Schrödinger-Lecture: Trinity College Dublin, Novemer 24, 1998

# Molecular Insights into Life and Evolution

Peter Schuster, Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Nature puzzles naturalists by unexpected and highly efficient solutions to complex problems. Her recipe of success is Darwin’s principle of variation and selection of best adapted variants. Simple experiments with nucleic acid molecules in the lab provide new tools to study Darwinian evolution in short times. Computer simulations complement these experiments by otherwise hardly accessible details. Exploration of evolution *in vitro* creates basic knowledge that can be applied to evolutionary design of new biomolecules which are tailored for predefined purposes.

Ladies and gentleman,

It is a great honor and pleasure for me to give the 1998 Schrödinger lecture. I have always admired the great Erwin Schrödinger and I have particularly enjoyed his early interest in the fundamentals of biology and his foresight that led to the famous series of lectures on the question „What is Life?“. Let me describe for you tonight, about fortyfive years later, the current understanding of the evolutionary process that shapes biology. In the last two decades insights were gained that allow to trace down evolution to the molecular level. Like Schrödinger I shall adopt the physicists‘methodology of reduction in order to discover the principles lying behind apparent complexity.

## 1. Adaptation of organisms and molecules

Everywhere in nature we observe adaptations to the finest degree one can think of. Examples are well known to each of us, and they cover an enormously wide range from host-parasite systems to symbiotic cooperations, from foraging strategies to mimicry. Adaptation and fine-tuning, however, are not confined to macroscopic biology dealing with entire organisms and their parts, they are evident at the molecular level as well where we find them to govern the whole machinery of life.

The major players in cellular metabolism are biopolymers, Erwin Schrödinger ingnorant of macromolecular chemistry called them „aperiodic crystals“ in his pioneering lectures. The two major classes are now well known, nucleic acids the carriers of biological information and proteins the enormously specific and highly efficient catalysts of nature. They are related by three processes called replication, transcription, and translation: (i) Nucleic acids have a chemical structure that can be copied or **replicated,** and (ii) nucleic acids, naturally appearing in two closely related subclasses called deoxyribonucleic acids (DNA) and ribonucleic acids (RNA), are **transcribed** into eachother, whereas (iii) RNA is **translated** into protein. The three processes are summarized in the following diagram which decribes the flow of biological information in the cell. Because of its fundamental nature and its direct deducibility from molecular structures it was called the central dogma of molecular biology:

$$\begin{tikzcd}
\text{DNA} \arrow[r, Rightarrow, "\text{Transcription}"] \arrow[d, Rightarrow, "\text{Replication}"'] & \text{RNA} \arrow[r, Rightarrow, "\text{Translation}"] & \text{Protein.} \\
\text{2 DNA} & &
\end{tikzcd}$$

Exploration of the biochemistry of viruses showed that the dogma had to be modified. RNA can be copied into DNA by **reverse transcription** and RNA can be copied without DNA being involved:

$$\begin{tikzcd}
\text{DNA} \arrow[r, Leftrightarrow, "\text{Transcription}"] \arrow[d, Rightarrow, "\text{Replication}"'] & \text{RNA} \arrow[r, Rightarrow, "\text{Translation}"] \arrow[d, Rightarrow] & \text{Protein.} \\
\text{2 DNA} & \text{2 RNA} &
\end{tikzcd}$$

Information flow is a one-way-street leading from nucleic acids to protein. It involves the genetic code which assigns three letters of the nucleic acid alphabet to one letter of the protein alphabet. Since 64 trinucleotide codons are assigned to 20 amino acids the protein language is redundant: several codons code for the same amino acid. In addition, and even more important, is the molecular structure of the cellular machinery for protein synthesis. It allows by using transfer-RNAs as adaptors to convert a sequence of nucleotides into a sequence of amino acid residues but it would never support the inverse process. The molecular biology of translation is the basis for the rejection of the Lamarckian mechanism of inheritance.

Biology is built upon the dichotomy of genotype, being the genetic information stored in a molecular carrier, and phenotype, representing the macroscopic appearence of the organism. Phenotypes originate from genotypes through a highly complex process. One can say that the genotype carries a program that is executed in a suitable environment and thereby produces the phenotype. The most relevant evolutionary consequence of this separation of legislation and executive power is that all variations through mutation and recombination occur on the genotype whereas selection operates exclusively on phenotypes. Mutations happen uncorrelated with the success of their phenotypes. A mutation, for example, does not occur more frequently because the organism carrying it has higher fitness.

