Manfred Eigen-Lecture, Göttingen 09.05.2018

Bridging from Chemistry to the Life Sciences – Evolution seen with the Glasses of a Physicist[a]

Peter Schuster, Institut für Theoretische Chemie, Universität Wien and Santa Fe Institute, Santa Fe

It is a great honor and pleasure for me to present the first Manfred Eigen Award Lecture and I am thankful to the prize committee for having me selected. Fifty years ago I was PostDoc here in Göttingen and I remember well Manfred, who had been awarded the Nobel Prize in Chemistry the year before and who was working on his theory of molecular evolution, coming one day to my desk in a small office still in Bunsenstraße. He was looking for somebody who could compute solutions to differential equations. I had brought a box of punch cards with me from Vienna that contained also a package for numerical integration of ODEs and it was straightforward to fulfill Manfred's task. This was the beginning of a wonderful scientific cooperation, which changed and determined further my entire scientific life, and which lasted for several decades.

At the beginning of my lecture I make two statements that are common place for scientists and would not have been necessary ten years ago. First, biological evolution is a scientific fact like gravitation and others, and second, the theory of evolution is a living scientific discipline. Like all science it is under construction, it becomes more and more complete over the years and will never be finished unless no new ideas are developed and no new experiments are made. What has changed in recent years? I mention only two issues, which apparently seem to require engagement in the fight for science: Biological evolution has been taken off the schedule of schools in a fairly large country near Europe, and to great dismay of the National Academies of the United States intelligent design and creationism see a revival overseas.

The lecture will consist of five parts: (i) a brief introduction to biological thinking exemplified by means of mathematical concepts before Darwin, (ii) theory and mathematics of molecular evolution, (iii) the influence of stochastic phenomena on evolutionary processes, (iv) the role of fitness landscapes in understanding evolution, and (v) a few sentences on the current state of the art of evolutionary theory in the light of modern molecular genetics.

1. Mathematical concepts before Darwin

People were aware of the special properties of reproduction already in medieval times, Fibonacci's multiplying rabbits may serve as an example. The English economist Reverend Thomas Robert Malthus was the first to analyze and discuss the consequences of multiplication on the population level. In his very influential book entitled "*An Essay on the Principle of Population*"[1] published 1798 he pointed out that population growth without birth control will readily consume any increase in a nation's food production. Reproduction without constraint leads to a geometric progression or exponential growth and will outgrow any finite resource and accordingly the struggle for resources is preprogrammed. Malthus thoughts were heavily debated but, nevertheless, very influential, in particular for the early development of the theory of evolution: Both, Charles Darwin and Alfred Russel Wallace were familiar with Malthus' book. The Belgian mathematician Pierre François Verhulst was prompted to conceive a

[a] This review has been presented on May 09, 2018 as the first Manfred Eigen Award Lecture at the Max Planck-Institute for Biophysical Chemistry in Göttigen, GE.

mathematical model for growth in a finite world by reading Malthus' "Essay". At first Verhulst's work found little response in the scientific community but seventy years later it saw many so-called "*rediscoveries*".[2,3,4,5] The logistic equation as Verhulst himself called his expression became particularly popular after the publication of a work by Raymond Pearl and Lowell Reed in 1920. Pearl and Reed applied the logistic curve to the growth of the population of the United States. It is remarkable that Pearl and Reed did not even mention the name Verhulst in their publication.[b] Despite its simplicity, which also caused some criticism in population theory, the Verhulst or logistic differential equation is still in use, for example in population ecology in order to develop models of microbial growth.[6]

Figure 1: **Exponential growth, logistic growth, and selection of the fittest**.

Verhulst[c] conceived a differential equation for growth on limited resources[7,8,9],

$$(1) \qquad \frac{dN}{dt} = rN\left(1-\frac{N}{K}\right) \text{ with the solution } N(t) = \frac{N(0)\,K}{N(0)+\left(K-N(0)\right)\exp(-rt)}$$

where N is the number of individuals X, and called it "*logistic equation*" without giving an explanation for this particular choice of notion. The logistic equation describes the growth of a homogeneous population of size $N(t)$ where all individuals X have the same fitness $f = r$ with r being the so-called Malthus parameter and K being the carrying capacity of the ecosystem. Verhulst models reproduction – the positive term in the differential equation – simply as an autocatalytic process $X \rightarrow 2X$. The negative term, $-r\,N^2/K$, takes care of overpopulation: the growth rate decreases when the population size becomes too large and approaches the value zero for $N = K$. In other words the population stops growing when the carrying capacity of the ecosystem is reached.

A minor generalization, which consists in substituting the homogeneous population, Π = {X}, by a structured population with a distribution of subspecies, Π = {X_1,X_2,…,X_n} with the particle numbers subsumed in the vector $\boldsymbol{N} = (N_1, N_2,\ldots, N_n)$, provides already a mathematical model for "*selection of the fittest*", which allows for a straightforward proof of fitness optimization during selection (figure 1). Solutions to this generalized Verhulst equation

$$(2) \qquad \frac{dN_j}{dt} = N_j\left(f_j - \frac{\sum_{i=1}^{n} f_i N_i}{K}\right) \quad \text{and} \quad \phi(t) = \frac{1}{C}\sum_{i=1}^{n} f_i N_i \quad \text{with} \quad C = \sum_{i=1}^{n} N_i\,; \quad j = 1,2,\ldots,n\,,$$

are readily obtained in terms of normalized variables: $\xi_i = N_i/C$ with $\Sigma_i\ \xi_i = 1$:

$$(3) \qquad \frac{d\xi_j}{dt} = \xi_j\left(f_j - \phi(t)\right) \quad \text{and} \quad \xi_j(t) = \frac{\xi_j(0)\exp(f_j\,t)}{\sum_{i=1}^{n}\xi_i(0)\exp(f_i\,t)} \quad \text{with} \quad \sum_{i=1}^{n}\xi_i = 1\,; \quad j = 1,2,\ldots,n\,.$$

[b] For a historical survey of the logistic equation and its multiple rediscoveries see, for example, P. J. Lloyd (Lloyd, 1967). Often people try to find explanations for names that were given to equations and other objects in science. Sometimes no explanations can be found and the logistic equation seems to be such a case.

[c] According to Sharon Kingsland (Kingsland, 1982, p.30) Verhulst was instigated by his mentor Adolphe Quetelet to work on the problem of population growth. In 1835 Quetelet had proposed that the resistance to population growth was proportional to the square of the speed with which the population size increases in analogy to the resistance a body experiences when it travels in a medium (Quetelet, 1835, volume 2 p.277).

The solution makes use of an integrating factor transformation.[10]

The fittest variant, X_m with $f_m = \max\{f_1, f_2, \ldots, f_n\}$ is picked out of the collection of initially present n subspecies and selected. The time dependence of the mean fitness of the population, $\phi(t)$, can be readily calculated and turns out to be the variance of the fitness values,

$$\frac{d\phi}{dt} = \sum_{i=1}^{n} f_i^2 \xi_i - \left(\sum_{i=1}^{n} f_i \xi_i\right)^2 = \mathrm{var}(f) \geq 0 \,, \tag{4}$$

and hence $\phi(t)$ is a non-decreasing function of time. Accordingly $\phi(t)$ approaches an optimum for long time: $\Pi(t) = \{X_1, X_2, \ldots, X_n\} \Rightarrow \{X_m\} = \Pi(\infty)$. Then the population becomes homogeneous and contains only the fittest variant. The fact that this simple access to an early mathematics of evolution remained unnoticed, although all mathematics needed was available already known twenty years before the publication of Darwin's "*Origin of Species*"[11] is remarkable or even puzzling. The principle of selection as a consequence of reproduction was and is easy to understand but nevertheless caused many debates. On the other hand, there were also more serious problems in evolutionary theory at Darwin's time. Among other things problems concerned inheritance and mutation, processes that are basic for understanding evolution, but no convincing concepts for explanations of these processes were at hand.

Figure 2: **Mutation in the Neo-Darwinian scenario.**

In contrast to the theory of evolution Isaac Newton's concept of gravity was pervaded by mathematics from the very beginnings. Although no contemporary of Newton has seen apples, sheets of paper and feathers falling at the same speed, the abstraction from multiple perturbations was readily accepted. One might ask why simplifying models are received so differently by the scientific audience in biology and in physics. Biologists are predominantly interested in field observations and single cases and not so much in generalizations. The concept of *natural selection* conceived independently by Charles Darwin and Alfred Russel Wallace[12] is rather an exception. Apart from others there is also a second reason for the enthusiastic welcome of Newton's mathematical theory by the scientific community: There is celestial mechanics where the laws of gravity can be seen in action without perturbation but there is no celestial biology.

Natural selection was not the only scientific achievement of Darwin. He made five fundamental contributions to the theory of evolution which are:[13,14]

(i) evolution is a historical fact, species have a finite lifetime and are subjected to change,
(ii) multiplication of species led to biological diversity,
(iii) all life had a common ancestor,[d]
(iv) all change happened gradually, and
(v) natural selection.

So far we were focusing on item (v), *natural selection*, and it will remain the major subject throughout the whole lecture. As we have seen natural selection follows immediately from multiplication through reproduction and finiteness of resources. In other words it is the question, "Does evolution optimize the reproduction relevant traits?", we shall be concerned with.

Figure 3: **From Malthus to the modern synthesis**

[d] The tree of life is a central issue of Darwin's „*Origin of Species*" and indeed the only illustration in Darwin's centennial book shows a sketch of a phylogenetic tree.

Charles Darwin's theory of evolution augmented by August Weismann's concept of the separation between germline and somatic cells is commonly denoted as the Neo-Darwinian theory, which dominated evolutionary thinking in the first two decades of the twentieth century. Mendel's laws of inheritance were "*rediscovered"* around 1900, became a discipline in its own right in the form of genetics, and did not seem to be reconcilable with Neo-Darwinism. The unification of the two concepts succeeded first only twenty years later in the theoretical, mathematical approach of the three famous population geneticists Ronald Fisher,[15] J.B.S. Haldane,[16] and Sewall Wright[17] and was achieved and completed later in form of the "Modern Synthesis" by several experimental evolutionary biologists.[12] A sketch of the "*Growth of Biological Thought*" during the first half of the twentieth century is shown in figure 3. The characteristic view of evolutionary processes at the end of the "Modern Synthesis" is characterized best as "*strong selection – weak mutation*" scenario. Mutations are considered as rare events and populations are almost always in a quasi-stationary state at which in the sense of selection of the fittest only the temporarily fittest variant is present. The role of genes was seen more or less in Ronald Fisher's view: Individual genes are largely independent and epistatic interactions between genes have the nature of perturbations.

