The Challenges of Sciences. A Tribute to Carlos Chagas
Pontifical Academy of Sciences, *Scripta Varia* 103, Vatican City 2002
www.pas.va/content/dam/accademia/pdf/sv103/sv103-arber.pdf

# MOLECULAR EVOLUTION: COMPARISON OF NATURAL AND ENGINEERED GENETIC VARIATIONS

WERNER ARBER

*Abstract*

Genetic engineering implies deliberate alterations of the genomic information of the considered organism. Such alterations may concern a local variation of the inherited nucleotide sequence, it can involve a structural rearrangement of genomic sequences, e.g. by bringing a given gene under a different expression control signal, and it can involve the addition of genetic information taken from another organism. These processes closely ressemble the spontaneous generation of genetic variations, which represent the substrate for natural selection. A critical comparison of similarities and differences in the formation of genetic variations in genetic engineering and under natural conditions of biological evolution can represent an essential help in the evaluation of conjectural risks of genetic engineering.

## *Introduction*

Genetic engineering was developed in the 1970's as a novel research strategy on the basis of scientific knowledge that had mainly been acquired in microbial genetics. In the meantime this strategy has proved to be an extremely powerful tool both in basic genetic research and in biotechnological applications (Sambrook *et al.*, 1989). The strategy includes steps of combining genetic information of sometimes different origin and/or steps of site-directed mutagenesis. These kinds of designed genetic variations are compared here to naturally occurring genetic variations known to be the driving force of biological evolution (Arber, 2000). Such comparison is a welcome contribution to the evaluation of conjectural risks of gene technology.

*Background knowledge*

Genetic information is contained in the linear sequence of nucleotides or base pairs in very long filamentous molecules of DNA. One can compare the genetic message with a written message in our common way of communication, by setting one nucleotide or base pair equal to one letter. In this comparison, the bacterial genome, i.e. the entire genetic information of a unicellular microorganism, corresponds to the content of a book. Depending on the kind of bacteria considered, the book can be more or less thick. Higher organisms have considerably larger genomes. The human genome, for example, corresponds to a library of roughly 1000 books.

The gene is the functional unit of the genetic information. Its size is quite variable from one gene to another. A gene may be as short as 100 nucleotides or as long as many thousand nucleotides, hence corresponding to a few lines or to one or a very few pages of the genetic library. The two essential parts of a gene are, on the one hand, the information serving for directing the synthesis of a gene product (usually a protein) and, on the other hand, the nucleotide sequences serving to control time and efficiency of gene expression, i.e. the synthesis of the gene product.

Following the definition generally used in molecular genetics, we define here as a mutation any alteration of the inherited nucleotide sequence. A mutation may affect a gene by changing either the reading frame serving in the biosynthesis of the gene product or else an expression control signal. In the former case the property of the gene product may become altered, while in the latter case the availability of the gene product may change.

*Major components of gene technology*

Due to its large size, the genome is generally too big to serve directly in analytical studies of gene structure and function. Hence, specific DNA segments must be purified from the bulk of DNA and ideally highly amplified to have enough material available for further studies (Sambrook *et al*., 1989). This can be accomplished as follows. Entire DNA is extracted from the organism under study. This DNA can be cleaved into specific fragments by using restriction enzyme preparations. If needed, the resulting DNA fragments can be specifically sorted out by electrophoretic separation. In a next step DNA fragments can be spliced into a so-called gene vector. This is often a viral DNA molecule or a bacterial plasmid. Both of these are able to replicate in an appropriate host cell. Therefore, if the product of *in vitro* recom-

bination between a vector and a DNA fragment is introduced into a host cell, this hybrid product will undergo replication, i.e. the passenger DNA becomes amplified together with its vector DNA. The reproduced DNA segment of interest can then easily be sorted out and used for further studies.

An efficient alternative method to amplify selected stretches of DNA is the polymerase chain reaction (PCR). This is carried out *in vitro* by using a DNA polymerase preparation. It requires as primers short oligomeric chains of DNA complementary to the two sites on the DNA between which the amplification will occur. Hence, for the PCR amplification short partial DNA sequences flanking the DNA to be replicated must already be known.

Highly amplified segments of DNA can serve for the determination of their nucleotide sequence by using appropriate chemical methods. If this process is carried out with a series of different DNA fragments prepared from a larger DNA molecule, the sequences of the fragments can be joined together like a puzzle. Upon reiteration, this process can eventually result in the availability of the sequence of an entire genome.

