-40, 2000.

Essentially, a cistron is a contiguous or uninterrupted piece of DNA encoding a particular protein. Cistrons turned out to be much larger than the units of mutation and recombination, thus confirming the possibility of multiple mutations within a single gene and of intragenic recombination.[22] Up to the present time, cistron is considered to be a synonym of gene, although its use is rather unfrequent.

New findings set hurdles to the gene concept

The development of DNA sequencing and of gene manipulation techniques in the 1970s allowed the rapid unveiling of the structure of genes and the detailed mechanisms involved in the regulation of their expression. Furthermore, the sequencing of large stretches of DNA and even of whole genomes led to the concept of *nominal gene*, to denote a sequence of DNA whose features allow the prediction of a protein sequence or of a known RNA product.[23] However, several novel situations related to the structure and function of genes that could not have been previously envisaged made evident that the classical molecular gene concept was completely improper. All at once, the gene seemed to have lost its identity both as a structural and as a functional unit. First of all, genomes from eukaryotes do not only consist of genes. They contain a significant fraction of non-protein coding and even highly repetitive sequences. Although originally labeled junk DNA due to its apparent lack of function, it has recently been shown that a large fraction of this DNA is transcribed (see later). Meaningless sequences in the genome did not necessarily challenge the classical molecular gene concept implying a stretch of DNA encoding a functional protein or RNA macromolecule. Neither did the fact that there are certain tandem repetitions of meaningful sequences, such as those encoding histones and ribosomal RNA. But other features widespread in genomes certainly did, especially some that entailed sequence elements external to the coding region, as well as dissection of the gene into smaller units. Some of the new findings contesting the classical gene concept (a stretch of DNA encoding a functional product) are the following:

a) Regulatory elements: already in 1961, F. Jacob and J. Monod had introduced the term promoter to describe a sequence located upstream of

[22] Benzer S. The elementary units of heredity. In: McElroy, W. and Glass, B. (eds) *The Chemical Basis of Heredity*. John Hopkins Press, Baltimore, pp. 70-93, 1957.

[23] Burian, R.M. Molecular epigénesis, molecular pleiotropy and molecular gene definitions. *Hist. Phil. Sci.* 26, 59-80, 2004.

the protein coding sequence that was responsible for controlling gene expression in bacteria. Later findings showed that all genes, both in prokaryotes and eukaryotes, require promoters for being transcribed into RNA. Since promoters can readily be recognized by their typical nucleotide sequences, they facilitate the identification of protein coding sequences in a genome. Thus, the term open reading frame (ORF) is commonly used to imply a DNA sequence that allegedly encodes a protein because it is flanked by a promoter (next to an initiation codon) and a stop codon. Often, promoters in eukaryotes are more effectively used when they are stimulated by cis-acting sequence elements called enhancers. These can be located either upstream or downstream of the promoter, sometimes thousands of base pairs away. Besides transcription, translation can also be regulated by sequence elements, which in this case are present in the transcript. A six nucleotide sequence located upstream of the initiating codon in bacterial mRNA, known as the Shine-Dalgarno element, contributes to positioning the initiating codon in the proper site of the ribosome. In mature eukaryotic mRNA, untranslated regions (UTRs) before the start codon (5' UTR) and after the stop codon (3' UTR) influence mRNA stability, mRNA localization and translational efficiency through proteins that specifically bind either one or the other, depending on the aspect to be regulated.

b) Intervening sequences: most eukaryotic genes are interrupted by non-protein coding sequences called introns, which are transcribed into RNA and thereafter removed prior to translation. Removal of introns and joining of the coding sequences (exons) is called splicing. Intron sequences largely exceed exons sequences with values of 20% versus 1.5%, respectively, in the human genome. On the other hand, many eukaryotic mRNAs can undergo alternative splicing, a process by which some exons are left out of the final transcript. In this case, a particular DNA segment in the genome can give rise to several variant proteins, thus expanding the coding capacity of the genome. It is estimated that 75% of the human genes are processed by alternative splicing. There is also the phenomenon of transplicing, mainly in lower eukaryotes, in which separate transcripts that may derive even from separate chromosomes are ligated to produce one mature mRNA. Splicing does not only occur at the RNA level. Intervening sequences in proteins (inteins) can be removed from a precursor protein and the flanking segments (exteins) can be ligated to generate a mature protein.

c) Transcripts including several genes: in bacteria it is widespread that genes involved in a particular biochemical pathway are clustered on the chromosome and transcribed together in a single polycistronic RNA. The gene cluster plus its single promoter is called an operon. Distribution of

genes in operon allows an efficient control of gene expression. Also, in bacteria, genes encoding 16S, 23S and 5S ribosomal RNAs are transcribed in a single pre-ribosomal RNA (30S), which is thereafter processed into its mature products. In turn, eukaryotes produce a 45S pre-ribosomal RNA that gives rise to 18S, 28S and 5.8S RNA. Studies on the human genome have also revealed the phenomenon called tandem chimerism, where two consecutive genes are transcribed into a single RNA. Differential splicing of this RNA can give rise to a fused protein containing domains encoded in both genes.[24,25]

d) Polyproteins: in this case, a transcript is translated into a protein that is subsequently cleaved to generate proteins with different activities. For example, transcription of retroviral DNA engenders one transcript comprising the *gag*, *pol* and *env* coding sequences. There are no intergenic sequences between them. The transcript is translated into a polyprotein corresponding to the *gag* and *pol* sequences that is cleaved into a total of six proteins: three viral structural proteins, an integrase, a protease and reverse transcriptase. On the other hand, splicing of the primary transcript gives rise to an mRNA encoding mainly the *env* gene. This is translated into another polyprotein that is processed to produce the viral envelope proteins.

e) Overlapping genes: in bacteria and viruses, as well in eukaryotes, genes sometimes overlap. Different proteins may be read from the same strand although in different reading frames. Reading frames may also be convergent or divergent, in which cases both DNA strands carry genetic information. When an entire coding sequence lies within the start and stop codon of another gene, typically in an intron, one speaks of a nested gene. There are also genes nested opposite the coding sequences of their host genes.

f) Genome rearrangements: immunoglobulins consist of two heavy and two light polypeptide chains. In turn, there are two types of light chains: kappa and lambda. Each of these chains, namely the heavy kappa and lambda chains, has a constant and a variable region. In all cases, the variable domain is encoded in a few hundred different gene sequences. Recombination of the latter with the sequences encoding the corresponding constant regions produces a wide diversity of light and heavy chains, which can in

[24] Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A. and Sorek, R. Transcription-mediated gene fusion in the human genome. *Genome Res*. 16, 30-36, 2006.

[25] Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, E.T., Castelo, R., Thomson, T.M., Antonarakis, S.E. and Guigó, R. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res*. 16, 37-44, 2006.

turn associate in all combinations. These DNA rearrangements explain how a mammal genome can literally produce millions of immunoglobulins.

g) Gene sharing: occurs when a single protein performs multiple functions. The most well-known example of gene sharing is eye lens crystallins. When expressed at low levels, the protein in many tissues functions as a housekeeping enzyme, but when expressed at high levels in eye tissue, it becomes densely packed and forms lenses.

h) Transcript editing: in eukaryotes, tRNAs, rRNAs and mRNAs may undergo chemical modifications that alter the information originally present at the DNA level. RNA editing mechanisms include cytidine to uridine and adenosine to inosine deaminations, as well as nucleotide additions and insertions. For example, in RNA transcripts coding for proteins in the mitochondria of trypanosomes, uridine nucleotides are inserted with the help of guide RNAs that hybridize to the transcript and direct the endonucleolytic cleavage of the RNA, the insertion of uridine nucleotides by uridylyl transferase and the subsequent ligation of the transcript, which now has both an altered sequence and reading frame. It is estimated that about one thousand human genes have an adenosine to inosine deamination. Editing is at odds with the classical gene concept because the RNA requires retrieving information from other genes to configure its final message.

Recently, as a result of the ENCODE project, new surprises complicated even further our understanding of the organization of the genome and of the gene concept itself. The *enc*yclopedia *o*f *D*NA *e*lements (ENCODE) consortium is an initiative launched by the National Human Genome Research Institute of the National Institute of Health (USA). It started with a pilot project aimed at thoroughly scrutinizing 30 mega bases (one percent) of the human genome, distributed in 44 genomic regions, with the goal to identify and map all the functional genetic elements.[26] Conducted between 2003 and 2007 by 35 groups from 80 organizations around the world, the ENCODE project confirmed what the Human Genome Project had anticipated, namely, that a genome entails much more than a mere collection of protein coding genes. One of the major findings of the ENCODE project was the realization that the majority (>90%) of the DNA is transcribed into primary transcripts that give rise to RNAs of various sizes. Most of them correspond to novel non-protein coding transcripts, some of which overlap protein coding sequences,

[26] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799-816, 2007.

whereas others derive from regions outside the previously annotated genes. Furthermore, there is abundant (>60%) overlapping between sense-antisense transcription across the genome. On average, 5.4 transcripts were found per annotated gene, showing some strand overlapping as well. Alongside, several alternative transcription start sites were identified for each protein coding sequence. About two thirds of the transcripts possess a tissue-specific 5 distal sequence, which can be located >100 kb upstream of the canonical start site.[27] Occasionally, transcripts span more than one annotated gene, as revealed by the presence of exons deriving from them. It is unclear how these fusion transcripts are generated. Possible mechanisms include trans-splicing and simply extended transcription. Neither is it known whether these fusion transcripts are translated. In addition, contrary to traditional views, regulatory regions were found both upstream and downstream transcription starting sites. In a subject which is beyond that of the gene concept, the ENCODE project also revealed that functional sequences do not seem to be evolutionary constrained, as shown by comparison with 28 other mammals. Moreover, about 40% of the constrained regions do not seem to play any functional role. In summary, the pilot step of the ENCODE project showed that the genome is a far more complex system than originally envisaged, with a variety of interconnected elements whose functionality we are only beginning to unravel.

Multiple efforts aiming at a consensus notion of the gene

As mentioned above, the gene concept was initially instrumental. It then turned into a material one, when the DNA macromolecule was identified as the carrier of the genetic message. The unit character of Mendelian genetics became a sequence in the DNA encoding a functional product, either a protein or RNA. But from the time this classical gene concept proved to be inadequate in the light of the complexity of the genome, there have been various attempts to improve it. For example, Fogle has proposed that as opposed to a unit, a gene is a construct resulting from the assemblage of embedded, tandem and overlapping domains in the DNA, a domain being a sequence that can be distinguished by virtue of its structural properties (exon, promoter, enhancer, etc).[28] Thus, although two organisms may have

[27] Denoeud, F., Krapanov, P. *et al.*: Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res*. 17, 746-759, 2007.

[28] Fogle, T. Are genes units of inheritance? *Biol. and Philos*. 5, 349-371, 1990.

a similar number of 'classical' genes, they may differ to a great extent in the way the domains combine to constitute *domain sets for active transcription* (DSATs, in Fogle's nomenclature). In turn, Brosius and Gould offer a new terminology for better understanding genome structure, function and evolution.[29] They recommend the term *nuon* for denoting any stretch of nucleic acid sequence that may be identifiable by some criterion. A nuon can be a protein-coding sequence, intergenic region, exon, intron, promoter, enhancer, terminator, pseudogene, telomere, etc. A sequence undergoing an adaptive change should be called *aptonuon*. On the other hand, it is well known that duplicated genes have the potential to give rise to a novel function, after passing through a silent stage. During this period, they should be called potogenes or *potonuons* (*pot*ential nuons). If these sequences appear to have become obliterated as genomic noise, they should be termed *naptonuons* (*non*a*p*tive nuons). If, in contrast, potonuons have been coopted for another function, they should be called *xaptonuons* or xaptogenes, since they constitute exaptation events.[30] This proposition by Brosius and Gould has not prevailed in the scientific community. A couple of years later, Waters proposed that the fundamental concept of the gene is that of a linear sequence in a product at some stage of genetic expression.[31] Thus, an intron is part of the gene if the focal point is the process of transcription itself, but it is not a gene if the focus of interest is the function of the protein encoded in it. Strange as it may seem, this concept allows a single gene at the DNA level to encode for several genes at the mRNA level (alternative splicing). Considering that this definition varies during different stages of the expression process, it does not contribute to clarification in language use.

Interestingly, Griffiths and Neumann-Held think that a univocal definition of the gene may not be necessary or even desirable, since different gene concepts may be useful in different areas of biology.[32] However, their opinion is that it is critical to be aware of the differences among the various concepts in order to use them properly in their corresponding domains. These authors

[29] Brosius, J. and Gould, S.J. On 'genomenclature': A comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proc. Natl. Acad. Sci. USA* 89, 10706, 10710, 1992.

[30] Gould and Vrba coined the term exaptation for designating functional features of the phenotype that were not built by natural selection as adaptations of the original function but were rather coopted from structures arising from adaptations for other functions.

[31] Waters, C.K. Genes made molecular. *Phil. Sci.* 61, 163-185, 1994.

[32] Griffiths, P.E. and Neumann-Held, E.M. The many faces of the gene. *BioScience* 49, 656-662, 1999.

are particularly concerned about the distinction between the molecular gene and the evolutionary gene. According to them, the difficulties with the classical molecular gene concept arise because it is centered in structure rather than in function (a stretch of DNA encoding a protein or RNA product). Therefore, they suggest replacing it by the *molecular process gene concept*, in which gene denotes the recurring process that leads to the temporally and spatially regulated expression of a particular polypeptide product. In this new definition, although the gene still implies a DNA segment encoding a polypeptide, the emphasis is placed on the process that allows this sequence to be expressed. Thus, if a transcript of a certain DNA segment undergoes differential splicing or editing that vary depending on the tissue or stage of development, such segment fits the proposed definition. Moreover, the latter takes into consideration other functions that participate in causing the sequence to generate its product. On the other hand, there is the *evolutionary gene concept*, first introduced by Williams[33] and then elaborated by Dawkins to denote any stretch of DNA that can be replaced by an alternative (alelomorphic) sequence in future generations.[34] Griffiths and Neumann-Held agree with Dawkins in that evolutionary genes need not necessarily be molecular genes, *e.g.* often do not correspond to specific stretches of DNA. However, as opposed to Dawkins, they lay emphasis on the fact that rather than being loosely defined segments in the DNA, evolutionary genes have particular roles in the expression of phenotypic traits. The evolutionary gene concept has also been worked by P. Beurton, who claims that the gene is the smallest collection of genetic elements that underlies a single adaptive difference and is thus a target of natural selection.[35] In this case, a collection refers to the fact that a phenotypic trait may involve several genetic elements, each of which is a target of selection. Another approach that is focused to function rather than to structure is the *developmental gene concept*, as advanced by Gilbert[36] and

[33] Williams, G.C. *Adaptation and natural selection*. Princeton NJ. Princeton University Press, 1966.

[34] Dawkins, R. *The extended phenotype*. Oxford: W.H. Freeman, 1982.

[35] Beurton, P.J. A unified view of the gene, or how to overcome reductionism, in *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, pp. 286-316, 2000.

[36] Gilbert, S.F. Genes classical and genes developmental: The different use of genes in evolutionary syntheses, in *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, pp. 178-192, 2000.

Morange.[37] They assign this name to genetic elements that play a leading role in shaping the phenotype through the development of parts or segments of organisms. Although these genes are widely distributed among metazoans, the concept of developmental gene is rather restricted to this particular aspect of biological function.

In 2003 Snyder and Gerstein defined the gene as a complete chromosomal segment responsible for making a functional product.[38] This definition encompasses the inclusion of both regulatory and coding regions, the expression of a gene product and the requirement that it be functional. Criteria to be used in order to identify genes in the DNA sequence of a genome include the identification of ORFs, specific sequence features (codon bias, splicing sites), sequence conservation among organisms, evidence for transcription and gene inactivation, the latter being aimed at ascertaining gene's function.

A different approach has been undertaken by Lenny Moss.[39,40] This author argues that there are two markedly distinctive meanings or senses of the gene. Although both are associated to the phenotype, neither indicates that the phenotype can be decomposed down to a compilation of genes. First there is *gene-P*, which has a predictable relationship with some feature of the phenotype. One speaks of the gene for muscular dystrophy, obesity or premature aging. In other words, every time we use the expression a 'gene for' a certain trait, we are referring to a gene-P. This concept is indeterminate with respect to the material gene, *i.e.* to the specific sequence of DNA. So indeterminate is this concept with respect to the DNA sequence that in common language one often speaks of the gene for a certain trait when such trait (a disease for example) is expressed due to the absence of the wild type or normal sequence. The P in the gene-P concept stands for 'preformationism', because it evokes the idea that all the traits are determined at the moment of birth. In contrast, there is the concept of *gene-D*,

[37] Morange, M. The developmental gene concept: History and limits, in *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, pp. 193-218, 2000.

[38] Snyder, M. and Gerstein, M. Defining genes in the genomics era. *Science* 300, 258-260, 2003.

[39] Moss, L. The question of questions: What is a gene? Comments on Rolston and Griffiths&Stotz. *Theor. Med. Bioethics* 27, 523-534, 2006.

[40] Moss, L. The meanings of the gene and the future of phenotype. *Genet. Soc. Policy* 4, 38-57, 2008.

which is specifically associated with a particular DNA sequence that can give rise to an RNA transcript. Gene-D is not determined with respect to phenotype, because it is unable to predict the appearance of a particular trait. Most often this is due to the fact that, as shown by studies at the molecular level, each DNA sequence contributes or is involved in the manifestation of several phenotypic outcomes, the resulting one depending on contextual factors. The D in gene-D stands for 'developmental resource', having therefore a more holistic scope that gene-P. According to Moss, a typical gene-D is NCAM (neural cell adhesion molecule), which in the fly *Drosophila* can give rise to about 38,000 proteins by differential splicing of a gene that possesses 19 exons. The different domains encoded in each of these exons will determine the cellular activity of the protein and hence the resulting phenotype.

On the other hand, Scherrer and Jost have proposed to preserve the concept of the gene as a basis of a function, that is to say, the sequence encoding a polypeptide within an mRNA, even though in most cases such sequence is not present at the DNA level as an uninterrupted sequence.[41] The gene in the mRNA is flanked by untranslated regions (5'- and 3'- UTRs). Superimposed onto the coding sequence is the *genon*, a program of oligomotifs that are eventual binding sites for regulatory proteins or small RNAs. At a higher level there is the *transgenon*, constituted by all the factors that influence gene expression by binding to the motifs in the genon. These factors are selected from the holotransgenon, which comprises all the factors (polypeptides and small RNAs) influencing gene expression in the cell. These concepts also apply when the gene product is RNA instead of a protein. A different approach is taken by Keller and Harel, which, according to these authors, is better grounded in biological findings than the gene has proven to be.[42] They define a *dene* as a DNA sequence plus all the elements that in a dynamic fashion make it functional (regulatory proteins and RNAs, epigenetic modifications, etc). The *bene* is the behavior of the organisms with which the dene is associated. In turn, the *genetic functor* or *genitor* is the logical relation that says whenever the organism's DNA is seen to satisfy the property expressed by the dene, its behavior satisfies the property expressed by the bene. This nomenclature offered by Keller and Harel is intended to em-

[41] Scherrer, K. and Jost, J. The gene and the genon concept: a functional and information-theoretic analysis. *Molec. Syst. Biol.* 3, 1-11, 2007.

[42] Keller, E.F. and Harel, D. Beyond the gene. PLoS ONE 2(11):e1231. doi:10.1371/journal.pone.0001231.

phasize the distinction between what an organism statically is (what it inherits) and what it dynamically does (its functionality and behavior).

After assessing the novel findings of the ENCODE project, Gerstein *et al*.[43] suggested five criteria to update the definition of the gene, namely: 1) the new description should comprise the former meaning of a gene; 2) it should be valid for any living organism; 3) it should be simple; 4) it should be straightforward, so anybody could distinguish the number of genes in a particular genome and 5) it should be compatible with other biological nomenclature. In addition, the new definition must take into account that the gene is a genomic sequence encoding a functional protein or RNA, it must consider the union of overlapping sequences when there are several functional products and it must be coherent in the sense that the union must be done separately for protein and RNA products, not being necessary that all the products share a common sequence. Gerstein *et al*. further put forward a new definition of the gene as a union of genomic sequences encoding a coherent set of potentially overlapping functional products. If there are no introns or no overlapping products, the new definition coincides with the classical one. Since this new definition covers only coding sequences, it does not include regulatory regions and untranslated regions (5' and 3' UTRs) in the RNA. In addition, it does not cover RNA editing.

There are two other recent attempts to define the gene that deserve to be mentioned because they represent collective efforts. One is that of the Human Genome Nomenclature Organization, which states that a gene is a DNA segment that contributes to phenotype/function. In the absence of a demonstrated function, a gene may be characterized by sequence, transcription or homology.[44] The other one, adopted by the Sequence Ontology consortium, was elaborated by 25 scientists and required two days to reach a consensus: a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions. The latter definition seems quite ample enough to accommodate most of the features challenging the classical gene concept, although it does not seem to accommodate well phenomena such as transplicing and gene rearrangements.

[43] Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S. and Snyder, M. What is a gene, post ENCODE? History and updated definition. *Genome Research* 17, 669-681, 2007.

[44] Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W. and Povey, S. Guidelines for human gene nomenclature. *Genomics* 79, 464-470, 2002.

All the aforementioned efforts to find a common norm to define the gene are undoubtedly worthwhile, but in some cases, perhaps overly elaborated. There have also been other straightforward propositions. For example, the highly respected textbook *Molecular Biology of the Cell,*[45] a bestseller among biology students throughout the world, defines a gene as a sequence that is transcribed as a single unit and encodes one set of closely related polypeptide chains. This definition has room for DNA segments that may give rise to various proteins due to alternative splicing, RNA editing, post-translational modifications, etc. Surprisingly, it doesn't make explicit the possibility that a gene may also generate a non protein coding RNA. Another definition, also with a molecular accent, is offered by Epp:[46] a gene is the nucleotide sequence that stores the information which specifies the order of the monomers in a final functional polypeptide or RNA molecule, or set of closely related isoforms. Epp stresses that regulatory sequences should not be considered part of a gene because there are too many types of them, they generally operate in complex combinations and often they influence the expression of several DNA segments. Besides, according to Epp, genes do not have to be expressed to be present.

Defining a gene remains an enduring endeavor

As it can be deduced from the aforementioned propositions for defining a gene, they are centered either in structure or in function. Interestingly, closely related definitions underlining the molecular quality of the gene, such as those offered by Gerstein *et al.*, the textbook *Molecular Biology of the Cell* and Epp, seem to be the most commonly accepted in the community of biological scientists. This is not only an impression based on subjective experience, but there are empirical signs that this is actually the case. For example, a few years ago, Stotz *et al.*[47] conducted a survey among Australian biologists from different areas (medicine, pharmacy, veterinary science, biochemistry, etc) to find out how they conceptualized the gene. Several questions were asked regarding the gene concept itself and the application of the gene concept to specific cases. The great majority of the responses ob-

[45] *Molecular Biology of the Cell*, 5th ed. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (eds) Garland Science, NY, 2008.

[46] Epp, C.D. Definition of a gene. *Nature* 389, 537, 1997.

[47] Stotz, K., Griffiths, P.E. and Knight, R. How biologists conceptualize genes: an empirical study. *Stud. Hist. Philos. & Biomed. Sci.* 35, 647-673, 2004.

tained favored the classical molecular concept, which is based more in structure than in function.

This way of conceptualization may reflect that in the era of genomics it is essential to be able to identify genes during the annotation of a newly sequenced genome. In addition, genetic engineering with academic or industrial purposes also requires a clear identification of the DNA segment that needs to be transferred in order to confer the desired phenotype to the recipient organism. These definitions assume the common criterion of leaving aside the concept of regulatory sequences controlling gene expression. However, they do not seem to encompass the phenomena of overlapping sequences, transplicing and RNA editing. Neither does Waters' more unified gene concept as a linear sequence in a product at some stage of genetic expression. The more comprehensive definitions to date appear to be those offered by ENCODE project and the Sequence Ontology Consortium, with the corresponding limitations already mentioned. Perhaps these could be overcome with a proper combination of both definitions. Thus, a gene could be defined as 'a union of genomic sequences encoding a coherent set of potentially overlapping functional products, these sequences being associated with regulatory regions, transcribed regions and/or other functional sequence regions'. However, this definition seems too intricate for everyday use. It is likely that once we get accustomed to the idea that a gene may comprise several segments dispersed throughout the genome and that it may also produce multiple transcripts that affect the same function, a definition such as the latter will prevail. In the meantime, other novel approaches may be worth considering. For example, taking into account the complex transcriptional organization of the genome, Gingeras contends that a simple operational unit linking a specific DNA sequence to phenotype/function is required.[48] According to this author, RNA transcripts are such fundamental operational units. Thus, each transcript could be catalogued according to the function it affects.

Raphael Falk, who has greatly contributed novel thoughts in the field, thinks that to arrive at a structural definition of the gene is a fruitless undertaking.[49] It may be even more difficult to merge the structural and func-

[48] Gingeras, T.R. Origin of phenotypes: Genes and transcripts. *Genome Res*. 17, 669-681, 2007.

[49] Falk, R. The gene – A concept in tension, in *The concept of the gene in development and evolution: Historical and epistemological perspectives*. Beurton, P.J., Falk, R. and Rheinberger, H.J. (eds) Cambridge University Press, Cambridge, pp. 317-348, 2000.

tional aspects in a single definition. Obviously, this state of affairs has not stopped scientists and philosophers to confront this task, simply because the gene concept represents a central issue in the biological sciences. Somewhere in the way, investigators have advanced reasons to declare the concept of the gene dead, to be thereafter refuted with arguments showing just the opposite.[50] Fortunately, finding a univocal definition of the gene persists as an ongoing intellectual challenge, because it gives us the opportunity to witness a fascinating display of thoughts and ideas at the boundary of knowledge. In the meantime, experimental molecular geneticists will continue to progress in the understanding of genome structure and expression. This situation evokes that of biology itself, in whose various branches scientists have been able to make paramount advances in spite of lacking a formal definition of living beings.

[50] Hall, B.K. The gene is not dead, merely orphaned and seeking a home. *Evol. &Dev.* 3:4, 225-228, 2001.

Molecular Darwinism and its Relevance for Translational Genetic Research

Werner Arber

Evolutionary biology and genetics have their roots around 1860 in fundamental publications by Charles Darwin and by Gregor Mendel, respectively. At that time, both of these fields of the life sciences had their experimental basis in the observation of inherited phenotypical traits of higher organisms, plants and animals. It is only around 1940 that microbial genetics made its start. Within a few years, bacterial transformation revealed that the nucleic acid DNA is the carrier of genetic information [1], and bacterial conjugation showed that genetic information of a donor bacterium became linearly transferred into a recipient bacterium upon close contact between the two cells [2]. This latter observation soon turned out to be consistent with the filamentous, double-helical molecular structure of DNA described in 1953 [3]. At that time, it became clear that genetic information is encoded by the linear arrangement of building blocks of DNA, i.e. nucleotides in a single strand of DNA or base pairs in the double-stranded helical DNA molecules. Already before this fundamental insight, bacteriophage-mediated transduction was discovered [4], in which a bacterial virus acts as a vector for bacterial genes which thereby can become horizontally transferred from a donor bacterium into a recipient bacterium. It is on the basis of these discoveries of research in microbial genetics and in structural biology that molecular genetics started and developed rapidly in the second half of the 20th century. It became thereby known that classical genes consist of linear sequences of nucleotides that encode in their reading frame a gene product that is mostly a protein and sometimes an RNA molecule. The average gene length is about 1000 nucleotides. Much shorter nucleotide sequences serve as expression control signals with which other gene products can positively or negatively interact. In view of this advanced knowledge it became clear that spontaneously appearing altered phenotypical traits of individuals must have their cause in an alteration in the nucleotide sequence of the genome. While classical genetics had defined a mutation as an altered phenotype that gets transmitted to the progeny, molecular genetics now defines the mutation by an altered nucleotide sequence. An experimentally based critical evaluation of this situation shows that by far not all spontaneously occurring nucleotide sequence alterations

result in altered phenotypes. Indeed, many nucleotide alterations in the genome remain without immediate influence on life processes. Some of these silent, neutral mutations may at some later times become of functional relevance together with still other mutational alterations of the genome. Among the spontaneously occurring nucleotide sequence alterations affecting a biological function, a majority is functionally unfavorable and provides a selective disadvantage. In contrast, favorable, 'useful' mutations providing a selective advantage are relatively rare. This situation can serve as an argument that spontaneous mutagenesis is, in general, not directed towards a particular, identified goal. Rather spontaneously occurring genetic variations must be largely contingent.

According to the theory of biological evolution spontaneously occurring genetic variation drives biological evolution. Without genetic variation there would be no evolution. The directions that evolution takes depend on the impact of natural selection and on the, at any time, available genetic variants. Natural selection is seen as the impact exerted by both physicochemical and biological, environmental constraints on the organisms living in ecosystems. A third pillar of biological evolution besides genetic variation and natural selection is reproductive and geographic isolation. This isolation modulates the evolutionary process.

Thanks to research strategies of molecular genetics, it has become possible to experimentally investigate molecular mechanisms of spontaneous genetic variation. Without going into experimental details, we will summarize here the available results of studies that were mostly carried out with microorganisms and then also validated for higher organisms. Relatively unexpectedly these studies revealed that a multitude of specific molecular mechanisms contribute to the overall genetic variation. These mechanisms can be assigned to three natural strategies of genetic variation, namely local sequence changes, intragenomic rearrangements of segments of the DNA filaments, and DNA acquisition by horizontal transfer of a DNA segment from another kind of living organism [5,6]. These natural strategies of genetic variation contribute with different qualities to the steady but slow progress of biological evolution.

Local nucleotide sequence changes can, for example, occur for various known reasons during DNA replication. This can result in a nucleotide substitution, in the deletion or the additional insertion of one or a few nucleotides or in a scrambling of a few adjacent nucleotides. These processes can occasionally lead to a stepwise improvement of an available biological function. The genetic variant in question may then profit from its selective advantage and eventually overgrow its parental population.

Intragenomic DNA rearrangements are often guided by genetically encoded recombination enzymes, such as for general recombination between largely homologous DNA segments, for transposition of mobile genetic elements, and for site-specific DNA reshuffling. By pure chance, a segment-wise DNA rearrangement may lead to an improvement or to novel combinations of available functional capacities. The fusion of two different functional domains of open reading frames, as well as the fusion of an open reading frame with an alternative expression control signal, may have their origin in such occasionally occurring DNA rearrangements. This might perhaps lead to a sudden emergence of novel properties, a phenomenon that had so far not found a satisfactory explanation.

Emergence of a novel property might also have its cause in the acquisition of a foreign DNA segment by horizontal gene transfer. A number of viruses are known to act occasionally as natural gene vectors, both in microorganisms as already discussed and in higher organisms. Horizontal gene transfer can serve in nature for the acquisition of a foreign functional domain, a single gene or a small group of genes. This strategy of sharing in successful developments made by others is quite effective and can provide to an organism a novel functional capacity in a single step of evolutionary progress. This process is facilitated by the universality of the genetic code [7].

The scientific insights into the mechanisms and natural strategies of genetic variation can validly contribute to our worldview and they have thus cultural, philosophical values. Particular genes that we now call evolution genes contribute as variation generators and/or as modulators of the rates of genetic variation to the evolutionary progress of populations. They do this together with non-genetic elements such as a limited chemical stability and structural flexibilities of biologically active molecules. Still other non-genetic elements involved in genetic variation are environmental mutagens and random encounter. We can conclude that the natural reality takes actively care of biological evolution. This represents an expansion of the Darwinian theory to the level of molecular processes that we call here molecular Darwinism.

We assume that the evolution genes exerting their activities in today's available living beings had become fine-tuned in their functions by second-order selection [8] in the course of their long past history. Biological evolution is governed by two natural antagonistic principles: on the one hand the promotion of genetic variation which is the driving force of biological evolution, and on the other hand a limitation of the rates of genetic variation. This provides a relatively comfortable genetic stability to most individual organisms, and it contributes to the longer-term preservation of species. Biological evo-

lution is the source of biodiversity. The natural potency to evolve guarantees for the future a steadily developing, rich biodiversity. This evolutionary progress could be considered as a permanent creation.

Since evolution genes belong to the genome of each living organism, the genomes show a conceptual duality: on the one hand, many genes work for the benefit of the individuals, for the fulfillment of their lives. The underlying genetic determinants are housekeeping genes, accessory genes of use under particular life conditions, and genes contributing in higher, multicellular organisms to the embryonic development of each individual. On the other hand, evolution genes contribute with their products to the occasional genetic variation in randomly involved individuals. This evolutionary driving force serves for an expansion of life and, as we have already discussed, for biodiversity.

We must be aware that life in natural environments is much more complex than under most experimental laboratory conditions. Nevertheless, the laws of nature discussed here guiding biological evolution are very likely to be of general validity. Most living species possess an evolutionary fitness by being equipped with evolution genes for each described strategy to generate genetic variants. We are more and more aware that symbiosis between different kinds of organisms plays important general roles in ecosystems. Plants, animals and human beings are full of microorganisms without being sick. Rather, these cohabitating organisms provide mutual help (symbiosis) to the partners in the communities. Their evolutionary potency also helps the populations to occasionally adapt to changes occurring in the composition of the ecosystems. This can include the possibility of horizontal gene transfer that might be favored under conditions of cohabitation.

Research strategies based on genetic engineering have been developed since the 1970s and they now serve both in fundamental and in applied research. In genetic engineering segments of DNA can be separated, purified and differently spliced together. Recombinant DNA often contains DNA segments from more than one genome. Genetic engineering can also produce local nucleotide sequence alterations by site-directed mutagenesis. All of these research methods are very similar to natural events of genetic variation that we have outlined above. As a matter of fact, both natural genetic variation and genetic variation directed by genetic engineering follow the same rules based on natural laws of biological evolution. This mainly involves relatively small steps of genetic alterations. In addition, any resulting genetic variants and hybrids become subsequently submitted to the laws of natural selection based on the requirement for a functional harmony and on the ability to deal with the encountered environmental conditions.

It follows from these considerations that longer-term evolutionary risks of genetically engineered organisms must be similar to comparable risks occurring in the natural processes of biological evolution. Similar risks are also expected for classical plant and animal breeding strategies. From long-time experience we know that such risks are quite small both for breeding techniques and for natural biological evolution. Thus, we can expect similarly low long-term evolutionary risks for genetic engineering. This holds as long as experimental procedures do not involve specifically designed, entirely novel DNA sequences which may, so far, have not been present in the biosphere. These reflections are of relevance for any project of translational genetic research. In addition, such research projects, particularly those involving human beings and higher animals, should pay full respect to ethical considerations on a case-by-case basis of specific projects.

As far as genetically modified food crops are concerned, a road map for agro-biotech applications has recently been proposed which would deserve to be followed for a functional improvement of nutritional values and for a more stable health of our most important food plants [9,10]. This could considerably improve the nutritional conditions and the food security for the world population. It has been reminded, however, that such a beneficial development in the next few decades should not be taken as a signal for a continued population growth. Rather, in view of improved health conditions and significant reduction of malnutrition and hunger, the human society should be reminded to attain a stable equilibrium of the population density by a responsible parenthood. Such an equilibrium could ensure a long-term sustainability of our cultural evolution, respecting the high diversity of forms of life and of the habitats for all living organisms on our planet Earth.

References

[1] Avery, O.T., MacLeod, C.M. and McCarty, M. (1944), Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.* 79, 137-158.

[2] Lederberg, J. (1947), Gene recombination and linked segregation in *E. coli*, *Genetics*, 32, 505-525.

[3] Watson, J.D. and Crick, F.H.C. (1953), Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid, *Nature*, 171, 737-738.

[4] Zinder, N. and Lederberg, J. (1952). Genetic exchange in Salmonella, *J. Bacteriol.*, 64, 679-699.

[5] Arber, W. (2003), Elements for a theory of molecular evolution, *Gene*, 317, 3-11.

[6] Arber, W. (2007), Genetic variation and molecular evolution, In: Meyers, R.A. (ed.), *Genomics and Genetics*, Wiley-VCH, Weinheim, vol. 1, 385-406.

(7) Arber, W. (2006), The evolutionary strategy of DNA acquisition as a pos-

sible reason for a universal genetic code, *Hist. Phil. Life Sci.*, 28, 525-532.
[8] Weber, M. (1996), Evolutionary plasticity in prokaryotes: a panglossian view, *Biol. Philos.*, 11, 67-88.
[9] Arber, W. (2009), The impact of science and technology on the civilization, *Biotechnology Advances*, 27, 940-944.
[10] Arber, W. (2010), Genetic engineering compared to natural genetic variations, *New Biotechnology*, 27, 517-521.

Evo-Devo: the Merging of Evolutionary and Developmental Biology

Edward M. De Robertis

Introduction

In the beginning of the 20th century, developmental biology was at the forefront of biology, but then declined and had a renaissance towards its end. The key to this revival were the techniques of molecular biology, which proved the great equalizer for all branches of biology. The fusion of molecular, developmental and evolutionary biology proved very fertile, and led to the birth of a new discipline, Evo-Devo. I would like to present a personal account on how this synthesis took place.

We will consider here three main points:

1) How are the mechanisms of self-regulation of cell differentiation observed in animal development explained at the molecular level?
2) How were conserved ancestral gene networks common to all animals – which pattern the Antero-Posterior (A-P) and Dorso-Ventral (D-V) axes – used to generate the immense variety of animal forms?
3) How has the use of a common tool-kit of genes present in the ancestral animal genome channeled the outcomes of evolution through natural selection?

The main conclusion that emerges from these genomic, developmental and evolutionary studies is that all bilateral animals – which comprise 30 of the 34 extant phyla – arose through gene mutation, duplication or deletion of the genome of a complex common ancestor, the *Urbilateria* (Ur: primeval; Bilateria: animals having bilateral symmetry).

1. Self-regulation of differentiating cell fields

1.1. Embryology at the forefront of biology

When biologists realized that it was necessary to take an experimental – rather than descriptive – approach to understand the mechanisms of development, embryology rapidly became the leading edge of biological studies. Embryos offer excellent material for experimental biology.

After fertilization, an amphibian egg – a large cell 1.2 mm or more in diameter – divides synchronously into 2, 4, 8, 16, 32, 64 and so on cells. At

these early stages, cells are dedicated to sensing their position within the embryo by signaling to each other without differentiating into particular tissues. At the 10,000 cell stage, cells on the dorsal side start to invaginate to the interior of what at this point constitutes a blastula or hollow ball. The cells that involute will form the endoderm and mesoderm of the body, while cells that remain on the outside give rise to ectoderm. By the end of this process – called gastrulation – a vertebrate embryo with defined A-P and D-V axes and differentiated tissue types is formed.

The beginning of experimental embryology can be traced back to 1891, when Hans Driesch separated the first two cells of a sea urchin embryo and obtained two complete larvae. At the turn of the century, in 1901, Hans Spemann obtained amphibian twins by gently constricting embryos with fine ligatures of hair from his newborn daughter. Much later, I found that identical twins can also be generated by simply bisecting an early embryo of the frog *Xenopus laevis* with a scalpel blade before gastrulation starts.

This tendency of the embryo to regenerate towards the whole is called self-regulation. This is not a property restricted to the early embryo. Most organs in the body start their development as 'morphogenetic fields' that are able to self-regulate their differentiation. This was discovered by Ross G. Harrison, who showed in 1918 that a circular region of flank mesoderm could induce the development of forelimbs when transplanted into a host embryo. When he cut this region in half, each half induced a limb. Not a half-limb, but rather a complete limb. From these transplantation experiments we learned that cells within the organism do not lead solitary lives, but are instead subsumed in larger fields of hundreds or thousands of cells that communicate to each other when to proliferate, differentiate, or die. We are only now beginning to understand the molecular mechanisms by which these cellular conversations take place.

1.2. Hans Spemann and embryonic induction

The way forward in the analysis of self-regulation of pattern came from a transplantation experiment carried out by a graduate student at Freiburg University, Hilde Mangold. Under the direction of Spemann, she transplanted the dorsal lip of the blastopore, the region in which the involution of mesoderm starts, and introduced it into the opposite (ventral) side of a host embryo. With a gentle push, the embryonic fragments heal together almost miraculously, and two days later perfect Siamese (conjoined) twins are formed. Spemann called this dorsal region of the embryo the 'organizer'.

Remarkably, the transplanted organizer cells themselves gave rise to notochord, yet were able to induce their neighboring cells to change their

differentiation into dorsal tissues such as central nervous system (CNS), somite (muscle), and kidney. Therefore, within the embryo, groups of cells (called organizing centers) are able to instruct their neighbors on the type of cell differentiations they should adopt.

Spemann was awarded the 1935 Nobel Prize for Physiology or Medicine for the discovery of embryonic induction by organizer tissue, which marked the apogee of experimental embryology. However, the isolation of the chemical substances responsible for embryonic induction proved impossible given the methods available at the time. After that, the genetics of Thomas Hunt Morgan became the pre-eminent biological discipline for most of the 20th century.

2. The ancestral A-P and D-V gene networks

2.1. Thomas Morgan, Edward Lewis and homeotic mutations

Morgan started his career as an embryologist. For example, he demonstrated that a 2-cell frog egg could self-regulate to form a whole embryo after killing one cell, but only when the dead cell was removed. He realized, however, that mechanistic progress using this type of experimental approach would be very difficult, and decided to study mutations in the fruit fly *Drosophila melanogaster* instead. Together with his graduate student Calvin Bridges, in 1923 Morgan isolated a mutant, *bithorax*, which gave rise to four-winged flies (flies normally have only two wings). This mutant was to provide the key that made possible the molecular analysis of development.

In 1946, a young student at Caltech, Edward B. Lewis, initiated studies on the genetics of the *bithorax* locus, which continued until his passing in 2004. He found that the *bithorax* region patterned the thorax and abdomen of the fruit fly and contained several genes. When mutated, these genes caused homeotic transformations, i.e., the transformation of one region of the body into the likeness of another region. For example, the third thoracic segment may become transformed into the second thoracic, thus generating the four-winged flies.

Remarkably, Lewis noted that the arrangement of homeotic genes in the DNA followed the same order in which they regulated the A-P identity of abdominal segments. He designated this surprising organization colinearity. Lewis proposed that homeotic genes had repressed thoracic identity in a centipede-like ancestor, and that recent duplications of these genes had further elaborated the identity of each abdominal segment.

When molecular biology became practical, the race to clone a homeotic gene began in several laboratories. It culminated with the isolation of *Anten-*

napedia, a homeotic gene that can transform antenna into leg, independently by Scott and Kaufman, and by Garber and Gehring in 1983. Searching for the hypothetical recently duplicated genes of Lewis, they discovered that many *Drosophila* homeotic genes crossreacted with a short region of DNA. This conserved segment of nucleic acid, called the homeobox, was found to encode a DNA-binding domain of 60 amino acids, designated the homeodomain.

2.2. Hox genes in vertebrates

At that time I was a professor in the same department as Walter Gehring at the Biozentrum of the University of Basel, Switzerland, and we shared group meetings. We decided to collaborate to test whether homeobox genes might be present in vertebrates. (The experiment was conceived for the wrong reasons: the first expression studies by Garber had shown *Antennapedia* expression in the CNS, and we suspected it might encode a peptide hormone, which were known at the time to have been conserved between Hydra and mammals). On the first try we cloned a gene, now called HoxC-6, from a *Xenopus laevis* genomic library which crossreacted with *Antennapedia* and *ultrabithorax* (Carrasco *et al.*, 1984). The sequence of the homeodomain was very similar to that of *Antennapedia*. Later gene knockout studies by Mario Capecchi and others showed that this gene, like the other 39 Hox genes, caused A-P homeotic transformations when mutated in the mouse. This was a good thing, because in our paper in the last sentence of the introduction I had written: 'If the frog gene cloned here eventually turns out to have functions similar to those of the fruit fly genes, it would represent the first development-controlling gene identified in vertebrates'. And so it was.

Vertebrate Hox genes are clustered in the genome. Work by other groups, mostly in mouse embryos, showed that vertebrate Hox gene expression in the body is colinear with their order in the DNA. The homeobox sequences and overall organization of the vertebrate Hox gene complexes were conserved with those of *Drosophila* and other invertebrates. Therefore, Lewis' hypothesis that homeotic genes were recently duplicated genes was not correct, yet provided the cornerstone for the new discipline of Evo-Devo. Edward Lewis received the Nobel Prize for Medicine or Physiology for his work on developmental genetics in 1995.

2.3. Whole-genome duplications in the vertebrate lineage

Many insects have eight or so Hox genes arranged in a single cluster. *Amphioxus*, a chordate closely related to the vertebrates, has a single cluster containing 14 Hox genes in a row. However, the situation is more complex in the vertebrates. This is because vertebrates underwent two rounds of

whole-genome duplications at the beginning of their evolution. Thus, for each gene humans may have up to four copies. Many of our genes are now present as single copies, but this only indicates that the other three were lost. Gene loss is easily achieved over evolutionary time. Duplicated genes are retained when a duplicated copy acquires a specialized function that makes it beneficial for the survival of the species. These two genome-wide duplications were probably a crucial event in the remarkable evolutionary success of vertebrate animals.

Humans contain four Hox gene complexes, called HoxA through HoxD. Each consists of about 100,000 base pairs of DNA and resulted from the duplication of an ancestral Hox complex containing 13 genes. However, instead of $13 \times 4 = 52$, humans retained a total of only 39 Hox genes. This is because some Hox genes were deleted. As will be discussed below, gene loss is an important force in shaping evolution.

The degree of conservation between these four mammalian Hox complexes and *Drosophila* is simply amazing. Not only homeobox sequences and colinearity of expression patterns were maintained, but even their regulation by an inhibitory microRNA (called *infra-abdominal-4* in *Drosophila* and miR196 in humans) was conserved.

This intricate genetic machinery that patterns the A-P axis could not have been assembled independently twice in *Drosophila* and vertebrates, let alone in all phyla. The only reasonable interpretation is that a Hox complex was already functional in *Urbilateria* and was inherited by its descendants. The discovery of conserved Hox gene complexes led to the realization that the gene networks that control the A-P axis share deep historical homologies. Before the discovery of the homeobox we did not imagine that the mechanisms of development would be so similar between fruit flies and humans. It was a great surprise.