Two examples of highly elaborate molecular adaptations will be mentioned here: (i) fine-tuning and sequence variation in the oxygen carrying protein hemoglobin and (ii) specificity and regulatory function of a viral protein called cro-repressor. Hemoglobins are the proteins in mammals and many other vertrebrates that supply cells with molecular oxygen. They are adapted to a certain altitude at which the animals or humans live. We have troubles, therefore, when we try to climb the high mountains in Asia or South America. To the great surprise of naturalists geese and other birds were observed flying above the Himalayas having apparently no breathing problems. The solution of the puzzle turned out to be very simple: these birds have two hemoglobins, one to live at ordinary altitudes and one for great heights. A comparison of protein sequences showed that three specific point mutations are sufficient to change optimal hemoglobin function from low to high altitudes. On the other hand, these sequence comparisons revealed also that the amino acid residues at the majority of positions in hemoglobin can be exchanged without measurable alterations of function.

The second example is a protein molecule called cro-repressor. It binds specifically to the DNA of phage λ and controls its life cycle by switching between the lytic process, where the resources of the bacterial cell are instantaneously exploited to produce phages, and the lysogenic phase, during which the genetic material of the phage stays dormant in the bacterium until it turns into the lytic phase later on. The specificity of the repressor in finding its binding site on the DNA is remarkable: Out of millions of nucleotides in the bacterial cell it finds a stretch of some ten nucleotides to which it binds with very high affinity. Regulation of genes in cellular metabolism is one of the most spectacular examples of adaptation and fine-tuning on the molecular level.

Relations between sequences, structures and functions of biopolymers are highly complex and in a way encapsulate the mysteries of life. Their exploration is a primary issue in biology. Essentially the same specificity of molecular recognition and fine adjustment of function can be created in the laboratory through suitable experiments. To observe adaptations is, however, not sufficient as the famous geneticist Theodosius Dobzhansky pointed out in his famous phrase, „Nothing in biology makes sense except in the light of evolution“, and thus we have to explain how these marvellous adaptations came about and what were the driving forces behind them.

## 2. Time and diversity

The evolutionary biologist like his colleagues from physics or chemistry would like to study his research object by means of carefully planned experiments. But, how can we investigate evolution by experiment? There are two nightmares for the experimental biologist which he encounters in evolution: time and diversity. The time scales of biological evolution are thousands and millions of years, much too long for any kind of experimental study. Diversities in the biosphere confront the scientist with a not less serious problems than the long time spans: When we ask for the numbers of possible mutations or recombinations in a genome, the aswers are discouragingly large. It will never be feasible to explore an appreciable fraction of the possible. This is true for both, the experimental biologist and for Nature herself as François Jacob points out in his essay on „The actual and the possible“. The reason for this overwhelming diversity simply is combinatorics. Whenever letters are combined to words similar huge numbers are obtained. We consider, for example, an RNA virus with a genome length of 5000 nucleotides. Biological information is written in the four-letter nucleotide alphabet {A,U,G,C} and this yields $4^{5000} \approx 10^{3000}$ different RNA molecules with this genome length. Needless to say, we have no imagination of the size of such large numbers.

Evolution experiments in the laboratory became accessible when Sol Spiegelman chose molecules rather than viruses or whole organisms as the target of evolution. He took viral RNA molecules and brought them into a proper medium which is suitable for replication. Under such conditions optimization *in vitro* and adaptation to changes in the environment were observed just as predicted by Darwin’s principle. Experiments with molecules provided a solution to both problems mentioned above: (i) Generation times are reduced from years to fractions of minutes and evolutionary phenomena can be observed within a few days or weeks, and (ii) RNA molecules capable of replication may be as short as some twenty nucleotides and then the diversity is reduced to about $10^{16}$ variants which is about the sample size that can be currently handled in molecular evolution.

