ELSEVIER

Biophysical Chemistry 66 (1997) 133–143

Biophysical Chemistry

On the Generation of Information as Motive Power for Molecular Evolution

Susanne Brakmann

Max-Planck-Institut für Biophysikalische Chemie, Abteilung Biochemische Kinetik, D-37077 Göttingen, Germany

Received 24 April 1997; accepted 24 April 1997

Abstract

Molecular evolution can be described as a learning process during which previously inanimate matter developed the ability to organize all the reaction pathways that establish a living system. Common to all natural self-organizing procedures is the ability of matter to store, process and evaluate the *information* achieved by learning. Genetic information which is stored in RNA or DNA is the object of natural evolution. With the recognition of nature's concepts, *evolutionary optimization* was applied to biopolymers that are not optimally adapted for particular technical or medical purposes. Information can also be stored in molecules with structures and chemical properties that are completely different from nucleic acids. Therefore, optimization processes that mimic the natural evolutionary strategies can also be applied to small organic molecules.

Much effort has been made theoretically and practically to find a certain optimized species within the (hyper)astronomical number of possible sequence alternatives. From a series of computer experiments it can be concluded that it is not necessary to search the entire sequence space in order to find a particular structure; this is advantageous because the diversity of *mutant libraries* that can realistically be achieved in the laboratory never extends to the number of theoretically possible sequences. Molecular mutant libraries that serve as starting populations for *in vitro selection* have been constructed for nucleic acids, proteins, peptides and small organic molecules. © 1997 Published by Elsevier Science B.V.

Keywords: Information; molecular evolution; selection; quasispecies; fitness; error-threshold; mutagenesis; recombination; libraries; error catastrophe

1. Introduction

Physicists strive to describe nature on the most fundamental level, and to find the logical consistencies between the principles on which our understanding is based; they leave the minute observations of more or less incidental details to others. And, at least since Schrödinger's famous lectures in Dublin, physicists have been inspired to tackle the problem "What is life?". Their ultimate goal is the integration of the knowledge obtained from research in molecular biology into the general conception and description of reality.

Manfred Eigen, who is one of the leading authorities on evolution, influenced our present understanding of the phenomenon of life by augmenting the classical physical concepts of energy, matter, space, and time by setting the stage for a molecular interpretation of biological *information* [1]. Together with Peter Schuster, he developed mathematical models

PII S0301-4622(97) 00066-5

for describing a fundamental natural principle that brings order into any random arrangement of autocatalytically replicating species [2]. With this principle, called *selection* in the Darwinian sense, information is generated successively, leading to the steady optimization of species, which can be either organisms or molecules. Not only did Eigen's ideas on molecular evolution shed light on the transition from inanimate to animate matter, but they inspired the imagination of modern chemists and biologists. This article describes the concept of information and its realization in the form of bioactive molecules.

2. What is information?

At the first glance, *information* is associated with the content and meaning of a "news"–message. From the viewpoint of a recipient, the news can be understood only if the recipient and the transmitter agree on the meaning of the symbols used to code the message. In other words, new information is gained by the recipient only when he can interpret the received symbols (the signals can be transmitted acoustically, optically, or by any other means). The symbols that make up the message can be understood as the elementary units of information.

The concept of information can be characterized more precisely by categorizing it in three dimensions [3,4]:

- The *syntactic* dimension of information reflects the relations between the symbols. This aspect is the central topic of the "classical" theory of information described by Shannon and Weaver [5]. Shannon's theory was originally developed for the mathematical description of communication problems, such as storage, transition, transmission and propagation of data (signs or symbols). The mathematical analysis of information processing requires a quantitative measure of the content of the information. Shannon assumed that the amount of information encoded by a message corresponds to the minimal number of symbols that formulate this message. The most concise expression of a message can be defined by a series of questions that must be answered with "yes" or "no" in order to recognize this message within a given set of alternatives. The number of binary yes/no decisions required for formulating the message then is a uniquely defined measure of the complexity. This becomes obvious by considering the number of possible alternative symbol sequences. The mathematical function that Shannon developed to describe this fact therefore corresponds to Boltzmann's H function, known as entropy.
- The *semantic* dimension of information reflects — beyond the relation between symbols — the meaning of the symbols and their combinations. This aspect of information is closely related to human language which serves to represent and transmit messages with meaning and significance. Information is gained by a recipient when a piece of news is deciphered. On an objective level, the decoding of any message involves a structural analysis as well as a comparison of the structural elements of a language: that is, with the determination of the frequency of using certain linguistic symbols, the probability of their sequence, the statistics of word lengths, and other objective assumptions. As a result of this analysis, it is possible to establish objectively some predictions of the sense of a message. The "most probable" meaning of the message can either be verified or completely altered under the influence of the recipient's subjective considerations. Depending on his previous knowledge (comprising prior experience and possibilities of evaluation) the recipient is influenced by his expectations concerning the meaning of the message; his interpretation of the message is not necessarily identical to the one intended originally by the sender. Therefore, a message and its recipient cannot be viewed as separable units. The recognition of information is an environment–dependent process; it is dependent on the starting and the boundary conditions.
- Lastly, the *pragmatic* dimension of information concerns the implicit instruction for some action that has been transmitted to the recipient. This dimension is beyond the relation between symbols, the meaning of the symbols and their combinations. This third aspect of information describes messages (or events) that cause either structural changes of the recipient or change the recipient's preparedness for directed action.