2.4. François Jacob's symposium on Evolution and Development

In 1991, a landmark meeting was held in Crete. Organized, among others, by academicians Nicole Le Douarin and Fotis Kafatos, it was entitled *Evolution and Development*. Its topic had been specifically requested by François Jacob, who was retiring. Jacob, a great geneticist, was very interested in evolution. In his excellent book, *The Possible and the Actual* (1982), Jacob explained why bringing these two separate fields together was important: 'For it is during embryonic development that the instructions contained in the genetic program of an organism are expressed, that the genotype is converted into phenotype. It is mainly the requirements of embryonic development that, among all posible changes in genotype, screen the actual

phenotypes'. The main argument of his book was that during evolution old components are retained and used again, comparing evolution to the work of a tinkerer or *bricoleur*. A tinkerer uses parts or materials that already exist to assemble objects having new purposes.

Jacob displayed great insight in bringing together developmental and evolutionary biologists as his swan's song. The symposium took place at the perfect time, when the conservation of the Hox system was already understood in general outlines. The star of the meeting was paleontologist Stephen Jay Gould. Wishing to learn more about evolution, I asked him to sit at my table during breakfast. Although he really wanted to read his newspaper in peace, I proved too eager and he reluctantly accepted. Gould recommended I should read two books. The first one was Gould's own *Wonderful Life*, which told the story of the Cambrian explosion in the fossil record.

The Cambrian explosion refers to the remarkable finding that all the body plans (34 phyla) of animals that exist today appeared as fossils over a narrow period of time, between 535 to 525 million years ago. Before that time a long line of Precambrian ancestors must have existed, but they left very few or no adult bilaterian fossils (except for tracks and trails in the ocean floor dating to 630 million years ago). We do not know why the appearance of body plans occurred so suddenly, and many possibilities have been proposed (Valentine, 2004). For example, in the 'snowball earth' scenario the diversification of body plans resulted from repeated bottlenecks of intense natural selection coinciding with several massive glaciation events that covered most of the earth between 750 and 550 million years ago. Even more mysterious than the sudden emergence of phyla, is the question of why no new animal body plans have evolved since then, for which we currently have no answers.

2.5. Geoffroy Saint-Hilaire and the unity of plan

The second book that Gould recommended was one by Toby Appel, on the historical debate that took place at the French Academy of Sciences between Georges Cuvier and Etienne Geoffroy Saint-Hilaire in 1830. Geoffroy held the view that a unity of plan existed among animals. In 1822, he dissected a lobster and placed it in an inverted position with respect to the ground. In this upside down orientation the lobster's normally ventral nerve cord was located above the digestive tract, which in turn was placed above the heart. In his own words: 'What was my surprise, and I add, my admiration, in perceiving an ordering that placed under my eyes all the organic systems of this lobster in the order in which they are arranged in mammals?'

Geoffroy went on to argue that there was a unity of plan, or design, among animals, so that the dorsal side of the vertebrates was homologous to the ven-

tral side of the arthropods. For historians of science the Cuvier-Geoffroy debate was of great interest because it took place decades before Charles Darwin published his *Origin of Species* in 1859. For our own work, reading this book was crucial, because when a few years later we isolated Chordin, we were prepared. Chordin was a dorsal protein secreted by Spemann's organizer that had a close homologue in the ventral side of the *Drosophila* early embryo.

At Jacob's symposium I presented the first investigations from our laboratory on the chemical nature of embryonic induction by Spemann's organizer. At that time, we had constructed libraries containing the genes expressed in dorsal lips manually dissected from the frog gastrula. We had just isolated a gene expressed exclusively in organizer tissue called *goosecoid*. It encoded a DNA-binding protein, but we knew from Spemann's work that embryonic induction required secreted factors able to change the differentiation of neighboring cells.

By continuing these explorations on the molecular nature of induction by organizer tissue, we isolated several secreted proteins such as Chordin, Frzb-1 and Cerberus, and other groups isolated Noggin, Follistatin and Dickkopf (De Robertis, 2006). Unexpectedly, all of these proteins turned out to function as antagonists of growth factors in the extracellular space. They prevent binding of growth factors to their receptors on the cell membrane, thus inhibiting signaling. Although we had hoped to isolate novel signaling growth factors from the organizer, what was discovered instead was that embryonic induction was mediated mainly through the secretion of a cocktail of inhibitory proteins.

2.6. Chordin, BMP and cell differentiation

Chordin proved to be the most informative of the organizer factors. Transplanted organizers in which Chordin expression is inhibited lost all embryonic induction activity. Thus, Chordin is essential for organizer function. Chordin induces the differentiation of dorsal tissues (such as CNS or muscle) by binding to Bone Morphogenetic Proteins (BMPs), which normally cause the differentiation of ventral tissues (such as epidermis or blood). Two BMP genes are expressed in the ventral region of the embryo, and Chordin is secreted in prodigious amounts by dorsal cells. In principle, this would suffice to establish a gradient of BMP activity, yet by further investigating the system we discovered much more complexity.

Dorsal-ventral tissue differentiation results from a biochemical network of proteins secreted by the dorsal and ventral sides of the embryo. For each action of the dorsal organizer there is a compensating reaction in the opposite side of the embryo. The expression of genes on the dorsal and the ventral sides are

under opposite control, which explains in part the self-regulation phenomenon. The dorsal side also expresses BMPs, which when bound to Chordin are able to flow towards the ventral side. There, a protease called Tolloid specifically degrades Chordin, liberating BMPs for signaling through its cell surface receptors. The flow of Chordin and its cleavage by this protease are key steps in maintaining the self-regulating gradient of BMP activity. A number of additional secreted proteins (called Sizzled, Crossveinless-2, Twisted gastrulation and Crescent) function as feedback regulators, providing additional resilience to the D-V patterning system (De Robertis, 2009).

Remarkably, other investigators found that this basic biochemical network is also used to regulate cell differentiation along the D-V axis in the early embryos of many other organisms, such as *Drosophila*, beetles, spiders, hemichordates, amphioxus, zebrafish and chick. This intricate molecular machinery is most unlikely to have evolved independently multiple times during evolution specifically to control D-V patterning. The reasonable conclusion is that the Chordin/BMP/Tolloid pathway patterned the dorsal-ventral axis of the last common bilaterian ancestor and was inherited by its descendants.

The conservation of the Chordin/BMP/Tolloid system provided strong molecular support for the hypothesis of Geoffroy Saint-Hilaire that the mammalian and arthropod body plans are homologous. An inversion of the D-V axis occurred during evolution. The ventral side of the arthropods is equivalent to the dorsal side of the vertebrate, and the entire Chordin/BMP/Tolloid pathway was inverted. In both vertebrates and invertebrates, the CNS is formed where the gradient of BMP signaling is lowest. A unity of plan, both for the A-P and D-V axes, exists among animals.

3. A conserved gene tool-kit generates variety in evolution

3.1. Urbilateria had considerable regulatory complexity

These deep homologies in the way all embryos pattern their A-P and D-V axes are having a profound impact on current evolutionary thinking. One might argue that the power of natural selection of the fittest, working on chance mutations over immense periods of geological time, is *per se* sufficient to explain the variety of animal forms. In the absence of any constraints, competition in crowded ecosystems, particularly among closely related species, would lead to new and improved animal designs in the victorious species, through the creative force of natural selection. Ever more adapted generations would be formed because the invisible guiding hand

of natural selection integrates useful mutational changes, forming ever fitter individuals and gradually generating new structures and species. On the other hand, what we are now learning is that a very important source of variation for specifying the arrangements of cells with respect to each other – which is what ultimately determines morphological change – resides in the ancestral developmental gene networks shared by all animals.

3.2. Eyes have a common origin

One might argue that while the Hox and Chordin/BMP gene networks are complex, they could have been used to pattern a very simple ancestral animal. However, there are reasons to think that *Urbilateria* was anatomically complex. One such reason is provided by the ancestral eye structures.

An important problem in evolution is whether adaptations arise through homology or convergence. Homology means that two structures are derived from an ancestral one present in a common ancestor. An example of homology could be the hoof of a horse and the middle digit of the ancestors from which it evolved. Convergence occurs when similar solutions are reached to resolve common functional needs. An example could be the wings of various animal groups, which evolved at very different times but represent similar solutions to a functional requirement. Distinguishing between homology and convergence in evolution can be very difficult. Now molecular biology gives us a historical record of how evolution took place. In the case of animal eyes, conventional wisdom was that animal eyes had arisen independently 40 to 60 times through convergent evolution to fulfill the need for vision.

In 1994 Walter Gehring's group isolated the *eyeless* gene from *Drosophila* and found it had homology to the mammalian *Pax6* homeobox gene. In the mouse, mutations in *Pax6* caused the *small eye* phenotype. In humans, the *Aniridia* gene corresponded to *Pax6*. When mouse *Pax6* was artificially expressed in the antenna or leg precursors of *Drosophila* embryos, it caused the formation of ectopic eyes (Gehring, 1998). Of course, these were *Drosophila* eyes, not mouse ones. In the reciprocal experiment, overexpression of *Drosophila eyeless/Pax6* induced eyes in microinjected frog embryos. The eyes of the clam *Pecten,* and even those of jellyfish, also express *Pax6*. The reasonable conclusion is that all eyes are derived from an ancestral eye that expressed *Pax6*.

3.3. The urbilaterian CNS was anatomically elaborate

One might argue that the eye of *Urbilateria* could have been a very simple photoreceptor cell. However, this does not seem to be the case. We now have a very detailed understanding of the molecular switches (called transcription factors) that control the differentiation of the different neurons of the retina,

which derives from the forebrain. The morphology of mammalian and *Musca domestica* eyes had been described in loving detail by Santiago Ramón y Cajal. In 1915, he noted that by simply displacing the cell body (soma) of two neurons in *Musca*, leaving the cell projections and synaptic connections in place, the entire arrangements of intricate neural connections was maintained, with only small variations, between flies and humans. Recent studies have shown that the transcription factors expressed by various mammalian retinal neurons (photoreceptors, bipolar, and retinal ganglion cells) are replaced in the predicted corresponding fly neurons by their *Drosophila* homologues genes. This has provided molecular confirmation for Cajal's homologies, which he had predicted from pure morphology (Sanes and Zipursky, 2010).

Extensive conservations in 'molecular fingerprints' of particular combinations of transcription factors have also been found between vertebrate and *Drosophila* nerve cord neurons. In addition, mammalian brain hypothalamic neurosecretory cells express the same combinations of transcription factors as their corresponding *Drosophila* or annelid counterparts, which are located within the CNS region traversed by the mouth in protostomes. These neurosecretory peptides, important for sensing and signaling the availability of food, are expressed in the infundibulum of the mammalian brain, through which the gut probably traversed in our hypothetical ancestors (Tessmar-Raible *et al.*, 2007). Thus, *Urbilateria* had a CNS, including eyes, that was sophisticated both from molecular and anatomical standpoints. Before this stage was reached, a long line of Precambrian ancestors must have existed, in which their brains, neural circuits, and eyes were gradually perfected.

3.4. Animals share a conserved genomic tool-kit

Until recently the history of animal life on earth had to be deduced from the fossil record. Rapid advances in DNA sequencing have now made available entire sequenced genomes from multiple animal phyla. Because the genetic code arose only once, evolutionary studies are now less dependent on paleontology. We will be able to reconstruct the history of life on earth, registered in the language of DNA, with a degree of precision that seemed impossible only a decade ago. For those interested on how animal evolution actually took place, comparative genomics offers the best of times.

The most important lesson we have learned so far from genome sequences is that animals from the most diverse phyla share a common ancestral tool-kit of genes (De Robertis, 2008). In particular, all the signaling pathways used by cells to communicate with each other – and therefore to regulate their anatomical position with respect with each other in the body – were already present in pre-bilaterian such as cnidarians (sea anemones,

medusae and Hydra). Therefore, evolutionary changes resulted from the shuffling of a full ancestral set of genes, rather than from the introduction of new genetic mechanisms from scratch. There was remarkably little biochemical novelty during animal evolution.

3.5. Adaptive mutations

DNA sequencing has given us the opportunity of identifying the adaptive mutations that were actually selected during the evolution of animal populations in nature. The main types of variations on which natural selection acted to select the adaptive ones were: cis-regulatory mutations, structural gene mutations, gene duplications and gene deletions.

Cis-regulatory mutations are those found in the regulatory regions – called enhancers – located in cis (in the same DNA molecule) near genes. Enhancers regulate in which tissues genes are expressed. Enhancer DNA sequences provide binding sites for combinations of transcription factors that turn genes on and off. By changing the tissue or region in which a gene is expressed, morphological change can be generated. For example, crustaceans such as shrimp and lobsters evolved a considerable diversity of feeding appendages; it has been shown that these changes repeatedly correlated with independent shifts in the border of expression of Hox genes. New enhancers can be readily generated by bringing together combinations of DNA binding sites. They can also be easily lost without paying a large penalty, because the protein encoded by the gene remains and can still be expressed in other tissues under the control of the remaining enhancer elements. Mutations in tissue-specific enhancers are a major source of variations in evolution (Carroll, 2005). However, because enhancers are not highly conserved in sequence, their mutations are rarely detected by automatic sequence comparisons.

Structural mutations affect the sequence of the proteins encoded by genes. Interestingly, adaptive changes many times result from selection of mutations in the same gene. Melanism can be a useful adaptation. Melanic leopards, jaguars, mice, birds and lizards all arose from amino acid changes that increased the activity of the Melanocortin-1 receptor (Hoekstra and Coyne, 2007). Conversely, decreased activity of this receptor is seen in yellow Labradors and human redheads. Thus, natural selection chooses the same solutions repeatedly.

Gene duplications are very powerful source of evolutionary variation because the duplicated gene can be used to fulfill new functions without loss of the original gene (Ohno, 1970). For the molecular biologist they offer the additional advantage that the duplication – or the deletion – of an

entire gene is easily recognized when comparing genomic DNA sequences, thus facilitating the reconstruction of the history of animals.

Gene deletions are a very effective, although generally underappreciated, source of adaptation. Many cave animals – such as salamanders, shrimp and fish – adapt to their new troglodyte environment by losing their eyes and skin pigment. In the case of Mexican Tetra fish, their entrapment in subterraneous caves has led to deletion events in the *ocular and cutaneous albinism gene-2* that occurred independently in different populations (Protas *et al.*, 2006). Natural selection tends to choose mutations in the same genes over and over again. Although gene deletions are an effective way of rapidly adapting to changes in the environment, this is achieved at the expense of limiting future evolutionary flexibility.

3.6. Gene losses in the ancestral tool-kit

There are 30 bilateral animal phyla with distinct body plans, which can be classified in two branches. In the protostomes (mouth-first), the mouth is formed near the blastopore – these animals include most invertebrates. In the deuterostomes (mouth-second), the blastopore gives rise to the anus and the mouth is perforated secondarily - these animals include the phylum *Chordata* to which we belong. For example, if a gene is found both in fruit flies and in humans, it was also present in their last common ancestor, *Urbilateria*, as well. Similarly, if a gene is found both in pre-bilaterian animals such as sea anemones as well as in humans, it follows that this gene was also present in *Urbilateria*.

The role of gene loss in the evolution of Phyla has been highlighted by the sequencing of a sea anemone genome. The bilaterian lineage separated from cnidarians, at least 650 million years ago, from a common animal ancestor designated *Ureumetazoa*. About 2.5% of sea anemone genes are not present in any higher animals but, interestingly, have homologues in fungi and plants. The human genome contains twenty-plus genes of the Wnt family of growth factors. These can be arranged into 13 subfamilies according to their sequence. The sea anemone has 12 Wnt genes, each corresponding to one of the human subfamilies. (Kusserow *et al.*, 2005). Therefore, *Urbilateria* had genes corresponding to at least 12 Wnt subfamilies. Sequencing of the nematode *C. elegans* showed that it has a grand total of five Wnts; the *Drosophila* genome contains only seven. Thus, our human lineage retained most of the ancestral Wnt genes, while worms and fruit flies lost a great many. There are also examples in the opposite direction, in which humans have lost genes present in other vertebrates such as fish, frog or chick. Comparative genomics indicates that gene losses, as well as duplications, may have played an important role in the evolution of body plans.

3.7. Historical constraints in animal evolution

A key question in Evo-Devo is to what degree the deep homologies in embryonic patterning networks have channeled the outcomes of evolution. Many body plans that could have been excellent functional solutions might not exist in nature because they cannot be constructed unless they are compatible with the developmental networks that control the blueprint of animal body form. The respective contributions of functional needs and structural constraints is of great interest in evolutionary biology (Gould, 2002). Paraphrasing François Jacob, not all that is possible finds its way into the actual animal world.

The deep homologies in the developmental tool-kit seem likely to have constrained animal evolution by natural selection. Constraints resulting from the obligatory use of these ancestral patterning networks should not be considered a negative influence. On the contrary, mutations in these gene networks may have been a positive influence that channeled effective adaptation responses to the strictures of natural selection. Adaptation tends to follow the channel of least resistance to ensure survival of the species and it seems likely that modifications in developmental networks have been used repeatedly to resolve related functional needs. Many anatomical structures now considered to result from convergent evolution may turn out to result from the deep homologies in the genetic structure of all animals. Evolution of animal forms involved tinkering with the conserved A-P, D-V, and other developmental gene networks.

3.8. Open questions in Evo-Devo

Three directions will be particularly important for the young discipline of Evo-Devo:

- First, the reconstruction of the ancestral genetic tool-kit from which all animals were built should be a priority. This is at present a bioinformatic computing challenge. Many complete genome sequences are available already. Ideally the DNA of at least one species for each one of the 34 phyla should be completed. The ancestral tool-kit of yeasts has been determined and has proven interesting. Several groups are close to assembling an ancestral mammalian genome. Reconstructing the hypothetical genome of our urbilaterian ancestors will be very informative concerning the origin of body plans – particularly with respect to the role played by gene duplications and deletions during evolution.
- Second, retracing the adaptive mutations that caused the actual anatomical changes selected by natural selection is another priority. Biology is a historical science, and it will be fascinating to unravel the successive molecular steps by which we evolved into our present human condition.

– Third, determining how cells read their positional information in the embryo and adult tissues within self-regulating fields of cells will have both medical and evolutionary implications. In the organism, cells receive a multitude of signals that must be integrated and transformed into well-defined cell behaviors. These responses include cell division, differentiation and death, and are ultimately the determinants of morphological change.

Conclusion

The merging of Evolution and Development at the end of the 20th century has already provided important insights into how animals evolved an immense variety of body forms. The astonishing realization that has already emerged from Evo-Devo is that all animal life on earth evolved by differential use of the same ancestral tool-kit. A crucial role was played by variations in ancestral developmental gene networks that are hard-wired within our DNA.

Bibliography

Appel, T.A. (1987), *The Cuvier-Geoffroy Debate*, Oxford University Press, Oxford.

Cajal, S.R. and Sanchez, D. (1915), Contribución al conocimiento de los centros nerviosos de los insectos, *Trab. Lab. Invest. Biol. 13*, 1-167.

Carrasco, A.E., McGinnis, W., Gehring, W.J. and De Robertis, E.M. (1984), Cloning of a *Xenopus laevis* gene expressed during early embryogenesis that codes for a peptide region homologous to *Drosophila* homeotic genes: implications for vertebrate development, *Cell 37*, 409-414.

Carroll, S. (2005), *Endless Forms Most Beautiful: The New Science of Evo-Devo*, W.W. Norton & Co., Inc., New York.

Darwin, C. (1859), *On the Origin of Species by Means of Natural Selection, or Preservation of Favored Races in the Struggle for Life*, Murray, London.

De Robertis, E.M. (2006), Spemann's organizer and self-regulation in amphibian embryos, *Nat. Rev. Mol. Cell Biol. 7*, 296-302.

De Robertis, E.M. (2008), Evo-Devo: Variations on Ancestral themes, *Cell 132,* 185-195.

De Robertis, E.M. (2009), Spemann's organizer and the self-regulation of embryonic fields, *Mech. Dev. 126*, 925-941.

De Robertis, E.M. and Sasai, Y. (1996), A common plan for dorso-ventral patterning in Bilateria, *Nature 380*, 37-40.

Gehring, W.J. (1998), *Master Control Genes in Development and Evolution: The Homeobox Story*, Yale Univ. Press, New Haven.

Geoffroy Saint-Hilaire, E. (1822), Considérations générales sur la vertèbre, *Mém. Mus. Hist. Nat. 9*, 89-119.

Gould, S.J. (1989), *Wonderful Life*, W.W. Norton & Company, New York.

Gould, S.J. (2002), *The Structure of Evolutionary Theory*, Harvard University Press, Cambridge, Massachusetts, Chapter 10.

Hamburger, V. (1988), *The Heritage of Experimental Embryology*, Oxford University Press, New York.

Hoekstra, H.E. and Coyne, J.A. (2007), The locus of evolution: Evo Devo and the genetics of adaptation, *Evolution Int. J. Org. Evolution 61*, 995-1016.

Kusserow, A., Pang, K., Sturm, C., Hrouda, M., Lentfer, J., Schmidt, H.A., Technau, U., von Haeseler, A., Hobmayer, B., Martindale, M.Q. and Holstein, T.W. (2005), Unexpected complexity of the Wnt gene family in a sea anemone, *Nature 433*, 156-160.

Jacob, F. (1982), *The Possible and the Actual*, University of Washington Press, Seattle.

Monod, J. (1971), *Chance & Necessity*, Alfred A. Knopf, New York.

Ohno, S. (1970), *Evolution by Gene Duplication*, Springer-Verlag, Heidelberg.

Protas, M.E., Hersey, C., Kochanek, D., Zhous, Y., Wilkens, H., Jeffery, W.R., Zon, L.I., Borowsky, R. and Tabin, C.J. (2006), Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism, *Nat. Gen. 38*, 107-111.

Ratzinger, J. (1995), *In the Beginning*, B. Eerdmans Publishing Co., Michigan.

Sanes, J.R. and Zipursky, S.L. (2010), Design principles of insects and vertebrate visual systems, *Neuron 66*, 15-36.

Tessmar-Raible, K., Raible, F., Christodoulou, F., Guy, K., Rembold, M., Hausen, H. and Arendt, D. (2007), Conserved sensory-neurosecretory cell types in Annelid and fish forebrain: insights into Hypothalamus evolution, *Cell 129*, 1389-1400.

Valentine, J.W. (2004), *On the origin of Phyla*, The Univ. of Chicago Press, Chicago.

New Developments in Stem Cell Biotechnology

Nicole M. Le Douarin

The subject of stem cells has attracted a great deal of interest in the public during the last twelve years. Indeed it brings about the hope of a novel medicine through which cells in the adult organism that are deficient or subjected to massive death could be replaced by healthy ones. With the increase in longevity in industrialized countries, such instances, resulting from degenerative diseases, are more and more common. This *regenerative medicine* would complement therapeutics relying on surgery, chemistry and antibodies, which are one of the most important legacies of the 20th century.

During the last four decades it has been recognized that stem cells are present in virtually all tissues in adult vertebrates and are a source of youth, since their role is to replace cells which regularly die during the lifetime of the individual. Moreover, vertebrate embryos are entirely made up of stem cells at the early stages of their development. This pluripotent state of embryonic cells is transitory, but can be captured thanks to the spectacular advances in the biotechnologies during the last decades. It is now possible to maintain this particular *stemness* state in culture, thus generating permanent cell lines, endowed with the properties of this pivotal and intriguing type of cells.

The term *stem cell* can be found in the scientific literature of the first half of the 20th century. However, its definition was not clear until it was based on rigorous experimental criteria. This was achieved in the 1960s thanks to a series of studies that demonstrated the mechanisms through which the replacement of blood cells, whose normal lifespan is short, takes place.

In this article I will review the seminal work that has led to the scientific definition of a stem cell and then go through the successive breakthroughs that have stood out as landmarks in the field over the years and have led to the state of the art of today.

The definition of stem cells

As a general rule, the cells of the body that differentiate to fulfil definite functions have a lifespan shorter than that of the organism and are, therefore, subjected to constant and periodic renewal. Cell turnover varies considerably from one type of tissue to the other. It is rapid for the epithelium lining the

intestinal lumen whose cells are replaced every three to five days, or for the skin epidermis that, in humans, is renewed every twenty-one to thirty days. In the blood, the erythrocytes survive one hundred and twenty days after they have reached their fully functional state.

In the nervous tissues in contrast, most neurons are not renewed during lifetime in Mammals except in some areas such as the olfactory bulb in rodents and regions of the brain associated to memory (e.g. the hippocampus).

The concept of stem cells and the demonstration of their properties emerged from the observation that victims of the Hiroshima and Nagasaki nuclear bombs, who did not die at the time of explosion, died ten to fifteen days later in a state of advanced anaemia, with severe depletion of the bone marrow and spleen. The bone marrow had previously been recognized to be a site of production, proliferation and maturation of blood cell progenitors, and this effect was attributed to the sensitivity of dividing cells to ionizing radiations.

Experiments were conducted in the mouse that reproduced this effect of irradiations on the blood cell lineage. It was shown that irradiated mice could be rescued if they received bone marrow cells from histocompatible donors. Rescue was complete provided that donor cells became stably engrafted within the recipient spleen and bone marrow, thus providing a long-term reconstitution of the irradiated recipient hematopoietic system by the injected cells.

One of the consequences of this treatment was the fact that the spleen size, which shrank after the irradiation, regained its normal volume after the hematopoietic reconstitution.

At that time, two views were held concerning the cells that were at the origin of the renewal of the blood cell lineages. One proposed that each type of blood cells (e.g., erythrocytes, and the different sorts of leucocytes) were produced by a distinct undifferentiated progenitor. This was held by the tenants of the *polyphyletic origin* of the blood cells. According to the other view (*monophyletic*), one single pluripotent progenitor was at the origin of the various types of blood cells. The problem was solved by experiments carried out in the early 1960s by two Canadian haematologists working in Toronto, James E. Till and Ernest A. McCulloch [1]. Their experimental design consisted in reducing as much as possible the number of bone marrow cells able to reconstitute the blood cell system of the irradiated recipient. This goal was attained with 10^5 bone marrow cells. This experimental protocol led to the formation on the recipient's shrunken spleen of individually distinguishable bumps instead of the general swelling of the organ observed after the injection of larger numbers of cells.

They could show that each of these bumps, which contained all kinds of blood cells (except lymphocytes), corresponded to the engraftment of one single progenitor of donor origin. They were subsequently able to demonstrate

that the progenitor cell at the origin of the colony had also produced in its progeny undifferentiated cells similar to itself, which were able to produce new colonies if subjected, *in vivo* or *in vitro,* to appropriate conditions.

These experimental data led to the denomination of these blood cell lineage progenitors as *Hemopoietic Stem Cells* (HSC). These HSC were endowed with the following characteristics:

HSC are undifferentiated, divide asymmetrically and give rise to a cell similar to themselves (which remains undifferentiated and slow dividing) and to another cell with high proliferative potential, which can yield various phenotypes of differentiated cells.

In other words, stem cells are undifferentiated, pluripotent and able to self-renew, thus forming a reserve of cells able to maintain *homeostasis* in adult tissues by renewing cells that disappear through normal cell death.

One can consider the characterization of the hematopoietic stem cells as the *first breakthrough* discovery in the history of the stem cell field.

Apoptosis or normal cell death

One of the major advances in the field of cell biology in the second half of the last century was the discovery of the genetic mechanisms leading to natural cell death, also designated as *Apoptosis*.

Studies carried out on a Nematode, *Caenorhabditis elegans,* revealed that all living cells possess a gene network that enables them to commit suicide. Thus, these *suicide genes* need to be inhibited for the cell to be able to survive. Environmental signals such as growth and survival factors counteract the intrinsic cellular apoptotic machinery.

Cell death by apoptosis is unobtrusive, it starts by fragmentation of its DNA and then of its cytoplasm, and the cellular debris of the dying cells is rapidly absorbed by the neighboring cells. This is the reason why apoptosis had not been described before.

This process plays a major role during development, which involves the production of cells in excess. It is one of the means through which shaping of the organs and of the body is achieved. Moreover, it is a natural barrier against the development of tumors since, when a cell becomes abnormal by mutations paving the way to cancer, its cell death program is most often activated. This role is further attested by the fact that mutagenesis targeted to genes involved in apoptosis in the mouse markedly increases the incidence of tumors.

Cell death by apoptosis is involved in *tissue homeostasis*, which is the equilibrium between elimination of aged or abnormal cells and their replacement by new cells. The latter role belongs to the stem cells present in virtually all adult tissues.

The origin of the adult stem cells

Experiments carried out on the mouse in the 1960s have demonstrated that, at its early stages of development (i.e. morula and blastocyst stages), the mammalian embryo is composed of a clump of cells that have stem cell characteristics: they are pluripotent and able to self-renew [2]. At the blastocyst stage, the germ is composed of an epithelium that becomes the placenta (after implantation of the conceptus in the uterus) and lines a cavity in which sits an inner cell mass (*ICM*) from which the embryo develops. The cells of the *ICM* are all equivalent and each of them is able to produce all the differentiated cell types present in the adult body. Thus, one single cell of the *ICM* of an 'A' strain of mouse (with black fur), introduced within the blastocyst of a 'B' (white colored) strain recipient at the same stage, yields a chimeric mouse all tissues of which are composed of a mosaic of A and B cells. This is evident from its fur, which exhibits black and white hairs.

This early stage, where all embryonic cells are pluripotent and equivalent, is *transitory* and ends with the process of *gastrulation*, which leads to the formation of the three germ layers: *ectoderm, mesoderm, endoderm*. In each of these layers the potentialities of the embryonic cells become restricted to a defined set of phenotypes that will characterize the organ and tissue that they respectively yield. In each of these organs and tissues a reserve of stem cells subsists. These will remain undifferentiated and, later on, ensure the renewal of the differentiated cells that have reached the end of their normal life span.

These stem cells will, in the adult, subsist as discrete populations located within a 'niche' in which they will be 'protected' and maintained in an undifferentiated, pluripotent state by environmental factors. These adult stem cells are very few and, to a certain extent, specified since they produce only cells of the same type as those of the tissues they belong to.

Adult stem cells have been found in virtually all types of tissues, even in the brain and spinal cord where no new neurons were supposed to be produced after birth in mammals and birds. In fact, a certain level of cell renewal exists also in the nervous tissue and neural stem cells have been characterized in both the central (brain, spinal cord) and the peripheral nervous systems (CNS, PNS).[1]

One can consider that, in the history of the stem cell subject, the discovery of the HSC is the conceptual acquisition upon which sits the whole field

[1] For more information see *Des chimères, des clones et des gènes*. (2000) Odile Jacob Ed. and *Cellules souches, porteuses d'immortalité*. (2007) Odile Jacob Ed., by Nicole Le Douarin.

that developed later on. Twenty years later a second step took place that considerably widened its interest owing to the perspective of the potential applications it offered. This step, which pertains to the biotechnologies, consisted in 'capturing' the transitory state of pluripotency exhibited by the early mammalian embryonic cells to make it permanent. This technology has enabled to immortalize embryonic cells in a normal state in which they remain still capable of differentiating in all the cell types encountered in the adult mammalian body if provided with appropriate conditions.

The generation of Embryonic Stem Cells

In 1981 two laboratories [3] published a striking result: cells of mouse embryos of the 129 strain could be cultured permanently while remaining in the same pluripotent and undifferentiated state they exhibited in the inner cell mass. This was achieved by the particular culture conditions provided by the co-culture on certain feeder layers. If withdrawn from this environment and subjected to regular culture conditions these cells were able to differentiate in various cell types, as do cells of normal *ICM*. They were also endowed with self-renewal capacities, were pluripotent and represented the *in vitro* capture of a transitory developmental stage. For these reasons, they were designated *ES cells* (standing for Embryonic Stem cells).

For many years, *ES cell* lines could be successfully established from embryos of a particular strain of mice, the 129 strain only. Various lines of *ES cells* available were used as tools for genetic experiments in the mouse. They were namely instrumental to produce gene targeted mutations through homologous recombination, a pivotal technique to investigate the functions of genes that were currently discovered and cloned at that time by genetic engineering.

For many years the numerous attempts made to obtain *ES cell* lines from embryos of other mammals failed. But, seventeen years after mouse ES cell lines were established, James Thomson of the University of Wisconsin succeeded in deriving *ES cells* from Rhesus monkey first and from human embryos, provided to him by an *in vitro* fertilization clinic [4].

Human ES cells and the perspective of a regenerative medicine

James Thomson's experiments were reproduced by other laboratories in the world, and their results aroused a great deal of interest among the general public. The characteristics of the mouse *ES cells* were shared by human ones: one could establish permanent, virtually immortal, cell tines of human *ES cells* that remained pluripotent and could be led to differentiate *in vitro* into a

large number of cell types, including neurons, cardiomyocytes, vascular endothelial cells, striated muscle fibers, tendons, bones, cartilages… The possibility of using them for regenerative therapy in patients was then open.

Several problems however were raised by the use of human *ES cells* for this purpose.

Some are biological while others are ethical in nature.

The former concern the fact that if differentiated cells obtained from human *ES lines* are introduced into a patient, they will be subjected to immune rejection from the recipient. Ideally, the grafted cells should be 'customized' for each patient and therapeutic cloning was proposed as a method to circumvent this difficulty. Therapeutic cloning involves the substitution of the nucleus of a human oocyte by the nucleus of one of the patient's somatic cells. This technique, also designated as 'nuclear transfer', turned out to be of extremely low efficiency in mammals (mouse, sheep, cow etc.) on which it has been practiced and was unsuccessful in the few cases in which it has been applied to a human oocyte.

Moreover, it raised ethical problems of two kinds: one is the fact that it necessitates a large amount of human oocytes taken from young women, a highly unethical practice. The second is that it was argued that the improvement of the cloning technique could lead to reproductive cloning, which is generally considered as unacceptable.

Another problem, biological in nature, resides in the fact that the cultures of differentiated cells derived from *ES cells* might be 'contaminated' by pluripotent stem cells at the time they are introduced into the patient. These cells are prone to develop tumors when subjected to an adult environment.

Finally, the derivation of ES cells from a human embryo is considered by certain people as unethical, since it interrupts the development of a human being. Such is the position of the Catholic Church for whom the human nature of the conceptus starts from the moment when the two gametes fuse and form a zygote. Such a position does not hold for other religions, such as the Jewish, for which 'humanity' is acquired by the embryo only when it has reached a certain stage of development, about 40 days after fertilization.

Researchers have proposed several possibilities to circumvent these problems. One of those, for example, was to remove one single cell from an 8-cell stage human embryo and, through a biotechnological 'tour de force', derive an *ES cell* line from it. The remaining 7-cell-embryo is able to safely pursue its development as shown in routine techniques used for antenatal diagnosis.

The most spectacular result in this area was the recent demonstration that adult differentiated cells can be reprogrammed and reacquire the characteristics and potentialities of embryonic cells.

Rejuvenating adult differentiated cells

The increasing interest devoted to stem cells has led researchers to investigate the genetic characteristics of the 'stemness' state. What are the genes activated in these cells and responsible for their unique properties: undifferentiated state, pluripotency and self-renewal capacities? Several laboratories have attacked this problem using diverse types of stem cell lines and, although some differences arose in the lists of genes, the results converged on about 20 that turned out to be activated in virtually all the stem cell lines studied.

The laboratory of Shinya Yamanaka, then at the Riken Institute in Osaka, produced its own list of 24 genes and transfected cultured mouse skin fibroblasts with these genes through retroviral vectors.

They used a selection system based on the insertion of a resistance to the neomycin gene under the control of the promoter of a gene expressed in *ES cells* but not in fibroblasts, in order to recognize the cells that had been reprogrammed by the factors. The remaining cells of the fibroblast type died. Rare events of reprogramming of the fibroblasts nuclei occurred which led to the growth of colonies with the morphology of *ES cells*.

Shinya Yamanaka, with his co-worker Kazutoshi Takahashi, could obtain the same reprogramming of the fibroblasts into *ES-like* cells by transducing only four of these genes which turned out to be necessary and sufficient to produce this effect: *oct4, Klf4, c-Myc* and *Sox2*.

These genes are all transcription factors, regulating the activity of other genes.

A first report on these results appeared in 2006 and was followed one year later by an article reporting that the same reprogramming could be obtained with human fibroblasts [5].

The stable cell lines resulting from these experiments were designated as *iPS cells* for induced Pluripotent Stem cells. The *iPS cells* were found to express all the 24 genes including *Nanog*, which is often used as a 'marker' for the ES cell state. The gene expression profile of the *iPS cells* was found to be very similar to that of ES cells – although not identical – but very different from that of the original fibroblasts.

All the tests which are known to characterize *ES cells* and cells of the *ICM* were positive in *iPS cells*: formation of teratomas in adults; *iPS cells* can be led to differentiate into all tissues type cells, they form viable germ line chimeras when introduced into blastocysts, can support the complete development of an organism as shown by their capacity to yield viable mice entirely constituted of *iPS derived cells* in the tetraploid complementation assay. In this assay, *iPS cells* are introduced into the blastocyst of a mouse embryo whose cells are tetraploid. The placenta of these mice survives up

to term but the embryonic cells die progressively during the course of development; only the diploid cells that have been introduced into the blastocyst survive, thus giving rise to a normal mouse.

Since 2007 many laboratories in the world have switched to this new research line and an impressive number of results have been obtained.

Reprogramming of a large variety of differentiated cells has been achieved. Hemopoietic cells, including T and B lymphocytes and hematopoietic stem cells, could be an attractive type of cells for the generation of iPS for therapeutic purposes, liver cells, stomach epithelium, pancreatic β cells, and human keratinocytes. For example, Juan Carlos Izpisúa Belmonte has been able to derive lines of *iPS cells* from a single human hair. Neural progenitor cells can be induced into *iPS cells* without *Sox2* that they already express.

It seems therefore that reprogramming is a universal process that can be obtained from differentiated cells belonging from the three germ layers.

iPS cells have also been derived from differentiated cells of various other species of rodents (rats) or Primates.

All these results are very encouraging as to the possibility of devising novel techniques for the onset of an efficient regenerative medicine. *iPS cells* can fulfill the requirements of 'customized' cells that will not trigger an immune response from the recipient in which they will be introduced, since they can be derived from the own cells of the patient. However, their use still raises certain problems. Experiments carried out in mice have shown that chimeric mice, which are made up of a mosaic of normal and iPS derived cells, often develop tumors. This was attributed to the retroviral vectors used for gene transduction. Such vectors become inserted randomly into the host cell DNA. They may be positioned in critical locations capable of activating endogenous oncogenes. Moreover, *c-Myc*, which is one of the four genes introduced into adult cells, is itself an oncogene, overexpressed in most spontaneous tumors. Its localization within the host DNA may result in its overactivation, thus also being a cause for tumor formation.

Several laboratories are now developing methods to reprogram adult cells which would avoid the difficulties presently encountered.

In conclusion, one can consider that, following the pioneering work of Martin Evans and Gail Martin in 1981, the work of Shinya Yamanaka and colleagues must surely be regarded as the single major advance in the stem cell field in recent time.

There is every reason to suppose that it will have widespread therapeutic applications for human diseases.

References

1. Till, J.E., McCulloch, E.A. (1961) A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Rad. Res.* 14, 213–222.
2. Gardner, R.L. (1968) Mouse chimeras obtained by the injection of cells into the blastocyst. *Nature* 220, 596-597; Papaioannou. V.E., McBurney, M.W., Gardner, R.L. and Evans, M.J. (1975) Fate of teratocarcinoma cells injected into early mouse embryos. *Nature* 258, 70-73; Tarkowski, A.K. (1961) Mouse chimaeras developed from fused eggs. *Nature* 190, 857-860; Mintz, B. (1962) Formation of genotypically mosaic mouse embryos. *Amer. Zool.* 2, 432; Mintz, B. (1962) Experimental recombination of cells in the developing mouse egg: normal and lethal mutant genotypes. *Amer. Zool.* 2, 541-542.
3. Martin G.R. (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells, *Proc. Natl. Acad. Sci. Paris*. 78, 7634-7638; Evans, M.J. and Kaufman, M.H. (1981) Establishment in culture of pluripotential cells from mouse embryos, *Nature* 440, 1199-1203.
4. Thomson, J.A., Itskovitz-Eldor, J., Shapiron S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S. and Jones, J.M. (1998) Embryonic Stem Cell Lines Derived from Human Blastocysts, *Science* 282, 1145-1147.
5. Takahashi, K. and Yamanaka, S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126, 663-676; Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., Yamanaka, S. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131, 861-872.

Genomic Exploration of the RNA Continent

Takashi Gojobori

Introduction

1.1. Susumu Ohno

Susumu Ohno was born in 1928 and died in the year 2000. He had spent more than 40 years in the City of Hope in Los Angeles, USA. One of his famous books is *Evolution by Gene Duplication*, which was published in 1970 (1). In this book, he pointed out that genome duplication and gene duplication are very important not only for evolution but also for function and structures of the genome. This is mainly because a duplicated copy of gene can enjoy the freedom of functional differentiation as long as the original gene can retain the original function.

1.2. Junk DNA

The term 'junk DNA' was coined by Susumu Ohno probably in 1972, a bit later than the time when the above-mentioned book was published, as long as my memory is correct. That was presented in the Brookhaven Symposium on Biology in the United States with his paper entitled '*So Much Junk DNA in our Genome'* (2). Already at that time, it was known that the human DNA genome may have had only 5% for protein-coding regions and the other 95% for non-coding or unknown functions. From this fact and also from other observations, he coined 'junk' DNA that literally represent the protein-non-coding regions of the DNA genome.

However, the term 'junk' really brought about intense controversy over its biological significance. This is because many people have believed that there are no regions of meaningless function in a human body.

According to the Merriam-Webster's online dictionary (3), the followings are given as definition of 'junk':

1) Pieces of old cable or cordage used especially to make gaskets, mats, swabs, or oakum.
1) Old iron, glass, paper, or other waste that may be used again in some form; second hand, worn, or discarded articles.
1) Something of poor quality, almost trash, something of little meaning, worth, or significance.

I do not know exactly which meaning Dr Susumu Ohno took when he coined such a term as the 'junk DNA'. According to my personal impression, however, because I had personal and intimate communication with him when he was alive, the second one of the above-mentioned definitions is probably the most appropriate. In fact, it is close to the meaning of trash, implying that human DNA contains a vast amount of trash. That is why it caused a lot of arguments.

1.3. Neo-Darwinism

The arguments on junk DNA are so important because it gives a paradigm question of whether all the genomic DNA regions are subjected to natural selection. According to Neo-Darwinism, all the traits and features of organisms are explained by natural selection and mutation with a law of inheritance. The premise is that most mutations are deleterious against the survival of organisms but all the other mutations would be very much adaptive and advantageous for survival. For the former mutations negative selection works whereas for the latter mutations positive selection operates.

1.4. Genetic drift

There is another mechanism of changing gene frequencies in a natural population of a given organism, which is called 'genetic drift'. Genetic drift is a kind of genetic phenomenon in which gene frequencies will change simply by mating. Natural selection can be a kind of second force, in particular when the population size is small and random mating dominates.

In population genetics in the 1930s, a heated controversy took place between so-called selectionists versus proponents of genetic drift (4). An American geneticist, Sewall Wright, was kind of a hero of genetic drift. On the other hand, R.A. Fisher, who was a British man, is a hero of selectionists. Somehow selectionists appeared to have won.

1.5. Motoo Kimura

A Japanese geneticist called Motoo Kimura proposed the neutral theory of molecular evolution [5,6]. He contended that at the DNA level or the genomic level, most mutations were selectively neutral. That means that mutations are not so much deleterious, not so much advantageous, either. This is now simply called as the neutral theory.

I will not go into the details of this controversy, but what I would emphasize is that this kind of controversy always exists in the background of a discussion on biological significance of transcribed RNAs in the non-protein-coding regions of the DNA genome.

1.6. TSS: Transcriptional Start Sites

So here in this paper, our specific question is how transcription start sites (TSSs) are distributed over the human genome. Of course, transcription depends heavily upon types of tissues and cells or even environmental conditions. Using over 60 different tissues and cell lines of humans, two Japanese governmental projects on transcription were conducted, in which we actively participated: One is called the 'H Invitational Human Transcript Project' (7) and the other is the 'Human Genome Network Project' (8). Using the outcome of both projects, we would discuss the genomic distribution of TSSs in conjunction with the biological significance of transcribed RNAs in the non-coding regions.

Materials and Methods

2.1. H-invitational Project

Almost eight years ago, we initiated the annotation jamboree on the human full-length cDNAs over the human genome (7). Note that cDNA is a kind of RNA information most of which are supposed to encode for proteins. Furthermore, this jamboree was conducted as an international cooperation, in which about 120 experts gathered in Tokyo, Japan and spent ten days making annotations for actual and possible human genes. This project has been called as the 'H-Invitational Project'. We still continue this endeavour, presently including not only all cDNAs but all available transcripts.

2.2. Transcription regulation network as a small universe in a cell

The transcription regulation in a given cell may be illustrated as follows. Let us suppose that there is a certain gene in a given genomic region. We now know that *prior to* the gene; there is a *cis*-acting element or a promoter region. When a certain protein such as a transcription factor binds to *cis*-acting elements in a regulatory region, it works like a switch to order the gene to be transcribed to produce messenger RNAs. Following information of messenger RNAs, a particular protein is made through splicing in the case of eukaryotes. The protein may have interaction with other proteins that are made in the same way. Then, interacted proteins may constitute a protein complex, which would bind to DNA again to switch on production of messenger RNAs of its own protein or the other proteins. This is exactly what we call network. This transcriptional network is extended over the entire human genome, which is now called the 'genome network'. It looks just like a small universe with a single cell.

2.3. Human Genome Network Project

We conducted a project called the 'Human Genome Network'. This project was carried out in collaboration with RIKEN. From messenger RNAs (mRNAs) in a given type of tissue or cell, we can obtain only the first 20-nucleotide segments from the start site of the mRNAs, using their cap sites, in the process of making cDNAs. Let us call those 20-nucleotide segments 'TSS tags'.