## 3. Molecular evolution and the RNA model

Accumulation of experimental data on evolution is only one side of the coin. As Peter Medawar has precisely stated, „No new principle has ever declared itself from below a heap of facts“, we need a theory that is in a position to bring order into observations and to make predictions. In his seminal works in the early seventies Manfred Eigen developed a theory of molecular evolution that starts out from chemical reaction kinetics and centers on the origin of biological information through replication and mutation. One major result of this and forthcoming studies is the concept of molecular quasispecies (figure 1) and the existence of error thresholds. Populations with asexual reproduction approach distributions of genotypes rather the homogenous states. After sufficiently long times and at mutation rates above a threshold to be discussed later the population approaches stationary mixtures of phenotypes called **molecular quasispecies**. A quasispecies consists of a fittest genotype, the master sequence, and its closely related variants. Relatedness refers here to evolutionary distance commonly counted as the minimal number of mutations converting two sequences into each other. In case of point mutations as the only source of variation this distance is called Hamming distance which is, at the same time, the common metric in sequence space. Quasispecies cover connected areas in sequence space (figure 1).

Increase in mutation rates leads to a decrease in the relative concentration of the master genotype until a sharp threshold is encoutered above which the quasispecies concept brakes down. The population of genotypes does not approach a steady state but migrates through sequence space in random drift like manner. The error threshold defines the maximal diversity of asexually replicating populations which is compatible with inheritance and evolution. At the same time it represents the condition of maximal variability that is needed for evolution in rapidly changing environments. RNA viruses that have to cope with highly efficient defense mechanisms of their hosts were indeed found to operate at mutation rates very close to the threshold. The notion of quasispecies turned out to be very useful in virology. Virus populations resemble quasispecies very closely. They represent the genetic repertoire of viruses from which evolution chooses during infections in order to escape the host’s immune system. Quasispecies theory provides a new and powerful concept for the development of new antiviral strategies.

For evolution, however, the concept of the molecular quasispecies is neither sufficient for understanding  the development of phenotypes nor does it allow predictions on the course of the process as a whole. In order to achieve this goal quasispecies theory, being a version of population genetics that accounts for the molecular machanism of mutation, has to be extended by a description of the migration of populations through sequence space and by an explicit consideration of the relation between genotypes and phenotypes (figure 2).

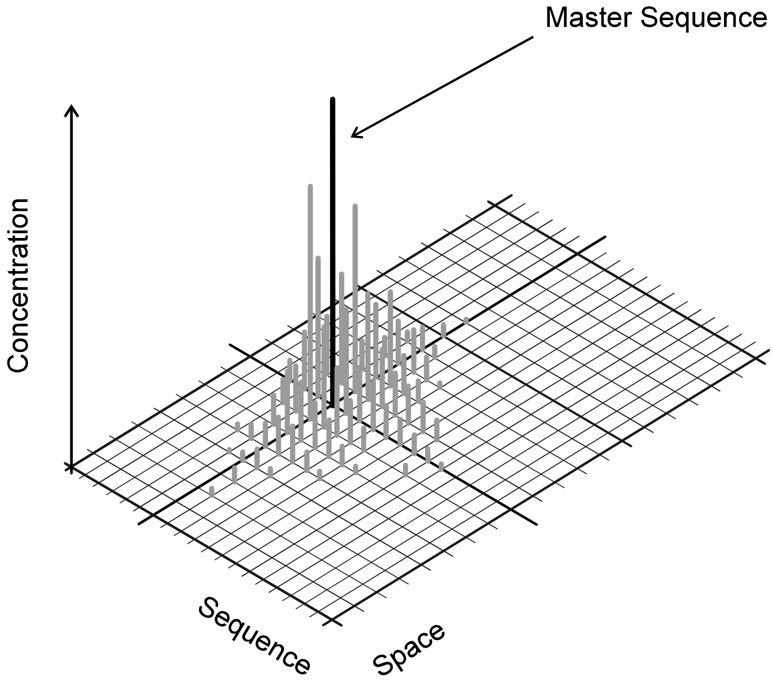


**Figure 1**: The molecular quasispecies. Mutation prevents populations from becoming homogeneous even when they contain the genotype producing the optimal phenotype. A typical population after sufficiently long type in order to equilibrate with respect to mutations consists of a master sequence, usually the most frequent genotype, and its variants. The concentrations at which individual mutants appear in equilibrated populations depend on their own fitness and on the mutational distance from the master sequence, i.e. the minimal number of mutations required to convert the master into the variant.