These three dimensions formally describe information in the context of communication with the aim of exchanging messages, or instructions, or portray-

ing structures of reality; in this sense information is a tool for composing and moulding, for thinking and learning. But what is the relevance of the above considerations for chemistry and biology?

The evolutionary self–organization of life can be seen as the result of a gigantic process of learning by matter. During this process of organization, previously inanimate matter developed the ability to organize all the reaction pathways that establish a living system. Every living system, whether bacterium, rose, or man is completely instructed by information: At the basic level, information controls the replication and the variation of species. But information not only directs the process of hereditary transmission, it also completely determines certain phenomena, such as the fate of single cells during the differentiation of organs, the development and modulation of the immune system, and the organization of the central nervous system. Common to all these organizational procedures is the ability of matter to store, process and evaluate the information obtained by learning.

Genetic information, challenging Manfred Eigen to develop his theory of the self–organization of matter, is stored in the linear copolymers of RNA and DNA molecules. The number of symbols and their sequence within the polymeric chains designate whether or not a molecule encodes a message, and the arrangement of the symbols determines the molecule's information content according to the syntactic aspect of information described above. The meaning of the sequence of symbols (or its semantic dimension) originates with the unique ability of RNA and DNA to recognize and "read" themselves. Reading is based on particular chemical interactions mediated by specific hydrogen bondings that link the four monomer bases A, T, G, and C into the complementary pairs "GC" and "AT". The encoded messages transmitted by RNA and DNA include instructions that initiate and control biochemical reaction cascades; this aspect corresponds to the final pragmatic dimension of genetic information.

However, information can also be stored in molecules with structures and properties that are completely different from nucleic acids. *Legibility*, whether by the molecules themselves or by an external operator (another molecular species, computer, or man) is the central requirement for identifying symbols, or combinations of symbols, as information. Therefore, small (organic) non–polymeric molecules can — in Shannon's sense — also store information (structural formulae, or the IUPAC names can be seen as "codes" for this information). However, — and this is the difference to molecular populations that store *genetic information* — populations of these low–molecular–weight compounds cannot "learn" by themselves. They need external help to achieve (evolutionary) optimization. How this is done will be discussed later.

3. How does genetic information originate?

Complementarity is the fundamental property of nucleic acids that guarantees the legibility of the message, and that is indispensable for genetic information. On the other hand, complementary interaction guarantees the formation of an encodable alphabet of information units independent of their (possibly differing) natural abundancies. Nucleic acids still are the only known natural biopolymers that enable the unique complementary base pairing due to their chemical structure, and by recognition of the symbols (bases) that are "read", they produce a negative copy of their sequence. The negative sequence can be converted into a copy of the positive original by the same procedure. This chemical copying, or *replication*, is based on a process that involves stabilizing conservative forces as well as selective dynamic ordering. Replication emerged most likely by the formation of RNA and RNA–like structures; only later was this passed on to DNA which allowed for much longer chains because of its error–correcting capacity [6,7].

The process of complementary reproduction of nucleic acids behaves very much like a simple autocatalytic reaction system that can be described in detail by applying the mathematical formalism of chemical kinetics. The exact solution [8–10] of the differential equations that include such autocatalytic processes demonstrates clearly that *selection* is a direct consequence of autocatalytic self–reproduction of any population of individuals. The natural conditions for the parallel replication of individuals within a population make this fact evident: (1) The individuals consume (and process) resources, i.e. they have an own metabolism that effects a dependence (of the individuals) on their environment. (2) Because the hydrogen

bonding energies that stabilize complementary base pairing are finite, and because the molecules undergo Brownian movement, fluctuations will occur and introduce errors into the genetic information. Thereby, variety will arise among the pool of individuals. (3) As a result of the first two considerations, the diverse population of individuals starts competing for the available resources. Their ability to replicate (and this, in turn, is instructed by their genetic information!) is assessed with respect to the requirements that are set by the environment. By this assessment the most efficient individual reproduction pathways, i.e. the "fittest" replicators, will be determined. These optimal mutants will be amplified while the others are removed due to the competition. The emergence of mutants that replicate more efficiently than their ancestors with the "original" genetic information, is a manifestation that new information has been generated and selected.