A next step in experimental investigations can be a search for open reading frames as potential parts of genes as well as for potential expression control signals for such genes. However, the presence of an open reading frame does not necessarily indicate the presence of a functional gene, nor could it generally reveal what kind of biological function the suspected gene would carry out.

In order to find answers to these questions, mutations are deliberately introduced into the DNA sequences under study. There is a number of particular strategies to do so (Sambrook *et al.*, 1989). For our present discussion it is sufficient to refer in general terms to the principles of such site-directed mutagenesis. By either *in vitro* or *in vivo* methodologies a given DNA sequence can be altered by introducing additional nucleotides, by provoking substitutions, by deleting inherited nucleotides, or by a combination of these strategies.

In any case, the consequences of such manipulations will then be investigated by comparative studies of the functional characteristics of the mutated genetic information and of the unaltered parental DNA. These studies are normally carried out in the concerned living organisms with appropriately altered genomes. Phenotypic differences point to an impairment or an improvement of particular biological functions by the introduced mutation. This can guide the researcher in the identification of the gene function under study.

*Research strategies of classical genetics and of molecular genetics differ*

Investigations of both classical genetics and modern molecular genetics are based on the availability of mutants. However, we must be aware that there are fundamental differences between these two approaches. These differences relate to the definition of a mutation and to the principle of the research approach.

In classical genetics, spontaneous or mutagen induced mutants are first identified by an altered phenotype, hence at the level of a biological function. Inheritance of an altered phenotype into the progeny indicates that an identified phenotypical alteration is due to a mutation on the genome. Genetic crosses between independent mutations serve to establish genetic maps. These are normally linear, but give no hint as to the chemical nature of the carrier of genetic information. Nevertheless, based on today's scientific knowledge, we can conclude that classical genetics starts from the biological function and ends up with a map of sites of mutations on the genome, i.e. on the DNA molecule.

In contrast, investigations in molecular genetics, which is sometimes also called "reverse genetics", start with a segment of DNA and they aim to identify the biological function encoded by that DNA. To do so, the DNA is isolated as we have already discussed, amplified and its nucleotide sequence is determined. The sequence is screened for potential open reading frames, expression control signals and other controlling elements. The inherited sequences are then deliberately altered at expected strategic sites. The altered sequences are compared with the unaltered parental sequences with regard to their effects on biological functions in the concerned organisms. Changes in the phenotype become manifest at that moment, and this can guide the researcher in the identification of the biological function encoded by the DNA sequence in question. Principally this strategy can be applied to any gene as well as to any living organism, although it will not always immediately lead to valid conclusions, depending on the possibility to identify the respective phenotypes.

While in the classical genetic approach the researcher cannot know the molecular nature of the mutation responsible for an altered phenotype, the investigator in reverse genetics designs himself the site-directed mutation and novel sequence combinations, and he can verify the result of his intervention by sequence analysis and other molecular genetic approaches. Any given genetic alteration introduced by genetic engineering into a genome may concern one to a few letters in the genomic library,

in other cases a few lines, a part of a page or at most a few pages. Some of such alterations are additions, others deletions and sometimes substitutions or reshufflings of sequences. In any case, the alterations generally affect a very small part of the genome. Remember that the genome of higher organisms represents a library of a large number of books. As we will discuss later, this strategy of genetic alteration in small steps is quite essential for the maintenance of viability. This has to do with natural selection.

*The impact of genetic engineering on biotechnology*

Efficient new impulses were given to biotechnology by gene technology as it was developed in the 1970's. As a matter of fact, genetic engineering led to strategic changes in biotechnological applications of scientific knowledge.

Classical biotechnology, including agricultural practicing, uses organisms as originally found in nature on the basis of some properties which were identified as useful for human applications. In more recent times breeding between related forms of organisms has served and still serves to obtain more productive and qualitatively more valid recombinants. Thereby, mutagens are sometimes used to increase the chance to find organisms with improved properties. In these investigations, systematic screening for phenotypes serves to identify rare derivatives responding to the expectation, while neither the relevant DNA sequence changes nor any other mutations resulting from the applied mutagenesis are usually identified.