The TSS tags can be actually sequenced by the so-called next-generation sequencers (NGS) in an enormous amount of numbers. Therefore, once we have a sample of tissues or cells, we can sequence the TSS tags as a form of cDNAs immediately and enormously.

The TSS tags obtained are mapped, by technologies of bioinformatics, onto the human genome, meaning that transcription start sites can be identified in a genomic scale. Thus, we can now raise unique and significant questions about how much transcription is taking place over the human genome and how these transcription start sites are distributed over the human genome. The Human Genome Network Project was conducted to answer those questions as one of the purposes.

2.4. Tissue-type and cell-type dependencies

Of course, the transcription depends heavily upon the types of tissues and cells and even upon environmental conditions. From this standpoint, it will be idealistic if we are able to examine TSSs in a single cell.

For example, we are now trying to examine TSS tags from a single cell such as a monocyte, which can differentiate to a macrophage. Because it takes more time to obtain finalized results, however, we decided to examine TSS tags for a mixture of 60 different human tissues and cell lines. We believe that it will still be useful for understanding the overview of transcription activities over the entire human genome, namely for conducting *genomic exploration of the RNA continent* of humans.

2.5. Quality control of sequence data produced by NGS

In order to elucidate a distribution of TSSs over the human genome, we have made great effort to clean the data. In particular, when a single TSS tag is tried to map on the genome, it sometimes happens to be mapped in more than one location. For other TSS tags, it also sometimes happens to be no matched locations in the human genome. Those observations are apparently due to artefacts of the experimental efforts in a process of producing TSS tags.

Thus, it is very important how much the TSS tag data can be cleaned. In this case of the next generation machine called 454, we know that a specific type of sequencing errors have been expected in a certain frequency.

By making an algorithm, we can rescue a portion of sequencing errors. Conducting computer simulations, the TSS tag data obtained has been evaluated that almost 10% of sequencing errors can rescued computationally.

Results and Discussion

3.1. Distribution of TSS tags over the human genome

When a distribution of about 47 millions of TSS tags for a mixture of 60 different types of human tissues and cell lines was examined for all the chromosomes, from chromosome numbers 1 to 22 and sex chromosomes X and Y, over the human genome, it immediately became clear that transcription for producing mRNAs are taking place actively at a tremendous numbers of locations in the human genome. Taking into account the fact that the number of human genes is about 23,000~24,000 in the genome, the number of TSSs far exceeded those numbers. Thus, we assure that transcription takes place, in an enormous number, in the protein non-coding regions of human genome. We call this situation *the RNA continent* of the human genome. Of course, the information on transcription activities of genes, such as typical disease-sensitive genes, is also very useful for understanding how and when these genes are transcribed.

3.2. Distribution of TSS tags in a liver tissue

In the previous section, we discussed the distribution of TSSs for a mixture of 60 different types of human tissues and cell lines. Here, we can focus on a single tissue. Now, we can have a distribution of the transcription start sites only for the human liver.

As long as we see, the TSS tags in the human liver are very scarcely distributed over the human genome. This is apparently due to the lack of a sufficient number of TSS tags. Thus, we point out that although this kind of study is certainly feasible now, it may take a bit more time to obtain a sufficient number of TSS tags. However, the acute developments of NGS (Next-generation sequencing machines) are expected to resolve this problem because of enormous speed and capacities of sequencing capabilities.

3.3. Examination of TSSs with known distribution: Two categories of human genes

We made comparisons of the transcription start sites obtained from genomic locations of the TSS tag with the already known transcription start signs of protein-coding regions of the human genome.

From the database such as RefSeq at NCBI/NIH in the United States, or the H Invitational database that we have constructed in the H-Invitational Project, we obtained information of genomic locations for all the protein-coding regions available. Then, we made comparisons.

If the transcription site is so sharply determined, then the distribution of TSS tags should be very sharp. On the other hand, if transcription start sites are so stochastic or if they are not really sharply determined, even though they give the right direction of transcription of a given coding region, the distribution of TSS tags should manifest a broad distribution.

As a result, we observed that there were two types of coding regions, depending upon transcription start sites. One type of coding regions has very sharp transcription start sites whereas the other type of coding regions has very broad start sites.

Although we should have understood how those transcription start sites are biologically determined, we do not know how the coding regions of having broad locations of transcription start sites are regulated. Anyway, it is very interesting to know that transcription start sites are not always sharply and uniquely determined. Therefore transcription start sites seem to have a stochastic nature, which we should keep in mind.

3.4. Susumu Ohno's Junk DNA

Let me go back to a story of Susumu Ohno. When he coined 'junk DNA', he predicted that even junk DNA would be transcribed. However, transcription itself does not mean any functional significance. In this sense, Susumu Ohno was right. In 1972 he actually and clearly showed that 'junk DNA' would be transcribed (2).

Now we are confronted by a very important question to answer. Is there any functional significance for the transcription activities observed in the protein non-coding regions occupying huge portion of the human genome? Yes, partly. We have known this answer because we know there are functional non-coding RNAs such as Micro RNAs and natural antisense RNAs.

However, the problem is whether a substantial portion of non-coding regions is subjected to the so-called 'transcriptional noise'. It is just like the engine of an old car. Once you start the engine, it cannot start immediately. You need idling. Just like this, we may be observing *transcriptional idling.*

In order to make transcription possible, opening of chromosomal structures may be prerequisite. This may cause transcriptional noise or idling because of preparation for appropriate changes of chromosomal structures.

The problem was whether junk DNA is really junk or not. We do not think this may be the right question. Because we know that there must be

functional non-coding RNAs such as micro RNAs among all transcripts, the right question should be to be asked in a way is how many are functional and what percentage are not functional. We believe that the question should be changed into the new question; otherwise the RNA continent cannot be explored in an appropriate way.

Summary

We have conducted two Japanese governmental projects: the H-Invitational Human Transcript Project and the Human Genome Network Project. Using the outcome of these two projects, we examined a distribution of transcription start sites over the entire human genome. We pointed out that tremendous transcription activities are taking place in a substantial portion of protein non-coding regions that occupy a huge portion of the entire human genome. Moreover, the transcription sites for some genes are not sharply and uniquely determined. Finally, the right question to ask should be in a way how many are functional and what percentage are not functional. We believe that the question should be changed into the new question; otherwise the RNA continent cannot be explored in an appropriate way.

Acknowledgements

First of all I would like to express my special thanks to the Holy See and also to the organisers, Dr Werner Arber and Dr Jürgen Mittelstrass. In particular I would like to extend my thanks to the Chancellor, Dr Marcelo Sánchez Sorondo for his never-changing support to me.

References

[1] Ohno, S. (1970) *Evolution by Gene Duplication*, Springer Verlag, Berlin.

[2] Ohno, S. (1972) *So Much Junk DNA in our Genome*, Brookhaven Symposium, New York.

[3] Merriam-Webster's online dictionary (2011), www.merriam-webster.com

Provine, W.B. (1971) *The Origins of Theoretical Population Genetics*, With A New Afterword.

[4] Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* 217 (5129): 624-626.

[5] Kimura, M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press.

[6] Imanishi, T., other 152 authors, Ikeo, K. Gojobori, T., and Sugano S. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2, 1-21.

[7] The FANTOM Consortium: Carninci, P., Gojobori, T., Ikeo, K. and other 158 authors and Hume, D.A., and Genome Network Project Core Group: Kai, C., and other 31 authors and Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. *Science* 309 (5740): 1559-1563.

Transgenic Crops and the Future of Agriculture

Peter H. Raven

One of the frustrations and joys of science is that we try to be objective, to offer hypotheses and test them, and to discover, to the extent of which we are capable, what is true and what is not true. It would be a serious mistake to claim that science is not affected by its context, as the case of Galileo Galilei illustrates so dramatically! Importantly, however, science does not in itself instruct us whether or not to jump off a tall building or provide moral judgments even about serious matters such as whether it is wise to pollute the atmosphere beyond the point at which our civilization can survive in something resembling its present form. What is does do is to tell us what is likely to happen as a consequence of particular actions. Given that information, we are free to take whatever course of action we wish.

It is often frustrating for scientists, however, when a situation is as clear as it can be from a scientific point of view, and yet counter opinions are offered without any basis in fact by non-scientists and regarded by the general public and the media as equal in value to scientific conclusions published in peer-reviewed journals. In science, not all opinions are equal, a relationship that the general public and the media all too often forget.

A particular case in point has to do with the adoption of GE crops as an important element in improving the efficiency and productivity of agriculture worldwide. Last year, our Academy held a study week on the use of GE crops in development worldwide, and concluded that, in general, they would be highly beneficial. This view basically reflected and recapitulated the findings of our Academy, other academies, and scientists generally in appraising the use of such crops to improve productivity and to help alleviate hunger throughout the world. What then is the rational basis for continuing to regard the use of such crops as questionable and even dangerous?

With some regional exceptions, virtually every bit of the world's land surface that can be cultivated is cultivated now, and it is exceedingly important – a matter of survival for many people – to make the productivity of this scarce resource as high as it can be, consistently with its sustainability. Of the 6.8 billion people in the world, one billion are malnourished to the point that their bodies and brains do not develop properly and are literally wasting away, with 100 million of them on the verge of starvation at any

given time. It is projected that 2.5 billion people will be added to the world population during the next four decades, and clear that they will join the poorest fringes of society. Nearly 99% of them will be born in countries that are considered to be in the 'developing' category. Global Footprint (see http://globalfootprint.org) currently estimates that we are using 140% of what the world can produce on a sustainable basis. As a result, the world is inevitably becoming more uniform, less interesting, less healthy, and with less potential than it has today. To counter this trend, we clearly need to achieve a stable human population, to adopt conservative and reasonable levels of consumption around the world; and to develop and use new technologies that will help to save us from reducing the quality of our civilization even more drastically over the decades to come.

In our attempt to feed people adequately, however, we need to adopt all the tools available to modern agriculture: more efficient use of water; limited use of pesticides and fertilizers; integrated pest management; precision farming; and the continued genetic improvement of our crops to fit the needs of the present and the future. Curiously, a particularly knotty problem has arisen around the use of the available modern methods to improve the characteristics of plants and animals. Called into question is the production of GE (genetically enhanced) plants and animals with traits by virtue of which they perform better than they would otherwise. At the same time, we are content to continue to use traditional, relatively imprecise breeding methods for plants, as for example to irradiate their whole genomes and let the parts of those genomes come together in any combination that they would. In contrast, we are afraid to use precise methods that involve transferring one or a few carefully selected genes from one kind of organism to another. Before they are released for us, the products of GE technology are tested much more carefully than any crop varieties we have adopted in the past, and we understand their features with much more precision; but strangely that does not seem to be sufficient to satisfy a host of critics of the methodology. Why is this so, particularly when the scientists of the world are essentially fully united in their conclusion that such crops are not only harmless to human health and to the environment, but that they will contribute substantially to the huge problem of feeding people adequately?

The potential of improving the characteristics of organisms through genome manipulation was opened up through the experiments of Boyer and Cohen in 1973, about a decade after we first began to understand the genetic code. These scientists transferred a gene successfully from the bacterium *E. coli* to the African clawed toad, the first time that a gene was moved successfully from one kind of organism to another that was unrelated to the donor. Scientists

were concerned with the consequences of producing this kind of newly-constituted organism, hurriedly organized a major conference at Asilomar, California. There they laid down rules for dealing carefully with the new kinds of organisms in laboratories until they were better understood. As our knowledge base improved over the subsequent years, we began to apply these techniques to the production of often-improved versions of various items that we use.

Thus, many of the drugs now used in industrialized countries are produced by GE organisms. For example, virtually all of our insulin is produced in this way, and it is cheaper; the derived product is dependably purer than insulin derived from collecting and extracting cow thymuses, the earlier practice, and much less expensive. Virtually all of the beer and all of the cheese produced in the world is produced using enzymes from GE organisms. Nobody worries about these processes at all! As Per Pinstrup Anderson has pointed out so well, however, while we who live in Europe and North America may use drugs from GE organisms because our lives are at stake, for a mother in Africa the disease she faces is starvation for her children, and the medicine she needs is food – food that we may deny to her as a result of our suspicious and irrational fears, and by a disregard for the underlying science. Pinstrup Anderson then went on to point out that the world's poor spend 60 to 80 percent of their incomes on food, and even then there often isn't enough to alleviate starvation. It seems morally wrong for rich people to block others from using GE crops when the evidence is so clear that they are helpful in elevating productivity and thus that they will contribute substantially to the welfare of poor people all over the world. We need to remember the unfortunate spectacle, played out a few years ago, of Zambia turning back hundreds of tons of maize in food aid from the U.S. because some of it was GE. At the same time, hundreds of millions of people in the world consume such maize with not a single case of sickness or any other problem ever having been detected; many thousands of people were starving to death in Zambia but were denied access to the food because of ill-founded concerns with its safety. The memory of that tragic event should become a moral burden for those who created the false impression on which the decision to deny the use of that food was based.

Let us now consider the facts about GE organisms that have been established clearly. First, the horizontal transfer of genes between different kinds of organisms, as our distinguished chair Werner Arber has continually stressed, is common in nature. Thus there is no rational basis for considering such transfers to be unnatural or avoiding there use for that reason.

Second, there is no known mechanism that makes gene transfer generally dangerous or potentially harmful. Obviously it is possible to transfer dan-

gerous genes from one kind of organism to another (for example, genes associated with the production of toxins), but why would anyone do so? Further, GE crops are more carefully tested than any other products of plant breeding before being released into the trade. This relationship seems a bit ironic since it is perfectly possible to produce, for example, poisonous tomatoes or poisonous potatoes by conventional breeding, but for obvious reasons we do not choose to do so. There is nothing generally dangerous about substituting one segment of DNA for another in the genome of another kind of organism. The genes that are introduced may or may not become incorporated and function well in their new environment, but if they fail to do so, they clearly could not be released for agricultural use.

Third, with about one-sixth of all of the world's cropland devoted to GE crops, and hundreds of millions of people consuming them every day, there has not been a single example of any problem for human health arising from eating such foods. Why then is this one of the great bugaboos posed by those who resist the introduction of GE crops in their own or other countries, regardless of how needy hungry people may be? How can anyone reasonably assume that something unexpected and awful would happen later, with no evidence whatever for such an assertion?

Fourth, the major crop strains that are now produced as a result of GE techniques have one of two features: they are either insect resistant or herbicide resistant. Many other products are in prospect for the future, including drought resistance or the production of higher yields, and many of them will be made available during the coming decade. There are certainly problems associated with industrial-scale agriculture, such as those arising from planting huge areas with a single crop genotype. When this is done, the whole crop may be susceptible to one kind of pathogen, which may harm or even kill it over wide areas. The problem here is, however, that of planting design, which has nothing to do with the choice of techniques used to produce the particular uniform crop strain in the first place. Planned variation in the genotypes of crops planted over large areas is an important strategy in any case, and should be employed generally. We should remember further that a large majority of the farmers who are using GE crops at the present day are smallholders, and not industrial-scale farmers; the idea that GE crops are inevitably planted on a large scale is a myth that should be removed from consideration. It is of great importance to consider how we can modify our crops relatively rapidly and precisely to adapt them to the altered climates of the future, an even more serious problem for feeding people than those we face now.

In a recent National Research Council study of the farm-level effects of the cultivation of GM crops in the U.S., where virtually all maize, soy-

beans, and cotton are genetically modified, we encountered a few instances of insect resistance arising in connection with properties of the GE crops. In a similar way, herbicide resistance had become widespread in some regions where particular herbicides were sprayed over wide areas. The study also demonstrated the substantial economic and ecological advantages associated with the use of such improved crops, advantages that account for their widespread adoption. Some reviews of our study highlighted the herbicide resistance, without mentioning the obvious fact that using any herbicide widely will result in the appearance of weed strains resistant to it.

If those weeds do not belong to the same species as the crop, they can be controlled by building resistance to other herbicides into the crop, or by building 'stacked' resistance to several different herbicides in a single crop strain. If as in a few cases, however, the weeds do belong to the same species as the crop – examples are sugarbeets, rice, and sunflowers – special care needs to be taken, because the weeds will tend to pick up the resistance genes as a result of hybridization with their cultivated relatives. This situation arises especially when the crops are grown in areas where their wild relatives or weedy relatives occur mixed with the crop or in the areas where it is being cultivated. Hybridization is a normal feature of the evolution of plant species and an important feature of their evolution. When no wild or weedy relatives are present, as in the case of most crops in the U.S. and Europe, this situation does not pose a problem. The recent legal rulings prohibiting the cultivation of GE sugarbeets and alfalfa in the U.S. are illogical for reasons that I shall discuss subsequently. In rice, GE technology has nothing to do with the appearance of the troublesome weedy varieties. And there is certainly no conceivable, much less demonstrated, rational basis for prohibiting the cultivation of GE maize in Europe, or GE brinjal in India: recent rulings barring them can only be attributed to the effects of runaway, effective propaganda.

Considering the problem of 'land races' as a whole, it is important to remember that in Mexico, for example, maize yields in the milpas of the southern part of the country amount to no more than one-fiftieth of the yield per hectare that is achieved on the intensively cultivated fields in the north. This, coupled with the rapid growth of the population of Mexico to its present 111 million people, with 18 million more people projected to be added by 2050, has forced to the country to import large amounts of maize from the U.S., much of it of GE origin.

The people who cultivate the 'land races' of corn on the milpas of southern Mexico are in general very poor. At the same time, the composition and nature of their 'land races' changes like the pattern in a slowly revolving

Kaleidoscope. If we want to save the strains that are being grown there today, we will either need to subsidize the people growing them so that they can overcome the poverty that is inherent in their situation, or else put samples of the seeds away in seed banks, or adopt both strategies. As is the case generally, there is no saving a way of life, tragic as that fact is for the survival of precious human diversity, by denying people the advantages of the modern world. The poor will find the means to change their lives anyway, or continue to live at a level that we should collectively reject purely for reasons of morality.

What about the general environmental effects of cultivating GE crops? Our farm-level study in the U.S. found cleaner soils with richer soil biotas and cleaner water occurred in areas where GE crops were cultivated than elsewhere. Additional sampling needs to be carried out, but there is certainly no sign of adverse environmental in these or other respects.

Taken together, these analyses bring us back to the question of why so many Europeans, particularly, are so concerned about the adoption of GE strains of crops that they are willing to cite a great deal of flawed science in support of their negative positions. Certainly some environmental organizations campaign on this issue, which apparently, presented as they do, provides a dependable source of cash to support their operations. Much of their argument seems to arise from an anti-corporate stance, which for various reasons is appealing to many people concerned with moral justice. Justifiable moral concern continues to be raised about a justifiable level of corporate profits, but these are problems for the whole array of products supplied from wealthy countries to poor ones, and not just food crops. Everyone seems to agree that desirable traits or products of all kinds need to be made available to the poor to the extent that they will prove helpful, but those also are considerations that lie beyond the realm of consideration of GE crops. About one-sixth of the world's farmland is now devoted to their cultivation, with no harmful effects related to the genetic traits involved having been demonstrated other than what I have just reviewed. Can we not stop using bad science to justify our anti-corporate inclinations?

A peculiar problem in the U.S. concerns our legal classification of GE crops as 'non-organic'. What this means is that while huge vats of *Bacillus thuringiensis* can be freeze-dried and the resulting substance broadcast, killing all of the target insects in the area whether they are harmful to crops or not, that is regarded as 'organic'. If the genes that produce the toxin are placed in the crop so that they will affect only the actual herbivores on that crop, that is classified as 'non-organic'. The logic eludes many of us, but what it means in practice is that if genes spread from better-producing GE

alfalfa, for example, to 'organic' crops of alfalfa by means of cross-pollination by insects, that the other crops would become 'non-organic' and therefore sold at a lower premium. The same is true of rapeseed, for example, where some weeds have also become herbicide resistant, causing further problems, but as far as 'organic' classification goes, the distinction is simply a legal one, and the 'problem' caused by the laws, not the biological facts of the matter.

Often countries are 'testing' GE strains prior to their 'release'. The problem is that we don't really know for which properties they should be tested. Will they escape? Will they provide higher yields? Why should they if farmers want them, and why should they alone be tested? There is simply no body of evidence that supports this extensive testing, and the harmful effects of not providing enhanced crops to those who really need them are evident. Isn't it time for the nations of the world to re-examine the Cartagena Protocol of the CBC and examine the facts on which it is based from a scientific perspective?

Let's look at some of the positive benefits of growing GE crops and improving the productivity of agriculture generally. The loss of biological resources has reached frightening levels and is highly significant for our future. Comparing the rate of loss of species observed in the fossil record with that documented now, we find that extinction rates have increased to thousands of times their historical rate. These losses, which are increasing rapidly, are resulting from habitat destruction, global climate change, selective hunting and gathering, and the unprecedented spread of invasive species to the extent that more than half of the species on Earth may disappear during the course of the 21st century. To a very large extent, we depend on these species for our opportunities to build sustainability throughout the globe, and have as yet recognized only a small fraction – perhaps no more than a sixth – of those we are losing. The loss of biological species and the productive systems of which they are a part is irreversible, and therefore, over the long run, is the most serious environmental problem that we confront. The more we encourage inefficient agriculture by discouraging the use of modern methods in the development of crop varieties, the faster biodiversity will disappear.

Another obvious benefit of GE crops is that their use has already achieved major reductions in pesticide applications, a highly desirable outcome for the environment in general and for human health in particular. Even by the year 2000, the use of GM soybean, oilseed rape (canola), cotton, and maize had reduced pesticide use by 22.3 million kilograms of formulated product, and the reductions have risen far above that level subsequently. Worldwide, there are at least 500,000 cases of pesticide poisoning and 5,000 deaths annually. Residues of pesticides are ubiquitous in foods

sold in markets throughout the world, and we should be striving to reduce them. The use of GE crops has already had a large effect on these levels in all regions where they are grown at a commercial scale.

For Europe, it has been estimated that if half of the corn, oilseed rape (canola), sugar beet, and cotton raised were genetically modified to resist pests instead of being treated by spraying that there would be an immediate reduction of about 14.5 million kilograms of formulated pesticide product (4.5 million kilograms of active ingredient). The reduction of 7.5 million hectares of crops sprayed as a result of growing GM crops would save approximately 20.5 million liters of diesel and prevent the emission of 73,000 tons of carbon dioxide into the atmosphere, thus driving global warming. Along with other methods to decrease the application of fertilizers and pesticides, such as Integrated Pest Management, the use of transgenic crops clearly can confer great benefits in our quest for sustainable, productive agriculture. Against this background, the choice of many Europeans to avoid the use of GE crops against all scientific evidence seems as bizarre as it is environmentally damaging.

As the global climate changes, the need for the rapid adaptation of our cultivated crops to the new conditions has become increasingly evident. Food production can be maintained only by the use of the best available methods, including those that lead to water conservation. We cannot achieve such changes by assuming that modern methods are inevitably bad, while the crops developed by, say, 1890, through genetic selection, are good. Political infighting about methods of selection leads directly or indirectly to the starvation of millions, and alleviates no known problem. Therefore, I consider it morally unacceptable, and await reasons as to why it is justified.

So Europe's strong stand against GM crops, which have the potential to produce more food available, may seem ill advised to hungry people in developing countries who need food and not unsupported arguments about why it might not be safe. Serious discussions of the appearance of large-scale agriculture, the corporatization of food systems, or the globalization of trade are clearly desirable, but it is not GM crops that are driving these trends, which they are sometimes used to represent. We badly need to develop transgenic cassava and other crops that are vital for feeding the people who live in the tropics, and do not have the right to play with their welfare for ideological reasons. Let resolve here to try to find ways to move forward for human welfare with the tools that science has developed for these purposes, a resolution that would have much in common with the aims of our Academy.

Genetic engineering of plants: my experience with the development of a key technology for food security

Ingo Potrykus

I have chosen a rather personal title for my presentation. Because of my age, I happen to be one of the pioneers of the development of this infamous GMO-technology and I considered it interesting to present you with a personal account of the development of this highly controversial technology (Genetically Modified Organisms) you all are familiar with to date. I am also responding to the prologue by Werner Arber and Jürgen Mittelstrass. I would like to show you my *personal testimony of the acquired new scientific knowledge including its application and the expected future impact especially for the welfare of human societies*, and I will include some personal recollections.

Since my youth I have been a zoologist by interest and it is surprising that I did my PhD at a Max Planck Institute for 'Plant Breeding Research'. The reason was, that I was impressed by the director of this institute, and that he encouraged the college teacher of sports and biology to work on a PhD thesis in his institute. At that time, it was in the early 60s, a hypothesis from the 1930s, that plant cells are potentially 'totipotent', by the Austrian botanist, Gottlieb Haberlandt, could be experimentally verified for the first time. This first evidence came from work with embryogenic carrot cell suspension cultures just during the time of my PhD thesis. Although working myself on chloroplast inheritance, I was very deeply impressed by this phenomenon of totipotency. Subsequently it could be shown that even highly differentiated plant tissues contain cells that have the capacity to develop into a complete fertile plant. During the course of my own first years in science I was able to add a few experimental examples, and in the course of a few years – in the early seventies – it was possible to take living cells from virtually every organ of a plant, including the germ cells (leading to haploid organisms), and allow them, under totally defined conditions, to regenerate to complete plants. I should stress that we had learned which experimental conditions we had to provide for the cells to embark onto the pathway to a complete plant. But we do not really understand – up to date – how the cells fulfil this miracle. So I was, and still am, fascinated by this capacity, but if I tell you why I was fascinated you will be disappointed. I was not fascinated by the scientific problem to be studied. I was fascinated

by the potential this phenomenon was offering for plant breeding. This indicates that I am not a 'scientist' in its true sense, but that I am rather an 'engineer'. My mind is primed towards solutions of concrete problems. If plant cells are totipotent, this would offer the possibility for plant breeding to work with millions of genetically identical single cells in the Petri dish (instead of thousands of plants in the field), to modify their genome and regenerate 'genetically modified' plants. In the early experiments with the model plant Petunia we explored all that would be technically possible. As the cell wall was to be considered an absolute barrier to virtually all genetic modifications we had in mind, we started to develop the first cases of cell wall-free 'naked' plant cells (protoplasts). We were interested in combining total genomes, in combining parts of genomes, in introducing complete nuclei, other organelles, such as chloroplasts or mitochondria, and we were interested in introducing pure DNA. As soon as it was possible to regenerate fertile plants from such cell wall-free cells, we tested all these novel genome combinations indicated above, and easily ended up with a few *Nature* publications out of this work. However, my motivation was to use this potential to contribute to plant breeding research, but to food security and model plants such as Petunia were not too promising in this respect. This was also in the early days of the Green Revolution. With the rapidly growing world population, we would need to continue on the path initiated by the work of Norman Borlaug, who became one of my heroes. I felt that I had to leave the easy work with model plants and shift to more important plants for food security, and this was the beginning of my work with cereals in 1972. In the subsequent decades I worked with wheat, barley, oats, maize and later with rice, cassava and sorghum. The concept for all our work was based on the well-documented fact that somatic plant cells are 'totipotent'. Well, I got a very tough lesson. I spent more than ten solid years of enormous experimental efforts in trying to convince cereal cells to behave as one could expect from them, but differentiated cereal cells refused to be 'totipotent' – and they still do so to date. After ten years of intensive experimentation and more than a hundred and twenty thousand variations in experimental culture conditions, using every possible growth factor and every possible media factor combination including up to seven factor gradients in a single Petri dish, I finally accepted that graminaceous plant species are obviously basically different from herbaceous plants with respect to 'totipotency'. The cause may have something to do with the strategy of how cereals defend themselves differently from mechanical attacks, compared to herbaceous dicots. If a herbaceous dicot is wounded, the cells adjacent to the wound dedifferentiate, re-embryonalise and replicate to close

the wound with newly formed wound tissue. If a cereal tissue is wounded, the response is totally different: the wound-adjacent cells in a cereal produce phenols and undergo a programmed cell death, and there is no wound healing. This wound healing reaction which is the biological basis for tissue culture – and totipotency – does not exist in graminaceous species. This was a big surprise and we were in trouble, because all our plans on the genetic engineering of cereals were based on the concept of totipotency. It took some time to forget about this concept. An alternative finally opened up through a development using meristematic (embryogenic) cells, preventing their differentiation, establishing embryogenic cell cultures (comparable to stem cell line research with animals) and using protoplasts from those embryogenic cells. On this rather 'unusual' basis for plants it was finally possible to also approach genetic engineering with cereals. There was, however another important consequence from this experience with cereals and it was that, most probably, *Agrobacterium* was no longer to be considered a useful vector for transformation. At that time, virtually all laboratories interested in the genetic engineering of plants were developing *Agrobacterium* as the gene transfer vector. From our experience with cereals it was obvious that the dicot-type wound response dependent transfer of a plasmid by *Agrobacterium* into plant cells would not function in cereals. As this meant that *Agrobacterium* was not an appropriate vector for gene transfer to cereals, we had to develop an alternative gene transfer technique on the basis of naked plant cells, allowing us to introduce naked DNA into naked plant cells independent from any biological vector. We had tried this already in the early 70s, a time when many laboratories worked on rather desperate experiments to demonstrate gene transfer into plants. To improve the situation we approached genetic evidence for putative integration of foreign DNA in contrast to those who looked for phenotypic data. Let me briefly describe an experiment – which failed – to give you a flavour of the situation around 1972: we hoped that naked plant cells would take up foreign DNA. To test whether DNA can be taken up and can be integrated we used a genetic system which was state-of-the art for this purpose at that time: we had a homozygote, recessive white flowering Petunia, the white flower colour representing a recessive, monogenic trait, and we had a dominant, monogenic and red flowering Petunia. We isolated total DNA from the red flowering petunia and treated protoplasts from the white flowering petunia with that DNA, hoping that, among thousands of offspring, we might find one with pink flowers (the sexual cross yielded pink flowers). This looks like a very rough experiment nowadays: at that time it was state-of-the-art and better than anything else. Well, the big surprise came when we finally had

a greenhouse full of Petunia plants regenerating from these DNA-treated protoplast: the first plant had pink flowers – fantastic! – the second plant flowered pink as well, the third plant flowered pink, the fourth plant flowered pink etc. At the end, the entire greenhouse was filled with pink flowering plants. This was, of course, no evidence for 100% transformation, but an artefact. In 1984 we did it better: we isolated a single defined microbial gene for antibiotic resistance and treated tobacco protoplasts using cell membrane modifying agents; we applied selection pressure for successful integration and selected among hundreds of millions of cells for developing cell colonies; we recovered fertile plants from those and we demonstrated the Mendelian pattern of inheritance for this single dominant trait, and we demonstrated the integration of this DNA into the host cell genome. This was the first clear-cut demonstration that genes can be introduced into naked plant cells without the contribution of any biological vector, thus finally opening the route for gene transfer to cereals. But this experiment was done with tobacco and not cereals. However, we had a technique at hand to introduce genes into naked cells and we applied this technique to cereals and our first transgenic cereal – it was rice – was published in 1988. Well, this was eighteen years from the time I was starting to work with these ideas. From then on we applied this technology to introduce agronomically important traits into cereals and other crop plants. We were determined to contribute to food security and tried in a first round of experiments to use this technology to rescue harvests which otherwise would have been lost to insects or destroyed by fungi, bacteria or viruses. We were introducing resistance genes into rice, and in 1991 we sent our first insect-resistant rice to our collaborating International Rice Research Institute in the Philippines. This GMO-rice did not reach IRRI. It was kidnapped by Greenpeace with the help of a sociology student from my university. This may indicate that by that time we already had a very radical opposition against this technology in Switzerland.

By then, from 1989 to 1990, I realized that food security does not only mean enough calories to avoid hunger. It also means having the right quality of food to avoid 'hidden hunger'. From then on I focused on this problem. Hidden hunger describes the fact that people who don't have a diversified diet are suffering from deficiencies in minerals, vitamins and essential amino acids – with most severe health consequences. Since by that time many laboratories, including powerful laboratories of large agbiotech multies, were working on resistance to any kind of biological or physical stress, and no laboratory was interested in the problem of hidden hunger – there was not much financial return to be expected – this became the field

of my lab. I started to focus on the problem of vitamin A deficiency. Vitamin A deficiency is a major public health problem and it affects 190 million preschool-age children and 19 million pregnant women around the world. Details from the WHO global database are given in Figure 1 (p. 368).

To reduce vitamin A-deficiency the World Health Organization (WHO) invests between 90 to 100 million dollars per year in the distribution of vitamin A capsules. We felt that a complementing intervention was a valuable task to test our technological possibilities. The distribution of vitamin A-deficiency around the world is given in Figure 2 (p. 368): exceptions are only Western Europe, North America and Australia, all the other countries are affected. The medical consequences from vitamin A deficiency are quite severe: irreversible blindness – every year we have about 250,000 children becoming blind due to vitamin A malnutrition; an impaired immune system – leading to the death of 2 million children from normal infectious diseases like measles; anaemia – because vitamin A plays an essential role in iron mobilisation and transport; impaired hematopoieses and maternal mortality during pregnancy – 19 million pregnant women at risk each year.

What was the scientific challenge we faced at the beginning of the 1990s? The status quo is the following. The rice plant produces large amounts of provitamin A in all green tissues (plants never produce vitamin A; plants produce provitamin A and our bodies convert provitamin A into vitamin A). Rice plants contain large amounts of provitamin A, but this is not accessible for our nutrition, because we can't eat the green parts; we eat the white starch-storing tissue in the seed, the 'endosperm' which doesn't contain any provitamin A. Therefore, poor people who can't afford to buy a diversified diet and depend upon rice as their major food source, are vitamin A-deficient. What alternatives were visible? One option was to try to find, within the entire gene pool of rice and its relatives around the world, a plant with 'yellow endosperm', indicating the presence of provitamin A. Such a plant, after the confirmation of the provitamin A nature of the yellow colour, could then be used as starting point for a breeding programme to transfer this trait into modern rice varieties. Well, the rice breeders had already been doing everything to find such a plant. They had studied more than 80,000 different genotypes but had not found any yellow endosperm and therefore had no possibility of initiating a breeding programme. Actually, the rice breeders were asking 'genetic engineering' for help and that's how I became aware of the situation. So what could we do on the basis of the knowledge about molecular biology and genetic engineering at that time? There were two alternatives and these were discussed in a brainstorming meeting at The Rockefeller Foundation in New York in 1991, organised in response to my request

for financial support. The foundation assembled 30 world experts of the biochemical pathway leading to provitamin A in any organism. The straightforward solution, as seen at this meeting, was trying to disclose the 'switch' that turns off the pathway in the white endosperm tissue. It was obvious that all necessary genes were present in rice, but they were selectively switched off in the endosperm. And there was good hope that this would be a relatively simple approach because there was a maize mutant known with a yellow endosperm, where such a switch had been identified. We – my partner Peter Beyer and I – proposed the alternative: to engineer the pathway. The assembled authority of these world experts felt (rightly) that this would be rather unfeasible, and they had very good arguments for their notion. Fortunately The Rockefeller Foundation decided to support both approaches. The group which had received funding to find the switch is still trying to find the switch and our 'totally unfeasible' approach – trying to engineer the entire biochemical pathway into rice endosperm – was successful. But this was, of course, not foreseeable in 1991. And we were fortunate that it worked. But it worked (Figure 3, p. 369).

Proof-of-concept was ready in February 1999. It came at the same date as my retirement as full professor from the Institute of Plant Sciences at ETH Zurich, and it came just one month before I had to leave. The rule says that you have to leave at the end of the semester in which you pass 65 years of age. But I was still able to present the results – including results on rice which had more iron to counteract iron deficiency – at my farewell symposium. Figure 4 (p. 369) shows what Golden Rice looks like. The left rice is yellow because it contains provitamin A and the right is white because it doesn't contain provitamin A. The colour is an indicator of the presence of provitamin A and, of course, we have all the necessary molecular evidence that this is the case.

Well, this was at the time of my retirement and, as a 'normal' scientist, I would have stopped there. The consequence would have been, however, that what I have been presenting to you about the vitamin A-rice would have remained an academic anecdote, but it would not have helped any vitamin A-deficient child. It has been stressed repeatedly during the few days of our Plenary that 'it is sufficient to do good science'; everything necessary will follow automatically. What we had done was definitely 'good science'. It became the most frequently cited plant paper for the three-year period from 2000-2003. If we had stopped there, it wouldn't have had any impact on vitamin A-malnutrition. The situation may be different in cases where there is an interest from the medical community or from industry to pick up a scientific novelty and to convert it into an economically viable product.

In our case however, there was no interest from industry because there was no foreseeable 'market' and consequently no chance for a return of the necessary investment. And there was no public institution ready to invest in the development of a 'humanitarian' product. Consequently, we decided to leave the convenient 'ivory tower' and we went into what turned out to be a very harsh environment. And we ran into many, many unforeseen non-academic problems that were not at all pleasant. For more information please see my paper on 'Lessons from the humanitarian Golden Rice project …' in the PAS Proceedings 2010, citation given at the end of this article. I won't refer here to the well-known problems with the professional GMO opposition. The first surprise came from the area of intellectual property rights. As long as one does basic science, patents don't play a negative role; they are a valuable source of technical information which can be used freely. But when one sets out to develop a 'product', patents suddenly play a key role. As typical scientists we didn't know how many patents we had been using with our technology. To find out, The Rockefeller Foundation commissioned two patent lawyers and the result was shocking: we had used 72 patents and a number of material transfer agreements. Since the concept of our 'humanitarian' project was to provide our 'Golden Rice' free of charge to subsistence farmers, this was a catastrophe, because it meant that we would have to bargain for free licenses for 72 patents. This appeared like an impossible task and the GMO opposition was certain that this was the end of our plans. However, thanks to our establishment of a 'public-private partnership' with agbiotech industry (Syngenta) we got help from experienced patent lawyers, who found out that we had to take care of only 12 patents, as the rest of the 72 patents were not recognized in those developing countries which were our target. So we had to get free licenses for 12 patents and, because of the popularity of our project, which was picked up by the press very readily – you may recall that it was even a cover story in *Time* magazine in 2000 – thanks to our colleague Peter Raven, who organised a press conference after inviting me to the 16th Botanical Congress in St Louis – the patent holders were very willing to provide us with free licences. The surprising outcome was, that whereas everybody had expected that the first insurmountable hurdle for our humanitarian project would be constituted by the problem of intellectual property rights, this didn't delay our project for a single day.

We then had to learn what it means to develop a product and that is basically very different from doing basic research. In summary, it requires solving many 'unacademic' tasks for which there is no funding and personnel in academia, including e.g. repetition of the same experiment hundreds of

times to find one transgenic event which is hopefully suitable for the development of a successful commercial product. That's very difficult in an academic environment, because there is no scientific novelty to be expected. Nobody in academia can invest the necessary time and nobody is willing to finance that. Another severe problem is the consequence of the GMO status. GMO plants are, as you all know, considered extremely dangerous plants. Nobody can tell why, but that's an established paradigm. The consequence is that work with GMO plants is restricted by numerous complicated and extremely restrictive hurdles. For a project aimed at using a GMO plant for improvement of a crop variety, e.g. to develop a vitamin A-rice variety, the fact that work in the field is prohibited inhibits possible progress to the extreme. Plant breeding is a numbers game; plant breeders need large numbers of plants to find an optimal variety: this is not possible in a growth chamber – as requested by law – where you can work with 50 instead of 500,000 plants. Also, plant breeding depends upon evaluation of agronomic traits in addition to the target trait, and it simply isn't possible to evaluate such traits in a growth chamber. Another very big hurdle was finding financial support for this work. It turned out that there's no public institution or funding agency in academia set up to support work beyond proof-of-concept. It was even very difficult to get modest bridging funds for the continuation of the project. Working on GMO product development requires (because of the regulation-caused costs) not the 'normal' EUR 100,000 to 500,000 like a 'normal' scientific project, or the exceptional one million euro. The costs for the development of a GMO-product accumulate to ca. USD 24 million. We had to spend much of our time during the last 11 years trying to acquire funding from year to year, from half year to half year, because, of course, nobody could provide 24 million USD for the completion of this project. We acknowledge gratefully all support received from altruistic sources such as The Rockefeller Foundation, USAid, Syngenta Foundation, Gates Foundation, and other foundations and this helped us to go on step by step. We established a public-private partnership because we learned very quickly that, as naïve academics, we had no idea what all this would involve to arrive at a GE product. We have built a board of experienced experts in many areas that are important to advance such a project to success. We had support from the private sector for our development. In order to develop local rice varieties we had to identify GMO-competent institutions in the developing countries that had the capacity to work with transgenic plants, not an easy and widespread capacity. We established collaboration on the basis of sub-license agreements (defining the 'humanitarian purpose' and the conditions for collaboration) with public rice research

institutes in our target countries India, Vietnam, China, Indonesia, the Philippines and Bangladesh. All this did not delay the progress of the project for long and it has been working fine since the early 2000s. These institutions are developing national varieties, they have the capacity, and part of them get sufficient funding.

A serious problem in this context is that, because of the expenses involved, regulation forces the entire breeding programme for all countries to be built on one selected lead event. This is very undesirable from the biological point of view. It would be far better to build on biological diversity also when breeding for different varieties. However, as the deregulation of one single transgenic event costs ca. 24 million USD, nobody can afford to build new varieties on several selected events. Prerequisite for the selection of such a 'lead event' amongst numerous transgenic events are reliable data on agronomic and target trait quality, which can't be collected in the growth chamber and absolutely require growth in the field. However, it took eight years to get the first permission for the first field release in the Philippines, our major partner country for testing Golden Rice in the field. Imagine what it means to select a lead event on the basis of agronomic traits which can't be studied because you are not allowed to work in the field! All these and hundreds of further hurdles are the consequence of GMO-regulation.

Despite of the numerous GMO-specific hurdles Golden Rice will reach farmers soon, however and unfortunately, with more than 10 years' delay compared to a novel non-GMO variety. The timeline for release is 2012 in the Philippines, 2013 in Bangladesh, 2014 in India and Vietnam, 2015 in China and Indonesia and further countries will follow. The figure below indicates our choice of collaborating partners in different countries: the countries highlighted in yellow are representing the actual programme and the grey ones are those into which I would very much like to extend the programme, but for which we have no financial support so far (Figure 5, p. 370).

In the following figure you see the expected impact for one representative country. According to a state-of-the-art socio-economic ex ante study for India, the annual burden of vitamin A-deficiency amounts to 71,600 lives lost per year: Golden Rice could save 39,700 of those lives. For those who may wonder why not more, the answer is very simple: only half of the Indian population depends upon rice, the other half depends upon wheat and, of course, Golden Rice cannot solve the problems of those who are vitamin A-malnourished but have wheat as their major staple. With regard to the rice-dependent poor, the success rate could reach an overwhelming 95%. Golden Rice interventions are extremely economic, because it could save one life year for 3 USD and, without the costs of regulation, Golden Rice could save one life year for 30 cents.

Golden Rice would substantially contribute to the UN Development Goal: eradication of extreme poverty and hunger (Figure 6, p. 370).

A World Bank study shows that the gain from the technology could be 15.6 billion dollars per year because of increased productivity of unskilled workers. It could lead to reduced child mortality (Golden Rice has the capacity to save India alone 40,000 lives), improved maternal health (vitamin A malnutrition is the prime important case for motherhood mortality, and Golden Rice could be of substantial help there). Golden Rice is followed by high iron, high zinc, high quality protein rice because there are these deficiencies as well and it is followed by the same traits – high vitamin A, zinc, iron, quality protein – in cassava, banana, sorghum, potato, to support those poor populations who are not dependent on rice but on other crops and Figure 7 (p. 371) indicates those countries that would benefit from transgenic cassava, banana and sorghum.

The examples given above demonstrate what potential genetic engineering with plants has to offer in the area of micro nutrient malnutrition or 'hidden hunger'. Golden Rice is the only case where scientific proof-of-concept has been carried through product development and deregulation and where the practical application will soon demonstrate the effectiveness of the concept of 'biofortification'. For all the other examples scientific proof-of-concept has been established, but product development and deregulation will delay use for at least ten years, as was the case with the vitamin A-rice – if the necessary funding (ca. USD 25 million per case) can be secured at all. GMO-regulation prevents use of the technology for public good and effective use of the potential of the technology will require a substantial change in public attitude and regulation.

This problem leads to the last theme of my presentation and to a few remarks about an important Study Week organized by the Pontifical Academy of Sciences on the topic of *Transgenic Plants for Food Security in the Context of Development*, to which about 40 renowned scientists from very diverse scientific backgrounds where invited to discuss, on the basis of peer reviewed literature, the recent advances in the scientific understanding of GMO plants and the social conditions under which GMO technology should be made available for the improvement of agriculture, especially for food security in developing countries. A short account has already been given by our colleague Peter Raven. The key message from this study week is the following: there is no scientifically valid argument justifying any specific concern about transgenic plants, and both practical experience of their use over more than twelve years on large acreages world-wide and by millions of small scale farmers, as well as all regulatory oversight and specific

biosafety research over 25 years, confirm this view. On the contrary, GMO-technology has been proven to be the safest and most predictable technique for producing new plant varieties. There is not a single documented incidence of harm, so far, to either consumer or the environment. Despite this overwhelming scientific evidence and practical experience, unjustified 'extreme precautionary' regulation, exclusively for GMOs, is maintained and enforced worldwide, with the consequence that GMO-technology is so expensive that it has led to a *de facto* monopoly in favour of a few financially powerful industries and to the exclusion of any possible altruistic application in the interest of public good. Golden Rice is the only exception and may be for a long time. There is, therefore, a moral imperative to change regulation from present ideology-based regulation to science-based regulation which would be based on novel traits instead of on the regulation of the technology used. The Proceedings of the Study Week have been published in parallel by Elsevier and the Pontifical Academy of Sciences. They are a rich source of science-based information on all aspects of this controversial but life-saving technology and should be studied by all who are interested in an unbiased view on the subject. They contain the full papers of all presentations, but more importantly also a 'Statement' endorsed by all participants, providing an authoritative and comprehensive summary. I would like to thank Peter Raven, who was instrumental in managing a draft and the formulation of the final Statement to which all forty participants agreed without exception, including the late President of this Academy. This Statement is available in 16 important world languages and it has been distributed to 200 countries. We hope that other academies will join and help distribute this information and that this statement and the publications will serve as a catalyst for a more rational attitude towards GMO-technology.