In conclusion, we can say: Not only is reproduction the basis of the conservation of information (that otherwise would be destroyed by hydrolytic or other degradation of its material carrier, RNA or DNA), but it is also the occasion for the generation of novelty through replication errors (i.e. mutations), and therefore it is the final cause of *natural selection*. Both reproduction and natural selection are necessary prerequisites of evolution.

4. No progress without errors!

Evolution continually improves the fitness–relevant properties and the complexity of animate matter. Without generating reproduction errors, i.e. without variation of the genetic starting material, the assessment and selection of new information could not occur. Therefore, the extent of erroneous copying decisively determines the evolutionary progress and its velocity.

Imprecise base pairing leads to replication errors that appear as stochastic events and produce distributions of mutants around a dominating species (master sequence). These mutant distributions — when stationary — are called *quasispecies* because they behave (both physically and mathematically) like a single species or, more precisely, like the selected wildtype of a species, and yet do not define a single species. A quasispecies usually has a defined consensus sequence, but otherwise represents a widely dispersed, not necessarily symmetrical, distribution of similar and related, but by no means identical, sequences. These species–like mutant distributions include many individual species that are "neutral" or "nearly neutral" with respect to the degree of fitness that is representative for the given distribution. They can be dominated by more than one, or even no, master sequence. Evolutionary progress of the quasispecies comes up with new mutants that appear on the periphery of the distribution.

In the case of fast replicating species, such as molecular replicators or viruses, mutations are not necessarily statistically rare events. In these systems many individual mutants occur often and reproducibly, but they appear with a frequency that is strongly influenced by their fitness. Advantageous mutants preferably populate regions of the mutant distribution that represent high degrees of fitness because these regions are more populated (due to high replicating efficiencies). As a result, the production of mutants is biased towards high fitness. This effect accelerates the evolutionary optimization by orders of magnitude compared to a purely stochastic trial and error strategy.

Evolutionary flexibility represents the pivot of the (natural) selection process. It is guaranteed by error rates that produce a sufficiently large variety of mutants. But, as the mutation rates increase, the accumulation of errors outweighs the selective narrowing–down. As a consequence, the information "melts" away into random combinations of symbols that correspond to nonsense messages without any ability to generate the necessary reaction pathways. This is, of course, a catastrophe for every living system because it leads to the evolutionary decay of information. Therefore, natural evolution processes are bounded by an *error threshold*. Far below this boundary, selection can be reduced to all–or–none decisions. Evolutionary progress is achieved at reasonable velocities only when the error rate is near the threshold value. Indeed, it has been established experimentally by several research groups that viruses (which, due to their short replication periods, are ideally suited for evolutionary studies) behave like quasispecies and operate close to their error thresholds [11,12]. In accordance with the theory, the reciprocals of the error

rates that have been experimentally determined for many viruses define the limiting information capacity of their genomes.

However, the complexity that life has developed by successive evolutionary optimization, does not only result from vertically transmitted point mutations (i.e. a hereditary transmission of errors to the direct offspring). Additional principles of organization exist that enable the horizontal transfer of genetic information. These additional principles are homologous genetic recombination (the basic mechanism of sexual heredity) and the transposition (and integration) of genes via mobile information–carrying elements, such as plasmids and viruses. It should be mentioned that recombination procedures contribute to the error threshold behavior without changing its nature [13,14].

5. How can a better species be recognized?

Finding a certain optimized species seems like a daunting undertaking when we consider the number of possible information–carrying sequences, particularly in the case of nucleic acids which comprise 4^n different polymers of chain length n. We could, for example, take the complete distribution of mutants and arrange all the sequences according to their correct kinship relations on defined loci (points) within a spatial model, the *sequence space*; this was first proposed by Wright [15], and later independently by Hamming [16] and Rechenberg [17]. Eigen [18] applied the idea to the genetic four–letter–alphabet. We can assign fitness values to every sequence present in the sequence space and look at a *fitness landscape* that consists of peaks (sequences with high fitness) which are connected by ridges and separated by saddles, valleys (sequences with low fitness) or planes. The topography of this fitness landscape then guides our view along the ridges towards the "mountainous" regions where we can find species with maximum fitness. Furthermore, we can monitor the fitness landscape over longer periods of time and determine the dynamics of the self–replicating population [19], i.e. the progress of evolutionary optimization.