In contrast, modern biotechnology relying on genetic engineering is scientifically much more straightforward and precise. First of all, biotechnological applications can profit of knowledge on specific biological functions as revealed by basic research in molecular genetics. Secondly, genetic engineering can help to improve a particular gene product, if relevant, by site-directed mutagenesis within the open reading frame of a particular gene. Similarly, by genetically altering an expression control signal, the yield of a gene product can be improved. In this latter case, this can, for example, imply the fusion of a reading frame with another, unrelated expression control signal. Finally, and most importantly for some biotechnological applications such as the biosynthesis of natural products for the use as medical drugs, the relevant genes can be introduced into the most appropriate organism for the synthesis of the gene product in question, rather than to use the natural host of the relevant gene.

*Conjectural risks of genetically modified organisms*

So far, strict natural laws which could enable the researcher to precisely predict the functional consequences of any given specific alteration in the genome sequences are not available. Gene technology and its application in biotechnology have thus to rely on experimental investigations of the outcome of any intervention in the genetic outfit of an organism. Predictions and expectations are often fulfilled, but with many exceptions. For example, a possible consequence of a genetic alteration is that the produced organism is not viable due to a major disturbance of the functional harmony causing a severe selective disadvantage. Most researchers working with genetic engineering have experienced this response of nature to one of their genetic constructs. Of more severe concern are those consequences of genetic alterations which might, for example, cause pathogenicity or more generally undesirable long-term consequences on the environment. To avoid such effects, guidelines to carry out the work with appropriate care and controls have been introduced at an early time in the 1970's. Still today, the scientific basis to give green light for some novel applications is sometimes relatively weak, particularly if an application involves the deliberate release of a genetically modified organism into the environment.

A better understanding of the process of biological evolution at the level of the biologically active molecules might be a welcome answer to these pending questions. As a matter of fact, any designed alteration of genetic information represents a reflected construction of a genetic variation. As we have seen, such variations can, in terms of the comparison already made, affect one or a few letters, a few lines, or one to several pages of added, deleted, substituted or scrambled genetic information in the context of the inherited genetic library. To which degree does this correspond to genetic alterations occurring spontaneously any time in genomes, upon the generation of natural genetic variations which form the substrate for natural selection in biological evolution?

*Bacterial genetics reveals the process of molecular evolution at work*

Bacteria are haploid unicellular organisms. They propagate by cell division and thus grow exponentially. As long as plenty of nutrition is available, growth is fast, typically with generation times in the order of 30 minutes. Occasionally, a cell may suffer a spontaneous mutation. Therefore, the propagating clone representing the progeny of a single cell contains an

increasing number of different genetic variants. In view of the haploid nature of the genetic information of bacteria, genetic variants express their phenotype rapidly. At any time the mixed population of parental and mutant cells is submitted to evolutionarily relevant natural selection. For these reasons bacteria are very well suited for population genetic and evolutionary studies. In particular, methodology now available in molecular genetics has made it possible to investigate the molecular nature of individual spontaneous mutants in bacterial plasmids, viruses and even in the chromosome.

On the basis of a wealth of genetic data made available in the literature it has become clear that a relatively large number of particular sources of spontaneous mutagenesis is at work in bacteria. It is possible, however, to group these sources into a few principal strategies followed by nature to ensure a steady, but slow influx of novel genetic variants into the microbial populations (Arber, 1995, 1999, 2000).

The three main strategies of genetic variation in bacteria are:

(1) *Small local sequence changes*, such as the substitution of one nucleotide by another, the deletion of one or a few nucleotides, the insertion of one or a few nucleotides, or a local scrambling of the inherited nucleotide sequences. The causes for such mutations are seen in replication infidelity, nucleotide instability and in the action of many known mutagens.

Genetically encoded repair systems keep the effects of this mutagenic strategy at low, tolerable frequencies. Such repair is of increasing importance for larger genome sizes. This has to do with the generally observed fact that spontaneous mutants are only rarely useful in terms of evolutionary adaptation and advance. More frequent are lethal mutations and variations providing selective disadvantage. Without repair activities this could lead to the eradication of organisms with large genomes. Of course, many DNA sequence alterations may remain without immediate effect on the organism's phenotype; they are neutral or silent.

(2) *Intragenomic DNA rearrangements*. These are mostly influenced by enzyme-mediated recombination processes, such as general (homologous) recombination between largely homologous sequences at different genomic locations, site-specific recombination (including DNA inversion) involving also deviations from the consensus crossover sites and last, but not least, transposition of mobile genetic elements.