References

Transgenic Plants for Food Security in the Context of Development. Proceedings of a study week of the Pontifical Academy of Sciences. Editors: Ingo Potrykus & Klaus Ammann. *New*Biotechnology, vol. 27 (5), 30 November 2010, pp. 443-717.

This 'open-source' publication is accessible via internet under www.ask-force.org/web/PAS-Studyweek-Leaflet-2010.pdf and under the Vatican homepage www.vatican.va/roman_curia/pontifical_academies/acdscien/2010/newbiotechnology nov2010.pdf.

Session V: Neuroscience and Immunology

Discovery of the Defensive System of the Endothelium, the Lining of the Arterial Wall

Andrzej Szczeklik

Introduction

Blood flows through the vessels that are tightly covered by a monocellular lining – 'a membrane' – endothelium. Is it really a membrane? 'Yes, indeed, a primitive membrane' – answered Rudolf Virchow, who first observed it at autopsy and described in 1860. A hundred years later, Sir Howard Florey expressed some doubts about this term, saying that 'endothelium could be more than a sheet of cellophane'. In 1996 Sir John Vane called the endothelium 'a maestro of blood circulation' [1].

Rudolf Altschul was the first to think that the endothelium might have a secretory function. He was a political émigré to Canada from Central Europe during World War II. With help of a simple microscope he perceived that the endothelium rises like a palisade to defend arteries against an approaching catastrophe, brought about by atherosclerosis. In his book published in 1954 he wrote: 'the secretory function of endothelium needs to be considered'. The book contains a moving dedication to 'Anni Caroline who was very brave when the ship went down' [2].

We know now that the endothelium is the main defensive barrier in the cardiovascular system. It achieves this goal by synthesizing several chemical compounds with powerful biological activity, of which prostacyclin and nitric oxide are most important. Other compounds, like heme oxygenase-1 [3], are also emerging, but they will not be discussed here.

The superfamily of eicosanoids

Prostacyclin belongs to the superfamily of eicosanoids [in Greek 'eicosa' (εικοσα) stands for twenty – in that case twenty carbon atoms in a molecule]. Indeed, prostaglandins derive from eicosa-all cis-5,8,11,14-tetraenoic acid, i.e. arachidonic acid (AA). AA may be subdued to a number of enzymic manipulations. Firstly, phospholipase A_2 cuts it out from the cellular phospholipids stores. Next, free AA is exposed to the enzymes available in various types of cells and in various compartments. For us, the most interesting are cyclooxygenases-1 and -2. They generate prostaglandin endoperoxides (PGG_2

and PGH_2), which are substrate for both thromboxane and prostacyclin synthesis, through action of the specific enzymes (synthases). Thrombogenic thromboxane A_2 is generated by COX-1 in blood platelets (Fig. 1). Aspirin at low doses is a pretty selective inhibitor of COX-1 [4] in blood platelets, hence aspirin is effective against myocardial infarction. However, in a special category of 'aspirin-sensitive' patients, aspirin itself and other non-steroidal anti-inflammatory drugs may precipitate asthmatics attacks interfering with COX-1 activity in the respiratory tract [5,6].

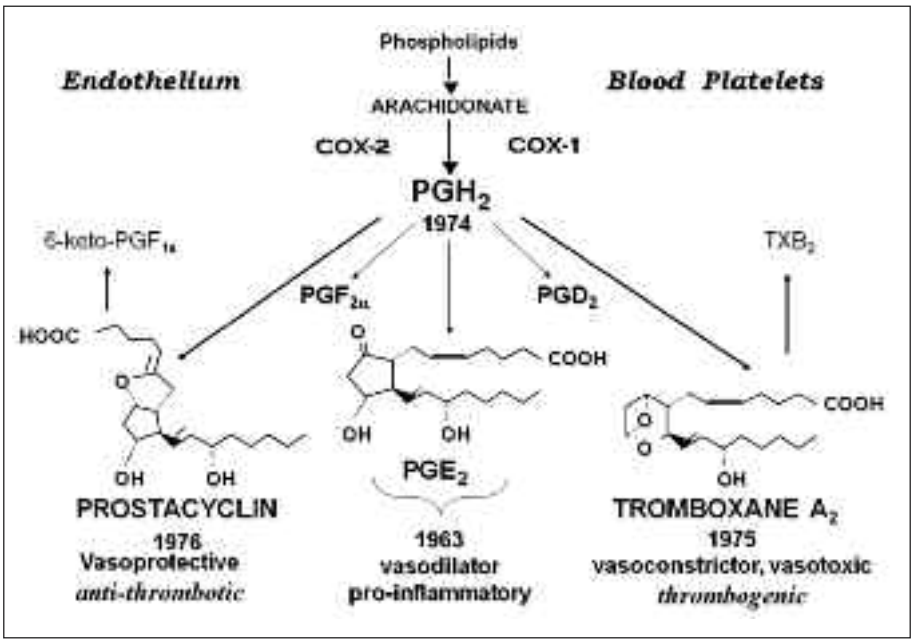


Figure 1. Arachidonate transformation via the cyclooxygenase (COX-1, COX-2) pathways. PGH_2 – prostaglandin endoperoxide H_2; PGD_2 PGE_2, PGF_2 – prostaglandins D_2 E_2 and F_2.

The early days of prostaglandins [7]

Ulf Svante von Euler was the first who in 1935 used the name *prostaglandin* for a lipid factor that he extracted from *glandula prostatae*; that factor contracted smooth muscles of various organs. In 1960 Sune Bergström and his coworkers isolated prostaglandins from the biological material and determined their chemical structures (cyclic lipids). The abbreviation PGs was introduced; the first established were: PGE_2 and $PGF_{2\alpha}$. Later other prostaglandins were dis-

covered. In 1964 David Van Dorp of Unilever in Holland and Sune Bergström of Karolinska Institutet in Sweden, with their coworkers, discovered that PGs were biosynthesized from polyunsaturated fatty acids. For the physiologically important PGs of the 2 series – a specific substrate is arachidonic acid (AA).

The discovery of prostacyclin

Prostacyclin was discovered in 1976 by Richard Gryglewski in collaboration with Salvador Moncada and a student, Stuart Bunting, in John Vane's laboratory. In the early 1970s Priscilla Piper, John Vane and Richard Gryglewski noticed that challenged, sensitized lungs release an activity which they called 'rabbit aorta contracting substance' [7]. Two years later Bengt Samuelson identified this activity as composed of prostaglandin endoperoxides (PGG_2, PGH_2) and thromboxane A_2 (TXA_2) [8]. He isolated these compounds and sent a sample of PGH_2 to John Vane. So when Richard Gryglewski, a young Polish pharmacologist, came for his third sabbatical to Vane's laboratory, John gave him these endoperoxides and asked him to look for their conversion to PGs or TXA_2 by ground-up cells (homogenates or microsomes) of various organs. He used the Vane Bioassay Cascade, equipped for the detector of PGs (mainly a rat stomach strip) and for TXA_2 (a special assembly of rabbit aorta). John McGiff of the Valhalla New York Medical College depicted the Vane Bioassay Cascade (Fig. 2) as 'the triumph of intellect over the technology'. So, with varying results, Gryglewski tested homogenates from different animal organs. The microsomes from the most studied organs converted PG endoperoxides to prostaglandins, exclusively. Of course, blood platelets converted PGG_2 and PGH_2 to TXA_2. When it came to the pig aortic microsomes – they behaved differently – since neither PGG_2 nor TXA_2 were produced and, even worse, the cascade detected no biological activity at all. At this point Gryglewski and his colleagues started to play around with their biological detectors within the Vane Bioassay Cascade (Fig. 3). They introduced alterations, incorporating rabbit celiac and mesenteric arteries as well as rat colon. Then they detected a unique set of contractions and relaxations (the unique set of fingerprints, as they called it) in response to a mixture of aortic microsomes incubated with PGG_2 and PGH_2. The responses were, however, variant and even elusive, so the jokes of 'an invisible Polish hormone' (PGX) appeared in the laboratory where we worked. Gryglewski had a brilliant thought. Maybe something so volatile was produced that it disappeared at room temperature? So he set up a trap for that 'something' by repeating the experiment on ice. This time the detector system showed – in a reproducible way – a compound

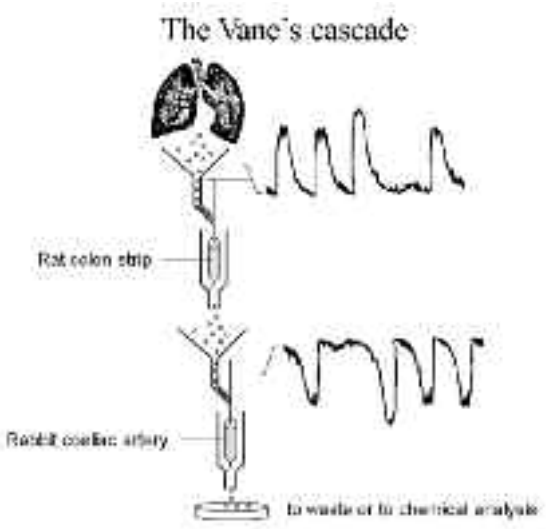


Figure 2. The Vane Bioassay cascade. The effluent from challenged lungs superfuses strips of various experimental organs. The registration system records contraction of rat colon and relaxation of rabbit coeliac artery.

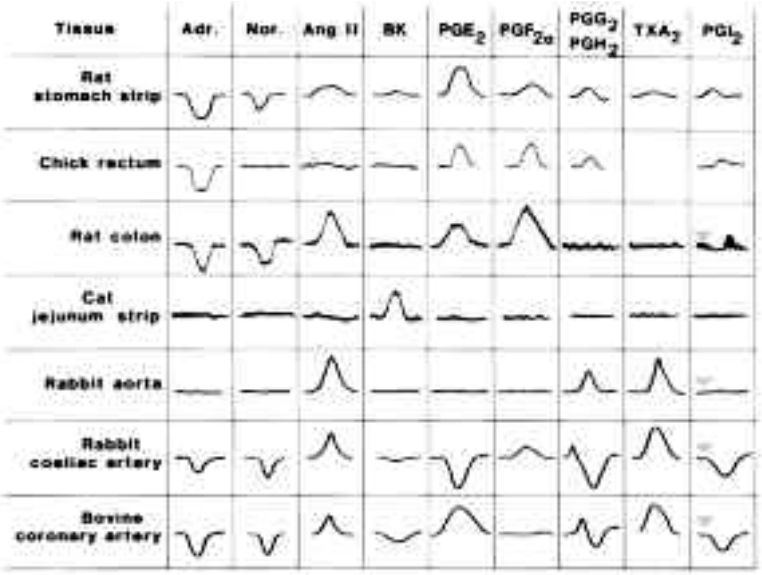


Figure 3. A slide presented by Sir John Vane during his Nobel Lecture in Stockholm 1982 shows set of 'fingerprints' for various biologically active compounds as they are registered in the bioassay cascade. (Adr = Adrenaline, Nor = noradrenaline, Ang II = angiotensin, BK = bradykinin).

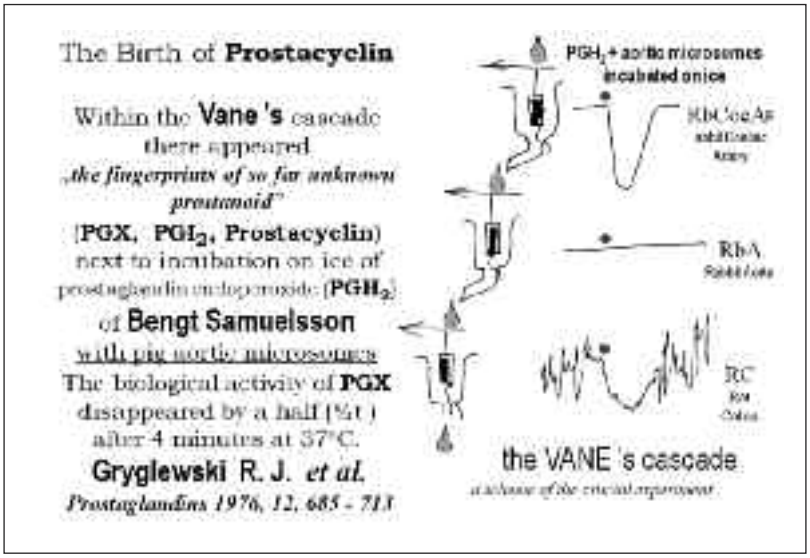


Figure 4. The crucial experiment showing generation of prostacyclin.

that was unknown (Fig. 4). It was prostacyclin (PGI_2). In a series of quick, ingenious experiments Gryglewski and his colleagues provided proof of the existence of prostacyclin [9]. Early studies [10,11] demonstrated that low density lipoproteins (LDL) inhibit prostacyclin biosynthesis, while high-density lipoprotein (HDL) exert an opposite, protective effect; results of these studies were later confirmed by several authors. A concept of inhibition of PGI_2 biosynthesis by lipid peroxides as a hypothetical step in development of atherosclerosis was proposed [12] (Fig. 5).

The chemical structure of prostacyclin was determined shortly after its discovery, followed by successful chemical synthesis. Soon it was given to man. Ryszard Gryglewski and the author of this paper were the first men to receive intravenous infusions of prostacyclin in 1977. These exciting experiments on ourselves, full of unexpected adventures, have been described [13]. Having established the safety of the procedure on ourselves and colleagues from our Department, we continued the observations on the action of prostacyclin on the volunteers, both healthy and patients. By the end of 1979 over 70 subjects had received PGI_2 either intravenously or by inhalation. These studies led to the following conclusions on the actions of prostacyclin in man:

1. PGI_2, administered either i.v. or by inhalation, exerted powerful anti-platelet effects. It prolonged bleeding time, suppressed platelet aggrega-

tion, dispersed circulating platelet aggregates and prevented formation of thrombin. On the contrary, it did not affect such plasma coagulation indices as prothrombin time or partial thromboplastin time [14,15].

2. PGI_2 produced profound circulatory effects [16]. Flushing of the face, spreading down to the neck in the form of a collar, were the first clinical symptoms which appeared in all the subjects after only a few minutes of the infusion at the low dose (2-5ng kg^{-1} min^{-1}). Erythema of the palms and feet was also observed in the majority of patients receiving PGI_2 by inhalation.
   There was a distinct fall in peripheral and total pulmonary vascular resistances. This was accompanied by a drop in intra-arterial blood pressure, and the acceleration of heart rate. Stroke volume, cardiac output, mean right arterial pressure, and left a ventricular end diastolic pressure showed no significant change. Prostacyclin appeared to act predominantly on resistance vessels (Fig. 6).
3. Prostacyclin simulated fibrinolysis without systemic degradation of fibrinogen [17].
4. PGI_2 and its stable analogs affected glucose metabolism, leading to a moderate hyperglycaemia upon i.v. infusion [18] and modulation of insulin secretion in isolated pancreatic islets [19].
5. Lung function studies revealed no changes following i.v. or inhaled administration of PGI_2 to healthy subjects and patients with asthma [14,15].

Figure 5. John Vane and Richard Gryglewski (at right) in 1976.

These insights into action of PGI_2 led to the following early clinical applications:

1. Pulmonary hypertension. In 1980, based on our pilot studies in primary and secondary pulmonary hypertension [16,20], we proposed that prostacyclin, administrated for an extended period of time either i.v. or by inhalation, may be a new useful therapy in these conditions.
2. Advanced peripheral artery disease, affecting middle- and low-caliber arteries (e.g. peripheral vasculopathies) [21,22].
3. Prinzmetal angina pectoris [23,24].

Clinical use of prostacyclin opened new revolutionary therapeutic possibilities. In pulmonary arterial hypertension it is now the treatment of choice. The synthetic stable analogues of prostacyclin, such as iloprost, treprostinil, epoprostenol, beraprost or cicaprost altered the approach to pulmonary arterial hypertensia, especially when combined with syldenafil (an inhibitor of phosphodiesterase-5) or bosentan (an antagonist of endothelin

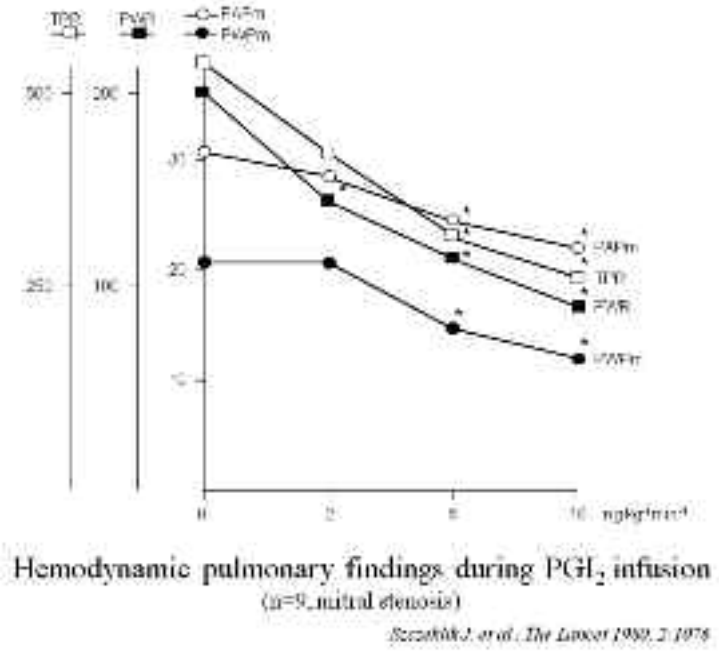


Figure 6. First demonstration of vasodilatory action of prostacyclin on pulmonary circulation. The graph represents mean values in 9 patients with mitral steriosis and moderate pulmonary artery hypertension, TPR = total pulmonary resistance, PWR = pulmonary wedge pressure, PAPm = mean pulmonary artery pressure, mPWP = mean pulmonary wedge pressure.

Figure 7. Robert Furchgott when visiting Cracow in 1994.

ET-1 receptor) [25]. Prostacyclin and its stable analogues proved to be a valuable therapeutic improvement in critical limb ischemia, where they show efficacy in rest-pain relief and ulcer healing; they also show favorable results regarding major amputations [26]. Finally, there are some well known drugs, which apart from their principal mechanism of action, perform also as 'pleiotropic' releasers of prostacyclin from the endothelium. The best known are lipophilic angiotensin converting enzyme inhibition (ACE-1, e.g. quinapril, perindopril, ramipril) and statins (e.g. atorvastatin). The long list of prostacyclin releasers include also some β-adrenoreceptor blocking agents (nebivolol, carvediolol), antiplatelet thienopyridines (ticlopidine, clopidogrel) and anti-diabetic drugs (e.g. glipizide, metformin) [27].

Discovery of nitric oxide

In the late 1970s and early 1980s, at the time when prostacyclin was an absolute hit, a scientist started to appear at medical meetings, claiming that the endothelium produces another biologically active compound, different from prostacyclin, which also dilates arteries. He called it Endothelium-Derived Relaxing Factor (EDRF) and not many of us, fascinated by prostacyclin, believed his story. But he was right. His name was Robert Furchgott (Fig. 7) and, in contrast to many self-promoting hyper-ambitious scientists, he was self-effacing with an ever mild manner, and generous to a fault. His

daughter called him 'a real Southern gentleman' (he was born and raised in Charleston, S.C.). He was professor of pharmacology in New York and spent most of his time studying in vitro the effects of acetylcholine (an important neurotransmitter) on strips of blood vessels of experimental animals. Acetylcholine was a well-known vasodilator in intact organisms. Furchgott was an expert on the arterial strips responses to acetylcholine. He noticed, for instance that one of his preparations, which sat beneath a sunlit window, dilated much more than preparations in a darker part of the laboratory (in retrospect, it seems that photorelaxation reflected the release of NO by blood vessels in response to light).

Furchgott showed, quite unexpectedly, that relaxation of blood vessels to certain substances depended on whether the endothelium was present or not. He made his clinical discovery in 1978 [28], when a technician failed to follow a standard protocol for preparing the rabbit aorta strips, and instead of contraction to acetylcholine, Furchgott saw relaxation. He was eager to troubleshoot this 'accident' and after several weeks realized that gentle rubbing of blood vessels transformed relaxation into contraction. One explanation was that acetylcholine acts on receptors on endothelial cells (removed by rubbing) to trigger the release of a substance with a relaxing activity – EDRF. Furchgott received direct evidence of this by making a 'sandwich' of a ring of aorta freed of endothelial cells to which he applied an endothelium of another aortic strip; the procedure transformed constriction into relaxation [28,29]. In the following years EDRF was shown to be nitric oxide (NO^0) [30] and in 1998 R. Furchgott, together with Ferrid Murrad and Louis Ignarro, received the Nobel Prize 'for their discoveries concerning nitric oxide as a signaling molecule in the cardiovascular system' (Fig. 8).

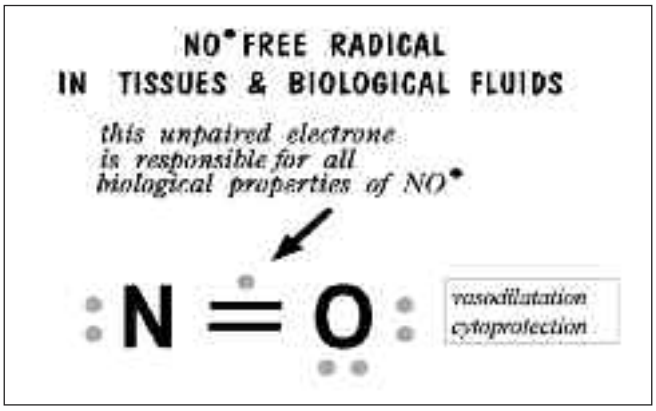


Figure 8. Nitric oxide.

Alfred Nobel was a chemist who, in 1866, discovered an explosive, dynamite, composed of nitroglycerin and a stabilizing absorbent. This discovery brought him fortune, which he used to create the famous award. In his last years of life, his physician prescribed Nobel nitroglycerin for angina pectoris. Nobel then wrote to a friend: 'It sounds like the irony of fate that I should be ordered by my doctor to take nitroglycerin *internally*'. Over a hundred years later Robert Furchgott, Louis Ingarro and Ferrid Murrad were recognized for showing that nitroglycerin produces long-lasting relaxation of cardiac muscle, because it breaks down, yielding a steady stream of NO. 'Today – said Furchgott, receiving the Nobel Prize in Stockholm – it seems like fate, but not the irony of fate' [31].

Nitric oxide and its relationship to prostacyclin

Over the last decades, appreciable knowledge has been acquired on the biological importance of nitric oxide. It is synthesized in the body from the amino acid l-arginine by the action of NO synthase enzymes (NOS). Nitric oxide is a gaseous free radical that serves multiple functions in human physiology (Fig. 9). It causes vasodilatation and inhibits platelet aggregation, when it is secreted from endothelial cells. It exerts antioxidant, antiproliferative and anti-inflammatory properties, thus playing an important role in inhibiting the atherosclerotic process. It modulates many reactions in the immune system. Produced by macrophages it combats bacteria directly and also signals other immune responses. Furthermore, it functions as a neurotransmitter by diffusing into surrounding cells rather than activating receptors. It also plays a role in reproduction, functioning as a vasodilator during penis erection [7].

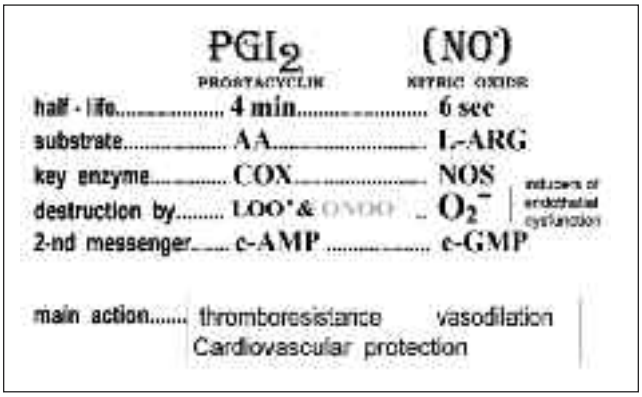


Figure 9. Prostacyclin vs. nitric oxide.

In pathological conditions a methylation of arginine to asymmetric dimethylarginine (ADMA) may occur. The latter inhibits eNOS. A toxic peroxynitrate ($ONOO^-$) is generated in a reaction between NO° and superoxide [32-34]. It selectively blocks the enzymatic activity of prostacyclin synthase, promoting development of atherosclerosis. Prevention of these disastrous processes opens a new avenue in cardiology (Fig. 10).

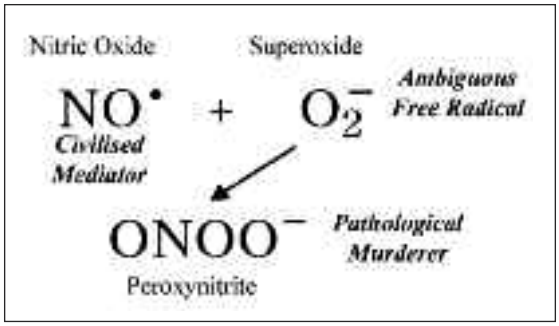


Figure 10. Reaction of nitric oxide with superoxide.

When measuring the value of scientific work, nothing is better than time. Time separates the wheat from the chaff. But we refuse to wait, we simply cannot, because we won't be here any more when the truth is revealed. We want it all here and now. However, there are no recipes for scientific discovery or for success. Max Delbrück, a brilliant physicist who introduced quantitative thought to biology, reckoned that in performing an experiment we should admit a certain degree of freedom, some flexibility, in order to perceive the unexpected, the surprise that is worth more than the expected result. He called this 'the principle of limited sloppiness'. To this principle, so well illustrated by discoveries of prostacyclin and nitric oxide, one may add a sentence: 'Never ignore the unexpected'.

References

1. Vane JR. The Croonian Lecture, 1993. The endothelium: maestro of the blood circulation. *Philos Trans R Soc Lond B Biol Sci*. 1994;343:225-46.
2. Altschul R. *Endothelium. Its Development, Morphology, Function, and Pathology*. New York, The Macmillan Company, 1954.
3. Dulak J, Loboda A, Jozkowicz A. Effect of heme oxygenase-1 on vascular function and disease. *Curr Opin Lipidol*. 2008;19:505-1.
4. Vane JR. Inhibition of prostaglandin syn-

thesis as a mechanism of action for aspirin-like drugs. *Nat New Biol.* 1971; 231:232-5.
5. Szczeklik A, Gryglewski RJ, Czerniawska-Mysik G. Relationship of inhibition of prostaglandin biosynthesis by analgesics to asthma attacks in aspirin-sensitive patients. *Br Med J.* 1975;1:67-9.
6. Szczeklik A. Nizankowska-Mogilnicka E., Sanak M. Hypersensitivity to Aspirin and Nonsteroidal Anti-Inflammatory Drugs. In: *Middleton's Allergy*, 7th Edition, Editors: Adkinson, Busse, Bochner, Holgate, Simons & Lemanske, Mosby Elsevier. 2009 pp. 1227-1243.
7. Gryglewski RJ. Pharmacology of vascular endothelium. *FEBS J.* 2005;272: 2956-67.
8. Hamberg M, Samuelsson B. Prostaglandin endoperoxides. Novel transformations of arachidonic acid in human platelets. *Proc Natl Acad Sci USA.* 1974;71:3400-4
9. Gryglewski RJ, Bunting S, Moncada S, Flower RJ, Vane JR. Arterial walls are protected against deposition of platelet trombi by a substance (prostaglandin X) which they make from prostaglandin endoperoxides. *Prostaglandins*. 1976; 12:685-713.
10. Szczeklik A, Gryglewski RJ. Low-density lipoproteins (LDL) are carriers for lipid peroxides and inhibit prostacyclin (PGI2) biosynthesis in arteries. *Artery*. 1980;7:488-95
11. Szczeklik A, Gryglewski RJ, Domagala B, Zmuda A, Hartwich J, Wozny E, Grzywacz M, Madej J, Gryglewska T. Serum lipoproteins, lipid peroxides and prostacyclin biosynthesis in patients with coronary heart disease. *Prostaglandins*. 1981;22:795-807.
12. Szczeklik A, Gryglewski RJ. Inhibition of prostacyclin formation by lipid peroxides in the arterial wall: hypothetical step in development of atherosclerosis. *Mater Med Pol.* 1978;10:338-41.
13. Szczeklik A. *Catharsis. On the art of medicine*. University of Chicago Press, 2005.
14. Szczeklik A, Gryglewski RJ, Nizankowska E, Nizankowski R, Musial J. Pulmonary and antiplatelet effects of intravenous and inhaled prostacyclin in man. *Prostaglandins.* 1978;16:651-660
15. Szczeklik A Gryglewski RJ. Actions of prostacyclin in man. *Prostacycline.* J Vane and S Bergström (eds). Raven Press, New York 1979. pp. 393-407.
16. Szczeklik J, Szczeklik A, Nizankowski R. Haemodynamic changes induced by prostacyclin in man. *British Heart Journal* 1980;44:254-258.
17. Szczeklik A, Kopec M, Sladek K, Musial J, Chmielewska J, Teisseyre E, Dudek-Wojciechowska G, Palester-Chlebowczyk M. Prostacyclin and the fibrinolytic system in ischemic vascular disease. *Thromb Res.* 1983;29:655-60.
18. Szczeklik A, Pieton R, Sieradzki J, Nizankowski R. The effects of prostacyclin on glycemia and insulin release in man. *Prostaglandins*. 1980;19:959-68.
19. Sieradzki J, Wolan H, Szczeklik A. Effects of prostacyclin and its stable analog, iloprost, upon insulin secretion in isolated pancreatic islets. *Prostaglandins*. 1984; 28:289-96
20. Szczeklik J, Szczeklik A, Nizankowski R. Prostacyclin for pulmonary hypertension. *Lancet*. 1980;2:1076
21. Szczeklik A, Nizankowski R, Skawinski S, Szczeklik J, Gluszko P, Gryglewski RJ. Successful therapy of advanced arteriosclerosis obliterans with prostacyclin. *Lancet*. 1979;1:1111-4.
22. Nizankowski R, Krolikowski W, Bielatowicz J, Szczeklik A. Prostacyclin for ischemic ulcers in peripheral arterial disease. A random assignment, placebo controlled study. *Thromb Res.* 1985; 37:21-8.
23. Szczeklik A, Szczeklik J, Nizankowski R, Gluszko P. Prostacyclin for unstable angina. *N Engl J Med*. 1980;303:881.

24. Szczeklik A, Nizankowski R, Szczeklik J, Tabeau J, Krolikowski W. Treatment with prostacyclin of various forms of spontaneous angina pectoris not responding to placebo. *Pharmacol Res Commun*. 1984;16:1117-30.
25. Chin KM, Rubin LJ. Pulmonary arterial hypertension. *J Am Coll Cardiol*. 2008;51:1527-38
26. Ruffolo AJ, Romano M, Cipponi A. Prostanoids for critical limb ischemia. *The Cochrane Library* 2000, issue 3. Wiley.
27. Gryglewski RJ. Prostacyclin among prostanoids. *Pharmacol Rep*. 2008;60:3-11.
28. Furchgott RF, Zawadzki JV. The obligatory role of endothelial cells in the relaxation of arterial smooth muscle by acetylcholine. *Nature*. 1980;288:373-6.
29. Cherry PD, Furchgott RF, Zawadzki JV, Jothianandan D. Role of endothelial cells in relaxation of isolated arteries by bradykinin. *Proc Natl Acad Sci USA*. 1982; 79:2106-10.
30. Furchgott RF. *Vasolidation: Vascular Smooth Muscle, Peptides Autonomic Nerves and Endothelium*, ed. Vanhoutte P.M. Raven, New York 1988, pp. 401-404.
31. Pincock S. Robert Francis Furchgott. *Lancet*. 2009, 373:2194.
32. Gryglewski RJ, Palmer RM, Moncada S. Superoxide anion is involved in the breakdown of endothelium-derived vascular relaxing factor. Nature. 1986 Apr 3-9;320(6061):454-6.
33. Vásquez-Vivar J, Kalyanaraman B, Martásek P, Hogg N, Masters BS, Karoui H, Tordo P, Pritchard KA Jr. Superoxide generation by endothelial nitric oxide synthase: the influence of cofactors. *Proc Natl Acad Sci USA*. 1998; 95:9220-5.

Intracellular Protein Degradation: From a Vague Idea thru the Lysosome and the Ubiquitin-Proteasome System and onto Human Diseases and Drug Targeting*

Aaron Ciechanover

Introduction

The concept of protein turnover is hardly 60 years old. Beforehand, body proteins were viewed as essentially stable constituents that were subject to only minor 'wear and tear': dietary proteins were believed to function primarily as energy-providing fuel, which were independent from the structural and functional proteins of the body. The problem was hard to approach experimentally, as research tools were not available. An important research tool that was lacking at that time were stable isotopes. While radioactive isotopes were developed earlier by George de Hevesy (de Hevesy G., Chemistry 1943. In: *Nobel Lectures in Chemistry 1942-1962*. World Scientific 1999. pp. 5-41), they were mostly unstable and could not be used to follow metabolic pathways. The concept that body structural proteins are static and the dietary proteins are used only as a fuel was challenged by Rudolf Scheonheimer in Columbia University in New York City. Schoenheimer escaped from Germany and joined the Department of Biochemistry in Columbia University founded by Hans T. Clarke (1-3). There he met Harold Urey who was working in the Department of Chemistry and who discovered deuterium, the heavy isotope of hydrogen, a discovery that enabled him to prepare heavy water, D_2O. David Rittenberg, who had recently received his Ph.D. in Urey's laboratory, joined Schoenheimer, and together they entertained the idea of 'employing a stable isotope as a label in organic compounds, destined for ex-

**Abbreviations used*: ODC, ornitihine decarboxylase; G6PD, glucose-6-phosphate dehydrogenase; PEPCK, phosphoenol-pyruvate carboxykinase; TAT, tyrosine aminotransferase; APF-1, ATP-dependent Proteolysis Factor 1 (ubiquitin); UBIP, ubiquitous immunopoietic polypeptide (ubiquitin); MCP, multicatalytic proteinase complex (26S proteasome); CP, 20S core particle (of the proteasome); RP, 19S regulatory particle (of the proteasome). *Keywords*: ubiquitin, proteasome, protein degradation, lysosome.

periments in intermediary metabolism, which should be biochemically indistinguishable from their natural analog' (1). Urey later succeeded in enriching nitrogen with ^{15}N, which provided Schoenheimer and Rittenberg with a 'tag' for amino acids and as a result for the study of protein dynamics. They discovered that following administration of ^{15}N-labelled tyrosine to rats, only ~50% was recovered in the urine, 'while most of the remainder is deposited in tissue proteins. An equivalent of protein nitrogen is excreted' (4). They further discovered that from the half that was incorporated into body proteins 'only a fraction was attached to the original carbon chain, namely to tyrosine, while the bulk was distributed over other nitrogenous groups of the proteins' (4), mostly as an αNH_2 group in other amino acids. These experiments demonstrated unequivocally that the body structural proteins are in a dynamic state of synthesis and degradation, and that even individual amino acids are in a state of dynamic interconversion. Similar results were obtained using ^{15}N-labelled leucine (5). This series of findings shattered the paradigm in the field at that time that: (1) ingested proteins are completely metabolized and the products are excreted, and (2) that body structural proteins are stable and static. Schoenheimer was invited to deliver the prestigious Edward K. Dunham lecture at Harvard University where he presented his revolutionary findings. After his untimely tragic death in 1941, his lecture notes were edited Hans Clarke, David Rittenberg and Sarah Ratner, and were published in a small book by Harvard University Press. The editors called the book *The Dynamic State of Body Constituents* (6), adopting the title of Schoenheimer's presentation. In the book, the new hypothesis is clearly presented:

> The simile of the combustion engine pictured the steady state flow of fuel into a fixed system, and the conversion of this fuel into waste products. The new results imply that not only the fuel, but the structural materials are in a steady state of flux. The classical picture must thus be replaced by one which takes account of the dynamic state of body structure.

However, the idea that proteins are turning over was not accepted easily and was challenged as late as the mid-1950s. For example, Hogness and colleagues studied the kinetics of β-galactosidase in *E. coli* and summarized their findings (7):

> To sum up: there seems to be no conclusive evidence that the protein molecules within the cells of mammalian tissues are in a dynamic state. Moreover, our experiments have shown that the proteins of growing E. coli are static. Therefore it seems necessary to conclude that the synthesis and maintenance of proteins within growing cells is not necessarily or inherently associated with a 'dynamic state'.

While the experimental study involved the bacterial β-galactosidase, the conclusions were broader, including also the authors' hypothesis on mammalian proteins. The use of the term 'dynamic state' was not incidental, as they challenged directly Schoenheimer's studies.

Now, after more then six decades of research in the field and with the discovery of the lysosome and later the complex ubiquitin-proteasome system with its numerous tributaries, it is clear that the area has been revolutionized. We now realize that intracellular proteins are turning over extensively, that this process is specific, and that the stability of many proteins is regulated individually and can vary under different conditions. From a scavenger, unregulated and non-specific end process, it has become clear that proteolysis of cellular proteins is a highly complex, temporally controlled and tightly regulated process that plays major roles in a broad array of basic pathways. Among these processes are cell cycle, development, differentiation, regulation of transcription, antigen presentation, signal transduction, receptor-mediated endocytosis, quality control, and modulation of diverse metabolic pathways. Subsequently, it has changed the paradigm that regulation of cellular processes occurs mostly at the transcriptional and translational levels, and has set regulated protein degradation in an equally important position. With the multitude of substrates targeted and processes involved, it is not surprising that aberrations in the pathway have been implicated in the pathogenesis of many diseases, among them certain malignancies, neurodegeneration, and disorders of the immune and inflammatory system. As a result, the system has become a platform for drug targeting, and mechanism-based drugs are currently developed, one of them is already on the market.

The lysosome and intracellular protein degradation

In the mid-1950s, Christian de Duve discovered the lysosome (see, for example, Refs. 8 and 9 and Figure 1). The lysosome was first recognized biochemically in rat liver as a vacuolar structure that contains various hydrolytic enzymes which function optimally at an acidic pH. It is surrounded by a membrane that endows the contained enzymes latency that is required to protect the cellular contents from their action (see below). The definition of the lysosome has been broadened over the years. This is because it has been recognized that the digestive process is dynamic and involves numerous stages of lysosomal maturation together with the digestion of both exogenous proteins (which are targeted to the lysosome through receptor-mediated endocytosis and pinocytosis) and exogenous particles (which are targeted via phagocytosis; the two processes are known as heterophagy), as well as digestion

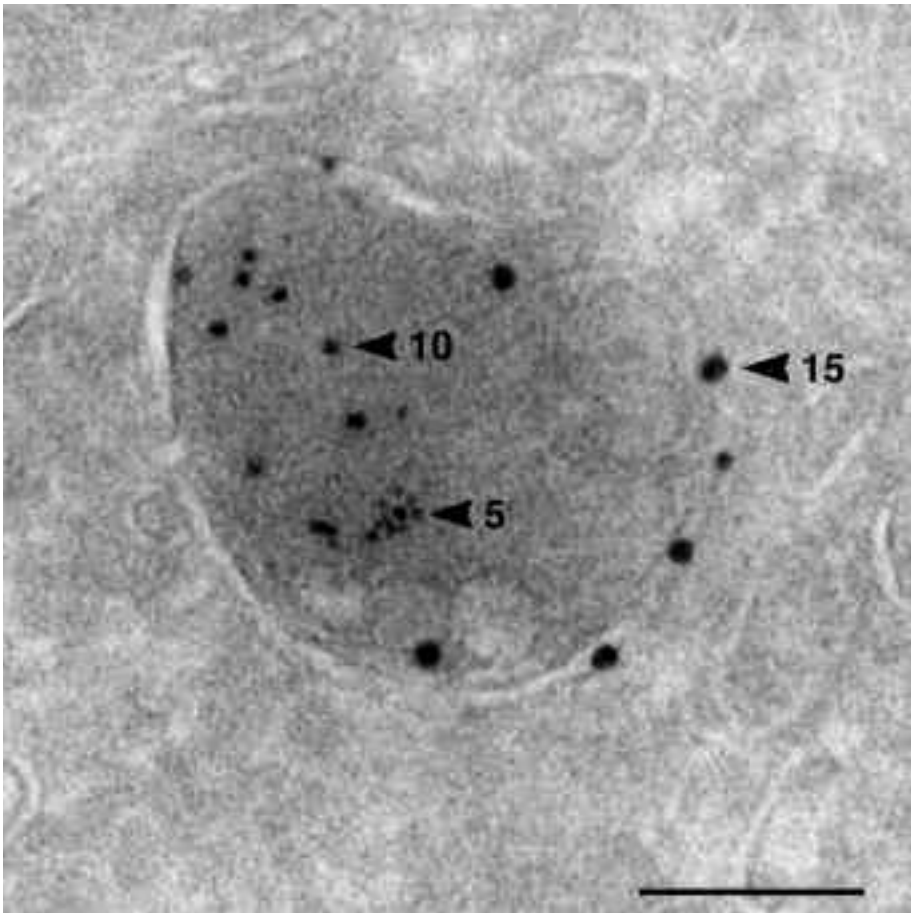


Figure 1. ***The lysosome:*** Ultrathin cryosection of a rat PC12 cell that had been loaded for 1 hour with bovine serum albumin (BSA)-gold (5 nm particles) and immunolabelled for the lysosomal enzyme cathepsin B (10-nm particles) and the lysosomal membrane protein LAMP1 (15 nm particles). Lysosomes are recognized also by their typical dense content and multiple internal membranes. Bar, 100 nm. Courtesy of Viola Oorschot and Judith Klumperman, Department of Cell Biology, University Medical Centre Utrecht, The Netherlands.

of endogenous proteins and cellular organelles (which are targeted by micro- and macro-autophagy; see Figure 2, p. 372). The lysosomal/vacuolar system as we currently recognize it is a discontinuous and heterogeneous digestive system that also includes structures that are devoid of hydrolases – for example, early endosomes which contain endocytosed receptor-ligand complexes and

pinocytosed/phagocytosed extracellular contents. On the other extreme it includes the residual bodies – the end products of the completed digestive processes of heterophagy and autophagy. In between these extremes one can observe: primary/nascent lysosomes that have not yet been engaged yet in any proteolytic process; early autophagic vacuoles that might contain intracellular organelles; intermediate/late endosomes and phagocytic vacuoles (heterophagic vacuoles) that contain extracellular contents/particles; and multivesicular bodies (MVBs) which are the transition vacuoles between endosomes/phagocytic vacuoles and the digestive lysosomes.

The discovery of the lysosome along with independent experiments that were carried out at the same time and that have further strengthened the notion that cellular proteins are indeed in a constant state of synthesis and degradation (see, for example, Ref. 10), led scientists to feel, for the first time, that they have at hand an organelle that can potentially mediate degradation of intracellular proteins. The fact that the proteases were separated from their substrates by a membrane provided an explanation for controlled degradation, and the only problem left to be explained was how the substrates are translocated into the lysosomal lumen, exposed to the activity of the lysosomal proteases and degraded. An important discovery in this respect was the unravelling of the basic mechanism of action of the lysosome – autophagy (reviewed in Ref. 11). Under basal metabolic conditions, portions of the cytoplasm which contain the entire cohort of cellular proteins, are segregated within a membrane-bound compartment, and are then fused to a primary nascent lysosome and their contents digested. This process was denoted microautophagy. Under more extreme conditions, starvation for example, mitochondria, endoplasmic reticulum membranes, glycogen bodies and other cytoplasmic entities, can also be engulfed by a process called macroautophagy (see, for example, Ref. 12; the different modes of action of the lysosome in digesting extra- and intracellular proteins are shown in Figure 2, p. 372).

However, over a period of more than two decades, between the mid-1950s and the late 1970s, it has become gradually more and more difficult to explain several aspects of intracellular protein degradation based on the known mechanisms of lysosomal activity: accumulating lines of independent experimental evidence indicated that the degradation of at least certain classes of cellular proteins must be non-lysosomal. Yet, in the absence of any 'alternative', researchers came with different explanations, some more substantiated and others less, to defend the 'lysosomal' hypothesis.

First was the gradual discovery, coming from different laboratories, that different proteins vary in their stability and their half-life times can span three orders of magnitude, from a few minutes to many days. Thus, the $t_{1/2}$ of or-

nithine decarboxylase (ODC) is ~10 min, while that of glucose-6-phosphate dehydrogenase (G6PD) is 15 hours (for review articles, see, for example, Refs. 13,14). Also, rates of degradation of many proteins was shown to change with changing physiological conditions, such as availability of nutrients or hormones. It was conceptually difficult to reconcile the findings of distinct and changing half-lives of different proteins with the mechanism of action of the lysosome, where the microautophagic vesicle contains the entire cohort of cellular (cytosolic) proteins that are therefore expected to degrade at the same rate. Similarly, changing pathophysiological conditions, such as starvation or re-supplementation of nutrients, were expected to affect the stability of all cellular proteins to the same extent. Clearly, this was not the case.

Another source of concern about the lysosome as the organelle in which intracellular proteins are degraded were the findings that specific and general inhibitors of lysosomal proteases have different effects on different populations of proteins, making it clear that distinct classes of proteins are targeted by different proteolytic machineries. Thus, the degradation of endocytosed/pinocytosed extracellular proteins was significantly inhibited, a partial effect was observed on the degradation of long-lived cellular proteins, and almost no effect was observed on the degradation of short-lived and abnormal/mutated proteins.

Finally, the thermodynamically paradoxical observation that the degradation of cellular proteins requires metabolic energy, and more importantly, the emerging evidence that the proteolytic machinery uses the energy directly, were in contrast with the known mode of action of lysosomal proteases that under the appropriate acidic conditions, and similar to all known proteases, degrade proteins in an exergonic manner.