2. Theory and mathematics of molecular evolution

Before the advancement of molecular biology no satisfactory mechanism of mutation was known. Indeed the pioneering work of James Watson and Francis Crick on the structure of DNA was a true milestone. It marked not only the beginning of the merger of chemistry and biology it demonstrated also impressively the explanatory power of structural biology. Watson and Crick expressed the first success of structural biology in the explanation of a biological process in their famous sentence: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."[18] The Watson-Crick structure at the same time provided a molecular mechanism for point mutations (figure 4): Assuming the incorporation of a wrong nucleotide through a base mismatch causes a change in the DNA sequence in the next generation and in all future generations, and thus establishes a mutant. The biochemistry of processing DNA, RNA, and protein was soon discovered and explored, the concept of genetic information and its dominant role in cellular biology was established. An enormous number of complex biological processes found a natural explanation on the molecular level. Genotypes were no longer abstract objects but concrete DNA molecules whose nucleotide sequences could be determined and analyzed.

Figure 4: **A molecular mechanism for mutation.**

Manfred Eigen's theory of evolution[19] combines evolutionary thinking with the insights gained from molecular biology. Correct reproduction and mutation are seen as parallel chemical reactions (figure 3), and in this way it becomes possible to analyze the "*weak selection – strong mutation*" scenario. Polynucleotide sequencing provided information on sequence heterogeneity of populations and established the fundamentals of virus and bacterial evolution. It is interesting to compare Eigen's replication-mutation process with a mutation model published by James Crow and Motoo Kimura about the same time:[20] Mutation and selection are considered as completely independent processes (figure 5), and in other words, mutations occur in Eigen's model during the replication process whereas they are caused in the Crow-Kimura approach by other events during the whole lifetime of the organism.

Remarkable is the fact that both mechanisms give rise to exactly the same mathematics in the analysis of the kinetic equations. What is different is the interpretation of the kinetic rate parameters.

Figure 5: **Mutation in Eigen's and in Crow-Kimura's mechanism of evolution.**

Evolution in Darwin's sense and likewise at the molecular level boils down to three basic requirements: (i) competition through reproduction or multiplication, (ii) variation of the inheritable traits of individuals, and (iii) finite resources. Variation in nature comes essentially in two forms: mutation and recombination. Mutation creates new biopolymer sequences (Figure 4) whereas recombination leaves individual parts of DNA sequences unchanged but combines them anew according to Mendelian rules. The inevitable result of (i), (ii), and (iii) is natural selection or, in other words, if the conditions are such that all three requirements are fulfilled, selection will happen, and the efficient combination of variation and selection leads to adaptation to the environmental conditions. Changing environments drive adaptive populations and lead to Darwinian evolution.

Natural selection alone and changing environments, however, are not sufficient to explain all periods of biological evolution on Earth. The occurrence of symbiosis frequently observed in nature, for example, requires the cooperation between competitors for which Darwinian evolution has no mechanistic explanation. Simple models may lead to cooperation through mutual dependence preferentially in the reproduction process. In biological evolution from the *origin of life* to our present day world there is clear evidence that the novel organisms originating during certain periods cannot be interpreted plausibly without the assumption of cooperation between species or subspecies.[21,22,23] Eörs Szathmáry and John Maynard Smith created the notion *major evolutionary transitions* for these periods and they define eight such major transitions.[21,24] A straightforward way to introduce cooperation at the molecular level is catalyzed reproduction as it is postulated in hypercycles.[15] Catalyzed replication involves two catalytic biopolymer molecules: One acts as template and the second one is a replication catalyst. Examples of catalyzed replication are found, for example, with RNA molecules.[25,26,27] Cooperation requires mutual support in the sense A helps B and B helps A. In order to form a hypercyclic organization mutual dependence has to form a closed loop, e.g., A helps B, B helps C, and C helps A. For a fixed number of members, n, hypercyles are the smallest catalytic networks that lead to cooperation.

All terrestrial systems are bound by finite resources and thus we remain with three classes of basic evolutionary processes: 1. competition, 2. cooperation, and 3. variation. Figure 6 shows an illustrative sketch of a three-dimensional Cartesian parameter space.[28,29] An intensity parameter is plotted on each coordinate axis. Fitness f is a measure for the success in evolutionary competition, the effectiveness of cooperation is expressed in terms of a cooperation parameter h, and the intensity of variation is expressed here by a mutation rate parameter p.[e] The three processes together shape evolution and, as sketched in figure 6 the triplet (f, h, p) defines a particular condition for evolution. Finally, we mention that evolution can be understood exclusively in the context of an environment, which is part of the evolving system. In the logistic equation the environment consisted only in the assumption of a constant carrying capacity of the ecosystem. A simple but structured environment is the flow reactor that will be used below in the more detailed discussion of a model for molecular evolution.

[e] Here we consider only mutation for two reasons: (i) In higher organisms recombination is directly coupled with reproduction and (ii) only mutation creates novel biopolymer sequences. The introduction of recombination gives rise to the equations of conventional population genetics, which lead to similar results, although the mathematical analysis is substantially more complicated.

Figure 6: **The parameter space of molecular evolution**.

Each coordinate axis in figure 6 is representative for one particular process: natural selection and survival of the fittest on axis 1 (f, 0,0), formation of symbiotic complexes on axis 2 (0, h, 0), and genetic drift or neutral evolution propelled by random mutation on axis 3 (0,0, p). The coordinate planes spanned by pairs of elementary processes give rise to specific evolutionarily relevant processes:
(i) Plane $\mathcal{A}$: Axes 1 and 3 together combine reproduction driven by fitness (f) and mutation (p). In the simplest case the corresponding equations model evolution of molecules *in vitro*, virus evolution and bacterial evolution without recombination.
(ii) Plane $\mathcal{B}$: The combination of axes 1 and 2 yields a scenario that goes from competitive selection to symbiotic cooperation. Here the availability of resources (a_0) determines evolutionary dynamics apart from fitness (f) and the cooperation parameter (h).
(iii) Plane $\mathcal{C}$: Cooperative dynamics of four and more species often leads to *hard* oscillations and stochastic extinction for small particle numbers. Mutation may reintroduce already extinguished subspecies and thus increase the lifetime of symbiotic systems.

Figure 7: **The flow reactor as a device for modeling evolution.**

A suitable environment of the evolving system can be modeled, for example, by a flow reactor in particular by a continuously fed stirred-tank reactor[30] (CSTR, figure 7). The major advantage of a CSTR is its suitability for both theoretical analysis and straightforward experimental implementation (for *in vitro* evolution see, e.g., Joyce,[31] Koltermann & Kettling[32] and Strunk & Ederhof[33]). We choose the flow reactor for modeling reproduction with consumption of a resource A. Instead of degradation giving rise to finite lifetime of individuals or molecules[19] we implement an unspecific dilution flow with a dilution rate d. A mean finite life time of molecules is then replaced by the mean residence time in the reactor, $\tau_R = d^{-1}$. The reaction system comprises three classes of reactions and *pseudoreactions*:[f]

(5)
$$\begin{aligned} &\text{inflow of resources, } a_0\,d\text{:} && * \rightarrow A \\ &\text{reproduction, } f_k\text{:} && A + X_k \rightarrow 2\,X_k \;;\; k = 1,2,\ldots,n \\ &\text{dilution, } d\text{:} && A \rightarrow \varnothing \quad \text{and} \quad X_k \rightarrow \varnothing \;;\; k = 1,2,\ldots,n \end{aligned}$$

The deterministic system consists of $n + 1$ simultaneous differential equations, one for the resource A and n for the subspecies X_k, $k = 1, 2,\ldots, n$:

(6)
$$\frac{da}{dt} = -a\left(\sum_{k=1}^{n} f_k x_k\right) + (a_0 - a)\,d \quad \text{and} \quad \frac{dx_k}{dt} = x_k\,(f_k\,a - d);\; k = 1,2,\ldots,n\,.$$

Concentrations are denoted by lower case letters, [A] = a , [X_k] = x_k , d is the parameter of the dilution rate, and a_0 is the concentration of the resource A in the stock solution (figure 7). The dynamical system in the flow reactor has two stationary states:[34] (i) the state S_0, the "state of extinction" with S_0 = ($a = a_0$, $x_k = 0\ \forall k = 1,\ldots, n$) at which only the resource A is present in the reactor, and (ii) the state $S_1^{(m)}$, the "state of selection". In the deterministic system selection implies selection of the fittest with $S_1^{(m)}$ = ($a = d/f_m$, $x_m = a_0 - d/f_m$, $x_k = 0\ \forall\ k \neq m$). Stability analysis yields asymptotic stability of $S_1^{(m)}$ or selection for

[f] Pseudoreactions are processes in chemistry that do not change the molecules involved but otherwise play the same role as reactions in reaction mechanisms. Inflow into and outflow from the reactor are pseudoreactions.

sufficiently small dilution rate parameters, $d < a_0\, f_m$, and stability of S_0 or extinction for $d > a_0\, f_{\mathrm{m}}$. The dilution rate determines the mean residence time of a reaction volume in the flow reactor, $\tau_R = d^{-1}$, and if τ_R is too short no reproduction can be completed before the volume element with the molecules leaves the reactor. In other words, when the flow is too fast or reproduction is too slow all autocatalytic molecules in the reactor are diluted out and the system goes extinct.