Genotype-phenotype relations are generally too complex to be studied by analytical techniques. The only currently known example is evolution of RNA molecules. In this case the phenotype is the molecular structure which under defined conditions, thermodynamic equilibrium for example, is fully determined by the sequence. Genotype-phenotype mapping then is reduced to the prediction of structure form known sequences. This is still a very difficult task but it can be solved for a coarse-grained version of RNA structure known as secondary structure. Modeling genotype-phenotype relations by folding RNA sequences into secondary structures lead to the RNA model that allows to study evolutionary phenomena by means of mathematical models and computer simulation. The structure is representative for all molecular properties which are assumed to be derivable from known conformations.

Systematic studies on the RNA model revealed several features which are apparently also relevant for evolution proper. A first result showed that there are many more sequences than structures: sequence-structure mapping of RNA is highly redundant. In addition, we found that relatively few common structures are formed by almost all sequences whereas the majority of structures is rare and corresponds to only one or a few sequences, and, hence,

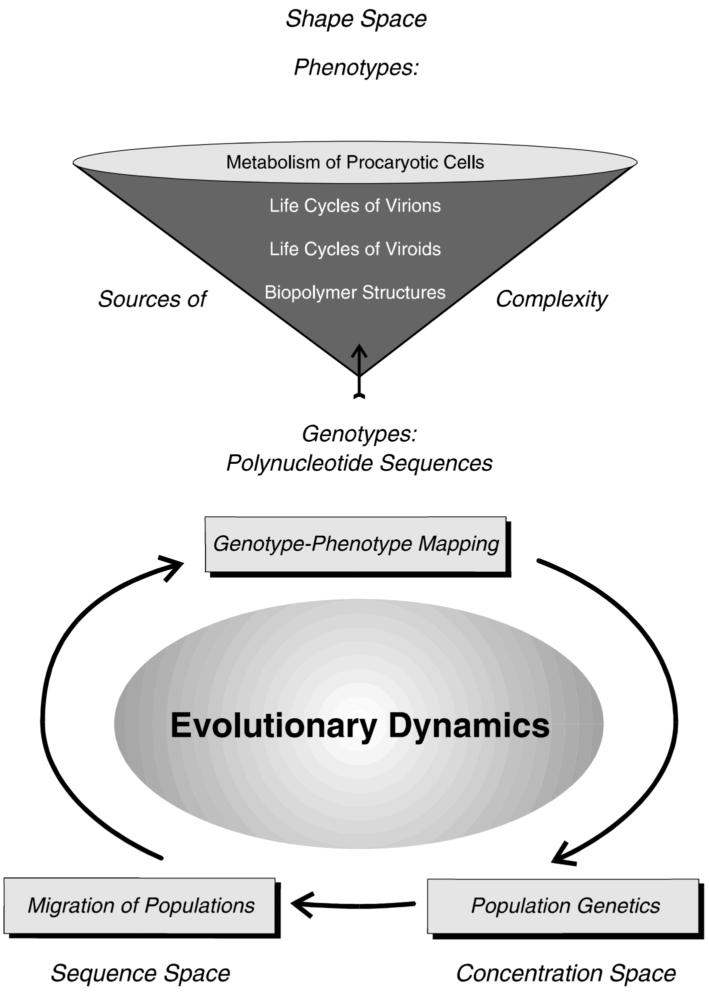


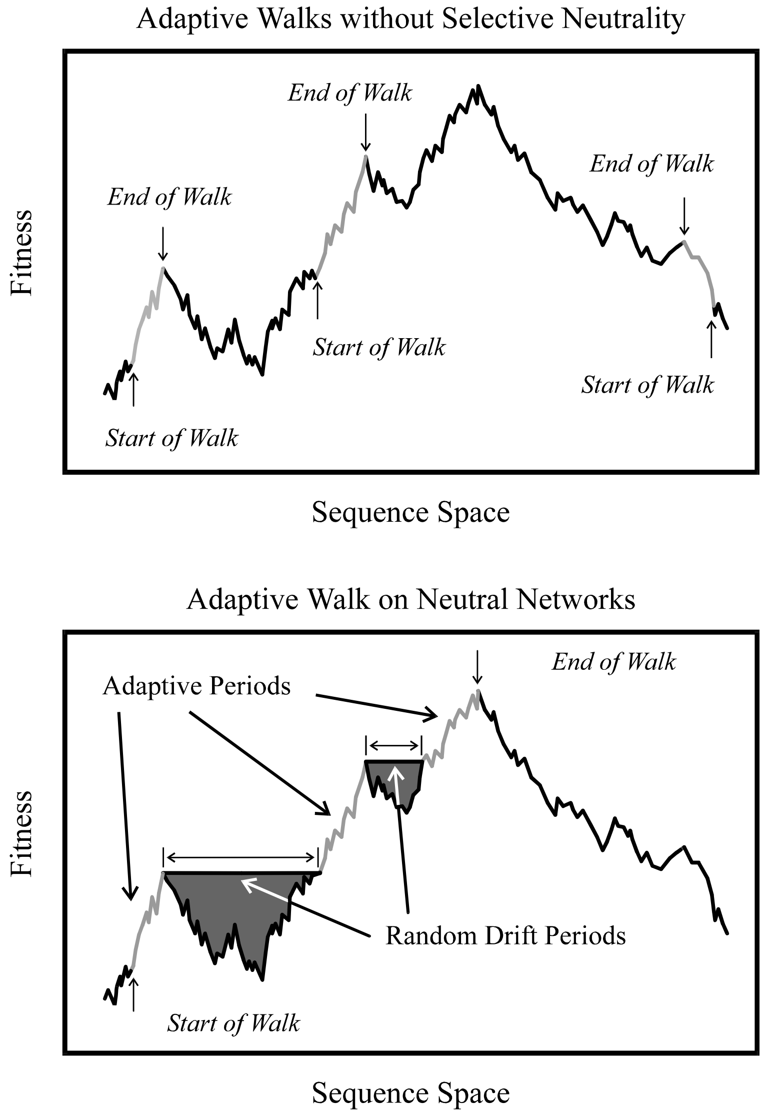
**Figure 2**: A comprehensive model of biological evolution. The highly complex evolutionary process is partitioned into three simpler and thus analyzable dynamical phenomena: (i) population genetics, (ii) migration of populations through sequence space, and (iii) genotype-phenotype mapping. Population genetics describes how optimal genotypes with optimal genes are chosen by natural or artificial selection from a given reservoir, a molecular quasispecies (figure 1) or, in case of higher organisms with sexual reproduction, a gene pool. The processes described by population genetics are derived from chemical reaction kinetics and consist of replication, mutation, and recombination. Gene or genotype frequencies are measured in particle numbers or concentrations. In essence, population genetics is concerned with selection and other evolutionary phenomena occurring on short