(3) *DNA acquisition* based on the horizontal transfer of genetic information between different bacteria. The well understood processes of horizontal gene transfer are: transformation by free DNA molecules, conjuga-

tion mediated by conjugative plasmids, and bacteriophage-mediated transduction in which the virus serves as a vector for bacterial host genes.

These three strategies of genetic variation differ in the quality of their evolutionary contributions. The local sequence change, on which the evolution biologists base the molecular clock to measure evolutionary times of separation of related organisms, is a slow, stepwise process from which novel biological functions can result. However, this process gains in efficiency only when expressed functions start to represent a substrate for natural selection. Hence, local sequence changes must mainly contribute to the amelioration of available biological functions and it can also serve for a steady adaptation to changing living conditions.

DNA rearrangements can bring about novel combinations of available capacities, particularly by the fusion of functional genetic domains of different origin and by the combination of a given reading frame with another, previously unrelated expression control signal. DNA rearrangements thus contribute both to the improvement of existing gene functions and - by assembling different functional domains and sequence motifs - to the innovative generation of novel gene functions.

DNA acquisition is a quite efficient evolutionary strategy. It is a sharing in successful developments made by other organisms. The chance that a particular novel function will also exert its actions in another genetic context is often quite high.

As to the extent of genetic information involved in intragenomic DNA rearrangements and in DNA acquisition it is in general relatively small, typically involving between a part of a page and a few pages of the genetic library. One may assume that very exceptionally larger texts may become acquired without seriously disturbing the functional harmony of the recipient. Such events could perhaps account for cases of a sudden emergence of complex novel properties which cannot easily be explained otherwise. But in general, DNA acquisition as well as intragenomic DNA reshuffling proceed in small steps involving one or a very few genes or even only part of a gene such as a functional domain.

*Products of evolution genes together with nongenetic elements generate spontaneous genetic variations*

As was already mentioned, genetic variants are the substrate for natural selection and they represent the driving force of biological evolution. The theory of molecular evolution postulates that the products of specific

evolution genes are important actors in the steady supply of genetic variations (Arber, 1997, 2000). In the outline of the three distinct natural strategies of the generation of genetic variations we have repeatedly encountered gene products involved in spontaneous mutagenesis. Many of these products are inessential for the normal bacterial life cycle from one cell division to the next. This, and the fact that such genes are widely present in microorganisms, is consistent with the postulate that they primarily serve biological evolution and exert their activities at the level of populations.

One can distinguish between two major types of action of evolution gene products. One class englobes the enzymes which actively generate genetic variants. Examples for this class are transposases and various other types of DNA recombinases involved in DNA rearrangements. Interestingly, this kind of gene products acting as variation generators do not and cannot conform with what one usually expects from gene activities; they work inefficiently and non-reproducibly. The second class is formed by those gene products which modulate the frequencies of genetic variation to low levels that are tolerable for the long-term maintenance of a given strain or species. Examples for this kind of evolution genes are those belonging to DNA repair systems. A good example for evolution genes limiting the frequency of horizontal transfer of DNA and at the same time stimulating occasional DNA acquisition to occur in small steps are restriction enzymes.

A number of non-genetic factors contribute each in its specific way to the generation of genetic variants. These factors include the chemical instability and the structural flexibility of biologically active molecules, environmental mutagens and random encounters (Arber, 2000).

In brief, nature makes use of naturally given factors that contribute to the generation of genetic variations and in addition it has developed enzyme systems to provide a steady influx of novel genetic variants of different qualities serving as substrate for natural selection. This principle ensures the maintenance and renewal of a rich genetic diversity in the living world.

At present time, there is ample evidence from microbial genetics for the views expressed here on evolutionary strategies. It is likely that precisely these microbial capacities might have been instrumental together with symbiotic associations (Margulis, 1981) for the evolution of higher eukaryotic and multicellular organisms. We can thus expect that the further evolution of higher organisms also benefits from activities that are exerted by evolution genes and that contribute together with non-genetic factors to the steady, but not excessive generation of genetic variations. An increasing number of circumstancial evidences is in support of this view (Caporale, 1999).