In the following paragraphs we choose three examples for processes confined to one of the three coordinate planes. We start by quasispecies formation, the best understood process on the selection-mutation plane $\mathcal{A} = (f, 0, p)$, which combines reproduction and mutation is described by Manfred Eigen's selection equation[15]

$$\frac{dx_j}{dt} = \sum_{i=1}^{n} Q_{ji}\, f_i\, x_i - \phi(t)\, x_j\,;\; j = 1,2,\ldots,n;\; \phi(t) = \frac{\sum_{i=1}^{n} f_i\, x_i}{\sum_{i=1}^{n} x_i} = \sum_{i=1}^{n} f_i\, \xi_i \quad , \tag{7}$$

wherein the mutation factor Q_{ji} is the frequency at which subspecies $\mathrm{X_j}$ is obtained as an error copy of $\mathrm{X_i}$, f_{i} is the fitness of subspecies $\mathrm{X_i}$ as before, and $\phi(t)$ is the mean fitness of the population. Equation (7) sustains only one asymptotically stable stationary state, which has been called the *quasispecies*,[35] because it represents the genetic reservoir of an asexually reproducing species as the *species* does in case of sexual reproduction with obligatory recombination. Applying some simplifications known as *uniform error rate model*[36] all n^2 mutation factors can be expressed by three parameters: (i) the mutation rate parameter p represented by the mutation rate per nucleotide and replication event, (ii) the chain length of the polynucleotide sequence l, and (iii) the Hamming-distance $d_{\mathrm{ij}} = d_{\mathrm{H}}(\mathrm{X_i},\mathrm{X_j})$ between the two sequences $\mathrm{X_i}$ and $\mathrm{X_j}$:

$$Q_{ij} = (1-p)^{d_{ij}}\; p^{l-d_{ij}} = \varepsilon^{d_{ij}}\,(1-p)^{l} \quad \text{with} \quad \varepsilon = (1-p)/p \tag{8}$$

In the non-neutral case – fitness values f_i are different – a typical quasispecies consists of a most frequent master sequence $\mathrm{X_m}$ and a cloud of mutants. The stationary state is unique, and so is the quasispecies. Instead of selection of the fittest we are now dealing with *selection of the fittest distribution of subspecies* (figure 8). In the *weak selection-strong mutation* scenario the concept of selection of the fittest is weakened: What is selected is not a fittest subspecies but a fittest distribution of sequences.

Figure 8: **The quasispecies as a function of the mutation rate parameter p.**

Considering quasispecies as a function of the mutation rate parameter p provides a surprising result. At not too large mutation rates the mutant distribution as a function of p behaves as expected: In the no mutation limit at $p = 0$, selection of the fittest subspecies takes place and after sufficiently long time the population consists exclusively of $\mathrm{X_m}$, $\Pi = \{\mathrm{X_m}\} = \{\xi_{\mathrm{m}} = 1,\ \xi_{\mathrm{k}} = 0;\ k=1,\ldots,n;\ k \neq m\}$, for increasing p-values, $0 < p \leq p_{\mathrm{thr}}$, the long-time population sustains the typical mutation distribution of quasispecies, which consists of a most frequent master genotype $\mathrm{X_m}$ and its mutants and becomes broader with larger mutation rates. This means the concentration of the master sequence goes down and the mutant cloud becomes larger with increasing p. At a critical mutation rate, the error threshold $p = p_{\mathrm{thr}}$, there exists a sharp transition from an ordered mutant distribution to a uniform distribution extending over sequence

space. With the assumption of the uniform error rate model[27] an analytical approximation of the critical mutation rate parameter can be derived:[19,37,38,g]

$$(9) \qquad p_{\mathrm{thr}} \approx \ln\sigma_m / l \text{ with } \sigma_m = f_m \big/ \bar{f}_{-m} \text{ and } \bar{f}_{-m} = \sum_{k=1,k\neq m}^{n} f_k \xi_k \Big/ (1-\xi_m) .$$

A straightforward interpretation of the error threshold phenomenon makes use of a plausibility argument from error propagation: Too many errors block transmission through making messages unreadable. The discovery of the error threshold phenomenon was the initiator for the development of novel antiviral strategies based on the increase of mutation rates by application of mutagenic pharmaceutical compounds. Two different mechanisms and their interplay were discussed: virus entry into error catastrophe[39] and lethal mutagenesis.[40] A mechanism using the fraction of lethal variants to distinguish between direct extinction or crossing the error threshold before extinction has been proposed.[41] Virus extinction is an excellent example where information on the local distribution of fitness values is required for definite conclusions on the mechanism (see section 4 and Schuster[42]).

The coordinate plane $\mathcal{B}$ spanned by the two axes 1 (f, 0,0) and 2 (0, h, 0) represents illustrative examples for transitions on the route from competition (h = 0) to cooperation (Δf = 0). Abundance of resources[23] is the most important prerequisite for the occurrence of these transitions and the evolution of cooperation is shown best along paths with increasing resources (a_0). In simple cases the positions in parameter space where the transitions occur are given by bifurcations and can be calculated analytically.[43] Formally the competition-cooperation transitions correspond to the major transitions in evolution of Maynard Smith and Szathmáry.[22] Most major transitions, however, involve complex organisms and require often quite elaborate interactions. Symbioses may serve as examples[44] and among symbiotic interaction the endosymbiosis integrating prokaryotic cells into eukaryotic organisms is particularly important.[45]

Figure 9: **Bifurcations along the path from competition to cooperation for three subspecies** (n = 3).

The cooperative unit built from three subspecies is a hypercycle with n = 3 in the simplest case with the catalytic interactions: …$X_3 \leftarrow X_1 \leftarrow X_2 \leftarrow X_3$… . Catalysis is introduced as an additional molecular factor in the rate equations:

$$(10) \qquad \text{catalyzed reproduction, } h_k\text{:} \qquad A + X_k + X_{k+1} \rightarrow 2\,X_k + X_{k+1} \text{ ; } k = 1, 2, 3; k \bmod 3.$$

The kinetic differential equations contain now uncatalyzed (5) and catalyzed reproduction (10):

$$(11) \quad \frac{da}{dt} = -a\left(\sum_{k=1}^{3}(f_k + h_k x_{k+1})x_k\right) + (a_0 - a)d \quad \text{and} \quad \frac{dx_k}{dt} = x_k\left((f_k + h_k x_{k+1})a - d\right); k = 1,2,3 .$$

Equation (11) cannot be solved analytically but stationary states can be calculated and discussed by qualitative analysis. In principle there are 2^3 = 8 steady states[h] We consider the sequence of stable stationary states as a function of increasing available resources represented by the inflow concentration

[g] We mention that on some smooth artificial fitness landscapes the transition from the quasispecies domain to the uniform distribution occurs without passing a sharp threshold (Wiehe, 1997; see also section 4).

[h] The condition $dx_k/dt = 0$ yields two possible solutions for each of the three subspecies and eight combinations are possible. For the cooperative state S_3 the stationary concentration of A is obtained from a quadratic equation. Below the critical dilution rate d* we are dealing with two solutions, S_3 and an unstable state S_3', and above $d = d$* the state S_3 does not exist (for details see Schuster, 2018).

a_0 and denote the states by $S_m^{(k)}$ with sub- and superscripts. The subscript gives the number of subspecies present at the steady state: S_0 implies extinction – no subspecies present, S_1 stands for selection, $S_1^{(1)}$ implies selection of subspecies X_1, $S_2^{(2)}$ is the exclusion state at which X_2 is excluded and X_1 and X_3 are present, and S_3 finally is the cooperative state – all three subspecies, X_1, X_2 and X_3, are present. The sequence of states with increasing resources a_0 then is

extinction $\leftrightarrow$ **selection** $\leftrightarrow$ **exclusion** $\leftrightarrow$ **cooperation** .

There are three selection and three exclusion states, which one of the three state is stable depends on the choice of the parameters f_1, f_2, f_3, h_1, h_2, and h_3, respectively. In the example presented here we choose $f_1 > f_2 > f_3$ and $h_1 = h_2 = h_3 = h$, and then the sequence of states is $S_0 \leftrightarrow S_1^{(1)} \leftrightarrow S_2^{(2)} \leftrightarrow S_3$. In table 1 analytical expressions are given for the individual states.

Table 1: **Asymptotically stable stationary states in the competition-cooperation system with** $n = 3$. The four stable stationary states are ordered with respect to increasing a_0 values of their asymptotically stable regimes. The relations $f_1 > f_2 > f_3$ and $h_1 = h_2 = h_3 = h$ between the rate parameters were assumed. For the cooperative state S_3 the stationary concentration of A is obtained as from the quadratic equation $a^2 + (a_0 + \psi)\, a + d\, \phi = 0$ with the two sums $\psi = \Sigma_i\, f_i/h$ and $\phi = \Sigma_i\, f_i/h_i = 3/h$, where the negative root, $\alpha = \left(a_0 + \psi - \sqrt{(a_0+\psi)^2 - 4d\phi}\right)/2$, represents the stable solution. The existence of the cooperative state S_3 requires a sufficiently small dilution rate: $d \leq (a_0+\psi)^2/(4\phi)$. The expressions with slightly modified notation are taken from ref.28.

name	symbol	stationary values				stability range
		$\bar{a}$	$\bar{x_1}$	$\bar{x_2}$	$\bar{x_3}$	
extinction	S_0	a_0	0	0	0	$0 \leq a_0 \leq d/f_1$
selection	$S_1^{(1)}$	d/f_1	$a_0 - d/f_1$	0	0	$d/f_1 \leq a_0 \leq d/f_1 + (f_1\text{-}f_3)/h$
exclusion	$S_2^{(2)}$	d/f_1	$(f_1\text{-}f_3)/h$	0	$a_0 - d/f_1 - (f_1\text{-}f_3)/h$	$d/f_1 + (f_1\text{-}f_3)/h \leq a_0 \leq d/f_1 + (2f_1\text{-}f_2\text{-}f_3)/h$
cooperation	S_3	α	$(d - f_3\alpha)/(h\alpha)$	$(d - f_1\alpha)/(h\alpha)$	$(d - f_2\alpha)/(h\alpha)$	$d/f_1 + (2f_1\text{-}f_2\text{-}f_3)/h \leq a_0$