Nevertheless, even for a gene encoding a typical functional protein with a (moderate) chain length of 100 monomers, the appertaining DNA sequence length of ≈ 300 monomers corresponds to $\approx 10^{180}$ different sequence alternatives, and hence, to a sequence space that is extended in the immense number of $2 \cdot 300 = 600$ dimensions. There is some consolation when we consider that two DNA sequences with quite distant kinship relations can code for protein sequences that fold into very similar or even identical structures (where the degree of similarity is also dependent on the resolution of the applied analysis), thus exhibiting the same function [20]. Although until now it has not been possible to study and classify the folding of protein mutant distributions, much effort has been spent analyzing the relation between the RNA sequence distributions and their corresponding minimum free–energy secondary structures. This relation is viewed as a mapping from sequence space into *shape space* [21]: It has been shown by Schuster [22] that the frequency with which a certain structure is realized in shape space is inversely proportional to some power $c > 1$ of the structure's frequency rank. More than 10^9 RNA sequences of chainlength $n = 30$ (GC only) were, for example, folded into the most frequent minimum free–energy structures. By comparison, it was found that the folding process in this case resulted in "only" ≈ 200000 different structures, i.e. on the average, more than 4900 separate sequences lead to a single shape. From a series of similar computer experiments it can be concluded that structures which are much more abundant than the average (*common structures*) can be found everywhere in the shape space. Even in the vicinity of an arbitrarily chosen starting point in the sequence space, a common structure will most probably exist. As a consequence, it is not necessary to search the entire sequence space in order to find a particular structure.

At this point of our theoretical considerations concerning the generation of information and the vast mutant distributions, we need to turn back to the phenomenon of *selection*. Since Manfred Eigen is pioneering work on the self–organization of matter [1], the process of selection has stimulated the development of a wide variety of strategies in the field of chemistry and biology that differ markedly from conventional concepts. In order to understand the various evolutionary approaches better, we must distinguish two separate strategies [23]:

- *Natural, self–organizing selection* that we have looked at so far, is due to the competitive growth

of a population of individuals. Thus, mutants with reproduction pathways that work most efficiently under certain environmental conditions have a substantial growth advantage. By adaptating these natural selection mechanisms to the methods of experimental optimization in the laboratory, the application of a precisely defined selection pressure can induce the emergence of a new (optimal) mutant with well–defined properties.

- *Artificial, non–self–organizing selection* is devoid of competition. Independently existing mutants with any desired property might be favored by external interference. Artificial selection, in contrast to natural selection, is based on the precise knowledge of the characteristics that define a desirable mutant. Therefore, this approach requires an appropriate detection method for screening large populations (mutant libraries) for individuals with promising features. The whole selection procedure, comprising the generation of mutant pools followed by suitable screening, is usually iterated with newly selected mutants until one or more optimal individuals are obtained.

The choice of a particular optimization strategy, whether natural or artificial, depends on the goal. However, natural selection is based on systems that can be characterized by inherent feedback–loops. Here, genetic information is strictly coupled to certain functions that influence the replication efficiency. In an ideal case, this kind of coupling can provoke an all–or–none decision that inevitably produces only those mutants that fulfill the applied selection constraints. Of course, the natural and the artificial selection procedures may be combined. The information–generating amplification (or self–amplification) of individuals is common to all of these experiments.

6. How does information originate *in vitro*?

As we have seen above, evolution — natural as well as artificial — could not exist without the generation of information. During the past decade, a wide repertoire of techniques has arisen for producing diversity *in vitro*. With these methods, evolutionary strategies in the search for novel effector molecules are greatly accelerated to much higher rates than can naturally occur. The resulting libraries have a much greater diversity than those found in nature and comprise a variety of compound classes, from natural biopolymers (nucleic acids and proteins) to low–molecular–weight organic molecules.

Since Wöhler's synthesis of urea in 1828, the organic chemist's view is focused on the most efficient and most specific synthesis of *single* products with defined structures. Searching for new effector molecules, e.g. pharmaceutics, the systematic approach starts with the synthesis, purification, characterization and examination of an appropriate target structure. The effects caused by this structure can then be improved either by systematic or by intuitive alterations, until through trial and error a structure with optimal function is finally found. Similar procedures can be applied to biological macromolecules, such as enzymes or receptors: The identification, preparation, crystallization and structural resolution of several proteins gave initial insights into the relations between a structure and its corresponding function, thereby enabling the systematic optimization of biopolymer structures known as *rational design*. Precise alterations in proteins can be engineered by *site–directed mutagenesis*. By substituting specific bases [24,25] within the coding gene, those amino acids of the protein that either are involved in the active site of the molecule, or contribute directly to the substrate binding or catalysis, can be determined and varied. However, with insufficient knowledge concerning the structure/function relationships of a certain protein, site–specific mutagenesis is not feasible: (1) In case more than one amino acid might contribute to a function, a series of single substitutions, together with all their possible combinations, would have to be examined. In a rather small region of 10 amino acids only, we would have to deal with $4^{30} \approx 10^{18}$ alternative DNA sequences, or $20^{10} \approx 10^{13}$ amino acid permutations. (2) We still lack the rules that govern the interchangeability of amino acids with respect to their intramolecular interactions. Without rules that relate structure and function, each amino acid at any position in a protein must be taken into account. Both limitations can be circumvented, even though we remain ignorant about the molecular basis of structure/function relationships, by just applying what has been learned about natural selection. With such *irrational design* strategies, "unnatural" molecules with

predefined properties that may be (completely) different from those selected by nature can be obtained.