time-scales. Migration of populations in sequence space is dealing with the change of genetic reservoirs when populations drift through the huge space of possible genotypes. Issues are the internal structure of the populations and the mechanisms by which the regions of high fitness are found in sequence space. Migration of populations deals with long-term phenomena of evolution, for example, with optimization and adaptation to changes in the environment. Genotype-phenotype mapping represents the core problem in evolutionary thinking since the dichotomy between genotypes and phenotypes is reflected by Darwin’s principle of variation and selection: all genetically relevant variation takes place on the genotype whereas only the phenotype is subjected to selection.

they can neither by found by random search nor by evolutionary methods *in vitro* or *in vivo*. Systematic searches will also fail unless they cover all sequences in sequence space which is prohibitive for molecules with chain lenght larger than 30. Accordingly, evolution is dealing only with common structures. The numbers of common structures, however, grow exponentially with chain length as do the numbers of all structures or the numbers of sequences. For moderate chain lenthgs we find already a very rich pool of common structures to choose from in evolutionary searches.

Careful analysis of the data from sequence-structure mapping of RNA revealed among other results the principle of **shape space covering** which is of direct relevance for evolution. In order to find a sequence forming a predefined structure with probability one we need not search the hyperastronomically large numbers of sequences in whole sequence space. Sequences for all common structures are found in relatively small neighborhoods around any arbitrarily chosen reference sequence. All sequences folding into the same structure form a **neutral network** in sequence space. Populations may migrate on neutral networks by random drift and thus can change genotypes without changing phenotype. Common structures form neutral networks which are connected and span whole sequence space. Provided there is sufficient time, a population on such an extended neutral network can reach every region of sequence space still remaining at the same fitness level.