The assumption that the degradation of intracellular proteins is mediated by the lysosome was nevertheless logical. Proteolysis results from direct interaction between the target substrates and proteases, and therefore it was clear that active proteases cannot be free in the cytosol which would have resulted in destruction of the cell. Thus, it was recognized that any suggested proteolytic machinery that mediates degradation of intracellular protein degradation must also be equipped with a mechanism that separates – physically or virtually – between the proteases and their substrates, and enables them to associate only when needed. The lysosomal membrane provided this fencing mechanism. Obviously, nobody could have predicted that a new mode of post-translational modification – ubiquitination – could function as a proteolysis signal, and that untagged proteins will remain protected. Thus, while the structure of the lysosome could explain the separation necessary between the proteases and their substrates, and autophagy could explain the

mechanism of entry of cytosolic proteins into the lysosomal lumen, major problems have remained unsolved. Important among them were: (i) the varying half-lives, (ii) the energy requirement, and (iii) the distinct response of different populations of proteins to lysosomal inhibitors. Thus, according to one model, it was proposed that different proteins have different sensitivities to lysosomal proteases, and their half-lives *in vivo* correlate with their sensitivity to the action of lysosomal proteases *in vitro* (15). To explain an extremely long half-life for a protein that is nevertheless sensitive to lysosomal proteases, or alterations in the stability of a single protein under various physiological states, it was suggested that although all cellular proteins are engulfed into the lysosome, only the short-lived proteins are degraded, whereas the long-lived proteins exit back into the cytosol:

> To account for differences in half-life among cell components or of a single component in various physiological states, it was necessary to include in the model the possibility of an exit of native components back to the extralysosomal compartment (16).

According to a different model, selectivity is determined by the binding affinity of the different proteins for the lysosomal membrane which controls their entry rates into the lysosome, and subsequently their degradation rates (17). For a selected group of proteins, such as the gluconeogenetic enzymes phosphoenol-pyruvate carboxykinase (PEPCK) and fructose-1,6-biphosphatase, it was suggested, though not firmly substantiated, that their degradation in the yeast vacuole is regulated by glucose via a mechanism called 'catabolite inactivation' that possibly involves their phosphorylation. However this regulated mechanism for vacuolar degradation is limited only to a small and specific group of proteins (see for example Ref. 18; reviewed in Ref. 19). More recent studies have shown that at least for stress-induced macroautophagy, a general sequence of amino acids, KFFERQ, directs, via binding to a specific 'receptor' and along with cytosolic and lysosomal chaperones, the regulated entry of many cytosolic proteins into the lysosomal lumen. While further corroboration of this hypothesis is still required, it explains the mass entry of a large population of proteins that contain a homologous sequence, but not the targeting for degradation of a specific protein under defined conditions (reviewed in Refs. 20,21). The energy requirement for protein degradation was described as indirect, and necessary, for example, for protein transport across the lysosomal membrane (22) and/or for the activity of the H^+ pump and the maintenance of the low acidic intralysosomal pH that is necessary for optimal activity of the proteases (23). We now know that both mechanisms require energy. In the absence of any alternative, and with lysosomal degradation as the most logical explanation for targeting all known classes of pro-

teins at the time, Christian de Duve summarized his view on the subject in a review article published in the mid-1960s, saying: 'Just as extracellular digestion is successfully carried out by the concerted action of enzymes with limited individual capacities, so, we believe, is intracellular digestion' (24). The problem of different sensitivities of distinct protein groups to lysosomal inhibitors has remained unsolved, and may have served as an important trigger in future quest for a non-lysosomal proteolytic system.

Progress in identifying the elusive, non-lysosomal proteolytic system(s) was hampered by the lack of a cell-free preparation that could faithfully replicate the cellular proteolytic events – degrading proteins in a specific and energy-requiring mode. An important breakthrough was made by Rabinovitz and Fisher who found that rabbit reticulocytes degrade abnormal, amino acid analogue-containing haemoglobin (25). Their experiments modelled known disease states, the haemoglobinopathies. In these diseases abnormal mutated haemoglobin chains (such as sickle cell haemoglobin) or excess of unassembled normal haemoglobin chains (which are synthesized normally, but also excessively in thalassemias, diseases in which the pairing chain is not synthesized at all or is mutated and rapidly degraded, and consequently the bi-heterodimeric haemoglobin complex is not assembled) are rapidly degraded in the reticulocyte (26,27). Reticulocytes are terminally differentiating red blood cells that do not contain lysosomes. Therefore, it was postulated that the degradation of haemoglobin in these cells is mediated by a non-lysosomal machinery. Etlinger and Goldberg (28) were the first to isolate and characterize a cell-free proteolytic preparation from reticulocytes. The crude extract selectively degraded abnormal haemoglobin, required ATP hydrolysis, and acted optimally at a neutral pH, which further corroborated the assumption that the proteolytic activity was of a non-lysosomal origin. A similar system was isolated and characterized later by Hershko, Ciechanover, and their colleagues (29). Additional studies by this group led subsequently to resolution, characterization, and purification of the major enzymatic components from this extracts and to the discovery of the ubiquitin signalling system (see below).

The lysosome hypothesis is challenged

As mentioned above, the unravelled mechanism(s) of action of the lysosome could explain only partially, and at times not satisfactorily, several key emerging characteristics of intracellular protein degradation. Among them were the heterogeneous stability of individual proteins, the effect of nutrients and hormones on their degradation, and the dependence of intracel-

lular proteolysis on metabolic energy. The differential effect of selective inhibitors on the degradation of different classes of cellular proteins (see above but mostly below) could not be explained at all.

The evolution of methods to monitor protein kinetics in cells together with the development of specific and general lysosomal inhibitors has resulted in the identification of different classes of cellular proteins, long- and short-lived, and the discovery of the differential effects of the inhibitors on these groups (see, for example, Refs. 30,31). An elegant experiment in this respect was carried out by Brian Poole and his colleagues in the Rockefeller University. Poole was studying the effect of lysosomotropic agents, weak bases such as ammonium chloride and chloroquine, that accumulate in the lysosome and dissipate its low acidic pH. It was assumed that this mechanism underlies also the anti-malarial activity of chloroquine and similar drugs where they inhibit the activity parasite's lysosome, 'paralyzing' its ability to digest the host's haemoglobin during the intra-erythrocytic stage of its life cycle. Poole and his colleagues metabolically labelled endogenous proteins in living macrophages with ^{3}H-leucine and 'fed' them with dead macrophages that had been previously labelled with ^{14}C-leucine. They assumed, apparently correctly, that the dead macrophage debris and proteins will be phagocytosed by live macrophages and targeted to the lysosome for degradation. They monitored the effect of lysosomotropic agents on the degradation of these two protein populations; in particular, they studied the effect of the weak bases chloroquine and ammonium chloride (which enter the lysosome and neutralize the H^+ ions), and the acid ionophore X537A, which dissipates the H^+ gradient across the lysosomal membrane. They found that these drugs specifically inhibited the degradation of extracellular proteins, but not that of intracellular proteins (32). Poole summarized these experiments and explicitly predicted the existence of a non-lysosomal proteolytic system that degrades intracellular proteins:

> Some of the macrophages labeled with tritium were permitted to endocytise the dead macrophages labeled with 14C. The cells were then washed and replaced in fresh medium. In this way we were able to measure in the same cells the digestion of macrophage proteins from two sources. The exogenous proteins will be broken down in the lysosomes, while the endogenous proteins will be broken down wherever it is that endogenous proteins are broken down during protein turnover (33).

The requirement for metabolic energy for the degradation of both prokaryotic (34) and eukaryotic (10,35) proteins was difficult to understand. Proteolysis is an exergonic process and the thermodynamically paradoxical energy requirement for intracellular proteolysis made researchers believe that

energy cannot be consumed directly by proteases or the proteolytic process per se, and is used indirectly. As Simpson summarized his findings (10):

> The data can also be interpreted by postulating that the release of amino acids from protein is itself directly dependent on energy supply. A somewhat similar hypothesis, based on studies on autolysis in tissue minces, has recently been advanced, but the supporting data are very difficult to interpret. However, the fact that protein hydrolysis as catalyzed by the familiar proteases and peptidases occurs exergonically, together with the consideration that autolysis in excised organs or tissue minces continues for weeks, long after phosphorylation or oxidation ceased, renders improbable the hypothesis of the direct energy dependence of the reactions leading to protein breakdown.

Being cautious however, and probably unsure about this unequivocal conclusion, Simpson still left a narrow orifice opened for a proteolytic process that requires energy in a direct manner: 'However, the results do not exclude the existence of two (or more) mechanisms of protein breakdown, one hydrolytic, the other energy-requiring'. Since any proteolytic process must be at one point or another hydrolytic, the statement that makes a distinction between a hydrolytic process and an energy-requiring, yet non-hydrolytic one, is not clear. Judging the statement from an historical point of view and knowing the mechanism of action of the ubiquitin system, where energy is required also in the pre-hydrolytic step (ubiquitin conjugation), Simpson may have thought of a two-step mechanism, but did not give it a clear description. At the end of this clearly understandable and apparently difficult deliberation, he left us with a vague explanation linking protein degradation to protein synthesis, a process that was known to require metabolic energy:

> The fact that a supply of energy seems to be necessary for both the incorporation and the release of amino acids from protein might well mean that the two processes are interrelated. Additional data suggestive of such a view are available from other types of experiments. Early investigations on nitrogen balance by Benedict, Folin, Gamble, Smith, and others point to the fact that the rate of protein catabolism varies with the dietary protein level. Since the protein level of the diet would be expected to exert a direct influence on synthesis rather than breakdown, the altered catabolic rate could well be caused by a change in the rate of synthesis (10).

With the discovery of lysosomes in eukaryotic cells it could be argued that energy is required for the transport of substrates into the lysosome or for maintenance of the low intralysosomal pH for (see above), for example. The observation by Hershko and Tomkins that the activity of tyrosine amino-

transferase (TAT) was stabilized following depletion of ATP (35) indicated that energy may be required at an early stage of the proteolytic process, most probably before proteolysis occurs. Yet, it did not provide a clue as for the mechanism involved: energy could be used, for example, for specific modification of TAT, e.g. phosphorylation, that would sensitize it to degradation by the lysosome or by a yet unknown proteolytic mechanism, or for a modification that activates its putative protease. It could also be used for a more general lysosomal mechanism, one that involves transport of TAT into the lysosome, for example. The energy inhibitors inhibited almost completely degradation of the entire population of cell proteins, confirming previous studies (e.g. 10) and suggesting a general role for energy in protein catabolism. Yet, an interesting finding was that energy inhibitors had an effect that was distinct form that of protein synthesis inhibitors which affected only enhanced degradation (induced by steroid hormone depletion), but not basal degradation. This finding ruled out, at least partially, a tight linkage between protein synthesis and degradation. In bacteria, which lack lysosomes, an argument involving energy requirement for lysosomal degradation could not have been proposed, but other indirect effects of ATP hydrolysis could have affected proteolysis in *E. coli*, such as phosphorylation of substrates and/or proteolytic enzymes, or maintenance of the 'energized membrane state'. According to this model, proteins could become susceptible to proteolysis by changing their conformation, for example, following association with the cell membrane that maintains a local, energy-dependent gradient of a certain ion. While such an effect was ruled out (37), and since there was no evidence for a phosphorylation mechanism (although the proteolytic machinery in prokaryotes had not been identified at that time), it seemed that at least in bacteria, energy is required directly for the proteolytic process. In any event, the requirement for metabolic energy for protein degradation in both prokaryotes and eukaryotes, a process that is exergonic thermodynamically, strongly indicated that in cells proteolysis is highly regulated, and that a similar principle/mechanism has been preserved along evolution of the two kingdoms. Implying from the possible direct requirement for ATP in degradation of proteins in bacteria, it was not too unlikely to assume a similar direct mechanism in the degradation of cellular proteins in eukaryotes. Supporting this notion was the description of the cell-free proteolytic system in reticulocytes (28,29), a cell that lacks lysosomes, which indicated that energy is probably required directly for the proteolytic process, although here too, the underlying mechanisms had remained enigmatic at the time. Yet, the description of the cell-free system paved the road for detailed dissection of the underlying mechanisms involved.

The ubiquitin-proteasome system

The cell-free proteolytic system from reticulocytes (28,29) turned out to be an important and rich source for the purification and characterization of the enzymes that are involved in the ubiquitin-proteasome system. Initial fractionation of the crude reticulocyte cell extract on the anion-exchange resin diethylaminoethyl cellulose yielded two fractions which were both required to reconstitute the energy-dependent proteolytic activity that is found in the crude extract: The unabsorbed, flow through material was denoted fraction I, and the high salt eluate of the adsorbed proteins which was denoted fraction II (Table 1; 38).

Table 1. ***Resolution of the ATP-dependent proteolytic activity from crude reticulocyte extract into two essentially required complementing activities*** (adapted from Ref. 38; with permission from Elsevier/Biochem. Biophys. Res. Commun.).

Fraction	Degradation of [^{3}H]globin (%)	
	-ATP	+ATP
Lysate	**1.5**	**10**
Fraction I	**0.0**	**0.0**
Fraction II	**1.5**	**2.7**
Fraction I and Fraction II	**1.6**	**10.6**

This was an important observation and a lesson for the future dissection of the system. For one it suggested that the system is not composed of a single 'classical' protease that has evolved evolutionarily to acquire energy dependence [although such energy-dependent proteases, the mammalian 26S proteasome (see below) and the prokaryotic *Lon* gene product have been described later], but that it is made of at least two components. This finding of a two-component, energy-dependent protease, left the researchers with no paradigm to follow, and in attempts to explain the finding, they suggested, for example, that the two fractions could represent an inhibited protease and its activator. Second, learning from this reconstitution experiment and the essential dependence between the two active components, we continued to reconstitute activity from resolved fractions whenever we encountered a loss of activity along further purification steps. This biochemical 'complementation' approach resulted in the discovery of additional enzymes of the system, all required to be present in the reaction mixture in order to catalyze the multi-step proteolysis of the target substrate. We chose first to purify the active

component from fraction I. It was found to be a small, ~8.5 kDa heat-stable protein that was designated ATP-dependent Proteolysis Factor 1, APF-1. APF-1 was later identified as ubiquitin (see below; I am using the term APF-1 to the point in which it was identified as ubiquitin and then change terminology accordingly). In retrospect, the decision to start the purification efforts with fraction I turned out to be important, as fraction I contained only one single protein – APF-1 – that was necessary to stimulate proteolysis of the model substrate we used at the time, while fraction II turned out to contain many more. Later studies showed that fraction I contains other components necessary for the degradation of other substrates, but these were not necessary for the reconstitution of the system at that time. This enabled us not only to purify APF-1, but also to quickly decipher its mode of action. If we had started our purification efforts with fraction II, we would have encountered a significantly bumpier road. A critically important finding that paved the way for future developments in the field was that multiple moieties of APF-1 are covalently conjugated to the target substrate when incubated in the presence of fraction II, and the modification requires ATP (39,40; Figures 3 and 4). It was also found that the modification is reversible, and APF-1 can be removed from the substrate or its degradation products (40).

The discovery that APF-1 is covalently conjugated to protein substrates and stimulates their proteolysis in the presence of ATP and crude fraction II, led in 1980 to the proposal of a model according to which protein substrate modification by multiple moieties of APF-1 targets it for degradation by a downstream, at that time a yet unidentified protease that cannot recognize the unmodified substrate; following degradation, reusable APF-1 is released (40). Amino-acid analysis of APF-1, along with its known molecular mass and other general characteristics, raised the suspicion that APF-1 is ubiquitin (41), a known protein of previously unknown function. Indeed, Wilkinson and colleagues confirmed unequivocally that APF-1 is indeed ubiquitin (42). Ubiquitin is a small, heat-stable and highly evolutionarily conserved protein of 76 residues. It was first purified during the isolation of thymopoietin (43) and was subsequently found to be ubiquitously expressed in all kingdoms of living cells, including prokaryotes (44). Interestingly, it was initially found to have lymphocyte-differentiating properties, a characteristic that was attributed to the stimulation of adenylate cyclase (44,45). Accordingly, it was named UBIP for *ub*iquitous *i*mmunopoietic *p*olypeptide (44). However, later studies showed that ubiquitin is not involved in the immune response (46), and that it was a contaminating endotoxin in the preparation that generated the adenylate cyclase and the T-cell differentiating activities. Furthermore, the sequence of several eubacteria and archaebacteria genomes as well as biochemical analyses in

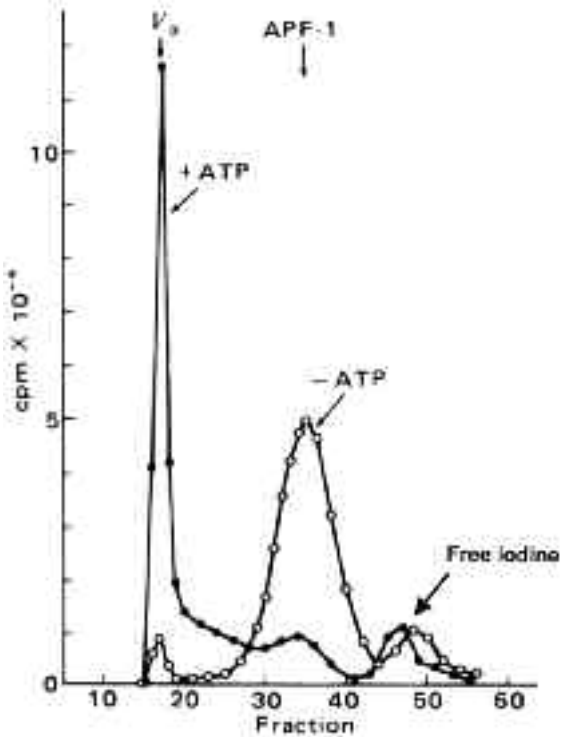


Figure 3: ***APF-1/Ubiquitin is shifted to high molecular mass compound(s) following incubation in ATP-containing crude cell extract.*** ^{125}I-labelled APF-1/ ubiquitin was incubated with reticulocyte crude Fraction II in the absence (open circles) or presence (closed circles) of ATP, and the reaction mixtures were resolved via gel filtration chromatography. Shown is the radioactivity measured in each fraction. As can be seen, following addition of ATP, APF-1/ubiquitin becomes covalently attached to some component(s) in fraction II, which could be another enzyme of the system or its substrate(s) (with permission from Proceedings of the National Academy of the USA; published originally in Ref. 39).

these organisms (unpublished) showed that ubiquitin is restricted only to eukaryotes. The finding of ubiquitin in bacteria (44) was probably due to contamination of the bacterial extract with yeast ubiquitin derived from the yeast extract in which the bacteria were grown. While in retrospect the name ubiquitin is a misnomer, as it is restricted to eukaryotes and is not ubiquitous as was previously thought, for historical reasons it has still remained the name of the protein. Accordingly, and in order to avoid confusion, I suggest that the names of other novel enzymes and components of the ubiquitin system, but of other systems as well, should remain as were first coined by their discoverers.

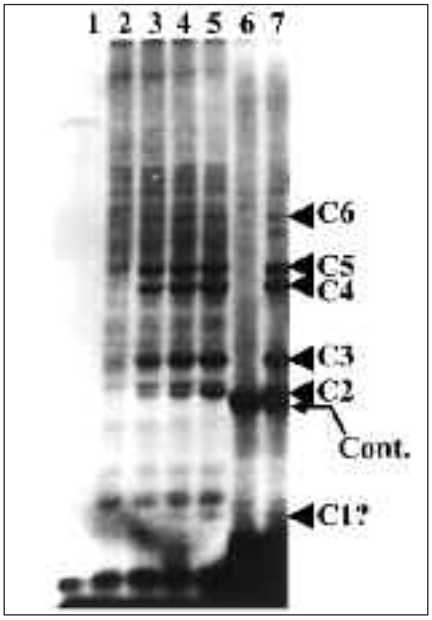


Figure 4: ***Multiple molecules of APF-1/Ubiquitin are conjugated to the proteolytic substrate, probably signalling it for degradation.*** To interpret the data described in the experiment depicted in Figure 2 and to test the hypothesis that APF-1 is conjugated to the target proteolytic substrate, ^{125}I-APF-1/ubiquitin was incubated along with crude Fraction II (Figure 3 and text) in the absence (lane 1) or presence (lanes 2-5) of ATP and in the absence (lanes 1,2) or presence (lanes 3-5) of increasing concentrations of unlabelled lysozyme. Reaction mixtures resolved in lanes 6 and 7 were incubated in the absence (lane 6) or presence (lane 7) of ATP, and included unlabelled APF-1/ubiquitin and ^{125}I-labelled lysozyme. C1-C6 denote specific APF-1/ubiquitin-lysozyme adducts in which the number of APF-1/ubiquitin moieties bound to the lysozyme moiety of the adduct is increasing, probably from 1 to 6. Reactions mixtures were resolved via sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and visualized following exposure to an X-ray film (autoradiography) (with permission from Proceedings of the National Academy of the USA; published originally in Ref. 40).

An important development in the ubiquitin research field was the discovery that a single ubiquitin moiety can be covalently conjugated to histones, particularly to histones H2A and H2B. While the function of these adducts has remained elusive until recently, their structure was unravelled in the mid 1970s. The structure of the ubiquitin conjugate of H2A (uH2A; was also designated protein A24) was deciphered by Goldknopf and Busch (47,48) and by Hunt and Dayhoff (49) who found that the two proteins are linked through a fork-like, branched isopeptide bond between the carboxy-

terminal glycine of ubiquitin (Gly^{76}) and the ε-NH_2 group of an internal lysine (Lys^{119}) of the histone molecule. The isopeptide bond found in the histone-ubiquitin adduct was suggested to be identical to the bond that was found between ubiquitin and the target proteolytic substrate (50) and between the ubiquitin moieties in the polyubiquitin chain (51,52) that is synthesized on the substrate and that functions as a proteolysis recognition signal for the downstream 26S proteasome. In this particular polyubiquitin chain the linkage is between Gly^{76} of one ubiquitin moiety and internal Lys^{48} of the previously conjugated moiety. Only Lys^{48}-based ubiquitin chains are recognized by the 26S proteasome and serve as proteolytic signals. In recent years it has been shown that the first ubiquitin moiety can also be attached in a linear mode to the N-terminal residue of the proteolytic target substrate (53). However, the subsequent ubiquitin moieties are generating Lys^{48}-based polyubiquitin chain on the first linearly fused moiety. N-terminal ubiquitination is clearly required for targeting naturally occurring lysine-less proteins for degradation. Yet, several lysine-containing proteins have also been described that traverse this pathway, the muscle-specific transcription factor MyoD for example. In these proteins the internal lysine residues are probably not accessible to the cognate ligases. Other types of polyubiquitin chains have also been described that are not involved in targeting the conjugated substrates for proteolysis. Thus, a Lys^{63}-based polyubiquitin chain has been described that is probably necessary to activate transcription factors (reviewed recently in Ref. 54). Interestingly, the role of monoubiquitination of histones has also been identified recently and this modification is also involved in regulation of transcription, probably via modulation of the structure of the nucleosomes (for recent reviews, see, for example, Refs. 55,56).

The identification of APF-1 as ubiquitin, and the discovery that a high-energy isopeptide bond, similar to the one that links ubiquitin to histone H2A, links it also to the target proteolytic substrate, resolved at that time the enigma of the energy requirement for intracellular proteolysis (see below, however) and paved the road to the untangling of the complex mechanism of isopeptide bond formation. This process turned out to be similar to that of peptide bond formation that is catalysed by tRNA synthetase following amino acid activation during protein synthesis or during the non-ribosomal synthesis of short peptides (57). Using the unravelled mechanism of ubiquitin activation and immobilized ubiquitin as a 'covalent' affinity bait, the three enzymes that are involved in the cascade reaction of ubiquitin conjugation were purified by Ciechanover, Hershko, and their colleagues. These enzymes are: (i) E1, the ubiquitin-activating enzyme, (ii) E2, the ubiquitin-carrier protein, and (iii) E3, the ubiquitin-protein ligase (58,59). The discovery of an E3 which

is a specific substrate-binding component, indicated a possible solution to the problem of the varying stabilities of different proteins – they might be specifically recognized and targeted by different ligases.

In a short period, the ubiquitin tagging hypothesis received substantial support. For example, Chin and colleagues injected into HeLa cells labelled ubiquitin and haemoglobin and denatured the injected haemoglobin by oxidizing it with phenylhydrazine. They found that ubiquitin conjugation to globin is markedly enhanced by denaturation of haemoglobin and the concentration of globin-ubiquitin conjugates was proportional to the rate of haemoglobin degradation (60). Hershko and colleagues observed a similar correlation for abnormal, amino acid analogue-containing short-lived proteins (61). A previously isolated cell cycle arrest mutant that loses the ubiquitin-histone H2A adduct at the permissive temperature (62) was found by Finley, Ciechanover and Varshavsky to harbour a thermolabile E1 (63). Following heat inactivation, the cells fail to degrade normal short-lived proteins (64). Although the cells did not provide direct evidence for substrate ubiquitination as a destruction signal, they still provided the strongest direct linkage between ubiquitin conjugation and degradation.

At this point, the only missing link was the identification of the downstream protease that would specifically recognize ubiquitinated substrates. Tanaka and colleagues identified a second ATP-requiring step in the reticulocyte proteolytic system, which occurred after ubiquitin conjugation (65), and Hershko and colleagues demonstrated that the energy is required for conjugate degradation (66). An important advance in the field was a discovery by Hough and colleagues, who partially purified and characterized a high-molecular mass alkaline protease that degraded ubiquitin adducts of lysozyme but not untagged lysozyme, in an ATP-dependent mode (67). This protease which was later called the 26S proteasome (see below), provided all the necessary criteria for being the specific proteolytic arm of the ubiquitin system. This finding was confirmed, and the protease was further characterized by Waxman and colleagues who found that it is an unusually large, ~1,5 MDa enzyme, unlike any other known protease (68). A further advance in the field was the discovery (69) that a smaller neutral multi-subunit 20S protease complex that was discovered together with the larger 26S complex, is similar to a 'multicatalytic proteinase complex' (MCP) that was described earlier in the bovine pituitary gland by Wilk and Orlowski (70). This 20S protease is ATP-independent and has different catalytic activities, cleaving on the carboxy-terminal side of hydrophobic, basic and acidic residues. Hough and colleagues raised the possibility – although they did not show it experimentally – that this 20S protease can be a part of the larger 26S pro-

tease that degrades the ubiquitin adducts (69). Later studies showed that indeed, the 20S complex is the core catalytic particle of the larger 26S complex (71,72). However, a strong evidence that the active 'mushroom'-shaped 26S protease is generated through the assembly of two distinct sub-complexes – the catalytic 20S cylinder-like MCP and an additional 19S ball-shaped subcomplex (that was predicted to have a regulatory role) – was provided only in the early 1990s by Hoffman and colleagues (73) who mixed the two purified particles and generated the active 26S enzyme.

The proteasome is a large, 26S, multicatalytic protease that degrades polyubiquitinated proteins to small peptides. It is composed of two subcomplexes: a 20S core particle (CP) that carries the catalytic activity, and a regulatory 19S regulatory particle (RP). The 20S CP is a barrel-shaped structure composed of four stacked rings, two identical outer α rings and two identical inner β rings. The eukaryotic α and β rings are composed each of seven distinct subunits, giving the 20S complex the general structure of $\alpha_{1\text{-}7}\beta_{1\text{-}7}\beta_{1\text{-}7}\alpha_{1\text{-}7}$. The catalytic sites are localized to some of the β subunits. Each extremity of the 20S barrel can be capped by a 19S RP each composed of 17 distinct subunits, 9 in a 'base' sub-complex, and 8 in a 'lid' sub-complex. One important function of the 19S RP is to recognize ubiquitinated proteins and other potential substrates of the proteasome. Several ubiquitin-binding subunits of the 19S RP have been identified, although their biological roles and mode of action have not been discerned. A second function of the 19S RP is to open an orifice in the a ring that will allow entry of the substrate into the proteolytic chamber. Also, since a folded protein would not be able to fit through the narrow proteasomal channel, it is assumed that the 19S particle unfolds substrates and inserts them into the 20S CP. Both the channel opening function and the unfolding of the substrate require metabolic energy, and indeed, the 19S RP 'base' contains six different ATPase subunits. Following degradation of the substrate, short peptides derived from the substrate are released, as well as reusable ubiquitin (for a scheme describing the ubiquitin system, see Figure 5, p. 373; for the structure of the 26S proteasome, see Figure 6, p. 374).

Concluding remarks

The evolution of proteolysis as a centrally important regulatory mechanism is a remarkable example for the evolution of a novel biological concept and the accompanying battles to change paradigms. The five-decade journey between the early 1940s and early 1990s began with fierce discussions on whether cellular proteins are static as has been thought for a long

time, or are turning over. The discovery of the dynamic state of proteins was followed by the discovery of the lysosome that was believed – between the mid-1950s and mid-1970s – to be the organelle within which intracellular proteins are destroyed. Independent lines of experimental evidence gradually eroded the lysosomal hypothesis and resulted in a new idea that the bulk of intracellular proteins are degraded – under basal metabolic conditions – via a non-lysosomal machinery. This resulted in the discovery of the ubiquitin system in the late 1970s and early 1980s.

With the identification of the reactions and enzymes that are involved in the ubiquitin-proteasome cascade, a new era in the protein degradation field began at the late 1980s and early 1990s. Studies that showed that the system is involved in targeting of key regulatory proteins – such as light-regulated proteins in plants, transcriptional factors, cell cycle regulators and tumour suppressors and promoters – started to emerge (see for example Refs. 74-78). They were followed by numerous studies on the underlying mechanisms involved in the degradation of specific proteins, each with its own unique mode of recognition and regulation. The unravelling of the human genome revealed the existence of hundreds of distinct E3s, attesting to the complexity and the high specificity and selectivity of the system. Two important advances in the field were the discovery of the non-proteolytic functions of ubiquitin such as activation of transcription and routing of proteins to the vacuole, and the discovery of modification by ubiquitin-like proteins (UBLs), that are also involved in numerous non-proteolytic functions such as directing proteins to their sub-cellular destination, protecting proteins from ubiquitination, or controlling entire processes such as autophagy (see for example Ref. 79)(for the different roles of modifications by ubiquitin and UBLs, see Figure 7, p. 375). All these studies have led to the emerging realization that this novel mode of covalent conjugation plays a key role in regulating a broad array of cellular process – among them cell cycle and division, growth and differentiation, activation and silencing of transcription, apoptosis, the immune and inflammatory response, signal transduction, receptor mediated endocytosis, various metabolic pathways, and the cell quality control – through proteolytic and non-proteolytic mechanisms. The discovery that ubiquitin modification plays a role in routing proteins to the lysosome/vacuole and that modification by specific and unique ubiquitin-like proteins and modification system controls autophagy closed an exciting historical cycle, since it demonstrated that the two apparently distinct systems communicate with one another. With the many processes and substrates targeted by the ubiquitin pathway, it is not surprising to find that aberrations in the system underlie, directly or indirectly, the

pathogenesis of many diseases. While inactivation of a major enzyme such as E1 is obviously lethal, mutations in enzymes or in recognition motifs in substrates that do not affect vital pathways or that affect the involved process only partially, may result in a broad array of phenotypes. Likewise, acquired changes in the activity of the system can also evolve into certain pathologies. The pathological states associated with the ubiquitin system can be classified into two groups: (a) those that result from loss of function – mutation in a ubiquitin system enzyme or in the recognition motif in the target substrate that result in stabilization of certain proteins, and (b) those that result from gain of function – abnormal or accelerated degradation of the protein target (for aberrations in the ubiquitin system that result in disease states, see Figure 8, p. 376). Studies that employ targeted inactivation of genes coding for specific ubiquitin system enzymes and substrates in animals can provide a more systematic view into the broad spectrum of pathologies that may result from aberrations in ubiquitin-mediated proteolysis. Better understanding of the processes and identification of the components involved in the degradation of key regulatory proteins will lead to the development of mechanism-based drugs that will target specifically only the involved proteins. While the first drug, a specific proteasome inhibitor is already on the market (80), it appears that one important hallmark of the new era we are entering now will be the discovery of novel drugs based on targeting of specific processes such as inhibiting aberrant Mdm2- or E6-AP-mediated accelerated targeting of the tumour suppressor p53 which will lead to regain of its lost function.

Many reviews have been published on different aspects of the ubiquitin system. The purpose of this article was to bring to the reader several milestones along the historical pathway along which the ubiquitin system has been evolved. For additional reading on the ubiquitin system the reader is referred to the many reviews written on the system, among them for example are Refs. 81,82. Some parts of this review, including several Figures, are based on another recently published review article (Ref. 83).

Acknowledgement

Research in the laboratory of Aaron Ciechanover has been supported along the years by grants from the US-Israel Binational Science Foundation (BSF), the Israel Science Foundation (ISF) founded by the Israeli National Academy of Humanities, Arts and Sciences, the German-Israeli Foundation (GIF) for Scientific Research and Development, the Israel Cancer Research Fund (ICRF) USA, the Deutsche-Israeli Cooperation Program (DIP), the

European Union (EU), the Israel Cancer Society (ICS), the Prostate Cancer Foundation (PCF) – Israel, the Foundation for Promotion of Research in the Technion and various research grants administered by the Vice President of the Technion for Research. Infrastructural equipment for the laboratory of Aaron Ciechanover and for the Cancer and Vascular Biology Research Center has been purchased with the support of the Wolfson Charitable Fund – Center of Excellence for Studies on *Turnover of Cellular Proteins and its Implications to Human Diseases*. Aaron Ciechanover is an Israel Cancer Research Fund (ICRF) USA Professor. This lecture is the written version of the Nobel Lecture delivered by Aaron Ciechanover on December 8, 2004, and is published with permission of the Nobel Foundation.

References

1. Clarke, H.T. (1958). Impressions of an organic chemist in biochemistry. Annu. Rev. *Biochem.* 27, 1-14.
2. Kennedy, E.P. (2001). Hitler's gift and the era of biosynthesis. *J. Biol. Chem.* 276, 42619-42631.
3. Simoni, R.D., Hill, R.L., and Vaughan, M. (2002). The use of isotope tracers to study intermediary metabolism: Rudolf Schoenheimer. *J. Biol. Chem.* 277 (issue 43), e1-e3 (available online at www.jbc.org).
4. Schoenheimer, R., Ratner, S., and Rittenberg, D. (1939). Studies in protein metabolism: VII. The metabolism of tyrosine. *J. Biol. Chem.* 127, 333-344.
5. Ratner, S., Rittenberg, D., Keston, A.S., and Schoenheimer, R. (1940). Studies in protein metabolism: XIV. The chemical interaction of dietary Glycine and body proteins in rats. *J. Biol. Chem.* 134, 665-676.
6. Schoenheimer, R. *The Dynamic State of Body Constituents* (1942). Harvard University Press, Cambridge, Massachusetts, USA.
7. Hogness, D.S., Cohn, M., and Monod, J. (1955). Studies on the induced synthesis of β-galactosidase in *Escherichia coli*: The kinetics and mechanism of sulfur incorporation. *Biochim. Biophys. Acta* 16, 99-116.
8. de Duve, C., Gianetto, R., Appelmans, F., and Wattiaux, R. (1953). Enzymic content of the mitochondria fraction. *Nature* (London) 172, 1143-1144.
9. Gianetto, R., and de Duve, C. Tissue fractionation studies 4. (1955). Comparative study of the binding of acid phosphatase, β-glucoronidase and cathepsin by rat liver particles. *Biochem. J.* 59, 433-438.
10. Simpson, M.V. The release of labeled amino acids from proteins in liver slices. (1953). *J. Biol. Chem.* 201, 143-154.
11. Mortimore, G.E., and Poso, A.R. (1987). Intracellular protein catabolism and its control during nutrient deprivation and supply. *Annu. Rev. Nutr.* 7, 539-564.
12. Ashford, T.P., and Porter, K.R. (1962). Cytoplasmic components in hepatic cell lysosomes. *J. Cell Biol.* 12, 198-202.
13. Schimke, R.T., and Doyle, D. (1970). Control of enzyme levels in animal tissues. *Annual Rev. Biochem.* 39, 929-976.
14. Goldberg, A.L., and St. John, A.C. (1976). Intracellular protein degradation in mammalian and bacterial cells: Part 2. *Annu. Rev. Biochem.* 45, 747-803.

15. Segal, H.L., Winkler, J.R., and Miyagi, M.P. (1974). Relationship between degradation rates of proteins *in vivo* and their susceptibility to lysosomal proteases. *J. Biol. Chem.* 249, 6364-6365.
16. Haider, M., and Segal, H.L. (1972). Some characteristics of the alanine-aminotransferase and arginase-inactivating system of lysosomes. *Arch. Biochem. Biophys.* 148, 228-237.
17. Dean, R.T. Lysosomes and protein degradation. (1977). *Acta Biol. Med. Ger.* 36, 1815-1820.
18. Müller, M., Müller, H, and Holzer, H. (1981). Immunochemical studies on catabolite inactivation of phosphoenolpyruvate carboxykinase in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 256, 723-727.
19. Holzer, H. (1989). Proteolytic catabolite inactivation in *Saccharomyces cerevisiae*. *Revis. Biol. Celular* 21, 305-319.
20. Majeski, A.E., and Dice, J.F. (2004). Mechanisms of chaperone-mediated autophagy. *Intl. J. Biochem. Cell Biol.* 36, 2435-2444.
21. Cuervo, A.M., and Dice, J.F. (1998). Lysosomes, a meeting point of proteins, chaperones, and proteases. *J. Mol. Med.* 76, 6-12.
22. Hayashi, M., Hiroi, Y., and Natori, Y. (1973). Effect of ATP on protein degradation in rat liver lysosomes. *Nature New Biol.* 242, 163-166.
23. Schneider, D.L. (1981). ATP-dependent acidification of intact and disrupted lysosomes: Evidence for an ATP-driven proton pump. *J. Biol. Chem.* 256, 3858-3864.
24. de Duve, C., and Wattiaux, R. Functions of lysosomes. (1966). *Annu. Rev. Physiol.* 28, 435-492.
25. Rabinovitz, M., and Fisher, J.M. (1964). Characteristics of the inhibition of hemoglobin synthesis in rabbit reticulocytes by threo-α-amino-β-chlorobutyric acid. *Biochim. Biophys. Acta.* 91, 313-322.
26. Carrell, R.W., and Lehmann, H. (1969). The unstable haemoglobin haemolytic anaemias. *Semin. Hematol.* 6, 116-132.
27. Huehns, E.R., and Bellingham, A.J. (1969). Diseases of function and stability of haemoglobin. *Br. J. Haematol.* 17, 1-10.
28. Etlinger, J.D., and Goldberg, A.L. (1977). A soluble ATP-dependent proteolytic system responsible for the degradation of abnormal proteins in reticulocytes. *Proc. Natl. Acad. Sci. USA* 74, 54-58.
29. Hershko, A., Heller, H., Ganoth, D., and Ciechanover, A. (1978). Mode of degradation of abnormal globin chains in rabbit reticulocytes. In: Protein Turnover and Lysosome Function (H.L. Segal & D.J. Doyle, eds). Academic Press, New York. pp. 149-169.
30. Knowles, S.E., and Ballard, F.J. (1976). Selective control of the degradation of normal and aberrant proteins in Reuber H35 hepatoma cells. *Biochem J.* 156, 609-617.
31. Neff, N.T., DeMartino, G.N., and Goldberg, A.L. (1979). The effect of protease inhibitors and decreased temperature on the degradation of different classes of proteins in cultured hepatocytes. *J. Cell Physiol.* 101, 439-457.
32. Poole, B., Ohkuma, S., and Warburton, M.J. (1977). The accumulation of weakly basic substances in lysosomes and the inhibition of intracellular protein degradation. *Acta Biol. Med. Germ.* 36, 1777-1788.
33. Poole, B., Ohkuma, S. & Warburton, M.J. (1978). Some aspects of the intracellular breakdown of exogenous and endogenous proteins. In: *Protein Turnover and Lysosome Function* (H.L. Segal & D.J. Doyle, eds). Academic Press, New York. pp. 43-58.
34. Mandelstam, J. (1958). Turnover of protein in growing and non-growing populations of *Escherichia coli*. *Biochem. J.* 69, 110-119.

35. Steinberg, D., and Vaughan, M. (1956). Observations on intracellular protein catabolism studied *in vitro*. *Arch. Biochem. Biophys*. 65, 93-105.
36. Hershko, A., and Tomkins, G.M. (1971). Studies on the degradation of tyrosine aminotransferase in hepatoma cells in culture: Influence of the composition of the medium and adenosine triphosphate dependence. *J. Biol. Chem.* 246, 710-714.
37. Goldberg, A.L., Kowit, J.D., and Etlinger, J.D. (1976). Studies on the selectivity and mechanisms of intracellular protein degradation. In: Proteolysis and physiological regulation (D.W. Ribbons & K. Brew, eds.). Academic Press, New York. pp. 313-337.
38. Ciechanover A., Hod, Y., and Hershko, A. (1978). A heat-stable polypeptide component of an ATP-dependent proteolytic system from reticulocytes. *Biochem. Biophys. Res. Common*. 81, 1100-1105.
39. Ciechanover, A., Heller, H., Elias, S., Haas, A.L., and Hershko, A. (1980). ATP-dependent conjugation of reticulocyte proteins with the polypeptide required for protein degradation. *Proc. Natl. Acad. Sci. USA*. 77, 1365-1368 (1980).
40. Hershko, A., Ciechanover, A., Heller, H., Haas, A.L., and Rose, I.A. (1980). Proposed role of ATP in protein breakdown: Conjugation of proteins with multiple chains of the polypeptide of ATP-dependent proteolysis. *Proc. Natl. Acad. Sci. USA* 77, 1783-1786.
41. Ciechanover, A., Elias, S., Heller, H., Ferber, S. & Hershko, A. (1980). Characterization of the heat-stable polypeptide of the ATP-dependent proteolytic system from reticulocytes. *J. Biol. Chem.* 255, 7525-7528 (1980).
42. Wilkinson, K.D., Urban, M.K., and Haas, A.L. (1980). Ubiquitin is the ATP-dependent proteolysis factor I of rabbit reticulocytes. *J. Biol. Chem*. 255, 7529-7532.
43. Goldstein, G. (1974). Isolation of bovine thymin, a polypeptide hormone of the thymus. *Nature (London)* 247, 11-14.
44. Goldstein, G., Scheid, M., Hammerling, U., Schlesinger, D.H., Niall, H.D.& Boyse, E.A. (1975). Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells. *Proc. Natl. Acad. Sci. USA* 72, 11-15.
45. Schlessinger, D.H., Goldstein, G., and Niall, H.D. (1975). The complete amino acid sequence of ubiquitin, an adenylate cyclase stimulating polypeptide probably universal in living cells. *Biochemistry* 14, 2214-2218.
46. Low, T.L.K., and Goldstein, A.L. (1979). The chemistry and biology of thymosin: amino acid analysis of thymosin α1 and polypeptide β1. *J. Biol. Chem.* 254, 987-995.
47. Goldknopf, I.L., and Busch, H. (1975). Remarkable similarities of peptide fingerprints of histone 2A and non-histone chromosomal protein A24. *Biochem. Biophys. Res. Commun*. 65, 951-955.
48. Goldknopf, I.L., and Busch, H. (1977). Isopeptide linkage between non-histone and histone 2A polypeptides of chromosome conjugate-protein A24. *Proc. Natl. Acad. Sci. USA* 74, 864-868.
49. Hunt, L.T., and Dayhoff, M.O. (1977). Amino-terminal sequence identity of ubiquitin and the non-histone component of nuclear protein A24. *Biochim. Biophys. Res. Commun*. 74, 650-655.
50. Hershko, A., Ciechanover, A., and Rose, I.A. (1981). Identification of the active amino acid residue of the polypeptide of ATP-dependent protein breakdown. *J. Biol. Chem*. 256, 1525-1528.

51. Hershko, A., and Heller, H. (1985). Occurrence of a polyubiquitin structure in ubiquitin-protein conjugates. *Biochem. Biophys. Res. Common.* 128, 1079-1086.
52. Chau, V., Tobias, J.W., Bachmair, A., Mariott, D., Ecker, D., Gonda, D.K., and Varshavsky, A. (1989). A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. *Science* 243, 1576-1583.
53. Ciechanover, A., and Ben-Saadon R. (2004). N-terminal ubiquitination: More Protein substrates join in. *Trends Cell Biol.* 14, 103-106.
54. Muratani, M., and Tansey, W.P. (2003). How the ubiquitin-proteasome system controls transcription. *Nat. Rev. Mol. Cell Biol.* 4, 192-201.
55. Zhang, Y. (2003). Transcriptional regulation by histone ubiquitination and deubiquitination. *Genes & Dev.* 17, 2733-2740.
56. Osley, M.A. (2004). H2B ubiquitylation: the end is in sight. *Biochim. Biophys. Acta.* 1677, 74-78.
57. Lipman, F. (1971). Attempts to map a process evolution of peptide biosynthesis. *Science* 173, 875-884.
58. Ciechanover, A., Elias, S., Heller, H. & Hershko, A. (1982). 'Covalent affinity' purification of ubiquitin-activating enzyme. *J. Biol. Chem.* 257, 2537-2542.
59. Hershko, A., Heller, H., Elias, S., and Ciechanover, A. (1983). Components of ubiquitin-protein ligase system: Resolution, affinity purification and role in protein breakdown. *J. Biol. Chem.* 258, 8206-8214 (1983).
60. Chin, D.T., Kuehl, L., and Rechsteiner, M. (1982). Conjugation of ubiquitin to denatured hemoglobin is proportional to the rate of hemoglobin degradation in HeLa cells. *Proc. Natl. Acad. Sci. USA* 79, 5857-5861.
61. Hershko, A., Eytan, E., Ciechanover, A. and Haas, A.L. (1982). Immunochemical Analysis of the Turnover of Ubiquitin-protein Conjugates in Intact Cells: Relationship to the Breakdown of Abnormal Proteins. *J. Biol. Chem.* 257, 13964-13970.
62. Matsumoto, Y., Yasuda, H., Marunouchi, T., and Yamada, M. (1983). Decrease in uH2A (protein A24) of a mouse temperature-sensitive mutant. *FEBS Lett.* 151, 139-142.
63. Finley, D., Ciechanover, A., and Varshavsky, A. (1984). Thermolability of ubiquitin-activating enzyme from the mammalian cell cycle mutant ts85. *Cell* 37, 43-55.
64. Ciechanover, A., Finley D., and Varshavsky, A. (1984). Ubiquitin dependence of selective protein degradation demonstrated in the mammalian cell cycle mutant ts85. *Cell* 37, 57-66.
65. Tanaka, K., Waxman, L., and Goldberg, A.L. (1983). ATP serves two distinct roles in protein degradation in reticulocytes, one requiring and one independent of ATP. *J. Cell Biol.* 96, 1580-1585 (1983).
66. Hershko, A., Leshinsky, E., Ganoth, D. & Heller, H. (1984). ATP-dependent degradation of ubiquitin-protein conjugates. *Proc. Natl. Acad. Sci. USA* 81, 1619- 1623.
67. Hough, R., Pratt, G. & Rechsteiner, M. (1986). Ubiquitin-lysozyme conjugates. Identification and characterization of an ATP-dependent protease from rabbit reticulocyte lysates. *J. Biol. Chem.* 261, 2400-2408.
68. Waxman, L., Fagan, J., and Goldberg, A.L. (1987). Demonstration of two distinct high molecular weight proteases in rabbit reticulocytes, one of which degrades ubiquitin conjugates. *J. Biol. Chem.* 262, 2451-2457.
69. Hough, R., Pratt, G., and Rechsteiner M. (1987). Purification of two high molecular weight proteases from rabbit reticulocyte lysate. *J. Biol. Chem.* 262, 8303-8313.