The transitions $S_0 \leftrightarrow S_1^{(1)} \leftrightarrow S_2^{(2)} \leftrightarrow S_3$ occur at transcritical bifurcations.[i,46] The bifurcation point indicates in each case the endpoint of the stability range (table 1): (i) $S_0 \leftrightarrow S_1^{(1)}$ at $a_0 = d/f_1$, (ii) $S_1^{(1)} \leftrightarrow S_2^{(2)}$ at $a_0 = d/f_1 + (f_1\text{-}f_3)/h$, and (iii) $S_2^{(2)} \leftrightarrow S_3$ at $a_0 = d/f_1 + (2f_1\text{-}f_2\text{-}f_3)/h$. In addition, there is an unstable point S_3' that correspond to the second root of the quadratic equation $\bar{a}(S_3') = \left(a_0 + \psi + \sqrt{(a_0+\psi)^2 - 4d\phi}\right)/2$. In the (d,a_0)-plane the two stationary points S_3 and S_3' originate from a saddle-node bifurcation at the critical dilution rate $d^* = (a_0+\psi)^2/(4\phi)$. The coordinates of the stationary points $S_k(\bar{a}\,, \bar{x_1}\,, \bar{x_2}\,, \bar{x_3}\,)$ are either constants or linear functions of the two external parameters a_0 and d. The only exception is S_3, where $\bar{a}$ is obtained from a quadratic equation. From the initially discussed general properties of the flow reactor follows that every state except extinction can be destabilized by raising the dilution rate parameter d above a critical value d^*. For the states $S_1^{(1)}$ and $S_2^{(2)}$ one coordinate becomes negative beyond d^*; the cooperative state S_3 passes the saddle-node bifurcation and vanishes. Generalization to arbitrary numbers of subspecies n is easily possible: The number of summands

[i] Transcritical bifurcations belong to the simplest class of bifurcations: Two fixed points, for example a stable and an unstable one, "collide" in parameter space and exchange their properties. At the bifurcation point the unstable point becomes stable and the stable point unstable.

in ψ and ϕ increases from 3 to n and the number of stable states goes up from 2 to n-1. The dynamics on the simplex S_3[j] can be sketched by diagrams showing only the fixed points and their stability (figure 9). In our example, n=3, the simplices are equilateral triangles, extinction is characterized by three unstable points at the three corners, selection shows one stable point at one corner, exclusion one stable point at one edge, and cooperation finally one stable point in the interior. Further unstable points may occur at the boundary of or outside the simplex.

Figure 10: **The mutation threshold in hypercycles with five members**.

The third and last coordinate plane combines cooperation and mutation.[29] We dispense here from all details and mention only that the lifetime of short-lived oscillating symbiotic systems can be extended when the mutation frequency exceeds a certain critical value. This threshold phenomenon is entirely stochastic – extinction of hypercycles does not occur in the deterministic system. In addition stochastic scatter is rather large and accordingly experimental detection will be rather hard. An example of such a mutation threshold in the mean lifetime of hypercycles with five members is shown in figure 10.

3. Stochasticity in evolution

The majority of biological and chemical models are formulated in the language of differential equations but applications to situations where small particle numbers are important and then stochasticity may lead to large errors. As an illustrative example we consider selection in a population of reproducing individuals. Two major effects of small numbers dominate: (i) Stochastic delay and (ii) the undermining of selection of the fittest. The first effect is very general and results from the discretization of the continuous variables in autocatalytic processes. It depends on the number of initially present autocatalytic particles and is largely independent of the total population size. The second effect is biologically more relevant, because the condition of *selection of the fittest* is seriously weakened. The deterministic system modeled by ODE (2) has only one asymptotically stable solution, the state $S_1^{(m)}$, that describes selection of the fittest: $\lim_{t\to\infty} \xi_m(t) = \overline{\xi_m} = 1$ and $\lim_{t\to\infty} \xi_k(t) = 0$ for $k = 1,2,\ldots; k \neq m$. To phrase it in popular language: All initially present subspecies, X_i with $\xi_i(0) > 0$, take part in the contest and the fittest always wins. Since the variables can become arbitrarily small they do not vanish at finite times. In stochastic processes the smallest values, which the variables for particle numbers can take on, are zero and one and therefore the variables N_k can vanish at all times. Then there is no reason why the fittest subspecies X_m could not disappear, although such an event might have quite low probability. In our simple system we have no mutation and then once a variant has disappeared, it is gone and it would never ever come back. This stochastic effect gives rise to Muller's ratchet named after the American geneticist Hermann Joseph Muller:[47,48] Asexually reproducing populations may lose their most advantageous genotypes by series of random events, whereas sexual recombination allows for keeping the best genes in the population. As a matter of fact Muller has used the ratchet as an argument for the advantage of sex. Returning to the outcome of selection we conclude that the prediction of survival of the fittest has to be replaced by a probabilistic result.

Stochastic selection is properly analyzed in the flow reactor for reproduction (figure 7, equations 5,6). Several models used for the deterministic simulations have marginal stability and the corresponding stochastic systems are unstable because of random drift. Examples are the linear birth and death process

[j] A simplex $S_k = \{\boldsymbol{x} \in \mathbb{R}^k: x_1+x_2+x_3=1, x_i \geq 0, i=1,\ldots,k\}$ is the set of all positive unit vectors of dimension k.

with equal birth and death rate parameters (see, e.g., Goel & Richter-Dyn[49] and Schuster[50]). The model equations (2) and (7) fall into the same class[51] as does the Wiener process but implementation in the flow reactor fixes the problem and leads to quasistationarity.[k] In the stochastic selection model implemented in the flow reactor (5) every pure state $S_1^{(k)}$ is a quasistationary state and the state of extinction S_0 is the only absorbing state of the system. The situation can be illustrated straightforwardly: If a giant fluctuation pushes the system sufficiently far away from the quasistationary state, it progresses further either to another quasistationary state or to an absorbing state. Only in the latter case the process has come to a definite end. Since the occurrence of giant fluctuations has extremely low probability, it may take very long time before such a large fluctuation happens.

Table 2: **Probability of selection of three subspecies with initial particle numbers** $X_1(0)=X_2(0)=X_3(0)=1$**.** The values are selection probabilities times 100 for the three subspecies $X_1(t_e)$, $X_2(t_e)$, and $X_3(t_e)$; $A(t_e)$ is the probability of extinction $X_1(t_e) = X_2(t_e) = X_3(t_e) = 0$. Choice of parameters: $\Delta f = f_1 - f_3$, $f = 0.1$ [$M^{-1}\cdot t^{-1}$], $f_1 = f + \Delta f/2$ [$M^{-1}\cdot t^{-1}$], $f_2 = f$ [$M^{-1}\cdot t^{-1}$], $f_3 = f - \Delta f/2$ [$M^{-1}\cdot t^{-1}$], and t_e is the computer time of the simulation.[l] The external parameters of the flow reactor are $d = 0.5$ [$V\ t^{-1}$], and $a_0 = N/2$.

$\Delta f/f$	t_e	Population size N = 100				Population size N = 200			
		$A(t_e)$	$X_1(t_e)$	$X_2(t_e)$	$X_3(t_e)$	$A(t_e)$	$X_1(t_e)$	$X_2(t_e)$	$X_3(t_e)$
0.0	600	**1.5** ± 1.3	**30.5** ± 3.9	**34.2** ± 4.6	**33.4** ± 4.1	**0.5** ± 0.9	**30.6** ± 4.6	**30.9** ± 5.0	**32.0** ± 4.7
0.02	600	**1.8** ± 1.4	**41.8** ± 4.8	**32.9** ± 3.8	**23.4** ± 4.0	**0.6** ± 0.8	**50.4** ± 5.7	**27.7** ± 4.9	**17.3** ± 2.6
0.04	400	**2.4** ± 2.1	**45.4** ± 5.0	**31.3** ± 4.5	**19.9** ± 2.5	**0.7** ± 0.8	**58.3** ± 4.6	**25.6** ± 4.5	**11.0** ± 2.9
0.1	400	**2.1** ± 1.7	**59.8** ± 5.5	**28.0** ± 4.1	**10.0** ± 2.9	**0.4** ± 0.5	**73.9** ± 4.1	**20.6** ± 3.5	**4.8** ± 1.9
0.2	400	**1.9** ± 1.1	**68.3** ± 4.5	**23.1** ± 3.7	**6.7** ± 2.8	**0.5** ± 0.7	**76.6** ± 4.1	**19.3** ± 2.8	**3.6** ± 1.7
0.4	400	**2.3** ± 1.8	**71.7** ± 6.0	**20.8** ± 5.2	**5.2** ± 2.4	**0.9** ± 0.6	**82.0** ± 4.2	**13.8** ± 3.8	**3.3** ± 1.7
1.0	200	**2.7** ± 2.4	**78.4** ± 4.7	**15.8** ± 3.3	**3.1** ± 1.5	**0.9** ± 0.9	**83.6** ± 4.0	**12.6** ± 3.2	**2.9** ± 1.5
1.8	200	**4.3** ± 1.1	**80.8** ± 2.9	**13.6** ± 3.1	**1.3** ± 1.2	**1.5** ± 1.3	**83.8** ± 3.3	**12.7** ± 2.5	**2.0** ± 1.7

At small discrete particle numbers individual subspecies may become extinct and each one, which was present initially, may be selected. The principle of selection of the fittest becomes obsolete and has to be replaced by a distribution of probabilities of selection, which are also known as *fixation probabilities of genotypes in populations*.[m] We illustrate by means of a simple example considering three competing subspecies in table 2: Probabilities of long-time appearance were calculated for the four states, S_0, $S_1^{(1)}$, $S_1^{(2)}$, and $S_1^{(3)}$, and compared for different population sizes and selection coefficients $\Delta f/f = (f_1\text{-}f_3)/f$.

[k] In stochastic processes absorbing states are distinguished from quasistationary states. When a trajectory of the system has reached an absorbing state it remains there forever. Trajectories approach also quasistationary states and fluctuate around them. Systems may stay in or near quasistationary states for very long even arbitrarily long times but in the limit $t \to \infty$ all trajectories must converge to one of the absorbing states.

[l] Unambiguous counting of selection scores requires the choice of a final time at which all selection processes have come to an artificial end.

[m] It is important to distinguish genotypes and genes: The genotype or genome contains the complete genetic information. It comes in different variants or subspecies and is the target of selection in asexual reproduction. Fixation is the process of selection that leads from distribution of genotypes to a homogeneous population of the selected variant. Genes are best visualized here as pieces of the genome, which have a defined function. *Alleles* are variants of genes and fixation of a given allele implies that all other variants have disappeared in the population.