Neutral evolution was discovered in the sixties through comparison of sequences coding for the same gene in different organisms. It fits perfectly into a theory of random drift in sequence space conceived and developed by the population geneticist Motoo Kimura. His neutral theory of evolution provided a proper frame for understanding selective neutrality as a by-product of replication and mutation, and it was turned into a valuable tool for the reconstruction of phylogentic trees from sequence comparisons. The neutral theory, however, did not indicate a positive role that neutrality might play in evolution. Based on the existence of extended neutral networks the course of optimization and adaptation changes with increasing degree of neutrality (figure 3). In the absence of selectively neutral variants Darwinian optimization always walking uphill will soon end on one of the minor peaks of the fitness landscape. The presence of extended neutral networks changes this scenario, because a sequence situated at a minor peak is then part of a network and the population does not get stuck but escapes the local optimum through random drift in other directions. As sketched in figure 3 the evolutionary optimization process then becomes an alternation of fast adaptive periods with strong increases in fitness and quasi-static phases of random drift at constant fitness. As populations can migrate on neutral networks through almost whole sequence space they will ultimately find the global optimum.

Optimization of RNA structures towards a predefined target has been studied by means of computer simulations. A population of some thousand sequences was subjected to replication and mutation in a kind of flow reactor that supplies the materials consumed in RNA synthesis. The population size was kept constant by means of a dilution flux removing randomly chosen excess molecules. A fitness function was defined which increases with decreasing distance from the target structure. As target we chose the structure of a biologically important RNA molecule: a transfer-RNA molecule that represents the link between a coding triplett on the messenger-RNA and an amino acid residue to be incorporated into the protein chain. Several optimization experiments were undertaken on the computer. With no exeption they confirmed the optimization scenario sketched above: the approach of the population towards the target structure occurrs in steps and random drift



Adaptive Walks without Selective Neutrality

Adaptive Walk on Neutral Networks

**Figure 3**: The role of neutral variants in evolution. Optimization of populations in sequence space through adaptive walks is sketched on fitness landscapes without and with selective neutrality. Realistic fitness landscapes are **rugged** and contain a great variety of local maxima at all heights. Adaptive walks allow to choose the next step arbitrarily from all directions where fitness is (locally) non-decreasing. Populations sustain variants covering a certain area of sequence space in the sense of molecular quasispecies (figure 1). Therefore, they can bridge over narrow valleys with widths of a few point mutations. They are unable, however, to span larger distances. In absence of selective neutrality (upper part) populations will approach the next higher fitness peak and therefore usually cannot reach the global optimum. At a sufficiently high degree of selective neutrality many genotypes produce a phenotypes with the same fitness and thus form an extensive neutral network in sequence space. Populations may migrate at constant fitness on neutral networks (lower part) and therby bridge over large low-fitness zones. By a combination of fast adaptive phases and periods of random drift populations can eventually reach the global optimum in a punctuated or stepwise process.