Large molecular libraries that serve as starting populations for natural or artificial *in vitro* selection have been constructed with (1) nucleic acids (and proteins), (2) peptides, and (3) small organic molecules. These libraries have different (historical) origins, and the generation of libraries for each of these three classes of compounds will be described separately.

6.1. Biopolymers

Because almost all phenotype characteristics, whether observed with nucleic acids or with proteins, are encoded and instructed by genetic information, the following considerations are limited to the problem of generating diverse biopolymer pools to nucleic acids:

The increasing interest in genetic elements with controlling activities (such as promotor sequences) and in protein regions that are resonsible for certain functions, instigated strategies for mutagenizing selected regions instead of single sites. New methods for the rapid and automatic chemical synthesis of DNA made it possible to produce oligonucleotides with randomly inserted bases at one or more positions. Double–stranded DNA sequences with randomized fractions comprising defined regions within a gene of interest can be produced by hybridizing two oligonucleotides (one or both of which contain random sequences). These *random cassettes* can then be inserted into gene regions supposed to be important for certain functions; the cassettes replace the original part of the sequence. The earliest random cassette libraries have been described by Horwitz and Loeb [26] for unnatural (random) sequences inserted into genes. A series of applications to nucleic acids (e.g. promotor sequences [27]) and proteins [28–31] followed this work. Likewise, random oligonucleotides have been used — via expression in phage display libraries [32] — for studies on binding proteins, for mimicking the vast diversity of the immunological repertoire [33] and for constructing catalytic antibodies [34].

In similar synthetic approaches, RNA sequences can also be randomized. In the first experiments of this type, single–stranded random RNA libraries were exposed to affinity screening in order to find molecules with ATP–binding activity [35]. Tuerk and Gold [36] exploited that work and applied repetitive cycles of amplification and screening to search for binding activity in RNA mutant libraries. Applying this *Systematic Evolution of Ligands by EXponential enrichment* (SELEX), they eventually selected RNA molecules with high affinity to a variety of small ligands and protein molecules. Using pools of random RNA sequences, new ribozymes have also been isolated that can mimic ligases, kinases, and isomerases and can even catalyze carbon–nitrogen bond formation [37].

However, with insufficient knowledge about structure/function relationships and increasing sequence lengths (up to the size of complete genes), it may be preferable to scatter random (point) mutations over the entire fragment, typically at a frequency of one or a few mutations per molecule, depending on the applied method. Passing cloned genes through bacteria that contain mutator genes was among the first approaches for random mutagenesis [38]. The first *in vitro* methods that were described included the treatment of single–stranded DNA with various mutagenic chemicals followed by enzymatic synthesis of the complementary strand [39–43]. Because these methods are quite laborious and result in incomplete libraries, alternative strategies have been developed that utilize the incorporation of dITP (deoxy–inosine–triphosphate, which forms base pairs with all of the "normal" bases, A, G, C, or T) [44,45], or of nucleotide analogs [46] in order to force the generation of random mutations. A major disadvantage of these methods is that they are not sufficiently efficient to produce libraries that are complex enough to encompass entire genes. Attempts were made to overcome this problem by either omitting one of the four nucleotides in each of four separate primer extension reactions [47], or vice versa, increasing the concentration of one nucleotide relative to the other three [48] to enhance the error rate of the polymerases involved. Because they are simple and verstile, the most attractive mutagenization strategies are those PCR methods that reduce the fidelity of the *Taq* polymerase. This can either be achieved by the addition of DMSO and manganese ions to the reaction mixture [49], or by just adding manganese ions and using unequal dNTP concentations [50]. If the reaction mixture contained DMSO, it was shown that approximately three times

more transitions than transversions were produced, and this method can hardly be applied to gene fragments longer than about 1000 base pairs. Therefore, the PCR mutagenesis described by Joyce [50] still seems to be the method of choice for genes encoded by more than 1000 base pairs. Under extreme conditions (i.e. highly biased dCTP and dTTP concentrations), the mutation frequencies achieved by PCR can be elevated to 20%, or even to 90% [51]. These *hypermutagenesis* reactions lead to greater than 20% amino acid replacements, and thereby mutants are produced whose kinship relationships are quite distant, and which could serve in the capacity of exploring the functional robustness of enzymes.