70. Wilk, S., and Orlowski, M. (1980). Cation-sensitive neutral endopeptidase: isolation and specificity of the bovine pituitary enzyme. *J. Neurochem.* 35, 1172-1182.
71. Eytan, E., Ganoth, D., Armon, T., and Hershko, A. (1989). ATP-dependent incorporation of 20S protease into the 26S complex that degrades proteins conjugated to ubiquitin. *Proc. Natl. Acad. Sci. USA.* 86, 7751-7755.
72. Driscoll, J., and Goldberg, A.L. (1990). The proteasome (multicatalytic protease) is a component of the 1500-kDa proteolytic complex which degrades ubiquitin-conjugated proteins. *J. Biol. Chem.* 265, 4789-4792.
73. Hoffman, L., Pratt, G., and Rechsteiner, M. (1992). Multiple forms of the 20S multicatalytic and the 26S ubiquitin/ATP-dependent proteases from rabbit reticulocyte lysate. *J. Biol. Chem.* 267, 22362-22368.
74. Shanklin, J., Jaben, M., and Vierstra, R.D. (1987). Red light-induced formation of ubiquitin-phytochrome conjugates: Identification of possible intermediates of phytochrome degradation. *Proc. Natl. Acad. Sci. USA* 84, 359-363.
75. Hochstrasser, M., and Varshavsky, A. (1990). *In vivo* degradation of a transcriptional regulator: the yeast α2 repressor. *Cell* 61, 697-708.
76. Scheffner, M., Werness, B.A., Huibregtse, J.M., Levine, A.J., and Howley, P.M. (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell* 63, 1129-1136.
77. Glotzer, M., Murray, A.W., and Kirschner, M.W. (1991). Cyclin is degraded by the ubiquitin pathway. *Nature* 349, 132-138.
78. Ciechanover, A., DiGiuseppe, J.A., Bercovich, B., Orian, A., Richter, J.D., Schwartz, A.L., and Brodeur, G.M. (1991). Degradation of nuclear oncoproteins by the ubiquitin system *in vitro*. *Proc. Natl. Acad. Sci. USA* 88, 139-143.
79. Mizushima, N., Noda, T., Yoshimori, T., Tanaka, Y., Ishii, T., George, M.D., Klionsky, D.J., Ohsumi, M., and Ohsumi, Y. (1998). A protein conjugation system essential for autophagy. *Nature* 395, 395-398.
80. Adams J. (2003). Potential for proteasome inhibition in the treatment of cancer. Drug Discov. Today. 8, 307-315.
81. Glickman, M.H., and Ciechanover, A. (2002). The ubiquitin-proteasome pathway: Destruction for the sake of construction. *Physiological Reviews* 82, 373-428.
82. Pickart, C.M., and Cohen, R.E. (2004). Proteasomes and their kin: proteases in the machine age. *Nature Rev. Mol. Cell Biol.* 5, 177-187.
83. Ciechanover, A. (2005). From the lysosome to ubiquitin and the proteasome. *Nature Rev. Mol. Cell Biol.* 6, 79-86.

Statements

Recent activities of the Pontifical Academy of Sciences

Statement of the 2010 Plenary Session

Jürgen Mittelstrass, Werner Arber, Marcelo Sánchez Sorondo

The 20th century was an important century in the history of the sciences. It generated entirely novel insights in all areas of research – often thanks to the introduction of novel research methods – and it established an intimate connection between science and technology. With this connection, science is dealing now with the complexity of the real world. In fact, it was in the 20th century that the human being landed for the first time on a heavenly body, the Moon, different from Planet Earth, this marvellous cradle that the cosmos, in its long and patient development, almost seems to have prepared for our existence. By leaving his habitat, man seems to have stepped on the threshold of infinity.

The members of the Pontifical Academy of Sciences were deeply involved in this development. In this year's Plenary Session with its subject *The Scientific Legacy of the 20th Century* they gave proof of the revolutionary changes in many areas of the sciences – in particular in physics and biology, but also in astronomy, in chemistry, in the neurosciences and in the earth and environmental sciences – and how they contributed to these changes. In this respect the Academy proved itself again to be the mirror of science and its development.

This is particularly true with respect to epistemological and methodological questions as well as to interdisciplinary aspects which become ever more important in scientific research. The Academy deals with these questions and aspects not only in the context of its plenary sessions, for example on predictability in science (2006) or on the evolution of the universe and of life (2008), but also in smaller conferences, workshops and study weeks, for example on the educated brain (2003) and on astrobiology (2009). As it was also shown in the meeting on paths of discovery (2004), the common denominator of the sciences is the notion of discovery, and discovery is an organised mode of observing nature. These meetings put the Academy right in the middle of the ongoing scientific research, especially in relation to the realities of nature, of the human body and of the human brain.

From the point of view of cosmology, the results of these meetings demonstrate that 20th century cosmology greatly improved our knowledge

of the place that man and his planet occupy in the universe. The 'wonder' that Plato and Aristotle put at the origin of thought, today extends to science itself. Questions now arise on the origin and on the whole, also thanks to the reflection of those who study the physical universe, its history and its laws. Physics has enabled us to understand the basic components of matter and we are well on the way to an ever more consistent and unitary understanding of the entire structure of natural reality, which we discover as being made up not only of matter and energy but also of information and forms. The latest developments in astrophysics are also particularly surprising: they further confirm the great unity of physics that manifests itself clearly at each new stage of the understanding of reality. Biology too, with the discovery of DNA and the development of genetics, allows us to penetrate the fundamental processes of life and to intervene in the gene pool of certain organisms by imitating some of these natural mechanisms. Information technology and the digital processing of information have transformed our lifestyle and our way of communicating in the space of very few decades. The 20th century has seen medicine find a cure for many life-threatening diseases and the beginning of organ transplants. It is impossible to list the many other discoveries and results that have broadened our knowledge and influenced our world outlook: from progress in computational logic to the chemistry of materials, from the neurosciences to robotics.

The Academy, however, is not only the mirror of science and research as well as a place where science deals with its problems and insights. It also engages in questions of the institutional role of science in society and issues of great social importance. Scientific research not only gives expression to the strength of rationality in explaining the world and the way in which this is done. The application of scientific knowledge can induce changes of environmental and thus living conditions. It is these aspects, the interrelations between scientific progress and social development, which together with insights into the epistemological structure and the ethical implications of science play an important role in the life and the work of the Academy. Recent meetings on science for man and man for science (1999), on the cultural values of science (2002), on the signs of death (2005) and on transgenic plants for food security in the context of development (2009) testify to this persistent engagement. Also in this respect, the Academy is unique in its structure, in its membership, in its aim, and in its efforts which are always directed at promoting the progress of the mathematical, physical and natural sciences, the study of epistemological and ethical questions and issues, the participation in the benefits of science and technology by the

greatest number of people as well as at the interaction between faith and reason, encouraging the dialogue between science and spiritual, cultural, philosophical and religious questions. The plenary session on the scientific legacy of the 20th century demonstrated afresh the strengths of these objectives and of the way the Pontifical Academy of Sciences in its constitution and activities is realizing them.

Appendix

Study Week on Astrobiology: Summary Statement

José G. Funes, S.J. & Jonathan Lunine

A Study Week on the subject of 'Astrobiology' was held under the sponsorship of the Pontifical Academy of Sciences at its headquarters in the Casina Pio IV in the Vatican from 6-10 November 2009. During the course of the meeting, a highly interdisciplinary group of more than two dozen scientists surveyed recent advances in the scientific understanding of the chemistry that supports life, the origin of planets and life, life's evolution over time in the context of the changing environment of the Earth, the search for life on other bodies within our solar system, the detection of ever smaller planets around other stars and how their habitability might be assessed. The prospect for detecting signals from extraterrestrial civilizations and the possibility of a second kind of life coexisting with known terrestrial biology were also discussed.

His Eminence Giovanni Cardinal Lajolo, in his opening remarks to the participants of the Study Week, referred to Astrobiology as a 'theme that is as new as it is difficult and fascinating...an intense and indispensable case of a vast multi-disciplinary research'. We find that, while still far from being a cohesive scientific field in the same sense that astronomy, biology, or geology are, Astrobiology is gradually turning into a well-defined discipline. Most importantly, 'astrobiologists' are united by their passion for answering the questions that define the discipline, and a strong sense that many exciting discoveries lie ahead.

To enhance the discipline of Astrobiology, we must establish a smoother pathway for education and career development of young scientists who wish to become 'astrobiologists'. For the foreseeable future, such young people will be educated so as to become experts in one of the traditional scientific disciplines underpinning Astrobiology. Further, it is important that students also be exposed in a meaningful way to the techniques and vocabularies of other astrobiological disciplines. How to do this most effectively in the context of already crowded university degree programs is an open issue that we cannot solve here, but we urge that this question be addressed.

Likewise, as scientists we are concerned with the state of the public's understanding of our field. In particular, we believe we must do a better job of explaining to the general public the scientific and logical underpinnings of the main areas of scientific research in Astrobiology. Optimally, we should do this

in the context of the most exciting discoveries of today. To that end, as a product of this conference, we will write a book describing the results of this Study Week that will be understandable to the educated public, especially those who teach science at the middle and high school levels. In this way, we hope to provide a clear and concise tool for understanding the latest discoveries about our planet, planets elsewhere, the origin and evolution of life, and the possible existence of life elsewhere in the cosmos. The book will also serve as a way to energize the public in support of the research goals of Astrobiology.

From the presentations and discussion during the Study Week, we have distilled a set of major scientific conclusions, as well as issues that are important but at the same time clouded by significant disagreements or uncertainties. Finally, we provide a set of recommendations potentially useful for scientists and planners in Astrobiology.

Main scientific conclusions

1. Prior to the evolution of present forms of life – in which proteins are the main catalysts, DNA (deoxyribonucleic acid) the information-storage molecule and RNA the transporter, transcriber, and regulator of information – a form of life existed having RNA as its only encoded biopolymer. This suggests that a key step in the origin of life was the abiological formation of RNA (ribonucleic acid), which could have served at one time as both catalyst and information-carrier. Future work is needed to understand how RNA might be formed in the absence of life, and how biologically useful behaviors might be found within a collection of random RNA sequences.
2. The earliest hint of life on Earth is at 3.8 billion years before present, just after the Late Heavy Bombardment, (3.8-3.9 billion years ago) during which there was a peak in the rate of large impacts on the Earth and the Moon. However, the origin of life may go back to when liquid water was first present on our planet's surface, sometime between 4.0 and 4.4 billion years ago; the rock record is at present too scanty to establish just when life first appeared. Overwhelming evidence of life is present in the geologic record after 3.5 billion years ago – simple cells having the characteristics of so-called 'prokaryotes' probably relying on chemical sources of energy and a form of photosynthesis that does not produce oxygen.
3. Life had an increasing influence on the terrestrial environment over time. Most dramatic is the evolution of oxygenic photosynthesis, which led to suffusion of the Earth's atmosphere and ocean with oxygen sometime near or after the end of Archean time (2.5-2.3 billion years ago), perhaps

triggering global glaciation as well as providing larger gradients in chemical oxidation state for life to exploit. The world that emerged from this transition, however, was not our own. Complex multicellularity (with its potential for intelligence) appears to have emerged only after a second round of environmental transition some 600 million years ago.

4. Serpentinization and chemistry involving reduced species in hydrothermal vents provide a rich source of possibilities for the emergence and maintenance of early ecosystems, on Earth and elsewhere. These ecosystems highlight the potential for remote exobiology that does not involve photosynthesis and does not depend on starlight as an energy source.
5. Life has an extraordinary ability to adapt to the most extreme environments on Earth. Places once thought to be sterile, such as the Atacama Desert and the deep reaches of the Earth's crust, contain life. The range of adaptation of terrestrial extremophiles implies that the traditional definition of the habitable zone within a planetary system is overly conservative; the microbial habitable zone could encompass planets and moons far from a star and not all microbial biospheres may produce remotely measurable biosignatures. Nonetheless, there are limits to the environmental distribution of life on Earth, suggesting that the presence of water on an exoplanet may not be sufficient to ensure its habitability.
6. The number of astrobiologically significant targets in the solar system has increased with the discovery or inference of liquid water in the interiors of icy moons such as Enceladus and Europa (and perhaps others in the outer solar system), of open bodies of liquid methane and ethane on the surface of Saturn's moon Titan, and methane emission from Mars' interior.
7. The so-called primitive bodies, i.e., comets, dust, trans-neptunian objects, and asteroids, also play an important role in Astrobiology; there are indications from dynamical simulations that scattering of bodies during planet formation resulted in the redistribution of organic- and volatile-rich materials from the outer to the inner Solar System. Analytical cosmochemistry, direct sampling, and theoretical research indicate that some of these planetesimals were affected by hydrothermal chemistry, possibly offering a context for organic chemistry.
8. From 1992 to today, more than 400 planets have been detected using several different techniques, the smallest of which is less than twice the mass of the Earth. Some of those exoplanets have atmospheres whose compositions are being studied. Over 60 display eclipses, from which the size and density can be measured in addition to mass and orbital parameters. The study of exoplanets is now firmly established as a discipline connecting astronomy with planetary science, atmospheric science and geophysics.

9. Mass, size and age are key parameters in the characterization of extrasolar planets. Ideally one should determine the mass to a few percent for a 10 Earth-mass rock-ice world to confidently characterize geological processes. While many future studies focus on spectra and brightness variability, knowing the mass, size and age of the parent star are necessary to understand a planet's evolution. Characterizing exoplanet atmospheres and searching for biosignatures is observationally very challenging, but it will be one of the most compelling uses of the next generation of large ground- and space-based telescopes.
10. Formation of planets like the Earth was the end result of largely stochastic processes of mutual gravitational interactions among rocky bodies, stimulated and hence accelerated by the gravitational influence of giant planets which formed earlier when the gas was still present in the form of a 'protoplanetary disk'. Bodies between 1 and 10 times the mass of the Earth divide into two classes: rocky-worlds and water-worlds. The latter have a deep liquid water or ice/liquid water layer. Size and mass of a planet provide density, which constrains composition, though not without significant ambiguity.

Major scientific puzzles and controversies coming out of the Study Week

What follows is a list of uncertainties. The fact that we can ask these questions reflects the mature status of the field. We are confident that great progress can be made on all of them.

1. The origins of life and the environmental context in which life appeared continue to elude us and the Earth no longer has rocks old enough to provide relevant information. By contrast, Mars may have these rocks and, even if life has not appeared on that planet, its ancient rocks may host essential traces of the prebiotic environment that existed.
2. While the participants agreed that atmospheric oxygen appeared late in Earth's history, between 2-3 billion years before present, rather than very early (3.5 billion years or before), the details are not agreed upon. There are competing views about the emergence of oxygen. One is that O_2 did not start being produced biologically until just before it became abundant in the atmosphere, around 2.4 billion years ago. Alternatively, biological production of oxygen in the oceans occurred 300 million years earlier (or even more) in Earth history, but levels in the atmosphere were kept low for many hundreds of millions of years due primarily to the outpouring of reduced materials from volcanoes and from serpentinization reactions in hydrothermal systems. (Serpentinization involves moderate-

temperature alteration of ultramafic (high-Mg) rocks to produce serpentine, magnetite, and reduced gases like H_2 or CH_4). At some point in time oxygenic photosynthesis became sufficiently active to overcome the various sinks of oxygen and oxidized the surface of the planet catastrophically. It is very likely that a further rise of oxygen levels around 600-800 million years ago was a crucial step in the emergence of complex multicellular life forms. Resolving the history of oxygen on Earth is a central problem in Astrobiology, as it has profound implications for the probability of complex multicellular life evolving on other Earth-like planets.

3. The tempo of events in biological evolution, including the origin of photosynthesis, the origin of eukaryotes, the origin of complex life, and the origin of intelligent life, remains a puzzle. How the genome and the environment interacted in each case remains unclear. Large differences between the past and present Earth – an atmosphere richer in methane and other reducing gases, a deep sulfidic ocean, Snowball Earth episodes – suggest a dynamic and perhaps precarious history of the Earth's biosphere. The implications of Earth's past history for its long-term future must be explored: how long can our planet remain habitable as the Sun continues to brighten, and geologic activity winds down?
4. The participants agreed that all astrobiologically relevant bodies deserve thorough future exploration, but the prioritization among them in the search for life is difficult, and various participants had different 'best chance' planets or moons. Titan, Mars, and Europa featured prominently in all the discussions.
5. The approach to planetary exploration – where on a body to sample and how to sample – is also the source of significant disagreement, reflecting the fact that there is no easy (or inexpensive) approach to access astrobiologically interesting terrains and detect life. Planetary protection considerations, both against contamination of planetary bodies and contaminating Earth, complicate the search as well, particularly for sample return.
6. The origin of water and other volatile elements on the Earth – the source regions and the process – remains an area of strong debate. Whether Earth acquired its water as part of a process that is general to planetary formation, or instead through an unusual series of stochastic events, has implications for whether Earth-sized exoplanets in their habitable zones commonly possess a large amount of water during their growth. How did Earth acquire its complement of carbon and in what form, organic or inorganic? How much organic interstellar chemistry survived inclusion into the protoplanetary disk and how much pre-biotic chemistry occurred in primitive bodies like comets before they

struck the Earth? A related question, because it pertains to the timing of the dispersal of nebular gas and the distribution of solid debris, is the origin in the same system of a small terrestrial planet like Mars along with large rocky planets like Earth and Venus.

7. Simulations and theory are still not good enough to confidently predict the properties of exoplanets; the surprising characteristics of young hot Jupiter-type planets found on eccentric orbits are sobering. It is also not possible, with the data at hand, to predict the abundance and spatial distribution of exoplanets as a function of the parent stellar type. More thorough surveys of exoplanetary systems, some of which are underway, are required.
8. To interpret observations of planets around other stars in terms of their habitability will require better understanding of the conditions on the Earth throughout its history, as well as environments on other planets in our solar system. It may not be necessary to reach Earth-mass detection thresholds to draw conclusions about the occurrence of habitable planets around other stars if the habitability of 'super-Earths' (rocky planets up to ten times the mass of the Earth) can be better understood. Further, planets around low-mass stars are generally easier to detect and study than are planets around stars like the Sun.
9. Did life on Earth begin on Earth or elsewhere? Where did life actually arise on Earth if it arose here – hydrothermal vents or somewhere else? How many times did life independently arise on the Earth? Are there multiple, non-overlapping biospheres on and in the Earth?
10. A form of life based on liquid methane and ethane rather than water might be possible. If so, life might exist on Saturn's moon Titan. How membranes might form in such non-polar fluids appears to be the key conceptual stumbling block; it is even conceivable that unusual forms of life might meet the demands of isolation without membranes in the terrestrial sense.

Recommendations regarding astrobiological research

1. Astrobiology relies fundamentally on exploration. Effort should be expended to explore multiple targets in our solar system of potential interest to Astrobiology. The pursuit of understanding of the present-day environments and environmental histories of these planets and moons is as important as the direct search for life. Exploration targets include (in alphabetical, not priority order) Ceres, Enceladus, Europa, Mars, Titan, and Venus (the last for its history which likely included loss of a habitable environment). These missions would have a very high likelihood of scientific payback at relatively modest cost. On Titan, it is cer-

tain that very interesting physics and chemistry can be readily accessed at the surface, especially in hydrocarbon lakes, and icy asteroids may provide a museum of early prebiotic chemistry. Europa and Enceladus may have liquid water ocean environments. On Mars, ancient rocks should be able to provide information about the early conditions and may host traces of early life preserved as fossils. Life may even survive to the present day below the surface. Cryogenic sampling of a comet nucleus, in which the delicate organics and ices are preserved, is important in understanding the starting material of life on Earth.

2. In view of the potential for ground-based detection of Earth-sized planets around nearby stars by enhancing existing techniques, in particular radial velocity, efforts should be made to cooperate on focused observing programs and construction of new facilities to advance the date by which study of Earth-sized planets becomes possible. It is important to probe the limit of the highly successful radial velocity technique. International collaboration will be of importance in this regard. Realistic costing, for example, for 8-10 meter telescope facilities dedicated to planet detection and characterization, should be included in these efforts. Exoplanet imaging and characterization will be a major science driver for proposed 25-40m ground-based optical telescopes. Observation of planets around young stars is a window to the early history of the Earth and its solar system environment; detection and characterization of such objects presents special challenges.
3. Although much work can be done from the ground, characterizing atmospheres of Earth-sized exoplanets and searching for signatures of life – for example, the chemical disequilibrium represented by the simultaneous presence of O_2 and CH_4 in a planet's atmosphere – will most likely require space-based telescopes. Such missions could operate at either at visible/near-IR wavelengths or in the thermal infrared. While space agencies have studied such missions, none has yet been selected for further development; technology development toward their realization should be pursued.
4. A renewed effort should be made to bring computational scientists together with chemists to undertake realistic simulations of self-organizing chemical systems. This could include evolution of metabolic cycles in exotic environments, development of templating molecules in standard (water-based) and exotic environments, and others. While such simulations are not definitive in indicating how life began, these provide potentially sophisticated guides for formulating hypotheses that might be tested by experiment. This work also has the potential to probe alternative biologies, and hence the limits of evolutionary innovation of car-

bon-based life. An example is the issue of a membrane capable of functioning in organic solvents with or without water and at low temperatures, with application to the Titan environment.

5. Chemists, geologists, planetologists, and astronomers should be encouraged to collaborate in developing and refining models of exoplanet atmospheres, surfaces, interiors, and evolution through time. Although a general theory of planetary evolution may not be possible, because of the complexity of planetary processes and their contingent dependence on external boundary conditions, progress can be made by ensuring involvement of experts from the relevant disciplines. The same can be said about planetary formation, where geochemistry should be more fully incorporated in formation models. The role of geophysical processes that result in the production of catalytic minerals – of which there are hundreds – should be investigated, in view of the indications from several speakers that the best chance of going from organic chemistry to some useful biochemistry is to find mechanisms/conditions that produce a limited number of different kinds of organic compounds that are known to be intermediates in metabolic networks or biochemical synthesis.
6. There are a number of ways that Mars early in its history could have had locally habitable climates without the entire planet being steadily warm and wet. While it seems unlikely that the amount of rainfall during a brief post-impact steam atmosphere could carve the observed fluvial features, there are many 'damp spot' scenarios – some in the literature and some yet to be proposed – that should be investigated.
7. We need more research focused on biosignatures, their identification, preservation and reliability, to test the hypothesis that with the evolution of key metabolic pathways there would be a period of 'enrichment' continuing until a key nutrient becomes limiting. The origin and succession of different anoxygenic photosynthetic pathways (or even oxygenic ones) could leave an atmospheric signature such as a decrease in atmospheric carbon dioxide and hydrogen sulfide, with potential application to the characterization of exoplanets.
8. Although we harbor a bias that the evolution of intelligent species elsewhere may be less frequent than the origins of other life forms, nevertheless it may be the actions of distant technologists, and technosignatures that provide an inexpensive shortcut to the detection of habitable worlds beyond Earth. For this reason as well as the chance to know whether there is anyone 'out there' to talk to, SETI should be encouraged and supported as a part of astrobiological research, and expanded beyond the private venture it is today.

9. Astrobiology addresses one of the most profound questions in science: is life on Earth unique? The subject fascinates the public, but we may not be able to convince them of what we think we know since evidence at the frontier of research is often fragmentary and ambiguous. We must develop clearly written materials to explain how science is done, what is known, how we ask and answer questions about what is unknown, and what are theories, hypotheses and speculation, so that our inferences can be properly understood.
10. Programs that involve astrobiologists from diverse disciplines in mono-disciplinary activities (astronomers on ocean research vessels, chemists at astronomical observatories) must be encouraged, especially at the doctoral and post-doctoral level but even for more senior researchers. Truly interdisciplinary meetings like this Study Week are essential to the health of Astrobiology. An example is the deep carbon initiative, which has several themes, the most important of which to Astrobiology is an understanding of all of the sources of abiotically-produced carbon compounds in the subsurface Earth and mechanisms that might lead to greater complexity of organic chemicals and biochemicals. Programs that permit members of the public to participate in analysis of astrobiological data (SETI and planet detection data are two pioneering examples) should be expanded.

Reflections on the demographic question and on pastoral guidance

Bernardo Colombo

On the July 1989 issue of *Science* was Carl Djerassi, a chemist at Stanford who liked to qualify himself as the mother of the pill – since G.G. Pincus was usually named the father of the pill – who, in Mexico City in 1951, had accomplished the first synthesis of an oral contraceptive, while he was associate-director of Syntex, a firm comprising research and business. In his article he outlines new approaches to birth control. These include a male pill, a once-a-month menses inducer and an antifertility vaccine, and an antiviral compound that associates birth control with the fight against the AIDS epidemics, and the measuring of certain biochemical changes which, if done accurately and simply, could reduce the fertile period in a cycle by more than 50% and thus significantly improve the poor image of natural family planning.

The Study Week on Resources and Population, held at the Pontifical Academy of Sciences, examined, by means of factual information and in-depth evaluations, both present situations and plausible developments, stressing current and future problems. The picture that emerged was a complex one. One fundamental aspect that was amply documented is that of the profound imbalances, from many points of view, among countries and among population categories. Resolving them will require an enormous effort at various levels, and this was also stressed in certain essential passages in the address of Holy Father John Paul II to the conference participants.

A careful examination of the topics presented and discussed at that meeting may serve to underline further elements of basic importance for the proper formulation of pastoral guidelines.

Above and beyond even the present severe difficulties, and those of the immediate foreseeable future, as well as the errors made in both creating and dealing with these difficulties, the reality of the situation which now confronts the whole of humanity requires careful reflection. Such reflection had already been initiated in Point 6 of the Papal address. This reality is one that must be measured against a norm that is perpetually valid, since it derives from natural law. To be precise, our present state of knowledge informs us that mankind in the future must limit itself on *average* to a little more than two children per couple. This is the inevitable consequence of the power that man has acquired over sickness and death, and which he will presumably further extend. Fol-

lowing another course would, within a few centuries, produce intolerable results bordering on the absurd. It is a matter of fate already foretold, if we exclude enormous catastrophes or drastic inversions of trends. Nuclear risks have been cited, as has the ozone hole; AIDS is spreading, with as yet poorly defined demographic consequences. However, regulating one's own actions on the basis of expecting unforeseeable and disastrous events to happen, does not seem to be a wise course to follow.

In substance, man is, unfortunately, *obliged to relinquish a large part of his procreative capacity*. One can of course comment that in numerous economically developed countries there are less than two children per couple. But it is one thing for behaviour to result from a choice – however mistaken it may be – and it is another to have such behaviour imposed by necessity.

The consequence and problems which derive from such a situation are of various types: there will be a general aging of the population; there will be many small families with all the attending psychological, educational and emotional implications, and also problems of inter-generational relationships; it will be necessary to find acceptable means of reconciling a couple's free and responsible choices with collective needs, and so on.

In the world today there is a considerable limiting of the number of births, though not everywhere, nor in equal measure, and not with identical means in the various countries and social classes. On the whole, one may say that at present only about one half of humanity's reproductive potential is fully utilized. Recent variations in individual choices regarding marital customs do not appear to have any appreciable influence on this figure. The causative factors which are by far the more relevant are: the limiting of conception and the limiting of birth.

With regard to induced abortion, figures in the area of tens of millions per year are being put forward, but it is difficult to check the reliability of such estimates.

The most recent calculations published by the United Nations indicate that circa 450 million married couples of reproductive age made use of family planning methods in 1987. Of these, 7.2% were said to be using their knowledge of the non-fertile periods of the female cycle. These figures and percentages are to be approached with great caution, as there are – for various reasons – significant margins of uncertainty, especially where so-called traditional methods are concerned, including what is known as the 'rhythm' method. On the other hand, now, after a period of five years, this figure will certainly have risen, especially – it would seem – with regard to female sterilization. In this context, a recent estimate speaks of 43 million additional cases in the five years following 1984.

The costs involved in limiting births by these means have also been calculated: in the developing countries alone the figure is put at more than *5 billion dollars* a year. Of this, three-quarters is said to come from local governments, 10% from users themselves, and some 15% from the so-called *donor agencies*: the United Nations Population Fund, the United States Agency for International Development, the International Planned Parenthood Federation, and a variety of international, governmental and private institutions. It should be kept in mind that the actual impact of the 'external' interventions is far greater than that implied by those figures. Indeed, these interventions strongly contribute to creating a cultural climate, and to steering the actions of local governments. The unscrupulousness, the amorality, the malice, the vested interests inherent in many initiatives are absolutely disgraceful. A strong condemnation of them would lose its force, however, if it were to call into question, or to minimise the seriousness of the demographic problems – a seriousness which, as already stated, was made very clear at the Study Week on Resources and Population.

Conversely, such denunciations would have more force if they were accompanied by firm support for the diffusion of natural fertility regulation methods. This is what *Humanae Vitae* has urged within Christian married couples (n. 25), to be promoted by doctors and members of the nursing profession (n. 27), in its recommendations to the clergy (n. 29), in its invitation to the Bishops (n. 30), in its encouragements to men of science. Similar expressions occur in *Gaudium et Spes* (n. 52) to recall the need for everyone to promote the good of marriage and the family. It is also implicitly referred to by the Holy Father John Paul II in his discourse addressed to the participants in the Study Week on Resources and Population (n. 6), recalling that the Church invites humanity to program its future, motivated not only by material concerns, but also and especially by respect for the order which God has infused into creation.

To try to ensure that one will not have children by applying knowledge of the infertile periods (*Humanae Vitae*, n. 16) is a very demanding responsibility, from several points of view.

First of all, concerning what has already been said, the 'circumstances' which were spoken of in *Gaudium et Spes* (n. 51) and the 'serious motives' mentioned in *Humanae Vitae* (n. 15) and previously specified by Pius XII in his Allocution to Obstetricians on October 29, 1951, may no longer be viewed as simply occasional and isolated events, but must be considered as a *human condition which is habitual and widespread*.

Further, these methods – in contrast to others, and more so that any other – cannot be reduced to purely individual initiatives, in no case what-

soever. Then too, it is clear how substantially different this way is from many others followed for similar ends. To voluntarily undergo sterilisation for anti-conceptional reasons certainly requires in the highest degree both a decision and a taking on of responsibility, once and for all. Conversely, the use of natural methods for regulating conception requires giving long-lasting, constant and continuous attention to responsible choices.

Further, these methods – in contrast to others, and more so than any other – cannot be reduced to purely individual initiatives, in no case whatsoever. They demand, and are capable of increasing, a complete, affectionate harmony of the spouses in their attitudes and their behaviour with regard to procreation. The married partners are continually being called upon to seek the proper balance between potential collaboration with that act of God's love which is Creation, and simple union in chaste intimacy with acts which, performed in a mode that is truly human, promote the mutual giving of which they are the symbol, enrich the spouses reciprocally in joyous gratitude (*Gaudium et Spes*, n. 49), constitute a support for mutual fidelity, and also contribute to the well-being of the children (*Gaudium et Spes*, n. 50 and n. 51; *Humanae Vitae*, n. 3 and n. 16).

The pursuit of such a balance is enhanced by a sure knowledge of the fertile phases of the female cycle, capable of responsible guiding behaviour in serene awareness of the potential consequences of one's own acts: knowledge which nowadays encounters considerable obstacles in the ordinary manifestations of biological variability, with negative effects on conditions of fertility regulation.

In the Holy Father's discourse (n. 5) it is stressed that humanity must confront the new circumstances by making use of all its intellectual and spiritual energies, recovering a sense of the moral significance of *setting limits* for oneself, learning to develop responsibility toward all forms of life. Here, the contribution which the Church can make is of fundamental importance. Concerning the natural methods for regulating conception, the Church says that 'For if with the aid of reason and of free will they are to control their natural drives, there can be no doubt at all of the need for self-denial. Only then will the expression of love, essential to married life, conform to right order. This is especially clear in the practice of periodic continence' (*Humanae Vitae*, n. 21), while also reminding us that 'God comes to their help with the grace by which the goodwill of men is sustained and strengthened' (*Humanae Vitae*, n. 20). This necessary self-discipline is of course not to be limited only to the sector in question, but must be encompassed in an integrated education covering the entire gamut of human activities. A coherent witness of this kind within the Catholic world could

also be of assistance in solving the enormous problems of peoples whose cultures are foreign to the idea of limiting the number of children and who are more likely to suffer from the 'new conditions': a large part of the Islamic world, for example, or sub-Saharan Africa, even India. For this actually to take place, it is of course necessary that the example be clearly visible and that the aid given be significant.

In this respect, one might make a few succinct observations. Mention has been made of the expenses involved in limiting births in developing countries. As a conjecture – supported by indicative evidence – it may be posited that the amount of funds, of whatever origin, earmarked for the diffusion of natural methods, is not much more than 1%. Additional estimates have also been made as to the future interventions considered to be necessary to keep pace with the medium variant projections of demographic development in the immediate future made by the United Nations; however, it is impossible to say how much the world in general, and the Catholic world in particular, is prepared to do in the future to promote natural methods in these countries, for there are no known programs of action, apart from very restricted initiatives.

There are enormous needs of all kinds in numerous countries, and Church institutions which are in the front line for giving aid must set priorities, given the limited resources which are available. Hunger, disease, tremendous poverty, all cry out for immediate attention. Facing these problems, Church bodies stand side by side with many other national, international, public and private institutions that are moving in the same direction. However, aid which is intelligent and respectful of the natural regulation of procreation by working in a way which so closely associates the rational creature with his or her Creator because it respects the order established by God (*Humanae Vitae*, n. 16) – such aid may be viewed as a proprium of the Church, which must give to it adequate space not only in moral teaching but in concrete action.

Certainly, every form of 'aid' which comes from outside runs the risk of being interpreted as a kind of masked imposition of others' interests. It is a risk run by every method which is introduced in this field, but it should be kept in mind that it is minimal where natural methods are concerned. What surely produces more negative consequences is the lack of coordination among such initiatives. In fact, at times, organizations which adhere to different 'schools' seem almost to be in competition with one another, and this severely damages their public image.

In practice, not only in economically developing countries but almost everywhere, instruction in and diffusion of the natural approaches to reg-

ulating conception are left to the voluntary sector. This has a few positive aspects: confidence that the advice given is good, and the generosity of working together for a common ideal. However, there are also a few serious negative sides: lack of permanent structures, precarious measures, too little professionalism. Very unpleasant situations arise when voluntary personnel feel themselves to be forgotten, victims of neglect on the part of Church authorities. On the other hand, the inefficiency of the voluntary services may produce loss of confidence and resignation on the part of those who are responsible for family pastoral work.

As the 50th anniversary of *Humanae Vitae* draws near, one would hope that the brief considerations put forward here will spur the Church to a renewed commitment of its enormous energies to furthering the spread of natural conception-regulating practices, as well to review the means which it employs for this purpose. In this regard, one might recall the 'pressing invitation' addressed in *Humanae Vitae* to all Bishops that they, together with the cooperation of priests and laity, should work 'with all zeal and without delay to safeguarding the holiness of marriage, in order to guide married life to its full human and Christian perfection' (n. 30). 'Consider this mission', the encyclical says, 'as one of your most urgent responsibilities at the present time. As you well know, it calls for concerted pastoral action in every field of human diligence, economic, cultural and social' (*ibid*.).

The strenuous and multiple efforts of the Catholic Church to meet the needs of suffering humanity are well known. But, whereas it is clear to all how much there can be done for the lepers, to take an example, or for educational opportunities in Catholic schools, there seems to be lacking an equally clear awareness of the millions and millions of persons who need to be wisely counselled, assisted and followed-up – including in situations where clinical opinions are required – regarding the everyday, fundamental life choices. It is not a question here of promoting chastity as part of a lifestyle of austerity – which, as has already been stated, is a basic option, and to which the Church can bring its immense spiritual riches. Here it is a matter of technical aspects, and these can be briefly outlined.

First of all, it should be resolved to make a far greater effort than is now the case in the field of research. There is still much to be desired, today, as to the applicability, the acceptability and the reliability – together with the simplicity of use – of natural ways of regulating conception. They must be considerably improved if this way is to be made available far beyond present limits, and if it is to be generally viewed as a valid alternative to other methods now being used which do not respect the dignity of human acts. Such research requires appreciably greater funding than the extremely small amount which a

few persons manage to obtain at present from public authorities that display little interest. This funding should preferably not be tied to the private sector which aims only at profits.

Further, it is indispensable that there be widely diffused services which provide information, counselling and assistance, rooted in professionally competent, stable structures, which operate in mutual collaboration, and which are staffed with qualified personnel. Coordination of this type can also be useful for setting up a system to record and document results, which would in turn make possible and facilitate critical evaluation of the work accomplished, thereby improving efficiency and efficacy.

In addition, one might consider promoting the establishment of one or more centres of excellence in every country. They would provide scientific support, contribute to a rational planning of activities, and could function as reference points for the professional training of workers at various levels. Such centres could also accomplish fruitful work in international solidarity.

These are merely sample proposals, to be considered along with others which might be put forward. What is needed in any event is that such things be done in a spirit of open, intelligent and active concord.

Bibliographic References

Lande R.E. and Geller J.S., Paying for Family Planning, *Population Reports*, Series J., Number 39, November 1991.

Church C.A. and Geller J.S., Voluntary Female Sterilization: Number One and Growing, *Population Reports*, Series C, Number 10, November 1990.

United Nations, *Levels and Trends of Contraceptive Use as Assessed in 1988*, New York, 1989.

United Nations, *Long-Range World Population Projections. Two Centuries of Population Growth 1950-2150*, New York, 1992.

Resources and Population, Bernardo Colombo, Paul Demeny and Max F. Perutz (eds), Clarendon Press, Oxford, 1996.

Carl Djerassi, The Bitter Pill, *Science*, 245, July 1989: 356-364.

Carl Djerassi, Fertility Awareness: Jet-Age Rhythm Method? *Science*, 248, 1061, 1990.

Carl Djrassi, The Pill: Emblem of Liberation, *British Medical Journal*, 334, Suppl. S.15, 2007.

How to Become Science? The Case of Cosmology

Michael Heller

1. Before the Beginning of Relativistic Cosmology

When Aristotle was writing 'the least initial deviation from the truth is multiplied later a thousandfold',[1] he was not fully aware of how much it was true and how much it referred to himself. His errors in establishing details of phenomena indeed multiplied later a thousandfold, but consequences of his error in choosing the method for investigating nature multiplied even more.[2] It is often said that Aristotle's errors and his misguided method of investigating natural phenomena blocked scientific progress for many centuries. However, long periods of blundering are, in certain conditions, an unavoidable price of the final success. If this was true as far as natural sciences were concerned, it was even more so in the case of cosmology. It seemed to be a helpless case. From the present perspective it is hard to say what was more reasonable in this field: Aristotle's seemingly precise, but in fact most often purely verbal, analyses, or Plato's openly metaphorical narrations. Eudoxian crystalline spheres and Ptolemaic epicycles rendered a service to positional astronomy, but from the cosmological point of view, being contradictory with each other, immersed the science of the universe in a persistent crisis.

The birth of modern science in the 17th century only slightly improved the situation in this respect. The notion of the universe, inherited from the Ancients, extended from the sublunar area to the sphere of fixed stars, and this is why polemics around the Copernican system, that strictly speaking referred only to the planetary system, had in fact a cosmological aspect. But in this aspect, it introduced more misunderstanding than real progress. The true promise of future successes was the discovery of the universal character of the law of gravity, but for the time being it generated, when applied to cosmology, more problems than solutions. Newton was not exaggerating when he claimed that the supposition that there should be a particle so accurately placed in the middle of stars 'as to be always equally attracted on

[1] *On the Heavens*, translated by J.L. Stocks, p. 271b.

[2] Aristotle, to be sure, was a great experimentalist of his time, especially in the field of life sciences, but controlled experiments played only a marginal role in his method.

all sides' is as difficult to implement as 'to make the sharpest needle stand upright on its point upon a looking glass'.[3] The problem of gravitational instability was one of the most difficult questions with which the physics of the universe had to cope.

Newton's conundrum with gravitational field instability, later known as the Seeliger paradox, led to some attempts at modifying the law of gravity, and quite unexpectedly surfaced in Einstein's first cosmological paper of 1917. In this paper, Einstein had to add to his equations the so-called cosmological constant to obtain a static model of the universe. It was a particularly malicious twist of history when some ten years later it turned out that the Einstein static model, in spite of this 'saving procedure', is in fact unstable.

The Seeliger paradox, in the 19th century, was paralleled by the optical Olbers paradox (in fact, this paradox was also known to Newton): if the infinitely extending universe is uniformly filled with stars (or galaxies, or galactic clusters, in the more modern version), the night sky should be as bright as the surface of the Sun, but this conclusion remains in sharp contrast with what can be seen with the naked eye.

In the second part of the 19th century there were strong reasons to believe that the universe as a totality cannot be made obedient to physical laws discovered in our local neighbourhood. Agnes Mary Clerke, the astronomer and historian of astronomy, was not an exception when, in 1890, she declared: 'With the infinite possibilities beyond [our Milky Way], science has no concern'.[4] Helge Kragh quotes an anonymous reviewer who in 1907 wrote: 'when there are no facts to be explained, no theory is required'.[5] In spite of these reservations, the universe is a challenge to the human mind. We prefer to create fancy hypotheses in order to tame the Unknown rather than to acknowledge our ignorance. Besides the 'cosmological skeptics', in the second half of the 19th century, there was a crowd of physicists, philosophers, amateurs, and sometimes also astronomers, who indulged their imagination and developed various cosmic scenarios.

[3] Newton claimed even more: since there is an infinite number of stars in the universe, to make the system of stars stable '…this is as hard as to make, not one needle only, but an infinite number of them stand accurately poised upon their points'. Newton's Letter to Bentley, in: *Isaac Newton's Papers and Letters on Natural Philosophy*, ed. by I.B. Cohen, Harvard University Press, Cambridge, Mass., 1971, p. 292.

[4] After H. Kragh, *Matter and Spirit in the Universe*, Imperial College Press, London, 2004, p. 20.

[5] *Ibid*.

History of science, especially as it is done by working scientists, is often highly selective. It focuses on ideas from the past that later on evolved into commonly accepted theories or trends, but it forgets or overlooks those side-branches that blindly ended with no consequences. It is instructive to read in this respect the second chapter of the above-mentioned book by Kragh: how rich the spectrum of conceptions and views was, in the 19th century, that attracted the general public and engaged some scientists, but left no traces in our mainstream scientific cosmology. However, there were also ideas, considered at that time highly exotic, that now belong to the standard conceptual tool-kit of our theories.

2. Many Dimensions and Non-Euclidean Geometries

The discovery of non-Euclidean geometries by Gauss, Bolyai and Lobachevsky opened a vast field of possibilities. Riemann and Clifford speculated about their eventual applications to physics, and the idea soon captured general attention. A popular book by Edwin Abbot Abbot, first published in 1884, presenting the adventures of a two-dimensional Square (inhabitant of Flatland) in three-dimensional Spaceland, soon became a bestseller.

Ernst Mach in his influential *The Science of Mechanics* relegated the problem of multidimensional spaces to a long footnote at the end of the book.[6] He considered the discovery of non-Euclidean geometries as an important mathematical achievement, but 'we must not hold mathematicians responsible for the popular absurdities which their investigations have given rise to'. The space of our sensual experience is doubtlessly three-dimensional. 'If, now, it should be found that bodies vanish from this space, or new bodies get into it, the question might scientifically be discussed whether it would facilitate and promote our insight into things to conceive experiential space as part of a four-dimensional or multi-dimensional space. Yet in such a case, this fourth dimension would, none the less, remain a pure thing of thought, a mental fiction'. He then develops the topic of 'popular absurdities': 'The forth dimension was a very opportune discovery for the spiritualist and for theologians who were in the quandary about the location of hell'. When writing this ironic sentence, Mach probably had in mind the German astronomer Karl Friedrich Zöllner who became engaged in spiritualism and claimed that the fourth dimension well explains spiritualistic phenomena.

[6] *The Science of Mechanics. A Critical and Historical Account of Its Development*, Open Court, La Salle, Illinois, 1974, pp. 589-591.

Besides this rather extravagant claim, Zöllner thought that non-Euclidean geometries were relevant for the study of the world as a whole.[7]

Zöllner was not alone to proclaim this idea. The first attempt to experimentally check the curvature of space should be attributed to Gauss himself who in his *Disquisitiones generales circa superficies curvas*, published in 1828, reported his experiment to survey a triangle formed by three peaks in the Herz mountains.[8] The result was of course negative. A more serious analysis of experimental possibilities in this respect was undertaken by Karl Schwarzschild. Before the Astronomische Gesellschaft in Heidelberg in 1900, he discussed four possible observational tests to detect space curvature: (1) the test from the minimal parallax of stars, (2) from the number of stars with different parallaxes, (3) from the possibility to see 'around the universe', and (4) from star count as a function of stellar magnitude. We should admire Schwarzschild's insight: having no help from a physical theory, such as later general relativity, he not only conceived four observational test, but also understood the necessity of taking into account various topological forms which could essentially modify the results.