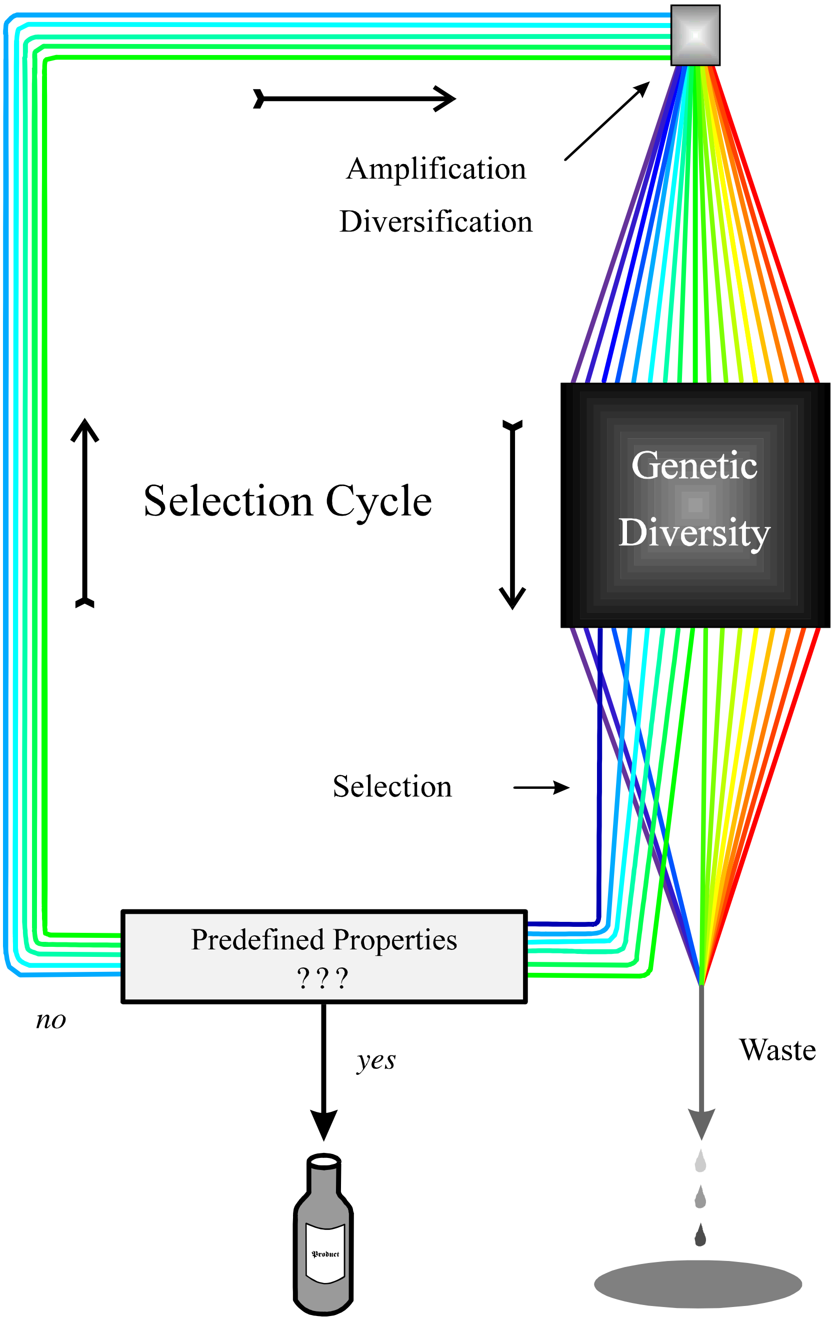


**Figure 4**: Evolutionary *in vitro* design of biopolymers. Properties or catalytic functions of biomolecules are optimized in iterative manner through selection cycles. Every cycle consists of three distinct phases: amplification, diversification, and selection. Amplification is achieved by one of the currently available polynucleotide replication assays. Diversification occurs through replication with artificially high error rates or random synthesis. Currently successful selection techniques apply one of two strategies: (i) selection from homogeneous mixtures using binding to solid phase target or reactive tags that allow to separate suitable molecules from the rest of the mixture and (ii) spatial separation of individual molecular genotypes and large scale screening.

phases are observed, which constitute characteristic plateaus in plots of mean fitness against time. Every plateau ends with a major change in the dominant phenotype that is caused by a single point mutation and leads to an increase in fitness. This abrupt changes are initiated by special genotypes and manifest themselves in punctuation of the evolutionary process. Accordingly, they were called **discontinuous transitions**. During quasi-static periods the population drifts neutrally until it encounters a special genotype. Along the plateaus we either observe a conserved phenotype or an apparently random sequence of related phenotypes with identical fitness. The first case is a typical example of conventional neutral evolution: changing neutral sequences that fold into the same structure. The second scenario is more elaborate. The changes in structures observed thereby occur readily and do not require special genotypes. We have called them therefore **continuous transitions**. Both, continuous and discontinuous transitions can be explained on the molecular level by means of their probabilities of occurrence as a consequence of a point mutation. Continuous transitions, for example, are the formation and the cleavage of single contacts in a structure and thus occur readily. Discontinuous transitions correspond to simultaneous changes of several contacts and therefore they have low probability. They represent true bottle-necks for optimization. Computer simulations have not only shown that the optimization scenario suggested in figure 3 is correct, they provided, in addition, a novel and phenotype related notion of continuity in sequence space which has been derived for molecules but seems to be generally valid for biological evolution.

## 4. Evolution experiments and evolutionary biotechnology

Sol Spiegelman used viral RNA in his evolution experiments and optimized them for high replication efficiency. Later experiments have shown that, in addition to replication rates, several other properties can by optimized by evolution *in vitro*. Examples are, resistance against inhibitors of replication or RNA degrading enzymes. Catalytic RNA molecules were trained to catalyze new reactions. For example, RNA cleaving molecules were turned into DNA cleaving molecules. In all these experiments evolution was carried out continuously with blending generations.