Since neither random cassette mutagenesis nor error–prone amplification procedures are combinatorial, their benefit in searching the sequence space for new properties is limited. Therefore, the arsenal of *in vitro* evolutionary procedures can — by analogy to nature — be augmented with (homologous) recombination procedures which result in the rapid accumulation of advantageous mutations. *DNA shuffling* was described as the first suitable method for achieving recombination *in vitro* [52]. This strategy involves the random (enzymatic) fragmentation of a pool of randomly mutated genes, followed by reassembly of the (purified) fragments by PCR. There is no reason that this concept should be limited only to mutant libraries derived from a single gene. In an extension of the idea, fragments from more than one gene pool or even random cassettes, together with mixtures of "wildtype" or random fragments, can be recombined [53]. A concept that is closely related to recombination, but different in practice, is the idea of *modular protein design* [54]. This method generates functional diversity rather than sequence variability; it is based on the fact that (1) a high degree of amino acid substitutions can be tolerated in proteins, and (2) proteins with almost no detectable sequence similarity can adopt similar three–dimensional folds. Suitable starting libraries of protein modules could be originated either from natural sources or from preselected combinatorial peptide libraries; the latter will be discussed next.

6.2. *Peptides*

Information–generating concepts were first applied in synthetic organic chemistry with the advent of methods for preparing peptide libraries. At that time, compound libraries received special interest because they could potentially provide a source for a wide variety of defined structures in one step. Based on the technique of the solid–phase synthesis of peptides that was conceived and developed by Merrifield, several methods for the synthesis of peptides in parallel were established [55]. By miniaturizing parallel reacting ensembles, a strategy was developed that enables the simple and efficient synthesis of compound libraries with almost any desired complexity. These methods are known as "portioning–mixing method" or "split synthesis" [56,57].

Using combinatorial peptide chemistry, libraries of separately produced single compounds or of defined mixtures can be produced. Keeping in mind that polypeptides with 20 amino acid residues provide about 10^{26} sequence alternatives, it is obvious that it is infeasible to synthesize all possible candidates for a certain effector molecule, even with fast parallel and combinatorial approaches. Therefore, the realistic diversity of the limited peptide libraries determines their power in finding new "lead structures". Highly diverse or, "random" libraries — comprising molecules that differ substantially with respect to their structural and functional features — are the most suitable libraries for this goal.

6.3. *Low–molecular–weight organic compounds*

Effector molecules that are based on peptide structures exhibit at least two principal disadvantages: (1) They are available within organisms on only a moderate to low scale, and they are quite sensitive to hydrolytic degradation. (2) The "translation" of a peptide lead structure into a non–peptide structure is time–consuming and not trivial. Therefore, the research efforts in the field of combinatorial chemistry have moved significantly from peptides to small organic molecules (not taking into account the various strategies comprised of oligonucleotides, oligosaccharides, and similar pseudo–biooligomers that result from the polymerization of the appropriate monomers).

In analogy to the combinatorial synthesis of peptides, two separate strategies can be distinguished for libraries of organic compounds, (1) the efficient parallel synthesis of a limited library (or "array") of single molecules, and (2) the synthesis of compound mixtures with varying complexity. The entire repertoire of organic chemistry is applied in order to produce diverse mutant pools [58]. The reactions span a wide range, starting from one–step syntheses (where two reactants, belonging to two different classes, react according to a defined mechanism), progressing to syntheses involving a sequence of two or three reactions, and to "one–pot syntheses", and finally extending to multi–component reaction systems that were first described by Ugi [59].

Returning to the world of molecular biology, the latest approach for synthesizing low–molecular–weight compounds must be mentioned. Polyketide libraries comprised of a family of structurally complex natural products, including a number of important pharmaceutics, can be generated via combinatorial biosynthesis [60]. Motivated by the values of these natural products, the biosynthetic pathway that enables the natural synthesis of this class of compounds has been elucidated. It was established that at least three architecturally different types of modular polyketide synthases are responsible for directing the biochemical reaction cascade. By combinatorial reassembly of the separately cloned enzyme modules, libraries of novel small molecules could be produced that can be utilized for screening new drugs.

7. The desirable error catastrophe — a perspective

Finally, a concept is presented that differs from all the methods presented above to obtain a "better species", insomuch as it entails a negative, rather than a positive, aim — the extinction of a species:

Viruses, which we have considered as quasispecies, operate close to their error thresholds. From the viewpoint of a virus, this flexibility is very advantageous because it supports the rapid adaptation to less favorable replication conditions, and, in the case of a mammalian host, it allows the escape from the organism's immune response. This fact has already been recognized as an opportune way to influence viral replication.

Recognizing the importance of the error threshold for self–replicating species, Manfred Eigen suggested that the error–producing capacity of a viral polymerase could be increased beyond its limiting threshold. Such a polymerase would theoretically induce an accumulation of errors within the viral genome and thereby catalyze the transition from readable genetic information to nonsense messages (without any ability to control the necessary replication pathways). As a consequence, the virus would die out.

For example, the catalytic efficiency and error–producing capacity of viral polymerases can be influenced by the addition of chemicals and divalent metals. However, directly changing a common viral polymerase into an error–prone enzyme is much more promising (and challenging), especially when considering an evolutionary approach. In order to favor an error–prone enzyme, a self–replicating system that couples mutant polymerase genes in a feedback loop to an appropriate selective constraint must be constructed. Thereby, desired mutants could receive a substantial growth advantage, eventually connected to an all–or–none decision. These ideas that were inspired by Eigen, are currently being investigated in his laboratory.