At the state S_0 all three species X_1, X_2, and X_3 are extinct and only the resource A is present, at $S_1^{(1)}$ the particle numbers of X_2 and X_3 have vanished and subspecies X_1 has been selected, and at the other two selection states, $S_1^{(2)}$ and $S_1^{(3)}$, X_2 and X_3 are selected, respectively. The probability of selection depends in essence on three factors: (i) the fitness of the variant relative to the rest of the population as expressed by the selection coefficient $\Delta f / f$ (table 2), (ii) the numbers of initially present individuals, $X(0)$, (table 3), and (iii) the total population size N (table 4). Table 2 covers also the neutral case in the sense of Motoo Kimura[52] where all fitness parameters are equal: $f_1 = f_2 = f_3 = f$. Then we expect to find and obtain almost equal probabilities for all subspecies and the deviations from the uniform distribution provide insight into the population size dependent natural fluctuations. As expected fitness differences are most strongly reflected by the distributions of selection probabilities. In table 3 we make a closer look on the dependence of selection probabilities on initial conditions Recalling that mutations start always from a single copy we conclude that the probability of selection of X_k from a single initial copy, $X_k(0) = 1$, is the relevant quantity for the fixation of mutants. As expected the dependence on initial conditions is strong and the largest difference is found between $X(0) = 1$ and 2. For a typical medium selection coefficient as applied here, $\Delta f / f = 0.1$, the probability of selection of the fittest goes up to about 85% when the initial values are raised to $X(0) = 10$.

Table 3: **Dependence of selection probabilities on initial conditions** $X_1(0)$, $X_2(0)$, and $X_3(0)$. The values are selection probabilities times 100 for the three subspecies $X_1(t_e)$, $X_2(t_e)$, and $X_3(t_e)$; $A(t_e)$ is the probability of extinction $X_1(t_e) = X_2(t_e) = X_3(t_e) = 0$. Choice of selection coefficient: $\Delta f / f = 0.1$,.and $f = 0.1$ [$M^{-1} \cdot t^{-1}$]. Further parameters see caption of table 2.

$X_1(0)=X_2(0)=X_3(0)$	t_e	N	**A**(t_e)	$X_1(t_e)$	$X_2(t_e)$	$X_3(t_e)$
1	400	100	**2.1** ± 1.7	**59.8** ± 5.5	**28.0** ± 4.1	**10.0** ± 2.9
2	400	100	**0.1** ± 0.3	**73.5** ± 4.2	**22.4** ± 4.4	**4.0** ± 1.4
3	400	100	**0**	**77.0** ± 4.5	**20.7** ± 4.1	**2.3** ± 1.6
4	400	100	**0**	**79.7** ± 3.1	**18.2** ± 4.1	**2.1** ± 1.3
5	600	100	**0**	**83.2** ± 4.8	**14.5** ± 4.6	**2.3** ± 1.8
10	600	100	**0**	**85.8** ± 3.8	**13.4** ± 3.7	**0.8** ± 0.8

The population size dependence is summarized in table 4. The probability of selection of the fittest increases from 72% to 91% for a selection coefficient $\Delta f / f = 0.4$ and from 81% to 94% for $\Delta f / f = 1.8$. A population size of $N = 800$ is sufficient to reduce the "misselection" to about 6%. Nevertheless, considering random effects shows appreciable deviation from selection of the fittest in the stochastic approach to evolution at small numbers.

Another important stochastic feature of evolution concerns the structure of "*discrete quasispecies*" as occurring in reality. The solution of the kinetic differential equation is given by the largest eigenvector of the selection-mutation or value matrix, $W = Q F$, which extends over the entire sequence space. Such a wide spreading is, of course, not possible in real systems since particle numbers have to be integers and the largest populations in test tube experiments with RNA molecules are about 10^{15}. This has the striking

implication that even large stationary populations hardly contain genotypes that are more than five mutations away from the master sequence.[n] What happens at the error threshold in finite populations? Certainly real populations can't fill sequence spaces uniformly. In a snapshot they will occupy only small connected areas. Then mutation creates new sequences whereas some of the old sequences become extinct. Above threshold populations are expected to migrate randomly through sequence space. Specific predictions on random drift are very difficult and commonly quite uncertain without detailed knowledge on the distribution of fitness values in sequence space (see section 4). Illustrative examples of computer simulation of RNA evolutions are found in the literature.[53,54,55] Fluctuating environments may impose additional random effects on quasispecies dynamics.

Table 4: **Dependence of selection probabilities on population size** N**.** The values are selection probabilities times 100 for the three subspecies $X_1(t_e)$, $X_2(t_e)$, and $X_3(t_e)$; $A(t_e)$ is the probability of extinction, which implies $X_1(t_e) = X_2(t_e) = X_3(t_e) = 0$. Initial conditions: $X_1(0) = X_2(0) = X_3(0) = 1$; mean fitness $f = 0.1$ $[M^{-1} \cdot t^{-1}]$. Further parameters see caption of table 2.

$\Delta f / f$	t_e	N	**A**(t_e)	**X$_1$**(t_e)	**X$_2$**(t_e)	**X$_3$**(t_e)
0.4	400	100	**2.3** ± 1.8	**71.7** ± 6.0	**20.8** ± 5.2	**5.2** ± 2.4
0.4	400	200	**0.9** ± 0.6	**82.0** ± 4.2	**13.8** ± 3.8	**3.3** ± 1.7
0.4	200	400	**0.1** ± 0.3	**86.3** ± 4.6	**12.4** ± 4.0	**1.2** ± 1.2
0.4	100	800	**0**	**90.6** ± 2.3	**8.5** ± 2.0	**0.9** ± 1.0
1.0	200	100	**2.7** ± 2.4	**78.4** ± 4.7	**15.8** ± 3.3	**3.1** ± 1.5
1.0	200	200	**0.9** ± 0.9	**83.6** ± 4.0	**12.6** ± 3.2	**2.9** ± 1.5
1.0	200	400	**0.2** ± 0.4	**88.9** ± 3.3	**9.5** ± 2.4	**1.4** ± 1.0
1.0	100	800	**0**	**91.8** ± 2.4	**7.5** ± 2.0	**0.7** ± 0.7
1.8	200	100	**4.3** ± 1.1	**80.8** ± 2.9	**13.6** ± 3.1	**1.3** ± 1.2
1.8	200	200	**1.5** ± 1.3	**83.8** ± 3.3	**12.7** ± 2.5	**2.0** ± 1.7
1.8	200	400	**0.6** ± 0.7	**88.8** ± 3.1	**9.0** ± 3.2	**1.6** ± 1.5
1.8	100	800	**0.1** ± 0.3	**93.7** ± 2.5	**5.7** ± 2.4	**0.5** ± 0.7

[n] For sequences of chain length 1000 in the natural alphabet we find 4.49×10^9 sequences up to Hamming distance three from a central master sequence. Hamming distance five requires 2×10^{15} molecules, which is about 3000 times larger than the pools of random sequences used in aptamer selection experiment (Keefe and Szostak, 2001).

4. Fitness landscapes and evolution

No theory of evolution is complete as long as it does not comprise a possibility to derive the fitness of a genotype within the method itself. Evolutionary dynamics as such is fairly simple but the relations between genotypes, phenotypes, and fitness values are highly involved and represent the source of complexity in evolution. This relation is often addressed as *fitness landscape* and commonly attributed to species. It is important to realize that fitness values and accordingly also fitness landscapes are strongly dependent on environments.[56] The idea of a *fitness landscape* has been conceived by the American population geneticist Sewall Wright.[57] In his illustration he considered different alleles at one locus of a diploid organism with sexual reproduction and recombination in the sense of Gregor Mendel, and he constructed a landscape by assigning a fitness value to every allele. Combining n alleles yields n^2 combinations out of which only $n(n+1)/2$ are different because in case of autosomes[o] it does – in a first approximation – not matter, which chromosome, the maternal or the paternal, carries the allele. Fitness landscapes for asexual reproduction are much more easily interpreted and modeled. Therefore, in the forthcoming discussion we shall restrict our model to asexual species although the generalization to sexual recombination is straightforward.

The concept of fitness landscapes was originally developed for the purpose of illustration only and most fitness landscapes are sketched in 3D-space, although all realistic sequence spaces are high-dimensional. Indeed, the support of a fitness landscape is sequence or genotype space $\mathbb{Q}$ (figure 11), which is a point space with dimension $\mathrm{M} = |\mathbb{Q}| = 2^l$ for binary or $\mathrm{M} = |\mathbb{Q}| = 4^l$ for natural four-letter sequences, and the appropriate metric upon sequence space is the Hamming distance d_{H}. The cardinality of sequence space is enormous[n] and exceeds all imagination, although distances are rather small (For a toy example that is illustrative see figure 11). In general many genotypes give rise to the same phenotype or, in other words, the relation of genotypes to phenotypes is many to one. The number of distinct phenotypes is still very large, although it is much smaller than the number of genotypes. The fact that many genotypes are related to a few phenotypes is the basis of Motoo Kimura's neutral theory of evolution.[48] The enormous size of the mappings and the complexity of phenotypes seem to be prohibitive for modeling and analysis of the landscape problem in real systems. Nevertheless, fitness landscapes and evolutionary landscapes for small RNA molecules were studied in detail[58,59,60,61] and led to the development of methods for computational analysis.[62] The fast development of new techniques, in particular polynucleotide sequencing and massively parallel screening and analysis, however, made it possible to deal with natural fitness landscapes and led to especially well studied cases of RNA viruses[63]. I mention here also an impressive example dealing with HIV-1 fitness,[64] a study where the environment is included in form of host species,[65] and a large scale study of adaptive landscapes.[66] Here I refrain from presenting and discussing extensive fitness landscapes. Instead we shall consider an illustrative toy example of a biopolymer fitness landscape and some computer simulations of evolution on model landscapes derived from RNA sequences and structures. As said before basic to the study of fitness landscapes is the relation between genotype and fitness that is visualized as a sequel of two mappings: (i) the genotype-phenotype mapping (figure 11) and (ii) a mapping from phenotypes into fitness values (figure 12). This combination of mappings is tantamount to the paradigm of structural biology. Biopolymer sequences – the genotypes – are folded into 3D molecular structures – the phenotypes, which in turn are evaluated, for example by evolution, to yield quantitative molecular properties like fitness.

[o] An autosome is a chromosome that is not a sex chromosome. In a diploid organism all autosomes are present in two copies.