As we can see, long before Einstein's special and general theories of relativity, some elements (such as multi-dimensionality and non-Euclidean geometries) that later on entered the very body of these theories, had already circulated among both scientists and amateurs. The point is, however, that Einstein, when time was ripe, did incorporate them into his way of thinking not by borrowing them from this circulation, but rather by distilling them from the logic of the evolution of physical problems.

Somewhere in the span of the 18th and 19th centuries, mainly due to the progress in astronomy, from rather fuzzy 'cosmological narratives' a core or a body of accepted views started to emerge that later became scientific cosmology. However, at that time these narratives were pushed to a fuzzy belt of speculations and hypotheses surrounding the core. When the science of the universe consolidates, some inhabitants of the belt assume more responsible forms and are absorbed by the core, some others are forgotten. The belt is important since it is an indispensable condition of progress. The situation becomes dangerous only when the subtle borderline between the core and the belt is regarded as non-existent.

[7] More about Zöllner in: H. Kragh, *op. cit.*, pp. 24-26.

[8] See M. Heller, P. Flin, Z. Golda, K. Maslanka, M. Ostrowski, K. Rudnicki, T. Sierotowicz, 'Observational Cosmology. From Gauss to Sandage', *Acta Cosmologica* 16, 1989, 87-106.

3. The Beginning of Relativistic Cosmology

The birth of General Relativity was a real breakthrough, and that of relativistic cosmology a spectacular application of Einstein's theory to the biggest physical system conceivable. After years of struggle and several dramatic months of painful coda, Einstein, in November 1915, finally wrote down his gravitational field equations. They were his response to the critical situation in which Newton's theory of gravity was deeply immersed (crisis in any of the major physical theories always has an echo in other areas of science). Some people were aware of this and tried to remedy the situation by modifying Newton's law, but only Einstein, owing to his work in special relativity, was able to see the connection between gravity and the spacio-temporal framework of physics, and understood that Newton's gravity should not be amended but suitably replaced.

Einstein's answer to the crisis was a piece of art of enormous beauty. Even if the final act is a sort of illumination, it certainly did not come as *deus ex machina*. Einstein was led to it by a chain of almost deductive reasoning, based on clearly formulated questions of deep physical significance. This does not mean that sometimes the chain did not need enormous effort to make the reasoning transparent. To change his ideas into the body of a physical theory Einstein had to use completely new mathematical theories, known only to some experts in pure mathematics but foreign to the community of physicists. As a result, he obtained a set of ten partial differential equations, the richness of which he was only dimly aware of. It is true that in the compact tensorial form they look quite innocent, and when applied to various physical situations they usually simplify to a tractable mathematical form. Only after acquiring a certain familiarity with them, one can guess their abysmal richness from the fact that, when applied to different problems, they reveal unexpected layers of their mathematical structure. If you read in popular books that Einstein's equations present the gravitational field as the curvature of a four-dimensional space-time, it is only a shortcut of a vast empire of mutual interactions between non-Euclidean (pseudo-Riemannian, in modern parlance) geometries and various aspects of physical reality.

In 1917 Einstein produced his first cosmological paper. We already know the beginning of the cosmological constant story. This constant was, so to speak, enforced upon Einstein by his equations. Some ten years later, when it turned out that the universe is not static but expands, Einstein proclaimed the introduction of the cosmological constant 'the greatest blunder of his life'. But in proclaiming this, it was Einstein who was wrong, not his equations. From the mathematical point of view, the most general (and therefore the most beautiful) form of Einstein's equations is the one with the cos-

mological constant. And today there are strong reasons to believe that this constant has an important physical interpretation. From recent observations of the Ia type supernovae we know, with a good degree of credibility, that the universe not only expands, but also accelerates its expansion. And to obtain agreement between these observations and theoretical models one should employ the equations with the cosmological constant (see below).

This is a typical story. The history of relativistic astrophysics and relativistic cosmology from Einstein up to about the seventies of the previous century, consisted mainly in solving Einstein's equations, analyzing and interpreting the structures of these solutions, and then looking at the sky to verify what the equations had predicted (of course, it is a highly idealized picture). In this process, various groups of people were engaged, there were many disputes and erroneous interpretations, in some of them other departments of physics had to be involved. In this way, young relativistic cosmology slowly crystallized. The new field of research was created, still full of question marks but ready for further developments.

Einstein's paper of 1917 was definitely wrong. Its static world model does not represent the world we live in. But the paper was epoch making; it opened a new way in our thinking of the universe.

4. Domination of Philosophy

The next move belonged to the Dutch astronomer Willem de Sitter. The story is well known; I shall only very briefly sketch its main stages. Yet in 1917 de Sitter found another cosmological solution of Einstein's equations which represented a world devoid of matter (with vanishing matter density) and, as it turned out later, with expanding space, thus ruining Einstein's hope that the 'cosmological problem' would have a unique solution. In 1922 and 1924, Russian mathematician and meteorologist Alexander Alexandrovitch Friedman published two papers in which he presented two infinite classes of solutions with the positive and negative curvature of space, respectively. In 1927 Georges Lemaître, for the first time, compared theoretical predictions of one of the expanding cosmological models with the results of galactic red shift measurements, and found that there was no contradiction between them. All so far considered solutions satisfied the postulate of maximal spacial symmetry (isotropy and homogeneity of space).[9] Mathematical discussion of such models was undertaken by Howard Percy

[9] This postulate was called by Edward Arthur Milne the Cosmological Principle.

Robertson and Arthur G. Walker. Probably the first ever attempt to fill in the geometric scene, as given by Einstein's equations, with physical processes reconstructing the history of the universe, belonged to Lemaître. It was called by him the Primeval Atom Hypothesis.

In the meantime extra-galactic astronomy gradually made a significant progress. After the discovery, in 1929, by Edwin Hubble of what is now known under the name of Hubble's Law (linear dependence of red shifts on distance for distant galaxies), the effect of the 'expansion of the universe' was reasonably well established, but the results of measurements were not enough to select a model, or a class of models, that would best fit the data. Another information important for cosmology, which should constitute its observational input, the uniform distribution of galaxies, was regarded as a simplifying postulate rather than the result of observation.

It goes without saying that it is a very sketchy picture of the situation. We should notice that cosmology at that time was regarded as a true science only by a very few, the main objections being: the lack of reliable experimental data, the huge degree of extrapolation and, quite often, the lack of confidence towards Einstein's theory of gravitation as a fully-fledged physical theory. No wonder that in such a situation philosophical prejudices played a greater role than is usually the case in the sciences. Mutual influences went in both directions: from philosophy to cosmological models, and from works in cosmology to philosophical views. For instance, philosophical views on the creation of the universe or its eternity influenced preferences of cosmological models with the initial singularity or with cyclic histories. On the other hand, since the majority of early works concerned spatially closed world models, the philosophical idea was favoured that space should be closed, otherwise it would not be cognizable to the human mind. The latter view is encapsulated in bishop Barnes' statement at the meeting of the British Association for the Advancement of Science in 1931, that 'infinite space is simply a scandal for human thought'.[10]

After the Second World War, cosmology resumed its progress more or less in the same style as before the war, but its theoretical environment slowly started to change. General relativity gradually became a fully acknowledged physical theory, more theoretical works related to this theory were accumulating, and its role in physics was increasingly important. These changes were not directly related to cosmology but, some two decades later,

[10] 'The Evolution of the Universe', *Supplement to Nature*, n. 3234, 1931, 704-722. This is a report of the discussion held at the British Association.

these processes significantly contributed to an acceleration in cosmological research. For the time being, cosmology continued to be dominated by philosophical speculations and polemics. One of the hottest topics was the problem of the beginning of the universe. Lemaître, after an early attempt to avoid this problem,[11] tried to incorporate the beginning ('natural beginning' as he called it) into the physical scenario. This aroused a strong reaction from some scientists who suspected in this move a hidden religious propaganda. In 1948, Herman Bondi, Thomas Gold and Fred Hoyle published their steady-state cosmology.[12] The clear motivation was to counteract the influence of the Big Bang theory.[13]

The universe is expanding. In the face of accumulating data this cannot be denied. The only way to save its eternity (no beginning) is to assume that matter is being created out of nothing to maintain world's constant density (this postulate was called the Perfect Cosmological Principle). It can be calculated that the creation rate necessary to obtain this goal is undetectable on the local scale. Initially, steady state cosmology met some resistance but later on, owing mainly to Hoyle's propaganda activity, discussions between steady state and Big Bang theories dominated the cosmological scene. Lemaître, highly disappointed, lost his interest in cosmology and turned to numerical computation.

There is no need to focus on this story; it was extensively studied by Helge Kragh, and the interested reader should be referred to his book.[14] It is a common conviction that it was the discovery of the cosmic microwave background (CMB) radiation that killed the steady state theory, although Kragh's study shows that at the time of this discovery the steady state theory was already practically dead.

5. Standard Model of the Universe

The great breakthrough in cosmology occurred in the sixties of the previous century. Although the existence of CMB was predicted already in

[11] In his work of 1927 he chose a model (later called Eddington-Lemaître world model) with no singularity to compare it with observational data.

[12] In Bondi and Gold's version the idea was based on speculative assumptions; Hoyle based his version on an unorthodox interpretation of de Sitter's solution to Einstein's equations.

[13] Somewhat later Hoyle used, for the first time, the term Big Bang as an ironic nickname of Lemaître's cosmology.

[14] *Cosmology and Controversy. The Historical Development of Two Theories of the Universe*, Princeton University Press, Princeton, 1996.

1948 by George Gamow and his coworkers and, quite independently of their work, there were some hints of its presence based on observed excitations of CN particles in clouds of interstellar space, nobody was able to appreciate the impact the CMB discovery was to make on cosmology. Its main merit was not so much that it contributed to eliminate sterile discussions around steady state cosmology but, first of all, that it has provided means to obtain access to the physics of the early universe. The discovery of quasars, almost at the same time, trigged interest in extra-galactic astronomy. Progress in radio astronomy and in optical astronomy was providing richer and richer data concerning the large-scale distribution of matter.

An insight into physical processes in the early universe gave momentum to the theory of nucleosynthesis, also initiated by Gamow and his team. The works by Margaret Burbidge, Geoffrey Burbidge, William Fowler and Fred Hoyle (initially done in the framework of the steady state cosmology) provided the part of the nucleogenesis theory missing in Gamov *et al.*'s works (formation of heavier elements in the interiors of massive stars). The observed abundances of chemical elements in the universe well agreed with theoretical models. Especially, the determination of the primordial deuterium abundance turned out to be a very sensitive cosmological test (it led to tight constraints for the baryon density a few minutes after the Big Bang).

All these achievements (and many others as well) created a basis for establishing the hot model of the universe (as it was then called) – a reconstruction of physical processes from the proverbial 'first three minutes' after the Big Bang, through the radiation era, and the origin of galaxies and their clusters to the present cosmic era. We should acknowledge the great contribution to this scenario of the Moscow school led by Yakov Borisovich Zel'dovich. With the incoming of observational data and the progress in theoretical works, the model of the hot universe slowly changed into what is now called the standard model of the universe – the commonly accepted scenario of the evolution of the universe. Of course, there were some opponents who tried to defend 'alternative cosmologies', but their works had only a marginal influence on mainstream cosmology.

When progress is made, details are seen more sharply which, in turn, allows one to see problems and difficulties previously invisible or dimly visible. Several such problems were identified in the otherwise very successful standard model. Let us enumerate some of them:

The horizon problem. The anisotropy of CMB is now known to be less than 10^{-4}. How to explain this degree of uniformity at two points in the sky that are separated by the causal horizon (the distance between such points is greater than that which light can cover since the initial singularity)?

The flatness problem. How to explain the fact that the parameter $\Omega=\rho/\rho_{cr}$ is very close to unity? Here ρ is the present density of matter, and ρ_{cr} is the density of matter characteristic for the flat cosmological model. It should be noticed that $\Omega=1$ is an 'unstable' value in the standard model.

The monopole problem. How to explain the fact that the standard model predicts an overproduction of magnetic monopoles, and they are not observed in the present universe?

The origin of structures problem. How to explain the generation of density fluctuations in the early universe (they are indispensable to form stars, galaxies, cluster of galaxies…)?

In 1981 Alan Guth proposed the first inflationary model (now there is almost infinite number of them). Enormously accelerated expansion (by the factor greater than 10^{30} in less than 10^{-32} sec.) blows up an inside-the-horizon part of the universe to a size greater than the present observable universe. This hypothesis almost automatically solves all the above enumerated problems,[15] and this is certainly its great merit, but the inflationary scenario also has its great problems. As nicely summarized by Andrew Taylor: 'There is no fundamental physics underlying inflationary models, unlike say the Big-Bang model which is based on General Relativity. Inflation takes many of its ideas from 'quantum field theory', but the fields it uses are not directly related to the known fields of nature. If it turns out to be true, inflation implies there is a wealth of physics beyond what we currently understand'.[16]

In the last decades of the 20th century we could observe two great streams in the study of the universe: the stream of audacious hypotheses and models aiming at solving all cosmological conundrums, and the stream of 'very responsible' investigations based on increasingly rich influx of observational data. The emerging image of the cosmic evolution is composed out of the network of very precise (in numerical sense) details with interspersed vast patches of unknown large-scale features. And an active interaction of both these streams is a promise of the future progress. Let us first focus on what is sometimes called the precision cosmology.

[15] See e.g., S.K. Blau, A.H. Guth, 'Inflationary Cosmology', in: *Three Hundred Years of Gravitation*, eds.: S.W. Hawking, W. Israel, Cambridge University Press, Cambridge, 1987, pp. 524-603.

[16] A Taylor, 'The Dark Universe', in: *On Space and Time*, ed. by Sh. Majid, Cambridge University Press, Cambridge 2008, pp. 1-55, quotation from pp. 16-17.

6. Precision Cosmology

New generations of surface telescopes, Hubble space telescope, technologically advanced radio observatories, and cosmic missions, especially those of COBE, WMAP and Planck satellites, are providing an avalanche of data of great cosmological interest. Owing to these new technologies the wavelength range is covered spanning from radio and optical waves, through microwave to X-rays and gamma rays. Let us enumerate a few fields in which progress is tremendous.

First, *CMB measurements*. Owing to CMB anisotropies on angular scales between 10° (Planck satellite is expected to go with the measurement precision dawn up to 0.1°) and 100° it is possible to determine, with great precision, the amplitude of mass fluctuations at the epoch before nonlinear structures are formed. The power spectrum of temperature fluctuations of CMB carries a lot of information about the early universe. The peaks observed in the CMB power spectrum are due to baryon-photon oscillations driven by the gravitational field. Their precise measurements help establishing (with smaller and smaller errors) important parameters characterizing the cosmological model (see below), and provide information concerning initial conditions for the structure formation.

Second, *mapping three-dimensional large-scale distribution of galaxies*. Two-degree-Field Galaxy Redshift Survey (2dF or 2dFGRS) was conducted by the Anglo-Australian Observatory between 1977 and 2002. The spectra of 245,591 objects were measured, including 232,155 galaxies, 12,311 stars and 125 quasars. Another, even larger program of this kind is the Sloan Digital Sky Survey (SDSS). In over eight years of its operation it has obtained deep, multi-colour images covering more than a quarter of the sky. Consequently, it was possible to create three-dimensional maps containing more than 930,000 galaxies and more than 120,000 quasars. Such programs are not only important for themselves but they also give, in combination with CMB measurements, two 'photographs' of the universe: one 300,000 years after the Big bang, and another at the present epoch.

Third, *gravitational lensing* which allows us to measure directly overdensities of dark matter at moderate red shifts. In a typical one square degree piece of the sky there are million of galaxies, and dark matter density perturbations systematically disturb images of these background galaxies along the line of sight. This turns out to be a very effective method of observing what is not observable otherwise. Numerical simulations are a bridge between these observations and theoretical models.

Fourth, *measuring cosmological parameters* (by using the above and other methods). They are extremely important for determining this cosmological

model (or this subclass of cosmological models) that best fit observational data. Here are some recent results:[17]

Hubble parameter $h = 0.72 \pm 0.03$ in units of 100 Km s^{-1} Mpc^{-1},
total matter density $\Omega h^2 = 0.133 \pm 0.006$,
baryon density $\Omega_b h^2 = 0.0227 \pm 0.0006$,
cosmological constant $\Omega_\Lambda = 0.74 \pm 0.03$,
radiation density $\Omega_r h^2 = 2.47 \times 10^{-5}$,
density perturbation amplitude $= 2.41 \pm 0.11) \times 10^{-9}$.

From the above results quite a detailed picture of the large-scale universe emerges; its age being equal to 13.73 billion years (with an uncertainty of 120 million years), and its space flat with only a 2% margin of error. The evolution of chemical elements and the evolution of structures are also well understood. The standard cosmological model is no longer reduced to the 'geometry of the Universe', i.e. to solving Einstein's equations and trying to select a solution (or a family of solutions) that best fits observational data (as it was in the first half of the 20th century), but becomes a true 'physics of the universe'. In fact, there are almost all physical theories that participate in reconstructing processes involved in shaping the structure and evolution of the universe. One often speaks about cosmology as an effective physical theory, i.e., as the result of cooperation of various physical theories or models that contribute effectively to a coherent image of the world.

5. Unsolved Problems in Cosmology

There are two sources of 'learned ignorance' in cosmology. One is the fact that the more details we know, the more penetrating questions we ask. The second follows from the very nature of cosmology – the science of totality. Facing totality it is hard not to ask questions that cannot be kept within the constraints of the empirical method. Let us start with those of a more technical character. As we shall see, they 'smoothly' go on to more metaphysical ones.

1. *The cosmological constant problem*. The essence of the problem consists in the fact that if we agree that the cosmological constant is related to the energy

[17] They are taken from: O. Lahav, 'The Cosmological Parameters', arXiv:1002.3488 [astro-ph.CO], 18 Feb. 2010. Interesting review: W.L. Freedman, 'Determination of Cosmological Parameters' (Invited review given at the Nobel Symposium, 'Particle Physics and the Universe'), Haga Slott, Sweden, August, 1998, *Physica Scripta*, T85, 2000, 37-46.

of quantum vacuum, then the current quantum field theories predict its value to be bigger by a factor of 10^{120} than the value implied by cosmological observations. This can be regarded as 'the worst theoretical prediction in the history of physics'.[18] So far there is no convincing explanation of this discrepancy. However, quite popular is the proposal made by Steven Weinberg to explain it by appealing to the anthropic principle.[19] If we suppose that vacuum energy assumes different values in different domains of the universe, then observers (such as ourselves) could only live in the domains in which the vacuum energy is similar to what we actually observe (such domain would be very rare). Essentially, large negative values of vacuum energy would imply a closed universe with too short a time for life to emerge; and large positive values would prevent galaxies and stars to form, which seem necessary for the evolution of life. A modification of Weinberg's proposal is to consider various 'parallel universes' instead of various domains of the same universe.

2. *The matter-antimatter asymmetry problem*. In the present universe we observe the strong asymmetry between baryonic and antibaryonic matter. Neither the standard particle model nor relativistic cosmology give us any explanation of this asymmetry. If the Big Bang produced equal amounts of particles and antiparticles (which seems to be a natural assumption), the universe would by now be a sea of photons (as the result of annihilation). How to explain that the universe is composed of matter rather than of only photons or antimatter? A natural reaction to this question is to look for the violation of the CP symmetry in the foundations of quantum field theories, but so far there is neither theoretical nor experimental evidence supporting this solution. Another explanation is again based on an anthropic reasoning and tries to restore the full matter-antimatter symmetry either on the level of various domains of our universe or on the level of the multiverse (a family of parallel universes).

3. *The dark matter problem*. The existence of a dark matter is deduced from the discrepancy between measurements of the mass of luminous objects (stars, galaxies, clusters of galaxies,…) and estimates made with the help of dynamical theories. According to present cosmological data, dark matter is responsible for 23% of the total mass-energy density of the observable universe, whereas ordinary matter only for 4.6%. The remainder is due to dark

[18] M.P. Hobson, G.P. Efstathiou, A.N. Lasenby, *General Relativity: An Introduction for Physicists*, Cambridge University Press, Cambridge, 2007, p. 187.

[19] S. Weinberg, 'Anthropic Bound on the Cosmological Constant', *Phys. Rev. Lett.* 59 (22), 1987, 2607-2610.

energy (see below). Dark matter (at that time called 'missing mass') was suggested, as early as in the thirties of the 20th century, by Fritz Zwicky. It plays an important role in current theories and models of structure formation in the universe and explanations of anisotropies observed in CMB. The very nature of dark matter and its interaction with electromagnetic radiation remain open questions.

4. *The dark energy problem*. Dark energy is a hypothetical form of energy responsible for the increase in the expansion rate of the universe inferred from the recent observations of supernovae of Ia type. It is supposed to account for 74% of the total mass-energy of the universe. The most probable form of dark energy is the cosmological constant corresponding to a constant energy density filling space in a homogeneous manner. Another possibility is provided by scalar fields, sometimes called quintessence; their energy density can vary in space and time. Determining the equation of state for dark energy (i.e., a relationship between energy density, pressure and vacuum energy density) is one of the biggest challenges of today's cosmology. The so-called Lambda CDM model, based on a FLRW solution to Einstein's equations with the cosmological constant, gives very good agreement with current observational data. The cosmological constant (Lambda) in this model is interpreted as responsible for dark energy.

5. *The horizon problem* (see above, section 5). The essence of the problem consists in the fact that even distant places in the universe exhibit the same (up to a high precision) characteristics, e.g., the same CMB temperature, in spite of the fact that they have never been in the causal contact (since the Big Bang). Some 300,000 years after the Big Bang photons decoupled from the other forms of matter. At that time the volume of a 'causally connected' region was about 900,000 light years across, whereas today the CMB temperature is the same over the entire sky, the volume of which is 10^{88} times larger. The most common explanation of this huge discrepancy appeals to the inflation in the very early history of the universe, which blew up a tiny causally connected region to the size containing the present observable universe. It was one of the reasons of the origin of the inflationary scenario.

6. *The inflation problem*. Inflationary models not only solve the horizon problem, but also some other problems of traditional cosmology (the matter-antimatter problem, the flatness problem, the horizon problem), but they themselves require rather special initial conditions, and still there is no decisive empirical evidence for the existence of inflation in the very early period of cosmic evolution. There are so many scenarios of inflation that – as it is sometimes claimed – at least one of them must fit empirical data. The truth is that inflation, when combined with the Cold Dark Matter (CDM)

scenario (which is today the most favoured one) predicts Gaussian perturbations, independent of scale, in the early universe. The model is characterized by at least ten parameters. This indeed creates serious problems as far as empirical testing is concerned, but with the abundance of high precision data, which continue to come, this also fosters hopes for solving riddles of the early universe. As noticed by George Ellis, 'It is this explanatory power that makes it [the inflationary scenario] so acceptable for physicists, even though the underlying physics is neither well-defined nor tested, and its major large-scale observational predictions are untestable'.[20]

7. The quantum cosmology problem. This is a subproblem of a major problem important for the entire physics, namely of the quantum gravity problem. The physics of the microworld is dominated by quantum mechanics and quantum field theories, whereas the physics of the world on the large scale (including cosmology) is dominated by general relativity. There is common agreement that on the fundamental level these two physical theories should be unified into one theory, called quantum gravity theory. In the very early stages of cosmic evolution, extremely strong gravitational fields had to reveal its quantum nature, and standard cosmology should be replaced by quantum cosmology. The most developed approaches to quantum gravity are superstring theory (and its newer incarnation, M-theory) and loop quantum gravity. The former attempts to unify quantum physics, gravitational physics and the physics of other fundamental interactions (electromagnetism, and strong and weak nuclear forces), whereas the latter simply quantizes the gravitational field, putting aside other interactions. Other attempts include: supergravity, causal dynamical triangulation, Regge calculus, twistor theory, causal sets, and approaches based on noncommutative geometry and quantum groups.

7. *The multiverse problem*. As we have seen, the concept of multiverse was invoked as a possible explanation for various problems. The concept of the multiverse itself changes depending on the problem that it is supposed to solve. The main concern related to the multiverse idea is whether it can be regarded as belonging to the realm of science. This concern comes from the fact that the concept itself is very fuzzy and has hardly any observational consequences. George Ellis justly remarks: 'Choices are needed here. In geometrical terms, will it [the multiverse] include only Robertson-Walker models, or more general ones (e.g. Bianchi models, or models without symmetries)? In gravitational terms, will it include only General Relativity, or also brane

[20] G.F.R. Ellis, 'Issues in Philosophy of Cosmology', arXiv:astro-ph/0602280 v2.

theories, models with varying *G*, loop quantum gravity models, string theory models with their associated possibility 'landscapes', and models based on the wave function of the universe concept? Will it allow only standard physics but with varying constants, or a much wider spectrum of physical possibilities, e.g. universes without quantum theory, some with five fundamental forces instead of four, and others with Newtonian gravity? Defining the possibility space means making some kinds of assumptions about physics and geometry that will then apply across the whole family of models considered possible in the multiverse, and excluding all other possibilities'.[21]

As far as the empirical testability is concerned, the problem arises of whether we may trade testability for the explanatory power of the multiverse idea. But the explanatory power of this idea is still highly debatable.

8. Summary and Perspectives

The history of cosmology in the last one hundred and fifty years or so is a beautiful example of how wild speculations and unverifiable hypotheses could change into reliable science. Before that time the only link of cosmology with reality was local physics. One thing was known for sure: that the universe at large must be such as to allow for what we know about here and now. An extrapolation from here and now was the main strategy of cosmological speculations.

There is a certain regularity in the history of science. In every epoch a hard core of scientific theories is surrounded by a vast band of speculations that are controlled mostly by common sense intuitions and philosophical prejudices. Some of these speculations may inspire fruitful ideas and, after suitable transmutations, become elements of genuine science, but the majority of them will turn out to be alleys leading to nowhere and will be forgotten or quoted in footnotes of the history of science.

In the last decades of the 19th century something started to happen in the bordering zone between the hard core of physics and the surrounding belt of cosmological speculations. Advances in astronomy and pressing questions from the Newtonian gravity put the case of cosmology on the market. Various speculations generated by the discovery of non-Euclidean geometries and adding more space dimensions to the traditional three turned the attention of some physicists and philosophers to the problem of the spatial arena for physical processes. The latter concept soon evolved, in the works of Einstein and Minkowski, into the concept of space-time, which in itself

[21] *Ibid*.

had cosmological connotations. The birth of Einstein's general theory of relativity created a new theoretical context for thinking about the universe. Some concepts so far dwelling in the belt surrounding science started to infiltrate the domains of physical theories. The process was slow and painful. Cosmology already had a consolidating conceptual basis, but was still dominated by philosophical presuppositions and prejudices. In the beginning of this process the observational basis of cosmology was rather fragile (only galactic red shifts with an uncertain Doppler interpretation and very rough data concerning the uniform distribution of matter), but as this basis gradually strengthened, some philosophical assumptions changed into working, observationally motivated models.

After the breakthrough of the sixties and the influx of observational data (trigged mainly by the discovery of CMB), the status of cosmology as a scientific discipline was established, and the standard model of the universe quickly started to emerge. This process was accompanied by a gradual coalescence of cosmological theories and models with other branches of physics. The fact that cosmological theories and models have to make use of physical phenomena studied in other departments of physical sciences is rather obvious, but the fact that also other physical theories (such as, for example, the theory of elementary particles) found this useful to conduct their investigation in a purposefully chosen cosmological context, was certainly a sign of the acceptance of cosmology into the family of empirical sciences.

This process substantially strengthened when, in the last decades, 'precision cosmology' entered the scene. Today, several programs of key significance for physics and astronomy could not even be imagined without the 'cosmological background'. It is enough to mention the program of unification of all fundamental forces and the theory of the origin and evolution of complex structures. Cosmology itself has become what is now called effective physical theory, i.e., the field of the effective cooperation of many physical disciplines.

The result is impressive: a very detailed, although still incomplete, scenario of cosmic evolution. Major problems plaguing cosmology (which are enumerated in sections 5 and 7) testify to this incompleteness. However, these problems could also be regarded as achievements of cosmology. Yet a few decades ago some of them could not even be suspected, and to replace an invisible hole with an open question is certainly a step in the right direction.

Since the majority of these problems are difficult and some of them balance on the verge of the scientific method, they easily give rise to new layers of the speculative belt surrounding current research in cosmology. Do some newer ideas inhabiting this belt have a chance of becoming full-fledged cos-

mological models some day? Let us compare the present situation with that from before the beginning of relativistic cosmology. At that time there were many ideas, which had to be regarded as inhabiting the belt, that by now are completely forgotten. For instance, the hypothesis invented and propagated by the Scottish physicist William Rankine who claimed that when radiant heat reaches a boundary of the interstellar medium (beyond which there is only empty space), it is reflected and could concentrate again, in this way averting the otherwise unavoidable thermal death of the universe.[22] On the other hand, some other bold ideas had matured and were fully incorporated into the body of science. Typical examples are vague speculations around non-Euclidean geometries and multidimensional spaces which have changed into indispensable tools of physical theories.

Something similar is to be expected as far as present speculations about 'unsolved problems in cosmology' are concerned. I think that such problems as the cosmological constant problem, dark matter and dark energy problems will sooner or later be solved, and their solutions will contribute to establishing new reliable cosmological scenarios. It could also be that some 'unsolved problems' will simply be liquidated by new more precise measurements, as it happened in the fifties of the last century with the age of the universe problem. The age of the universe as computed from the Hubble law was at that time shorter than the age of some rocks on Earth, meteorites and certain stellar systems. This was one of the main difficulties in relativistic cosmology in first decades of its existence. In 1955 Walter Baade discovered a systematic error in determining distances to galaxies, and when that error had been corrected the problem disappeared.

The fate of some other speculations is more debatable. I have in mind the whole collection of speculations related to the multiverse idea. I think two factors should be taken into account. First, as remarked in section 7, the multiverse concept is very vague, and in order to be scientifically productive it must be suitably narrowed (this in fact happens in the works of many authors). Some more precisely defined 'parallel universes' possibly have a chance to fulfil a positive role (at least a heuristic one) in cosmology,[23] provided that – and it is the second factor – some methodological standards

[22] See, H. Kragh, *Matter and Spirit in the Universe*, p. 47. Chapters 1 and 2 give more examples of this kind.

[23] George Ellis in his paper 'Multiverses: Description, Uniqueness and Testing' (in: *Universe or Multiverse*, ed. by B. Carr, Cambridge University Press, Cambridge 2007, pp. 387-409) discusses possible restrictions that could render the multiverse concept scientifically more acceptable.

in cosmology and in fundamental physics will somehow be relaxed. The present paradigm in cosmology is that of physics, and physical methodology is based on empirical testability. If we stick to this paradigm, the multiverse methodology has negligible chances. If we admit that, in some cases, an explanatory power (in a non-empirical sense or in a relaxed empirical sense) could supplement accepted strategies in the physical sciences then some versions of the multiverse strategy could possibly be viewed more favourably in the domain of science. This is a philosophical option, and there are signs that such an option is indeed slowly gaining acceptance. This is understandable in the sense that when theoretical curiosity reaches the limits of experimental possibilities, there is no other way round than to look for help in bold speculations. Although some people claim that when this happens, science starts to decline.

Tables

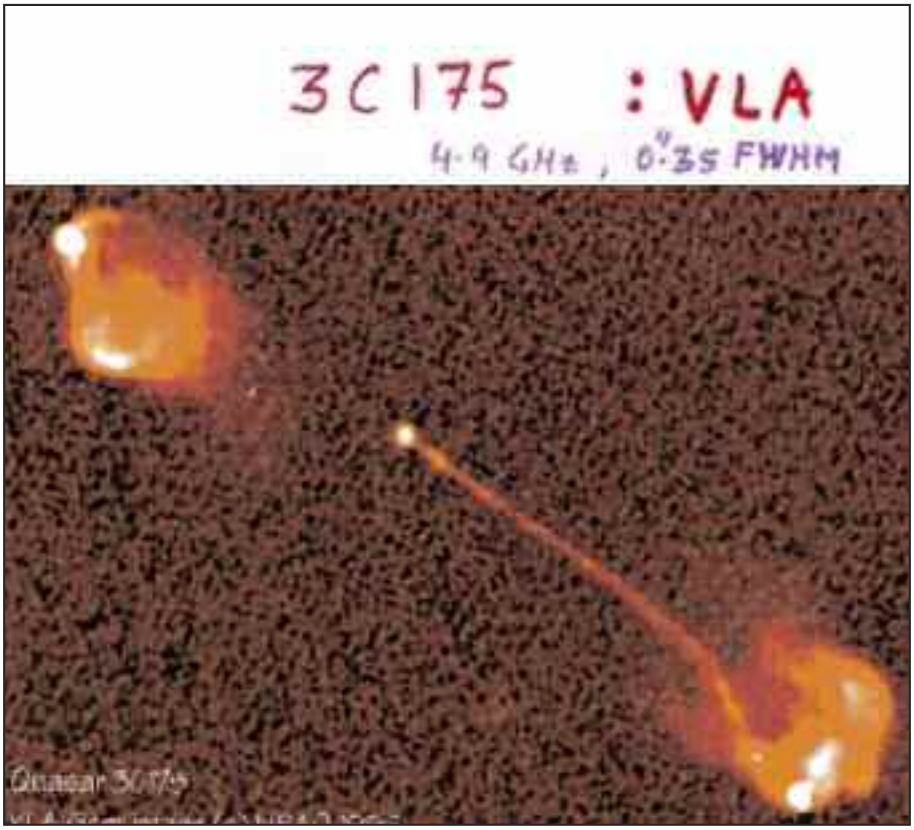


Figure 1. A radio map of the Quasar 3C175 made by Bridle *et al*. [16] with a resolution of 0.35 arcsec, using the Very Large Array (VLA), showing a central compact radio source coincident with an optical galaxy at a redshift z =0.768, a jet of radio emission on one side (most likely towards the observer) and the two outer radio lobes. The lobes occur as the relativistic plasma ejected by the central active galactic nuclei in two opposite directions suffers shock by the intergalactic medium. The overall linear size of 3C175 is about 1 million light years.

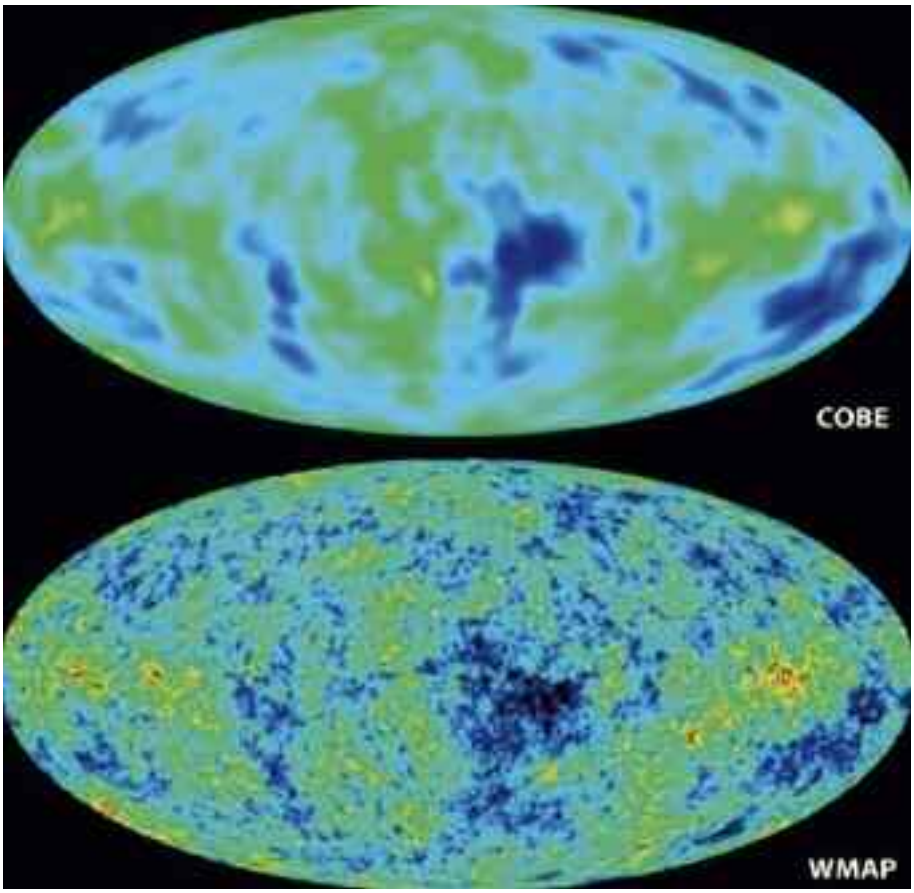


Figure 4. Fluctuations were detected in the Cosmic Microwave Background Radiation of about 1 part in 100,000 by the COBE satellite in 1992, as shown in the upper panel with an angular resolution of ~10 degree [24]. The WMAP observations shown in the lower panel have sufficient angular resolution of ~0.3 degree that clarifies detailed structure up to sub-horizon scale [25].

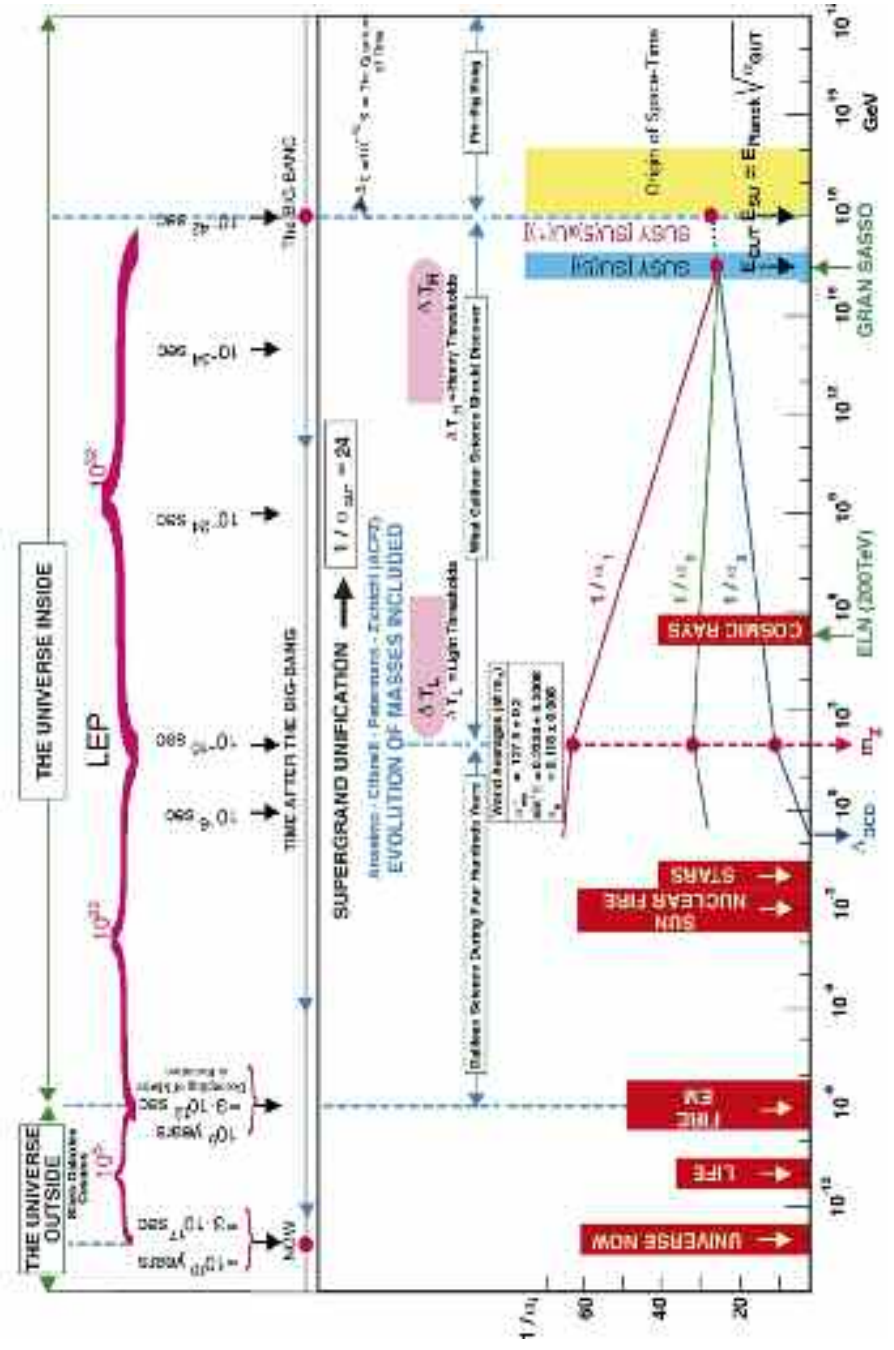

THE UNIVERSE INSIDE
THE UNIVERSE OUTSIDE
LEP
THE BIG BANG
NOW
TIME AFTER THE BIG-BANG
SUPERGRAND UNIFICATION
EVOLUTION OF MASSES INCLUDED
Origin of Space-Time
SUSY (SUSY-GUT?)
SUSY (susy!)
GRAN SASSO
COSMIC RAYS
SUN NUCLEAR FIRE STARS
UNIVERSE NOW
LIFE
TIME
GeV

Figure 9.

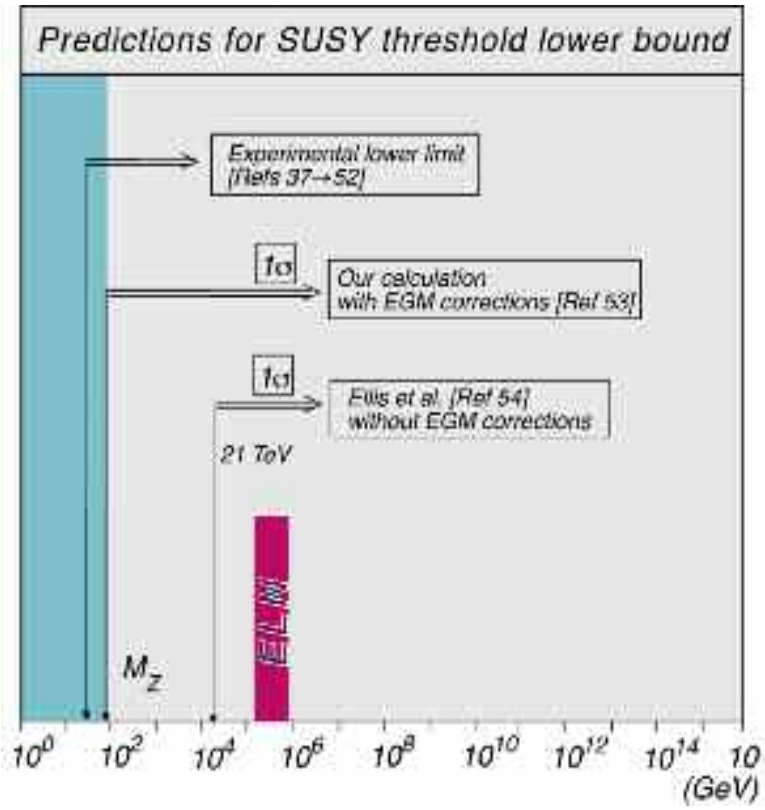

Predictions for SUSY threshold lower bound
Experimental lower limit
[Refs 37→52]
1σ
Our calculation
with EGM corrections [Ref 53]
1σ
Ellis et al. [Ref 54]
without EGM corrections
21 TeV
EGM
M_Z
10^0 10^2 10^4 10^6 10^8 10^{10} 10^{12} 10^{14} 10
(GeV)

Figure 10.

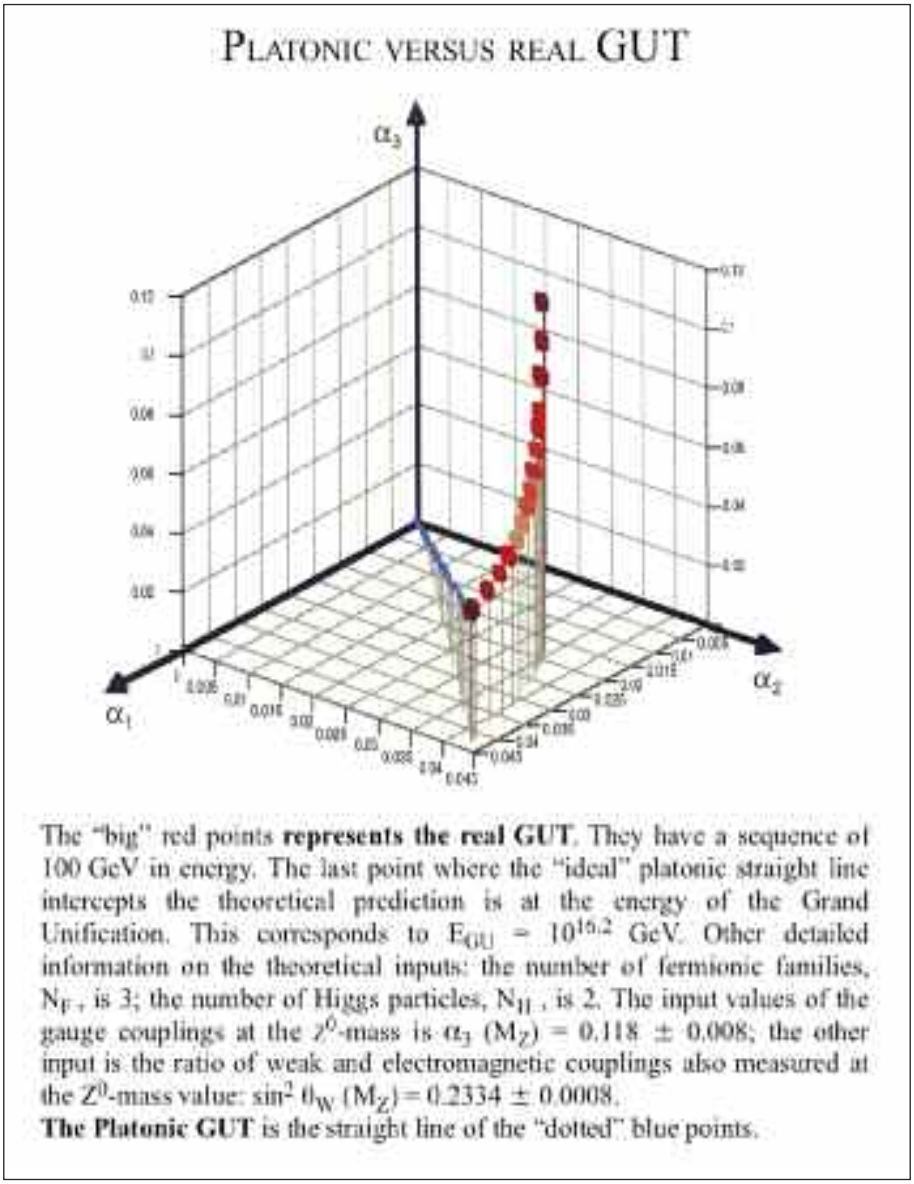

PLATONIC VERSUS REAL GUT
α_3
α_1
α_2
The "big" red points **represents the real GUT**. They have a sequence of 100 GeV in energy. The last point where the "ideal" platonic straight line intercepts the theoretical prediction is at the energy of the Grand Unification. This corresponds to $E_{GU} = 10^{15.2}$ GeV. Other detailed information on the theoretical inputs: the number of fermionic families, N_F, is 3; the number of Higgs particles, N_H, is 2. The input values of the gauge couplings at the Z^0-mass is $\alpha_3 (M_Z) = 0.118 \pm 0.008$; the other input is the ratio of weak and electromagnetic couplings also measured at the Z^0-mass value: $\sin^2 \theta_W (M_Z) = 0.2334 \pm 0.0008$.
The Platonic GUT is the straight line of the "dotted" blue points.