8. Conclusion

Information completely instructs the machinery of life: It directs the process of hereditary transmission as well as it determines certain phenomena such as the differentiation of organs or the organization of the immune system. Genetic information originates in a dynamical process of self–evaluation within any population of self–replicating species. The generation of information is connected with the principle of selection which — in the Darwinian sense — has been shown to be a category of behavior reaching down to the molecular level. The ability to store, process and evaluate the information obtained during selection characterizes the outstanding feature of nucleic acids.

More generally, information is a structural correlate of *function*. Any combination of symbols can be recognized to be information as soon as it becomes *readable*. Therefore, information can also be stored in

molecules that are completely different from nucleic acids. With artificial selection procedures that mimic the process of natural evolution, molecules which are incapable of self–replication and self–evaluation can also achieve evolutionary optimization.

The progress of natural evolution is guaranteed by replication error rates that produce a sufficiently large variety of mutants. Artificial evolution procedures, by analogy, profit from highly diverse mutant libraries. Beyond that, the error threshold which limits the natural selection processes could be the pivot of a new antiviral strategy. After all, it is information which continually pushes the evolutionary progress, in nature as well as in the laboratory. Information will continue to provide surprising improvements in the never–ending learning process of nature!

Acknowledgements

Manfred Eigen is the one who ingeniously influenced our present comprehension on molecular evolution. I will take this opportunity to thank Manfred Eigen for introducing me into this exciting field of the sciences, for providing me with an intellectual and inspiring atmosphere, and especially, for being my mentor. I am grateful to Peter Schuster for his critical comments on this manuscript. I wish to thank Timm Wetzel for critically discussing this text and for his patient and expert assistence in preparing the final layout.

References

[1] M. Eigen, Naturwissenschaften 58 (1971), 465–523.
[2] M. Eigen, P. Schuster, Naturwissenschaften 64 (1977), 541–565.
[3] C. F. von Weizsäcker, Die Einheit der Natur, Deutscher Taschenbuch–Verlag, München, 1971.
[4] B.–O. Küppers, Der Ursprung biologischer Information, Piper, München, 1986.
[5] C. E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Pres, Urbana, 1971.
[6] M. Eigen, Advances in Chemical Physics 38 (1978), 211–262.
[7] M. Eigen, Gene 135 (1993), 37–47.
[8] C. J. Thompson, J. L. McBride, Mathematical Biosciences 21 (1974), 127.
[9] B. L. Jones, R. H. Enns, S. S. Ragnekar, Bulletin of Mathematical Biology 38 (1976), 12.
[10] M. Eigen, J. S. McCaskill, P. Schuster, Advances in Chemical Physics 75 (1989), 149–263.
[11] E. Domingo, D. Sabo, T. Taniguchi, C. Weissmann, Cell 13 (1978), 735–744.
[12] E. Domingo, E. Matínez–Salas, F. Sobrino, J. C. de la Torre, A. Portela, J. Ortín, C. López–Galmidez, P. Pérez–Breña, N. Villanueva, R. Nájera, S. van de Pol, D. Steinhauer, N. de Polo, J. Holland, Gene 40 (1985) 1–8.
[13] T. Wiehe, E. Baake, P. Schuster, Journal of Theoretical Biology 177 (1995), 1–5.
[14] M. C. Boerlijst, S. Bonhoeffer, M. A. Nowak, Proceedings of the Royal Society B (London), in press.
[15] S. Wright in D. F. Jones (Editor), The roles of mutation, inbreeding, crossbreading and selection in evolution (International Proceedings of the Sixth International Congress on Genetics, Vol. 1), Ithaca (N.Y.), 1932, 356–366.
[16] R. W. Hamming, Coding and Information Theory, Prentice–Hall, Englewood Cliffs, 1973.
[17] I. Rechenberg, Evolutionsstrategie, Problemata frommann–holzboog, Stuttgart–Bad Cannstadt, 1973.
[18] M. Eigen, R. Winkler–Oswatitsch, A. Dress, Proceedings of the National Academy of Sciences of the USA 85 (1988), 5913–5917.
[19] P. F. Stadler, P. Schuster, Bulletin of Mathematical Biology 52, 1990, 485–508.
[20] T. L. Blundell, R. F. Doolittle, Current Opinion in Structural Biology 1 (1991), 319–320.
[21] P. Schuster, W. Fontana, P. F. Stadler, I. L. Hofacker, Proceedings of the Royal Society of London — Series B: Biological Sciences 255, 1994, 279–284.
[22] P. Schuster, Journal of Biotechnology 41 (1995), 239–257.
[23] S. Brakmann, M. Eigen in W. Gilbert, G. Tocchini–Valentini (Editors), Evolution in the Test Tube (Frontiers in Biology, Vol. 1), Rome, 1997, in press.
[24] T. A. Kunkel, J. D. Roberts, R. A. Zakour, Methods in Enzymology 154 (1987), 367–382.
[25] W. Ping Deng, J. A. Nickoloff, Analytical Biochemistry 200 (1992), 81–88.
[26] M. S. Horwitz, L. A. Loeb, Proceedings of the National Academy of Sciences of the USA 83 (1986), 7405–7409.
[27] A. R. Oliphant, K. Struhl, Methods in Enzymology 155 (1987), 568–582.
[28] D. K. Dube, L. A. Loeb, Biochemistry 28 (1989), 5703–5707.
[29] A. R. Oliphant, K. Struhl, Proceedings of the National Academy of Sciences of the USA 86 (1989), 9094–9098.
[30] J. Reidhaar–Olson, R. T. Sauer, Science 241 (1988), 53–57.
[31] J. F. Reidhaar–Olson, J. U. Bowie, R. M. Breyer, J. C. Hu, K. L. Knight, W. A. Kim, M. C. Mossing, D. A. Parsell, K. R. Shoemaker, R. T. Sauer, Methods in Enzymology 208 (1991), 564–586.
[32] J. K. Scott, G. P. Smith, Science 249 (1990), 386–390.
[33] G. Winter, C. Milstein, Nature 349 (1991), 293–299.
[34] R. A. Lerner, S. J. Benkovic, P. G. Schultz, Science 252 (1991), 659–668.