Figure 11: **The mapping from RNA sequence space into RNA shape space.**

Figure 12: **Fitness as the result of a mapping from RNA shape space into the real numbers.**

RNA-folding into secondary structures can be calculated fairly reliable by means of fast dynamic programming algorithms that are based on empirical thermodynamic and other data derived from RNA oligomers.[67,68,69,70,71,72] The relations between RNA sequences and RNA structures has been studied for small RNAs – **GC**-alphabet up to chain length $l = 30$ and **AUGC**-alphabet up to chain length $l = 15$ – by exhaustive folding of whole sequence space and enumeration.[73,74] Several results of exhaustive enumeration are particularly relevant for evolution: (i) Typical environments in sequence space are rugged in the sense that nearby neighboring sequences may form identical or very different minimum free energy structures. (ii) Folding sequences into minimum free energy structures yields relatively few common and many rare structures. (iii) The pre-images of common structures in sequence space are neutral networks,[p] which span large parts of sequence space or even whole sequence space. (iv) In a defined neighborhood of every sequence sequences are found for all common structures, which form them as their minimum free energy structure (shape space covering). Apart from few exceptional cases evolution is dealing exclusively with common structures.

Landscapes are constructed from shape space trough assigning a quantifiable property in form of a value to every shape. Landscapes derived from small polynucleotide structures and properties reveal two properties that are shared by all larger and more complex examples: (i) biopolymer landscapes are rugged in the sense that small changes in the sequences like point mutations may have no, small or drastic effects, and (ii) there is a relatively high fraction of neutral mutations. Neutral variants are different genotypes, which form the same structures or have indistinguishable fitness. Ruggedness is at least in part the result of non-locality of interactions. This is most clearly seen in case of polynucleotides: The stems in secondary structures combine stretches of nucleotides from distant regions and mutations may shift optimal combinations of base pairs to entirely different positions along the sequence. The combination of ruggedness and neutrality makes it possible to find optima or near optimal solution in highly irregular fitness landscapes.[53,54] Irregularity and neutrality in a natural fitness landscapes is illustrated by means of a simple example, a small four letter RNA sequence of chain length $l = 17$: **AGCUUAACUUAGUCGCU** and its 51 one error mutants (figure 13). The free energy of folding into the minimum free energy structure, ΔG_0, is chosen as a typical property. The fraction of neutral sequences is near 30% and the appearance of the landscape is highly rugged (figure 14).

Figure 13: **Minimum free energy structures of an RNA sequence and its 51 one error mutants.**

Figure 14: **Free energy of folding (ΔG_0) of an RNA sequence and its 51 one error mutants.**

Finally, we discuss computer simulations of evolutionary optimization in populations. Two different evolutionary scenarios are considered: (i) the optimization of a property – here the reproduction rate parameter or fitness value, which is calculated as the difference between a replication and a degradation rate parameter – in a logistic scenario (2) with the carrying capacity K and (ii) the approach to a predefined structure – here a tRNA cloverleaf – through minimization of the mean structure distance d_S between population and target in the flow reactor (figure 7). Despite different environments the individual simulation trajectories have a characteristic appearance in common: The optimization target is

[p] A neutral network is a graph in sequence space, which has all sequences folding into the same structure as nodes. All connections of Hamming distance one within this set are the edges of the neutral network.

approached in steps built by short periods of large progress, which are separated by long quasi-static stretches or plateaus with little improvement (figure 15). The stepwise approach towards the optimum in the result of population sizes that are tiny compared to the possible numbers of genotypes (figure 16).

Figure 15: **Evolutionary optimization in steps.**

In scenario (i) an evolving population of binary (**GC**) sequences takes place in a logistic environment with carrying capacity K in the sense of equation (2). The fitness of a sequence is calculated as the difference of a replication and a degradation rate parameter.[75] The replication rate is obtained from a model equation that relates it to the free energy of structure opening and the degradation rate parameter assumes attack of degrading agents in single-stranded regions of the molecule. Fastest degradation is assigned to the open chain. Long stretches of base paired regions so-called stacks require large energies for opening and give rise to slow replication. The optimal structures shown in figure 16 illustrate the evolutionary compromise: The structures contain many small stacks and as few unpaired nucleotides as possible. The origin of the complexity of the fitness landscape can be traced back to two opposite optimization goals: (i) a maximal replication rate parameter is achieved by the open chain and (ii) a minimal degradation rate parameter results from the structure with the longest hairpin. Each of the three optimization runs ends up with a quasispecies-like distribution of sequences.[q] All three trajectories lead to different structural solutions with respect to the compromise mentioned above. The three quasispecies are pairwise disjoint in the sense that they contain no common member. In other words and as shown in figure 16 the three independent optimization trajectories (A,B, and C) progress in three different regions of sequence space and are orthogonal to each other: The initial genotype – the open chain – and the three master sequences resulting from optimization span an almost regular tetrahedron.

Figure 16: **Evolutionary optimization of mean excess productivity in RNA populations.**

A typical trajectory of scenario (ii) is shown in figure 17 and reveals the stepwise approach of a population of RNA molecules towards a predefined target structure. The population is penalized by lower fitness values for increased mean structure distance[r] from the target structure and consequently evolution optimizing fitness migrates towards target. The population consists of one thousand RNA molecules in the flow reactor (figure 7) and it approaches the target in steps interrupted by rather extended plateaus. The steps are characterized by large shifts in shape space towards the target structure. On the plateau the population spreads in sequence space, the genotype sequences change without substantial alterations of the structure that dominates the population temporarily. A typical sequel of snapshots describing the migration of the population in sequence space shows extension, further spreading through splitting in several clones, jumping some distant position in sequence space and population size contraction at the new location (figure 18).

Figure 17: **Minimization of the distance between an RNA population and a target structure.**

Figure 18: **Images of an RNA population during evolutionary optimization.**

[q] The distribution is not a quasispecies in the conventional sense, because the stochastic process has not come to an end and we are not dealing with a (quasi)stationary state. Sequence space is so large that random drift goes on for all finite times.

[r] There are several definitions of the distance between pairs of structures, $d_S(S_i,S_j)$. The one we apply here counts the number of base pairs that have to be changed in order to convert structure S_i into structure S_j.

The sequence of images of an evolving population in sequence space has been analyzed in great detail together with an evaluation of the shapes along the trajectories and the development of a formal topological theory of evolutionary change.[50,76,77] Along the plateaus populations are migrating in essence on a neutral network or on a family of neutral networks belonging to closely related structures. Transitions from one structure to the next involve minor changes in the sequences and have therefore reasonable probabilities. Nevertheless, the population size is small compared to size of the neutral network and the population breaks up into several clones, which spread in sequence space thereby increasing the diameter of the population. The probabilities for jumping from one structure to another that is closer to the target and has higher fitness are small but different on different position on the neutral network. They range from (i) extremely low, and hence have vanishingly small probabilities of realization, to (ii) sufficiently small for such an event to happen in reasonable time. Accordingly, the population drifts and spreads until one of its members has reached a point of class (ii). From there a jump in sequence space to a sequence of higher fitness may occur. If it happened the entire population is drawn by the favorable fitness difference, the population shrinks rapidly, and the next step is initiated on a neutral network, which is now closer to target.

Coming back to our initial question of evolutionary optimization in nature we summarize our findings in table 5. Optimization of mean fitness during evolution is the exception rather than the rule in real systems. Small population sizes, small initial particle numbers in particular, cause the dominance of stochastic effects. In the beginning, every mutant – we should not forget – is present in a single copy only, and accordingly stochastic phenomena are inescapable. Manfred Eigen's theory of molecular evolution is highlighted in table 5 underlining the fact that stochasticity is the major source preventing evolution from reaching optima.

Table 5: **Evolution and optimization**

scenario	selection	mutation	population size	initial values	selected object	optimum
Darwinian	strong	weak	large	large	single variant	yes
molecular	**strong or weak**	**strong**	**large**	**large**	**quasispecies**	**yes**
stochastic	strong or weak	strong or weak	large or small	large or small	single variant	no
random drift	weak	strong	small	large or small	drifting clones	no

5. Mastering the complexity of present day molecular genetics

Molecular genetics provided a fairly consistent picture in the nineteen-nighties. An apparent problem, however, was that most experimental data had been derived from prokaryotes, viruses and bacteria. Beginning about ten years earlier experimental molecular geneticists realized that eukaryotic cells are not giant bacteria. Other than prokaryotic regulation mechanisms of gene expression in higher organisms were discovered and many principles and notions taken for granted before became questionable. As an illustrative example we refer to the notion of the gene,[78,79,80] which had a fascinating historical development: Starting out from an abstract unit of inheritance the gene became more and more concrete during the early development of molecular biology, it was seen as a kind of autonomous unit of inheritance represented by a piece of DNA encoding a protein with its own regulation, and finally within the last thirty years the clear notion was more and more blurred leaving only the coding property for

polypeptides in a certain often rather small fraction of the full DNA length. Many biologists think one should avoid the term *gene*, because its definition is often unclear and consider it as obsolete.

The simplest conceivable evolutionary system based on processes in prokaryotes has been sketched in figure 6 and can be readily made more complex by introducing essential features of eukaryote evolution: (i) On the reproduction axis we may add recombination to competitive selection since it is inseparably related to sexual reproduction through meiosis, (ii) cooperation may involve a great variety of regulatory as well as structural interactions that give rise to much more complex dynamics than hypercycles, and (iii) point mutations are the simplest changes in genotypes and they can be expanded by more complex phenomena like more involved mutations, gene duplication, genome rearrangements and others.

Today a rich repertoire of different regulatory mechanisms for gene expression is known, most of them subsumed under the general heading of epigenetics for which a operational definition was given at a Cold Spring Harbor-Symposium in December 2008: *An epigenetic trait is a stably inheritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence*.[81] Epigenetics covers a wide range of processes from covalent modifications of DNA nucleotides and histone proteins, which are altering gene expression patterns, to various forms of activities of transcribed RNAs, for example RNA silencing in which small RNAs interfere with transcription or translation. Evolutionarily relevant is the inheritance of epigenetic markers.[82,83] In mammals epigenetic marks are commonly erased after fertilization and during the development of the primordial germ cells but some marks, particularly those coming from maternal DNA methylation may escape erasure and then constitute inheritable epigenetics. Among other mechanisms they are considered by some biologists to be responsible for *soft* inheritance of acquired traits that usually last only for a few generations.