Later on, new experimental techniques also based on variation and selection but with clear separation of generations were developed. The experimenter interferes with the course of evolution by controling the selection process (figure 4) just as the animal breeder selects the individuals which are considered suitable for further breeding. Optimization is carried out in individual selection cycles consisting of three phases: (i) amplification, (ii) diversification, and (iii) selection. The selection process is the one that requires most experimental skill. At present two different strategies are applied. In the first technique suitable molecules are selected either through binding to target molecules or they are marked with tags that can be used to separate them from the rest of the population. The second technique is more elaborate but can be applied universally. The reaction mixture is partitioned into small samples containing single molecules and the molecules are handled in small reaction vessels. They are screened for the desired properties and suitable candidates are selected for the next selection cycle. The latter approach requires high-tech equipment and automation in order to be able to process millions of samples simultaneously.

The concepts of molecular evolution, apparently, are very attractive for application in biotechnology. The evolutionary design of new biopolymers with predefined properties and functions has become reality within the last ten years. In particular, the experiments with

RNA molecules were very successful. Highly specific RNA binders to molecular targets, so-called aptamers, were produced and they are already in use for pharmaceutical and biomedical applications. New RNA molecules were created with novel catalytic functions. The evolutionary design of molecules with new functions is an excellent example to demonstarte the power of the evolutionary approach. Only the desired function has to be known and it has to be incorporated appropriately into selection or screening. It is not necessary to know suitable sequences or structures. They are produced automatically by the evolutionary process.

## 5. Concluding remarks

Tonight I tried to show that complex biological phenomena like evolutionary optimization can be reduced in such a way that the essential features remain still observable whereas most of the sophistication commonly obscuring the underlying principles is lost. Reproduction of organisms has been reduced to replication of molecules in the test tube, to a process that can be studied by the conventional techniques of chemical kinetics and simulated on the computer. Cellular life, thus, is no prerequisite for evolution. One the other hand, we have seen that the dichotomy of genotypes and phenotypes is essential for the Darwinian mechanism: all variation occurs on the genotype while only the phenotype is the target of selection. This separation is the ultimate origin of the uncorrelatedness between occurrence and consequence of mutations that turned out to be a very powerful strategy in optimization. It is, for example, the recipe of success of Monte-Carlo and other optimization techniques with random elements. In case of RNA evolution in the laboratory the phenotype is represented by a molecular structure, and genotype-phenotype mapping boils down to folding RNA sequences into three-dimensional structures.

Redundancy of genotype-phenotype mapping is the basis of neutrality in evolution. Although neutral evolution has been found to occur in nature already about thirty years ago, it was not clear whether it is just an unavoidable by-product of molecular mechanisms or whether it can be exploited to improve the efficiency of evolution. Current studies have shown that the latter is the case: at sufficiently high degress of neutrality populations are not caught in evolutionary traps of mediocre fitness but may proceed towards higher peaks and eventually reach global optima. As a consequence of neutrality, optimization occurs in steps where short adaptive periods of fast increase in fitness are interrupted by long quasi-stationary phases of random drift.

Evolution at molecular resolution revealed a new concept of continuity that allows to explain the occurrence of steps in straightforward manner by the simple mechanism presented here. In general, evolutionary trajectories are not reproducible in detail: a repetition of an evolution experiment always leads to different sequences of genotypes and phenotypes on the way from initial structure to target. But there are also conserved and hence reproducible features like the number of steps required to the reach the goal and the regularities found in continuous and discontinuous transitions. This reproducibility, in essence, is the basis for the success of evolutionary biotechnology, an entirely new discipline that allows to tailor molecules for predefined purposes. Although we have so far learned to understand many aspects of life at the molecular level, many more details wait to be discovered in the future as evolution is an open and never ending process.

## References to more extensive treatises

Peter Schuster. *Evolution in an RNA world*. In: M.G. Ord and L.A. Stocken, eds. Foundations of Modern Biochemistry. Vol. IV: More Landmarks in Biochemistry, pp.159-198. JAI Press Inc. Stamford, CT, 1998.

Peter Schuster. *Genotypes with phenotypes: Advertures in an RNA toy world*. Biophys.Chem. 66:75-110, 1997.

Walter Fontana & Peter Schuster. *Continuity in evolution. On the nature of transitions*. Science 280:1451-1455, 1998.

Manfred Eigen & Peter Schuster. *The hypercycle - A principle of natural self-organization*. Springer-Verlag, Berlin, 1979.