*Conclusions and Outlook*

The strategies used in genetic engineering for the production of genetic variations resemble quite closely the natural strategies of generating genetic variants. In both instances, three major strategies with qualitative differences can be distinguished, namely small local sequence changes, intragenomic sequence reshuffling and DNA acquisition involving horizontal gene transfer. Recombinant DNA technology was, as a matter of fact, developed on the example of naturally occurring horizontal gene transfer involving vector DNA molecules. Similarly, site-directed mutagenesis simulates the natural strategies of either local sequence change or DNA rearrangement. Similarities are also seen in the extent of sequences involved in these processes, either natural or designed in genetic engineering. One might thus argue that what is done by design in the laboratory must have occurred already sometimes in nature. However, this rapid conclusion is not strictly correct, certainly not for any very specific DNA sequences resulting either from a genetic manipulation or from the natural generation of a genetic variant. This can be seen in the following reflexions on the uniqueness of DNA sequences longer than about 80 base pairs.

We can estimate today's number of living cells on our planet to be roughly $10^{30}$ (Arber, 1993). In the course of time this number may somewhat fluctuate but empty places due to cell death become rapidly reoccupied. Assume that this was so for the last three billion of years, which is $10^{15}$ minutes. Let us also assume that during that long time period each genome would have explored 1000 novel sequences each minute. This corresponds to a mutation rate much higher than what is usually observed. Under these conditions the total number of already explored sequences would be $10^{30} \times 10^{15} \times 10^{3} = 10^{48}$ different sequences. With the four different nucleotides of DNA available, $10^{48}$ corresponds to the number of different closed sequences that can be written with the length of only 80 base pairs. Therefore, novel DNA sequences longer than 80 base pairs have a fair chance to be universally unique and many possible sequences are likely to have never existed. This is surprising, since 80 bp is much shorter than the length of an average gene of about 1000 base pairs!

In view of these considerations it appears justified to maintain the introduced precautions and to carefully explore properties of novel genetic constructions before they are introduced into the environment, although the chance of undesirable effects is very low. Indeed, the same kind of uncertainty with regard to long-term effects must apply not only to

designed genetic variants but also to spontaneous mutations and, at least as far as local sequence changes and some DNA rearrangements are concerned, to mutagen induced genetic variants. In the past, these processes have only very rarely been noted to have caused undesirable effects such as increased pathogenicity of the resulting variant subclones.

In this context, it is relevant to note that up to now, not a single completely novel gene has been invented by genetic engineers. All functional DNA sequences so far involved in experimentation have been taken from natural, living organisms. It is plausible that at some future times entirely novel genes can be designed on the basis of by then increased scientific knowledge. Special care should be taken at that time to widely explore properties and effects of such biological functions before they become introduced into the environment.

To finish, let us briefly discuss a difference between natural biological evolution and evolutionary steps due to genetic engineering. As far as we know, nature has not developed any specific sensory organ that would enable an organism to identify any evolutionary need under particular living conditions and to react by preparing the appropriate mutations which could overcome novel selective disadvantages. Rather, nature follows the principle of producing more or less randomly enough different genetic variants to ensure evolutionary progress. In contrast, the genetic engineer reflects on possible developments which could serve his expectations. Although not all designed genetic modifications keep their promise, some biotechnological applications may guide a branch on the tree of biological evolution to grow into a specific direction.

This kind of reflexions would deserve to be deepened as well as extended into considerations on possible effects of various human activities on natural selection which largely influences the directions taken by biological evolution.

## REFERENCES

Arber, W., Evolution of prokaryotic genomes. *Gene,* 135, pp. 49-56, 1993.

Arber, W., The generation of variation in bacterial genomes. *J. Mol. Evol.*, 40, pp. 7-12, 1995.

Arber, W., *The influence of genetic and environmental factors on biological evolution.* In: Plenary Session on the Origin and Early Evolution of Life (Part I). Pontifical Academy of Sciences, *Commentarii* Vol. IV, N.3, pp. 81-100, 1997.

Arber, W., *Involvement of gene products in bacterial evolution*. In: Molecular Strategies in Biological Evolution (Caporale, L.H., Ed.) *Annals of the New York Academy of Sciences*, Vol. 870, pp. 36-44, 1999.

Arber, W., Genetic variation: molecular mechanisms and impact on microbial evolution. *FEMS Microbiol. Rev*., 24, pp. 1-7, 2000.

Caporale, L.H. (Ed.), Molecular Strategies in Biological Evolution. *Annals of the New York Academy of Sciences*, Vol. 870, 1999.

Margulis, L., *Symbiosis in Cell Evolution*. Freeman, San Francisco, USA, 1981.

Sambrook, J., Fritsch, E.F. and Maniatis, T., Molecular Cloning. *A Laboratory Manual* (2nd. ed.), Cold Spring Harbor Laboratory Press, 1989.