A presumably much more serious problem for the simple genetic view comes from gene interactions. The picture of essentially independent genes interacting weakly through epistasis is true only in a minority of cases. Monogenic diseases are well studied examples and the estimate is that about 10 000 diseases are caused by a single gene. The majority of genes, however, operate in strongly coupled genetic networks. Then the "*single gene view*" is no longer helpful. The traits of phenotypes are created by the gene clusters rather than by single genes.

How is efficient handling of the enormously complex genotype-phenotype maps possible when gene clusters instead of genes are the targets of selection? The answer, in principle, is simple: Neither the fitness landscapes discussed in section 4 nor the parameter spaces shown in figure 6 require genes. A gene network just involves a longer DNA sequence corresponding to a larger part of sequence space but the basic approach remains valid. Of course, there is no reduction in complexity but the formal considerations may help in classifications. Manfred Eigen put this aspect of the relation between theory and experiment in an illustrative sentence that I quote at the end of my talk: "*Theory cannot remove complexity, but it shows what kind of regular behavior can be expected and what experiments have to be done to get a grasp on the irregularities*".[84]

My profound thanks go to Manfred Eigen for decades of wonderful cooperation and friendship. I am thankful to the jury of the Manfred Eigen Award for having me selected.

Thank you for your attention!

Figure captions

Figure 1: **Exponential growth, logistic growth, and selection of the fittest**. The upper part of the figure compares solution curves for the logistic equation (1) (black) and exponential growth being a solution of the equation $\mathrm{d}x/\mathrm{d}t = r\,x$ of the form $x(t) = x(0)\exp(rt)$ (red). In the early phase the difference between the logistic and the exponential curve is small before the former levels off and converges to the carrying capacity K. The lower part of the figure shows logistic growth in a heterogeneous population consisting of four different subspecies, X_1 (yellow), X_2 (green), X_3 (red), and X_4 (blue). The total population (black) adapts fast to the carrying capacity K and the internal population shows selection of the fittest. Parameter values, upper part: $N(0)/K = x(0) = 0.01$, $r = 0.2$, and lower part: $x_1(0) = 9 \times 10^{-5}$, $x_2(0) = 1 \times 10^{-5}$, $x_3(0) = 2 \times 10^{-6}$, $x_4(0) = 4 \times 10^{-7}$; $f_1 = 1.75$, $f_2 = 2.25$, $f_3 = 2.35$, and $f_4 = 2.80$ $[t^{-1}]$ where $[t]$ is an arbitrary time unit.

Figure 2: **Mutation in the Neo-Darwinian scenario.** Before the path-breaking discoveries of molecular structures in biology no mechanism of mutation was known. In the figure at time $t = 6$ an advantageous mutant appears in the system like a *deus ex machina*. We remark that in the Darwinian scenario a mutant with higher fitness will replace the previously selected subspecies no matter how small the injected quantity is. In reality stochastic phenomena as discussed in section 3 become important. Initial conditions and parameters: $x_1(0) = 0.90$, $x_2(0) = 0.08$, $x_3(0) = 0.02$, $x_4(0) = 0$; $f_1 = 1$, $f_2 = 2$, $f_3 = 3$, and $f_4 = 7$ $[t^{-1}]$ where $[t]$ is an arbitrary time unit. At time $t = 6$ a small quantity of $X_4 - x_4(6) = 1 \times 10^{-4}$ – is injected into the reactor and becomes instantaneously selected.

Figure 3: **From Malthus to the modern synthesis.** The picture shows a sketch of the development of biological thinking from the beginning of the nineteenth century to the *modern synthesis* around the end of World War II. Robert Malthus was presumably the first to recognize the important role that is played in evolution by limited resources. In the second half of the nineteenth century two revolutionary ideas shaped biological thought: (i) natural selection introduced by Charles Darwin and Alfred Russel Wallace, and (ii) genetic variation and Mendelian inheritance initiated by Gregor Mendel but not immediately recognized in its importance. Later, around the end of the century came August Weismann's fundamental discovery of the separation of the germline from the somatic cells. Natural selection and germline-soma separation were the two most important intellectual ingredients of the *Neo-Darwinian t*heory. Two at first hostile lines of thought in evolution were put forward by the *selectionists* (red) and *geneticists* (blue). The *Modern synthesis* has been completed just a few years before molecular biology changed the biological view of the world again.

Figure 4: **A molecular mechanism for mutation.** The centennial note by James Watson and Francis Crick in *nature*[18] did not only show how genetic information can be copied (upper part of the sketch), it provided also a possible mechanism for mutation: Each misincorporation of a nucleotide – here a **G** instead of an **A** in the lower part of the figure – gives rise to an inheritable change of the nucleotide sequence. The enzyme (blue) symbolizes the DNA polymerase of *Thermus aquaticus*, which catalyzes a rather simple mechanism of plus-minus-DNA replication and which is used in PCR copying of polynucleotides.

Figure 5: **Mutation in Eigen's and in Crow-Kimura's mechanism of evolution.** In Manfred Eigen's theory of molecular evolution correct replication and mutation are parallel reaction channels of one and the same replication mechanism (upper part). A replication event involving X_j as template is initiated with a rate parameter f_j and the probability to enter the channel leading to X_i as product is Q_{ij} with $\sum_{i=1}^{n} Q_{ij} = 1$. Mutation is a replication error and occurs strictly together with the replication process. In the mechanism proposed in the monograph on population genetics by James Crow and Motoo Kimura[20] replication and mutation are completely independent processes (lower part). Replication is only a one channel mechanism with $Q_{jj} = 1$ and mutation is an independent event and may occur at any instant during the lifetime of the organism.

Figure 6: **The parameter space of molecular evolution**. Evolution is understood as the interplay of three basic process: (i) selection driven by fitness differences Δf, (ii) cooperation regulated by one or more cooperation parameters h, and (iii) mutation controlled by a mutation parameter p as well as random events giving rise to neutral evolution. On the coordination planes we find characteristic evolutionary processes. Plane $\mathcal{A}$: competition and mutation leading to quasispecies as longtime solutions, plane $\mathcal{B}$: competition and cooperation as discussed in case of *major evolutionary transitions*,[22,24] and plane C cooperation and mutation, where we find mutation driven reintroduction of extinguished species or subspecies. Further parameters contain also external factors imposed on evolution by the environment. In case of the flow reactor shown in figure 7 there are two external or environmental parameters: (i) the flow rate d and (ii) the availability of resources expressed by the concentration of A in the stock solution, a_0.

Figure 7: **The flow reactor as a device for modeling evolution.** The figure sketches a continuously fed stirred tank reactor (CFSTR or CSTR):[30] A stock solution with a resource concentration [A] = a_0 flows into a well-stirred reactor with a (volume) flow rate r $[V\,t^{-1}]$. The volume of the reactor is V_R $[V]$ and the inflow of stock solution is compensated by an outflow of reaction mixture of exactly the same volume. The reaction mixture contains the resource A and the reaction products here symbolized by X_k, $k = 1,\ldots,5$. Instead of the flow rate we shall use here the dilution rate $d = r \cdot V_R^{-1}$ $[t^{-1}]$, which is the flow rate divided by the reactor volume and has the advantage of being independent of the size of the reactor. The mean residence time of a volume element in the reactor is the reciprocal dilution rate: $\tau_R = d^{-1}$ $[t]$.

Figure 8: **The quasispecies as a function of the mutation rate parameter p.** The stationary frequency of the master sequence X_m denoted by $\bar{x}_m$ is shown as a function of the mutation rate p. In the phenomenological approximation shown here in the upper part of the figure[s,19] the function $\bar{x}_m(p)$ is almost linear in the particular example shown here. In the insert the approximation (black) is shown together with the exact solution (red). The mutation rate parameter p has two limitations: (i) The physical accuracy limit of replication sets a lower bound, and (ii) the error threshold sets an upper bound to the mutation rate parameter p. At error rates above threshold the solution is close to the uniform distribution, which can't exist in reality. Instead we observe joint or disjoint population clones migrating in sequence space (section 4). The sketch in the lower part considers lethal mutations[40] occurring at certain positions, θ in number. Mutations at these positions lead to the extinction of the mutant. With increasing mutation rate the population may either go through the error threshold or become directly extinct because of accumulation of lethal variants. As expected a satisfactory description of the extinction phenomenon requires information on the distribution of fitness values in sequence space (section 4). Choice of parameters, upper part: $l = 6$, $\sigma = 1.4131$ and lower part: $l = 20$, $\sigma = 5$, $f_m = 15$, and $D = 1$.[t]

Figure 9: **Bifurcations along the path from competition to cooperation for three subspecies** ($n = 3$). The stationary concentration values $\bar{a}$ (black), $\bar{x}_1$ (red), $\bar{x}_2$ (yellow), and $\bar{x}_3$ (green) are shown a functions of the resource concentration a_0. The stationary concentration values of the asymptotically stable points are drawn in full color as thick lines, those of unstable points as thin faint lines. On top we show sketches of the fixed point diagrams for the individual scenarios with increasing availability of resources expressed as the concentration of the stock solution a_0 (stable points are shown as full circles, unstable points as open circles; analytical expressions see table 1):

extinction (S_0) $\Leftrightarrow$ selection of X_1 ($S_1^{(1)}$) $\Leftrightarrow$ exclusion of X_2 ($S_2^{(2)}$) $\Leftrightarrow$ cooperation .

Choice of parameters: $d = 5.0$ [t^{-1}], $f_1 = 0.12$ [$M^{-1}t^{-1}$], $f_2 = 0.10$ [$M^{-1}t^{-1}$], $f_3 = 0.08$ [$M^{-1}t^{-1}$], and $h = 0.0006$ [$M^{-2}t^{-1}$].