[35] M. Sassanfar, J. W. Szostak, Nature 364 (1993), 550–553.
[36] D. Irvine, C. Tuerk, L. Gold, Journal of Molecular Biology 222 (1991), 739–761.
[37] C. Wilson, J. W. Szostak, Nature 374 (1995), 777–782.
[38] E. C. Cox, Annual Reviews in Genetics 10 (1976), 135–156.
[39] D. Shortle, D. Nathans, Proceedings of the National Academy of Sciences of the USA 75 (1978), 2170–2174.
[40] J. T. Kadonaja, J. R. Knowles, Nucleic Acids Research 13 (1985), 1733–1745.
[41] R. M. Myers, L. S. Lerman, T. Maniatis, Science 229 (1985), 242–249.
[42] J. J. Diaz, D. D. Rhoads, D. J. Roufa, BioTechniques 11 (1991), 204–211.
[43] J. O. Deshler, Genetic Analysis, Techniques and Applications 9 (1992), 103–106.
[44] H. Gram, L.-A. Marconi, C. F. Barbas III, T. A. Collet, R. A. Lerner, A. S. Kang, Proceedings of the National Academy of Sciences of the USA 89 (1992), 3576–3580.
[45] J. H. Spee, W. M. de Vos, O. P. Kuipers, Nucleic Acids Research 21 (1993), 777–778.
[46] M. Zaccolo, D. M. Williams, D. M. Brown, E. Gherardi, Journal of Molecular Biology 255 (1996), 589–603.
[47] P. M. Lehtovaara, A. K. Koivula, J. Bamford, J. K. C. Knowles, Protein Engineering 2 (1988), 63–68.
[48] X. Liao, J. A. Wise, Gene 88 (1990), 107–111.
[49] D. W. Leung, E. Chen, D. V. Goeddel, Technique 1 (1989), 11–15.
[50] R. C. Cadwell, G. F. Joyce, PCR Methods and Applications 2 (1992), 28–33.
[51] M. A. Martinez, V. Pezo, P. Marlière, S. Wain-Hobson, The EMBO Journal 15 (1996), 1203–1210.
[52] W. P. C. Stemmer, Proceedings of the National Academy of Sciences of The USA 91 (1994), 10747–10751.
[53] A. Crameri, W. P. C. Stemmer, BioTechniques 18 (1995), 194–196.
[54] M. A. Fuchs, C. Buta, this issue.
[55] G. Jung, A. G. Beck-Sickinger, Angewandte Chemie, International Edition in English 31 (1992), 367–383.
[56] K. S. Lam, S. E. Salmon, E. M. Hersh, V. J. Hruby, W. M. Kazmiersky, R. J. Knapp, Nature 354 (1991), 82–84.
[57] M. A. Gallop, R. W. Barrett, W. J. Dower, S. P. A. Fodor, E. M. Gordon, Journal of Medicinal Chemistry 37 (1994), 1233–1251.
[58] F. Balkenhohl, C. von dem Bussche-Hünnefeld, A. Lansky, C. Zechel, Angewandte Chemie, International Edition in English 35 (1996), 2289–2337.
[59] I. Ugi, C. Steinbrückner, Chemische Berichte 94 (1961), 734–742.
[60] C. Khosla, R. J. X. Zawada, Trends in Biotechnology 14 (1996), 335–341.