Figure 12.

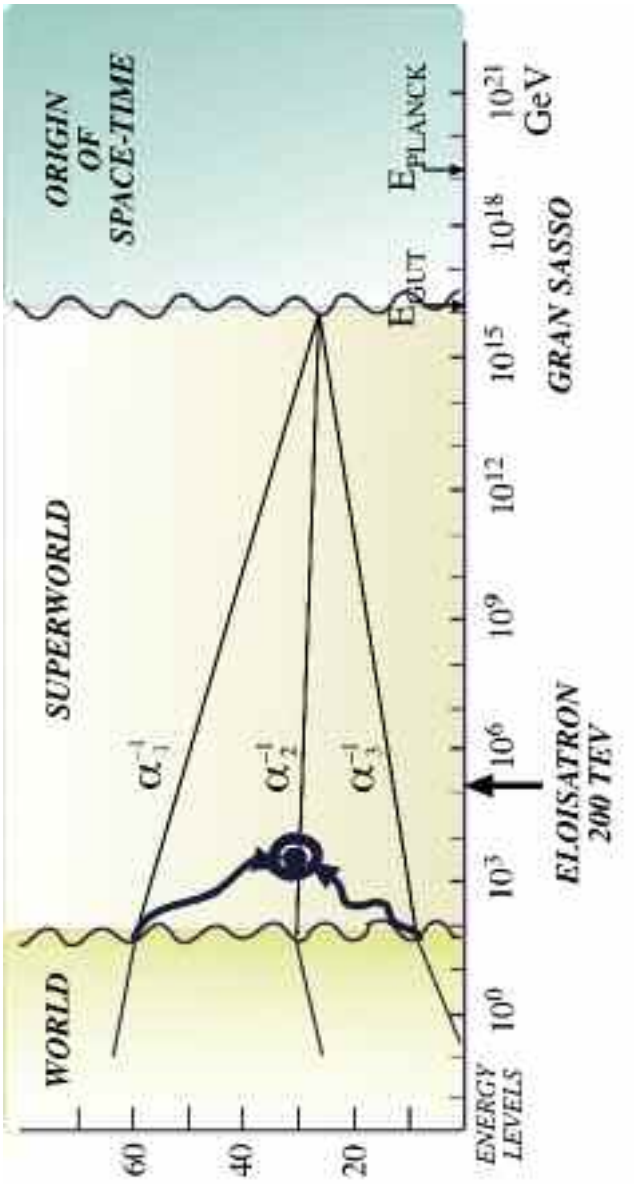

ORIGIN
OF
SPACE-TIME
SUPERWORLD
WORLD
α_1^{-1}
α_2^{-1}
α_3^{-1}
60
40
20
ENERGY
LEVELS
10^0
10^3
10^6
10^9
10^{12}
10^{15}
10^{18}
10^{21}
GeV
E_{GUT}
E_{PLANCK}
ELOISATRON
200 TEV
GRAN SASSO

Figure 18.

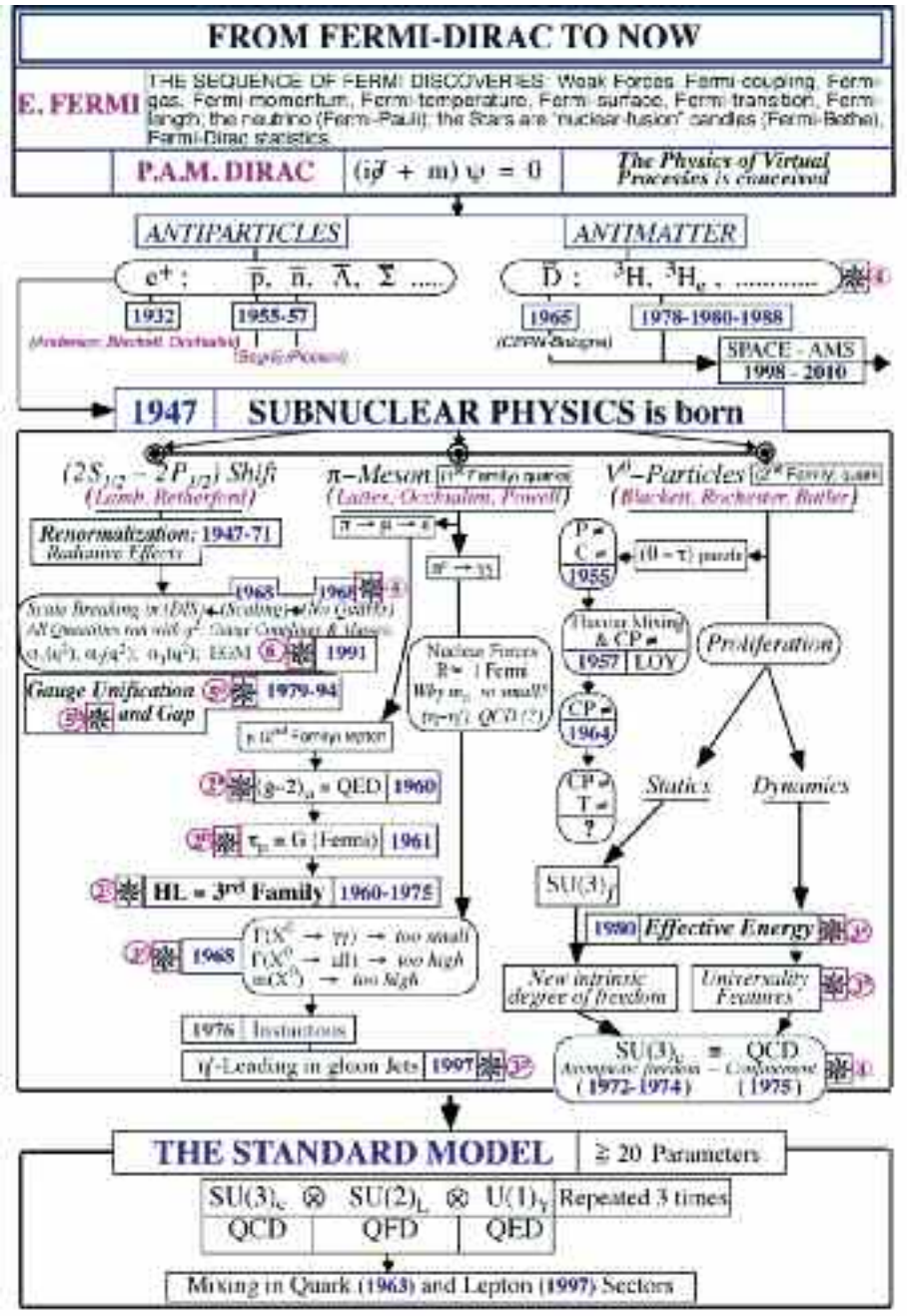

FROM FERMI-DIRAC TO NOW
E. FERMI
THE SEQUENCE OF FERMI DISCOVERIES: Weak Forces: Fermi coupling; Fermi gas; Fermi momentum; Fermi temperature; Fermi surface; Fermi transition; Fermi length; the neutrino (Fermi-Pauli); the Stars are "nuclear-fusion" candles (Fermi-Bethe); Fermi-Dirac statistics.
P.A.M. DIRAC
$(i\partial\!\!\!/ + m)\psi = 0$
The Physics of Virtual Processes is conceived
ANTIPARTICLES
ANTIMATTER
e^+ ; $\bar{p}$, $\bar{n}$, $\bar{\Lambda}$, $\bar{\Sigma}$ ....
$\bar{D}$; $^3\bar{H}$, $^3\bar{H}_e$, ...
1932
1955-57
1965
1978-1980-1988
SPACE - AMS
1998 - 2010
1947
SUBNUCLEAR PHYSICS is born
$(2S_{1/2} - 2P_{1/2})$ Shift
(Lamb, Retherford)
Renormalization; 1947-71
Radiative Effects
π−Meson
(Lattes, Occhialini, Powell)
V⁰−Particles
(Blackett, Rochester, Butler)
Nuclear Forces
R ≈ 1 Fermi
Why m_π so small?
$(\eta - \eta')$: QCD (?)
Flavour Mixing & CP
1957 LOV
Proliferation
CP
1964
CP T ?
Statics
Dynamics
$(g-2)_\mu$ = QED 1960
τ_μ = G (Fermi) 1961
HL = 3rd Family 1960-1975
1968
1976 Instantons
1980 Effective Energy
New intrinsic degree of freedom
Universality Features
SU(3)c ≡ QCD
Asymptotic freedom – Confinement
(1972-1974) (1975)
THE STANDARD MODEL
≥ 20 Parameters
$SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$ Repeated 3 times
QCD
QFD
QED
Mixing in Quark (1963) and Lepton (1997) Sectors

Figure 22.

Figure 1. Map of the Everglades National Park (ENP) study area. Figure taken from Todd *et al.* (2010).

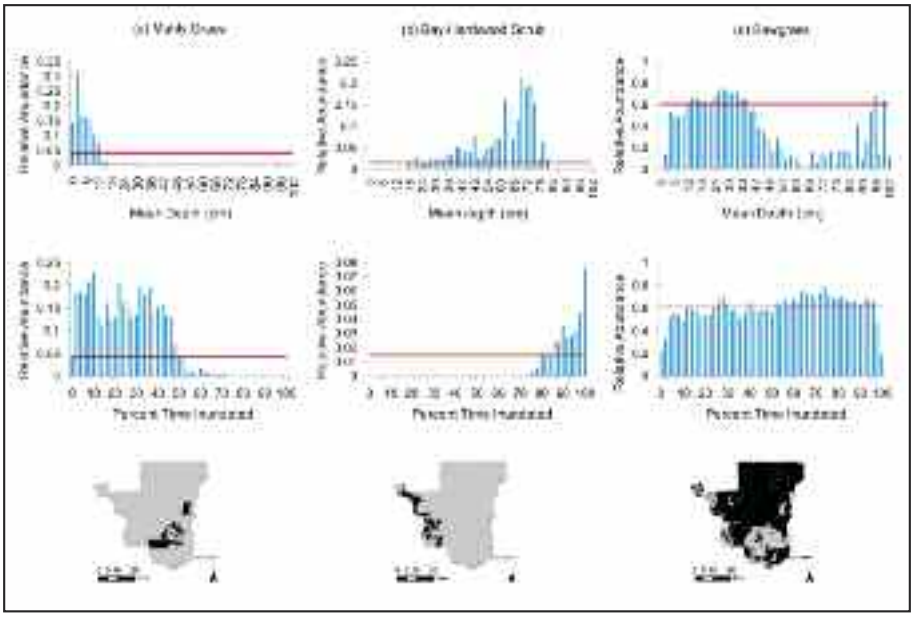


Figure 2. Relative abundance of mean depth, relative abundance of percent time inundated, and spatial distribution of three vegetation types: (a) Muhly grass; (b) Bay-Hardwood scrub; and (c) Sawgrass. The red line indicates the relative abundance of the given vegetation community across the entire landscape. Figure adapted from Todd *et al.* (2010).

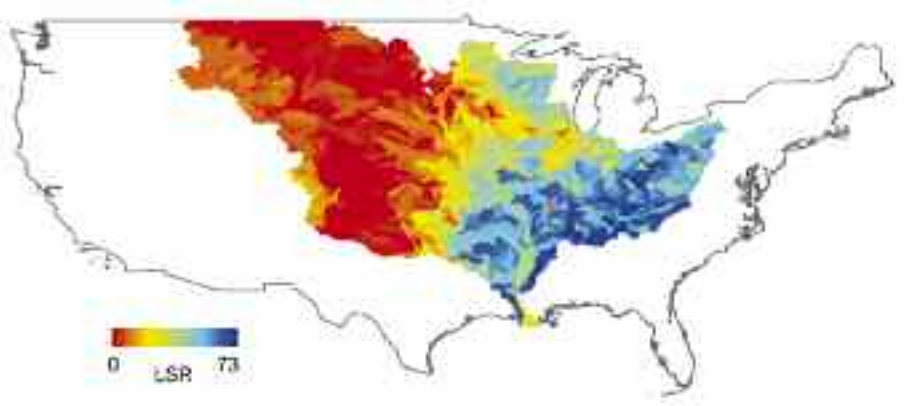


Figure 3. Map of local species richness (LSR) of trees in each direct tributary area (DTA) (that is, at the USGS HUC-8 scale; refer to text) of the MMRS. Taken from Konar *et al.* (2010).

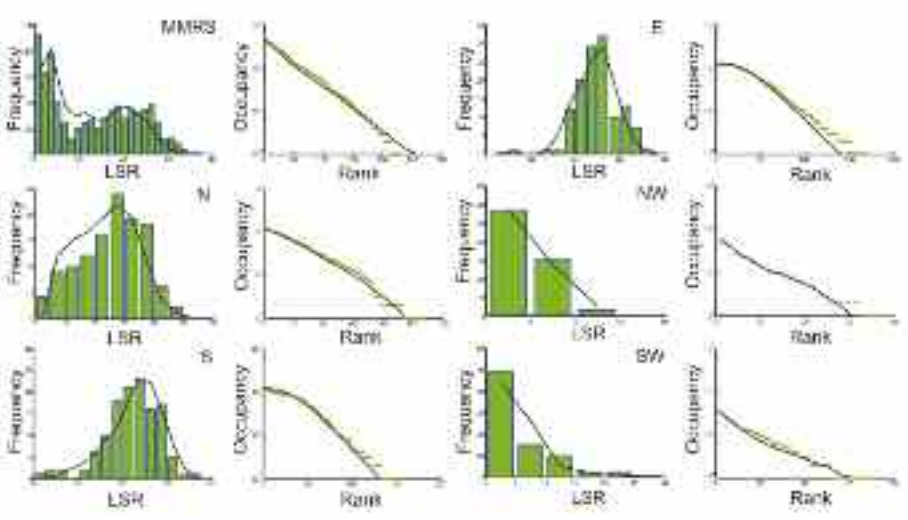


Figure 4. Model fit to empirical patterns of each system. Green shows empirical data; black curves model results. The first and third column illustrate the LSR histogram. The second and fourth column illustrate the rank-occupancy graph. 'MMRS' represents the Mississippi-Missouri River System; 'E' the East subregion; 'N' the North sub-region; 'NW' the Northwest sub-region; 'S' the South sub-region and 'SW' the Southwest sub-region. Refer to Fig. 7 for the spatial extent of each system. Taken from Konar *et al.* (2010).

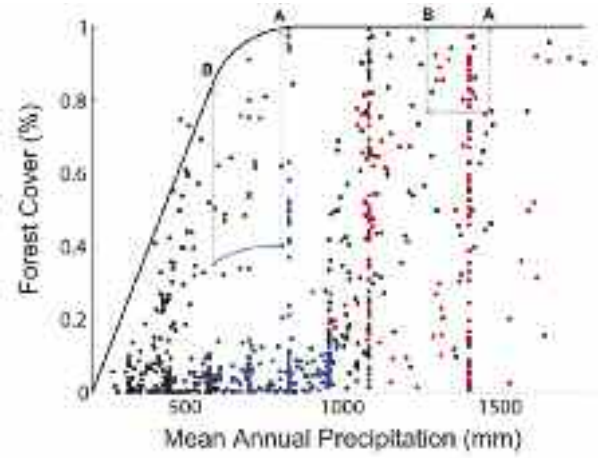


Figure 5. Schematic of how habitat capacity was calculated under climate change. The mean annual precipitation (MAP) for each DTA under every scenario was located on the graph; only data points from the current climate scenario are shown here. The corresponding potential forest cover (Pi) was determined as the upper bound of the function. As an example, points A on the figure indicate the potential forest cover under the current climate scenario, while points B indicate the new potential forest cover under climate change. This new potential forest cover was then multiplied by the forest cover index (Ii) to calculate the habitat capacity under each climate change scenario. This was done for all 824 DTA data points in all 15 climate change scenarios. Blue points indicate DTAs in the North regions; red points the South region; and black points the rest. Taken from Konar *et al.* (2010).

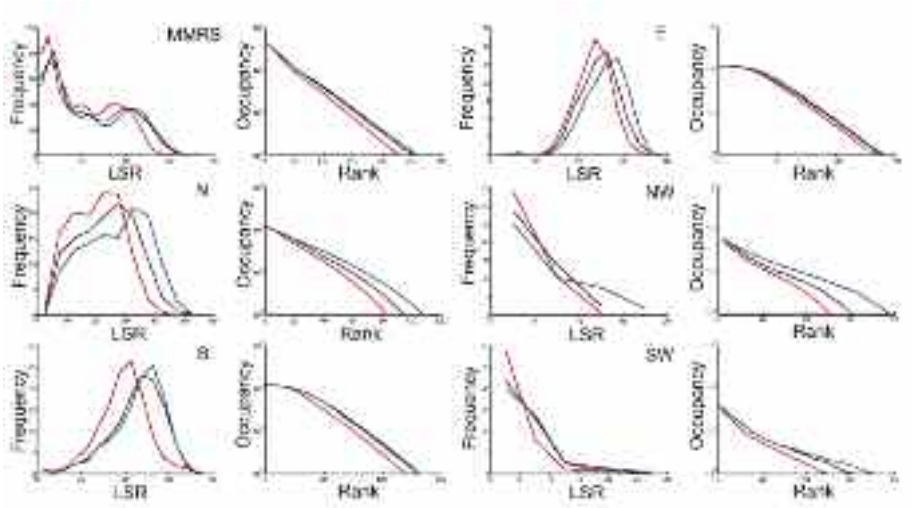


Figure 6. Impact of climate change on the biodiversity patterns of each system. The acroynyms are the same as in Fig. 4. The first and third column illustrate the LSR histogram. The second and fourth column illustrate the rank-occupancy graph. Black curves show model results under the current climate scenario; red curves show the species-poor scenario, and blue curves show the species-rich scenario. Taken from Konar *et al.* (2010).

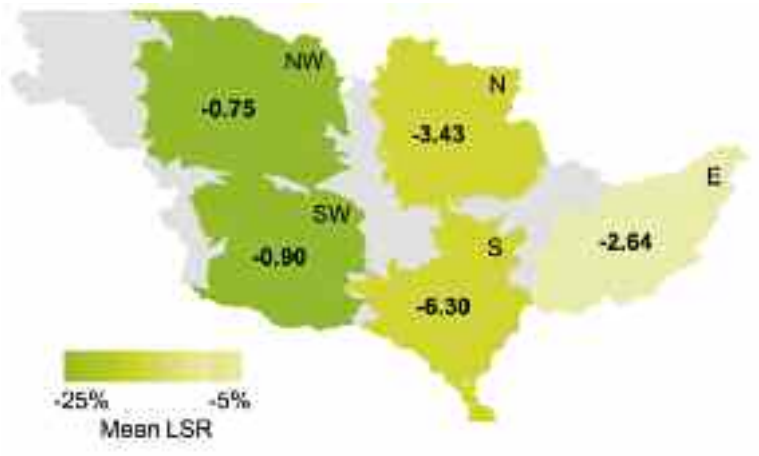


Figure 7. Impact of climate change under the species-poor scenario on region-averaged LSR in sub-regions of the MMRS. The acroynyms are the same as in Fig. 4. Shades of green indicate the percentage change in the region-averaged LSR under climate change, with dark green indicating a higher percentage lost. The general trend is that a higher percentage of species are lost in the west with a decreasing trend to the east. The change per DTA in region-averaged LSR under climate change is indicated for each region by the bold numbers. The species-rich regions east of the 100oW meridian lose more species, though these species represent a smaller percentage of species in these regions. The mean LSR in the South is anticipated to decrease by 6.3 species under climate change, the largest loss of all sub-regions.

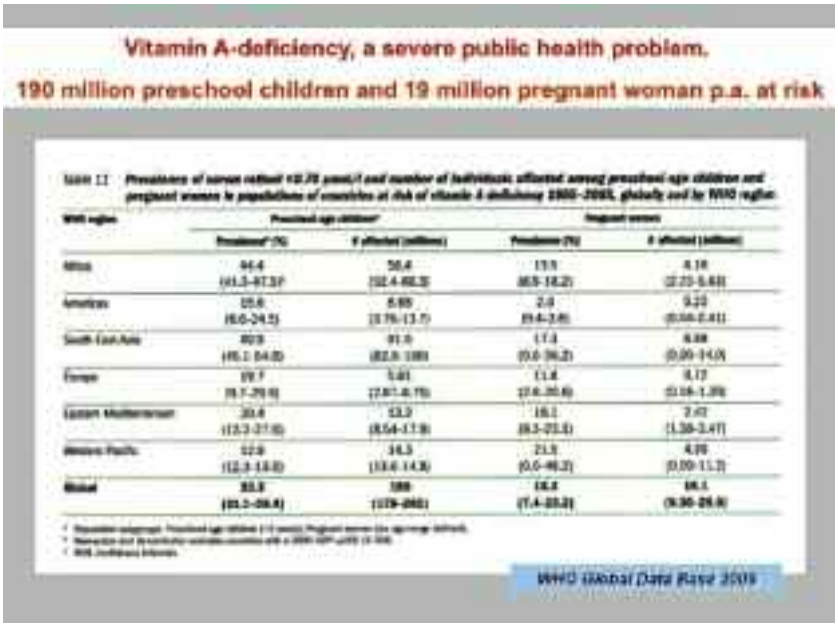


Figure 1.

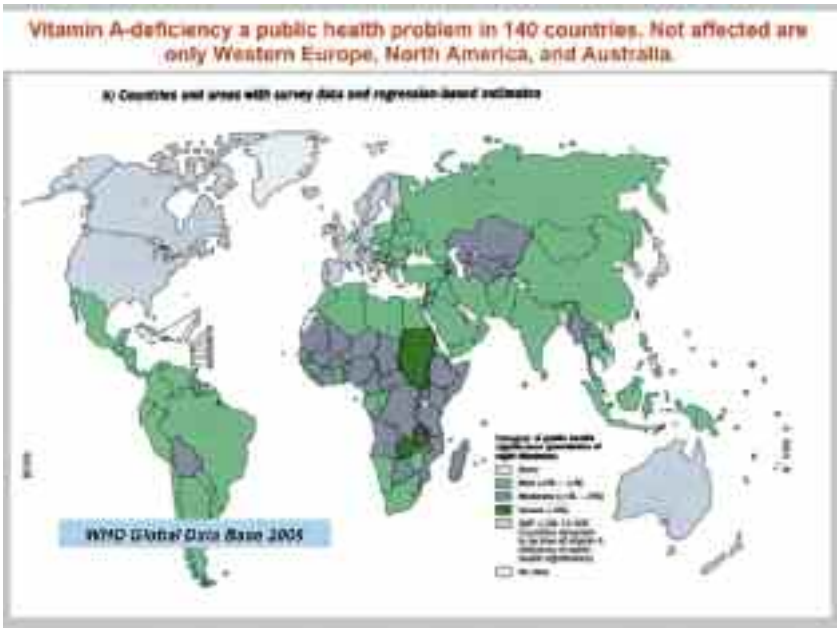


Figure 2.

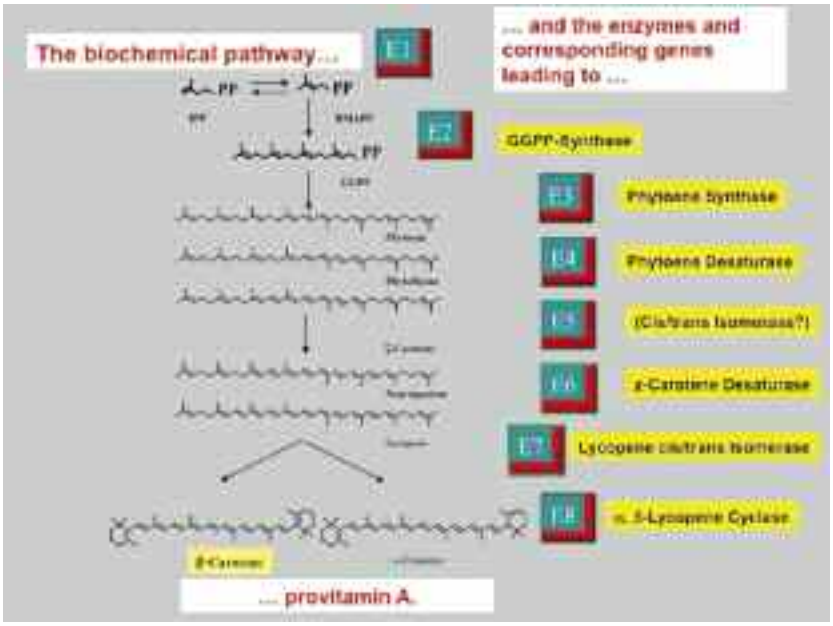


Figure 3.

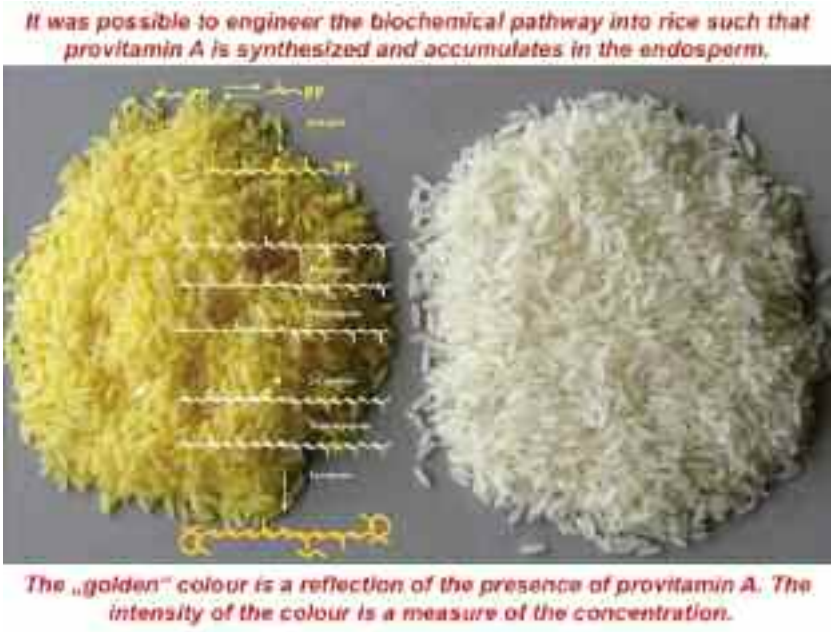
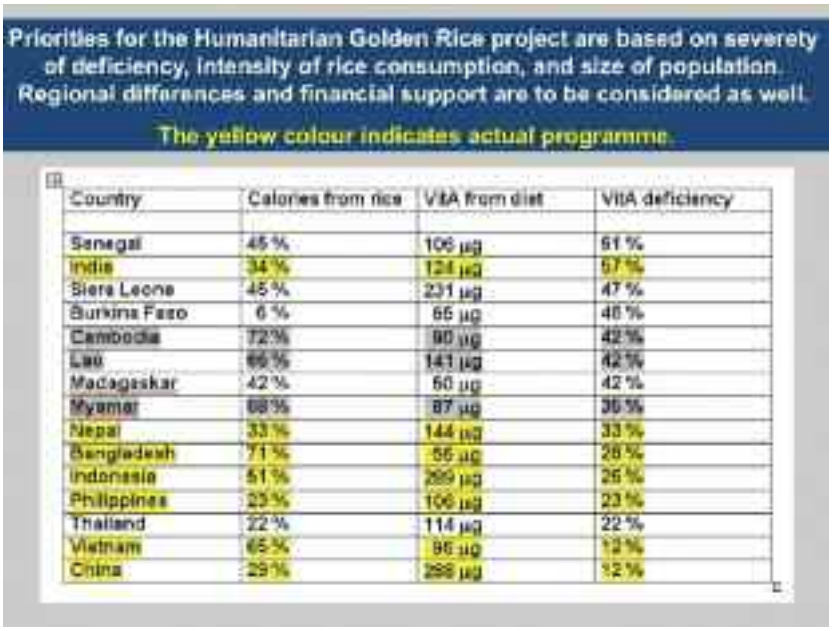


Figure 4.



Country	Calories from rice	VitA from diet	VitA deficiency
Senegal	45 %	106 µg	61 %
India	34 %	124 µg	57 %
Siera Leone	45 %	231 µg	47 %
Burkina Faso	6 %	66 µg	46 %
Cambodia	72%	80 µg	42 %
Lao	69 %	141 µg	42 %
Madagaskar	42 %	60 µg	42 %
Myamar	68 %	87 µg	36 %
Nepal	33 %	144 µg	33 %
Bangladesh	71 %	56 µg	26 %
Indonesia	51 %	289 µg	26 %
Philippines	23 %	106 µg	23 %
Thailand	22 %	114 µg	22 %
Vietnam	45 %	96 µg	12 %
China	29 %	288 µg	12 %

Figure 5.

Figure 6.


Countries where vitamin A-deficiency is a severe public health problem
Afganistan Angola Azerbaijan Bangladesh Benin Bhutan
Bolivia Cameroon Burundi Burkina F. Cambodia Cameroon
Chad Congo Cote Ivoir Korea DR Eritrea Ethiopia
Gambia Georgia Ghana Guinea Haiti Jamaica
Iraq Kazakhstan Kenya Kyrgystan Lao Lesotho
Liberia Madagascar Malawi Mexico Morocco Mozambique
Myanmar Nepal Niger Nigeria Philippines Senegal
Sierra Leone Somalia Sudan Swaziland Sri Lanka Tajikistan
Togo Tuvalu Uganda Ukraine Uzbekistan Yemen
Zambia Zimbabwe
Countries where vitamin A-deficiency is a moderate public health problem
Albania Algeria Argentina Belarus Brazil China
Ecuador Egypt Eq. Guinea Gabon Guatemala Honduras
India Indonesia Jordan Libya Lebanon Lithuania
Mongolia Namibia Pakistan Papua NG Paraguay Peru
Romania Russia Serbia South Africa Suriname Syria
Thailand Tonga Tunisia Turkey Uruguay Viet Nam
Countries with high rice consumption (200-400 g per day)
Countries with high consumption of Cassava, Sorghum, Banana, Potato
All countries would benefit from high iron, zinc, quality protein
Extreme precautionary regulation will prevent all possible solutions
Data from WHO Global Data Base 2009
Biofortification offers the potential to reduce micro-nutrient malnutrition worldwide, substantially, and in a sustained manner.
To be effective, will require a change in public attitude and GE-regulation.

Figure 7.

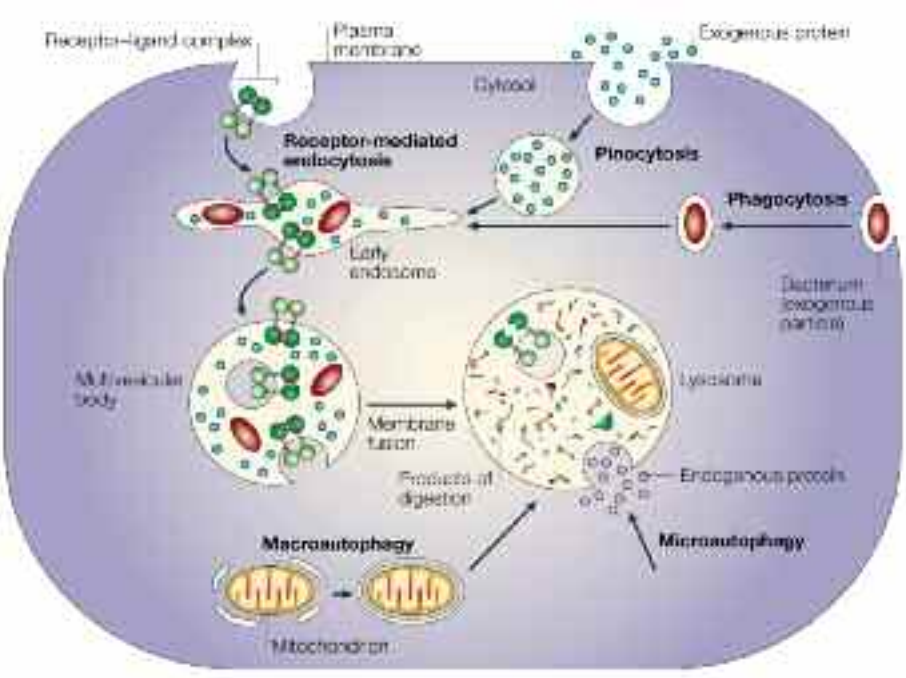


Figure 2: ***The four digestive processes mediated by the lysosome:*** (i) specific receptor-mediated endocytosis, (ii) pinocytosis (non-specific engulfment of cytosolic droplets containing extracellular fluid), (iii) phagocytosis (of extracellular particles), and (iv) autophagy (micro- and macro-; of intracellular proteins and organelles)(with permission from Nature Publishing Group. Published originally in Ref. 83).

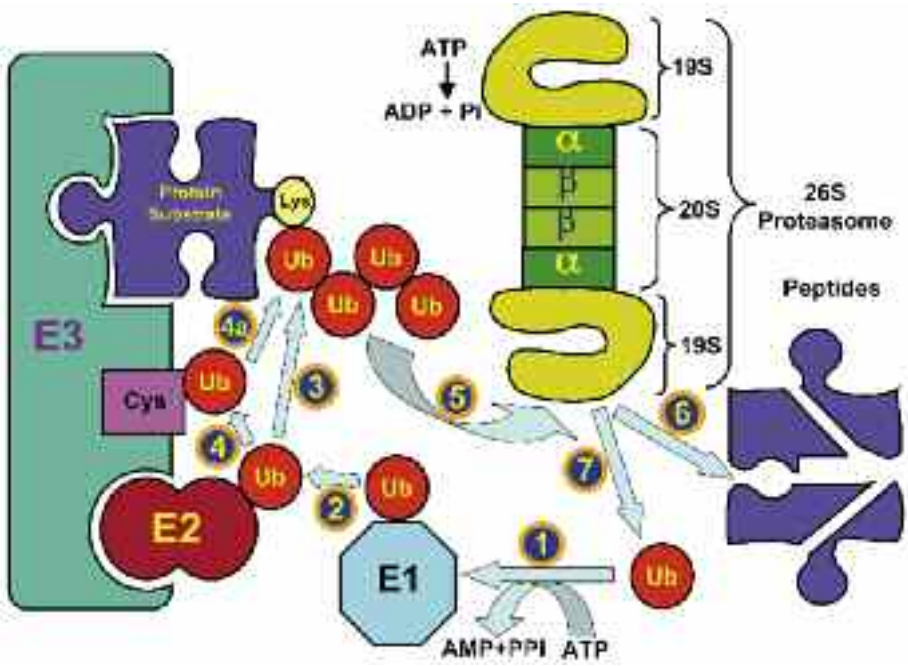


Figure 5: ***The ubiquitin-proteasome proteolytic system.*** Ubiquitin is activated by the ubiquitin-activating enzyme, E1 (1) followed by its transfer to a ubiquitin-carrier protein (ubiquitin-conjugating enzyme, UBC), E2 (2). E2 transfers the activated ubiquitin moieties to the protein substrate that is bound specifically to a unique ubiquitin ligase E3. The transfer is either direct [(3) in the case of RING finger ligases] or via an additional thiol-ester intermediate on the ligase [(4, 4a) in case of HECT domain ligases]. Successive conjugation of ubiquitin moieties to one another generates a polyubiquitin chain that serves as the binding (5) and degradation signal for the downstream 26S proteasome. The substrate is degraded to short peptides (6), and free and reusable ubiquitin is released by de-ubiquitinating enzymes (DUBs)(7).

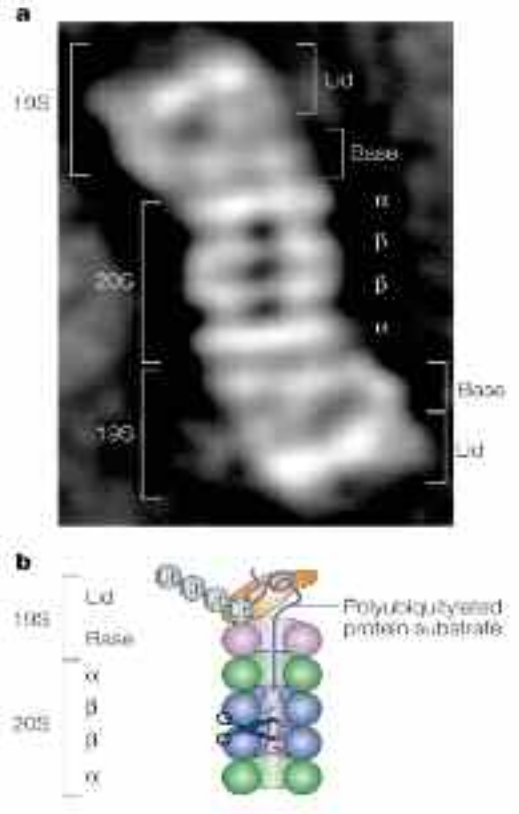


Figure 6: ***The Proteasome***. The proteasome is a large, 26S, multicatalytic protease that degrades polyubiquitinated proteins to small peptides. It is composed of two sub-complexes: a 20S core particle (CP) that carries the catalytic activity, and a regulatory 19S regulatory particle (RP). The 20S CP is a barrel-shaped structure composed of four stacked rings, two identical outer a rings and two identical inner β rings. The eukaryotic α and β rings are composed each of seven distinct subunits, giving the 20S complex the general structure of $\alpha_{1\text{-}7}\beta_{1\text{-}7}\beta_{1\text{-}7}\alpha_{1\text{-}7}$. The catalytic sites are localized to some of the b subunits. Each extremity of the 20S barrel can be capped by a 19S RP each composed of 17 distinct subunits, 9 in a 'base' sub-complex, and 8 in a 'lid' sub-complex. One important function of the 19S RP is to recognize ubiquitinated proteins and other potential substrates of the proteasome. Several ubiquitin-binding subunits of the 19S RP have been identified, however, their biological roles mode of action have not been discerned. A second function of the 19S RP is to open an orifice in the a ring that will allow entry of the substrate into the proteolytic chamber. Also, since a folded protein would not be able to fit through the narrow proteasomal channel, it is assumed that the 19S particle unfolds substrates and inserts them into the 20S CP. Both the channel opening function and the unfolding of the substrate require metabolic energy, and indeed, the 19S RP 'base' contains six different ATPase subunits. Following degradation of the substrate, short peptides derived from the substrate are released, as well as reusable ubiquitin (with permission from Nature Publishing Group. Published originally in Ref. 83). a. Electron microscopy image of the 26S proteasome from the yeast *S. cerevisiae*. b. Schematic representation of the structure and function of the 26SA proteasome.

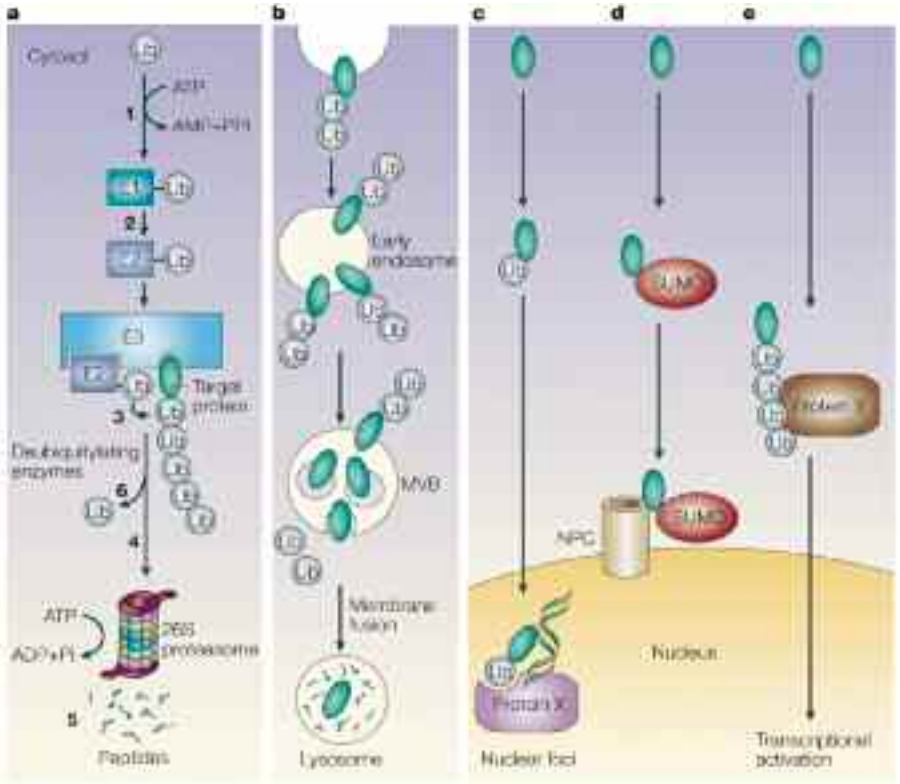


Figure 7: ***Some of the different functions of modification by ubiquitin and ubiquitin-like proteins.*** a. Proteasomal-dependent degradation of cellular proteins (see Figure 4). b. Mono or oligoubiquitination targets membrane proteins to degradation in the lysosome/vacuole. c. Monoubiquitination, or d. a single modification by a ubiquitin-like (UBL) protein, SUMO for example, can target proteins to different subcellular destinations such as nuclear foci or the nuclear pore complex (NPC). Modification by UBLs can serve other, non-proteolytic, functions, such as protecting proteins from ubiquitination or activation of E3 complexes. E. Generation of a Lys^{63}-based polyubiquitin chain can activate transcriptional regulators, directly or indirectly [via recruitment of other proteins (Protein Y; shown), or activation of upstream components such as kinases]. Ub denotes ubiquitin, K denotes Lys, and S denotes Cys. (with permission from Nature Publishing Group. Published originally in Ref. 83).

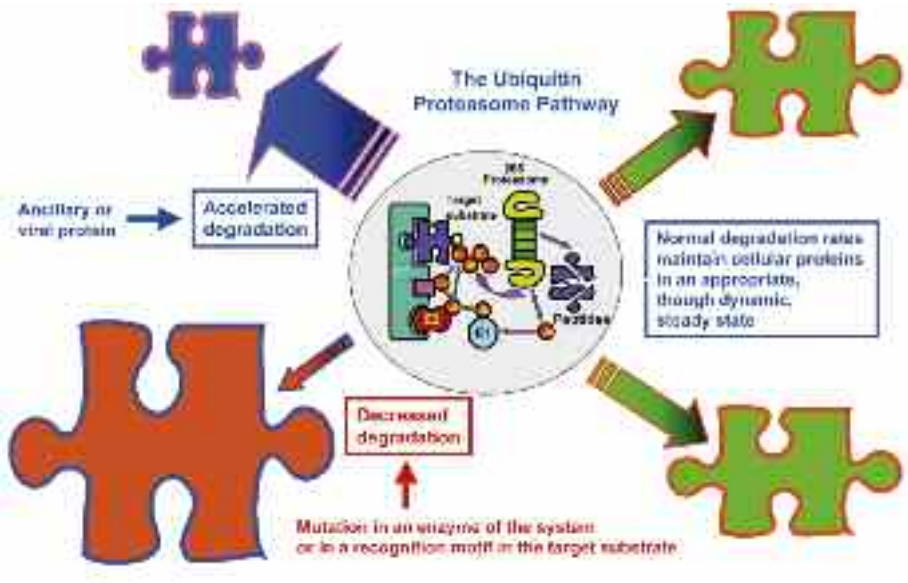


Figure 8: ***Aberrations in the ubiquitin-proteasome system and pathogenesis of human diseases.*** Normal degradation of cellular proteins maintains them in a steady state level, though this level may change under various pathophysiological conditions (upper and lower right side). When degradation is accelerated due an increase in the level of an E3 (Skp2 in the case of p27, for example), or overexpression of an ancillary protein that generates a complex with the protein substrate and targets it for degradation (the Human Papillomavirus E6 oncoprotein that associates with p53 and targets it for degradation by the E6-AP ligase, or the cytomegalovirus-encoded ER proteins US2 and US11 that target MHC class I molecules for ERAD), the steady state level of the protein decreases (upper left side). mutation in a ubiquitin ligase [such as occurs in Adenomatous Polyposis Coli – APC, or in E6-AP (Angelmans' Syndrome)] or in the substrate's recognition motif (such as occurs in β-catenin or in ENaC) will result in decreased degradation and accumulation of the target substrate.