Figure 10: **The mutation threshold in hypercycles with five members** ($n = 5$). In the stochastic approach cooperative oscillatory systems may have a finite lifetime, because escalating oscillations can lead to extinction of individual components and entire systems. Mutations can increase the lifetimes through reintroducing extinguished subspecies into the system. The upper part of the figure shows a trajectory of a five-membered hypercycle ($n = 5$) where escalating oscillations lead to extinction around $t_0 = 800$. In the lower part extinction are recorded as a function of the resource concentration a_0 for four different mutation rate parameters $p = 0.0$ (red), 0.0005 (yellow), 0.0010 (green), and 0.0020 (blue)**.** Increased mutation rates lead to a substantial increase in the lifetimes of the systems. Despite enormous scatter of the individual values it is possible to recognize threshold-like behavior. Initial conditions: $a(0) = 0$, $x_1(0) = 10$, $x_2(0) = x_3(0) = x_4(0) = x_5(0) = 5$. Choice of parameters: $a_0 = 200$, $d = 0.05$, $a_0 \cdot d = 4000$, $h_1 = h_2 = h_3 = h_4 = h_5 = h = 0.1$; color code, upper part: A black, X_1 red, X_2 yellow, X_3 green, X_4 blue, and X_5 cyan.

[s] The phenomenological approximation is based on neglect of mutational backflow which, however, is not applied fully consistently. Nevertheless, is shows remarkably good agreement with the exact results for sufficiently long sequences.[36,41]

[t] D is a degradation rate since the process is considered in a batch rather than in a flow reactor[19,30]. The conclusions drawn here are the same both reactor types.

Figure 11: **The mapping from RNA sequence space onto RNA shape space.** The relation between sequences and minimum free energy secondary structures of RNA molecules is sketched as a mapping from sequence space $\mathcal{Q}$ into shape space $\mathcal{S}$.[41] The sizes of sequence spaces escape all imagination: The possible number of sequences for a tiny RNA molecule with chain length $l = 17$ is $|\mathcal{Q}_{17}^{(4)}| = 4^{17} \cong 17 \times 10^9$ and already $|\mathcal{Q}_{17}^{(2)}| = 2^{17} = 131\ 072$ in a two-letter alphabet, e.g. GC, whereas the number of different secondary structures with minimum free energy is only 516. Both spaces, $\mathcal{Q}$ and $\mathcal{S}$, are metric spaces with the Hamming distance d_{H} and a suitable structure distance d_{S} playing the role of metrics. The mapping from genotype space onto phenotype space is many to one and accordingly many sequences form the same minimum free energy structure. A given RNA sequence, on the other hand, forms in general a great number of suboptimal structures, which may play a role at equilibrium in case they lie energetically close to the ground state.

Figure 12: **Fitness as the result of a mapping from RNA shape space into the real numbers.** The fitness landscape is the result of a second mapping from shape space $\mathcal{S}$ into the real numbers $\mathbb{R}_{\geq 0}$. A non-negative fitness value is assigned to every structure and the evolutionary process performs the evaluation on the level of populations. In general we are dealing with a mapping from a metric structure or shape space into a parameter space of non-negative parameter values.

Figure 13: **Minimum free energy structures of an RNA sequence and its 51 one error mutants.** The figure sketches the 16 secondary structures S_k ($k = 0,1,\ldots,15$) with minimum free energies of all 51 single point mutations of the sequence X_0 = (**AGCUUAACUUAGUCGCU**)**.** The structure S_0 in the center of the figure is the structure of the reference sequence X_0, occurs 15 times in the one error mutant spectrum, and is with $15/51 = 0.294$ also the most frequent mutant structure.[u] The structures at the periphery are ordered according to their first appearance in the series of consecutive mutations (figure 14). Inserted in the arrows pointing from S_0 to individual structures S_k are (i) the numbers of occurrence (color) and (ii) the structure distances d_{S} from the reference structure (larger numbers in gray). All drawings of structures begin at the 5'-end of the RNA, which is always the left end of the graph or string (in upright positioning), nucleotides are shown as beads and base pairs are connected by colored thick lines. Colors encode the numbers of base pairs: red 7, black 6, green 5, blue 4, pink 3, and lavender 2.[41]

Figure 14: **Free energy of folding** (ΔG_0) **of an RNA sequence and its 51 one error mutants.** The plot shows the folding free energies ΔG_0 at 0°C of the one-error mutants of X_0. At each position from 1 to 15, the sequence of mutants is **N** $\rightarrow$ **A**, **N** $\rightarrow$ **U**, **N** $\rightarrow$ **G**, and **N** $\rightarrow$ **C**, where the trivial replacement **N** $\rightarrow$ **N** leaving the sequence unchanged is omitted (**N** = {**A**,**U**,**G**,**C**}). The folding energy of the reference sequence is shown as dotted line, and the color code refers to the number of base pairs (see caption of figure 14).[42]

Figure 15: **Evolutionary optimization in steps.** The figure sketches a typical trajectory of evolutionary optimization in a population. The target is approached in steps. Short periods of efficient optimization are interrupted by plateaus along which only little progress in optimization is achieved. The beginnings of eight optimization steps are marked by vertical broken black lines, three steps are indicated by black arrows, and three plateaus are marked by red arrows.

[u] A degree of neutrality of about 0.3 is typical for RNA minimum free energy secondary structures.

Figure 16: **Evolutionary optimization of mean excess productivity in RNA populations**.[74] The figure shows three different RNA secondary structures that were obtained as master sequences of populations with optimized mean excess productivities, $\bar{E} = f - D$, of the populations. Here f is the fitness as before and D is a structure dependent degradation rate parameter.[74] The average population size is 2000 molecules of chain length $l = 70$, a homogeneous population of 2000 all-**C** open chain molecules is chosen as initial condition, and three trajectories with different seeds of the random number generator as the only difference are recorded until a predefined stopping time of the simulation is reached. A mutation rate parameter $p = 0.001$ was applied. The three trajectories led into orthogonally different directions in sequence space, the final populations consisted of quasispecies-like mutant distributions, which were centered around the master sequences A, B, C and had almost the same Hamming distance from the initial sequence **0**, d_{H}(**0**,K) ≈ 25 (K=A,B,C), and the final populations were pairwise disjoint with respect to their members. The three master sequences A, B, and C span an almost equilateral triangle and indicate that optimizations in the high dimensional sequence space occur in "maximally" different directions.

Figure 17: **Minimization of the distance between an RNA population and a target structure.**[53] The figure analyzes a typical trajectory of an evolutionary optimization of a population of $\bar{N} = 1000$ RNA molecules of chain length $l = 76$. The population size is regulated by the flow through the reactor: parameter d and $N(t) = \bar{N} \pm \sqrt{\bar{N}}$. The goal of the optimization is the approach towards a predefined target structure that was chosen to be a cloverleaf like t-RNA structure. The topmost plot presents the mean structure distance of the population from the target (black). The plot in the middle shows the width of the population expressed by the mean distance between sequences (blue), and the plot at the bottom contains a measure of the velocity with which the center of the population migrates through sequence space (green). Diffusion on neutral networks causes spreading of the population in the sense of neutral evolution.[52] When the population size is not sufficient to support a coherent area in sequence space the population breaks up into several individual clones (figure 18). A remarkable synchronization is observed: At the end of every quasi-stationary plateau a new adaptive phase in the approach towards target is initiated that is accompanied by (i) a drastic reduction in the population width and (ii) a jump of the population center towards a new position (jumps are marked by green arrows; the jump near 12×10^6 replications occurs with a velocity of $d_{\mathrm{H}} \cong 25$ per time unit). The time is given by the number of replications that have occurred in the population. Choice of parameters: fitness $f_k = \gamma/(\alpha + \Delta d(S_k, S_\tau))$ where S_τ is the target structure and α and γ are two empirical parameters[54]; $p = 0.001$.

Figure 18: **Images of an RNA population during evolutionary optimization**. Spreading, jump, and contraction of a population of RNA molecules during an evolutionary plateau and the following adaptive phase. Shown are the populations on a plane spanned by the two orthogonal directions of largest extension in sequence space. The red arrow at time $t = 0$ indicates the position of the population at the beginning of the plateau. The small population of 1000 molecules breaks up into clones between $t = 150$ and $t = 200$. It is important to note that spreading is relatively slow – it takes 800 units of time to reach the maximum extension of the population – whereas jump and contraction occur in less than 30 units of time.

[1] Thomas Robert Malthus. *An Essay of the Principle of Populations as it Affects the Future Improvement of Society*. J. Johnson, London 1798.

[2] T. B. Robertson: On the Normal Rate of Growth of an Individual and its Biochemical Significance. Arch. Entwicklungsmechanik Org. 25:581-614, 1908.

[3] A. G. McKendrick, M. Kesava Pai. Proc. Roy. Soc. Edinb. 31:649-655, 1911.

[4] Raymond Pearl, Lowell J. Reed. On the rate of growth of the population of the United States since 1790 and its mathematical representation. Proc.Natl.Acad.Sci.USA 6:275-288, 1920.

[5] P. J. Lloyd. American, German and British antecedents to Pearl and Reed's logistic curve. Population Studies 21:99-108, 1967.

[6] Sharon Kingsland. The refractory model: The logistic curve and the history of population ecology. Quart.Rev.Biol. 57:29-52, 1982.

[7] Pierre-François Verhulst. Notice sur la loi que la population pursuit dans son accroisement. Corresp.Math.Phys. 10:113-121, 1838.

[8] Pierre-François Verhulst. Recherches Mathèmatiques sur la loi d'accroisement de la population. Nouv. Mèm. de l'Academie Royale des Sci. et des Belles Lettres de Bruxelles. 18:1-41, 1845.

[9] Pierre-François Verhulst. Deuixèmme mèmoire sur la loi d'accroisement de la population. Mèm. de l'Academie Royale des Sci., des Lettres et de Beaux-Arts de Belgique 20:1-32, 1847.

[10] David Zwillinger. Handbook of differential equations. 3rd edition. Academic Press, San Diego, CA, 1998, p.322ff.

[11] Charles Darwin. On the Origin of Speciesby Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. John Murray, London 1859.

[12] Alfred Russel Wallace. Contributions to the Theory of Natural Selection. 2nd edition. Macmillan & Co. London 1870.

[13] Ernst Mayr. *The Growth of Biological Thought.* The Belknap Press of Harvard University Press. Cambridge, MA, 1982.

[14] Ulrich Kutschera. Tatsache Evolution. Was Darwin nicht wissen konnte. Deutscher Taschenbuch Verlag. München 2009. In German.

[15] Ronald A. Fisher. The genetical theory of natural selection. Oxford University Press, Oxford, UK, 1930.

[16] John B. S. Haldane. A mathematical theory of natural and artificial selection X. Some theorems on artificial selection. Genetics 19:421-429, 193.

[17] Sewall Wright. Evolution in Mendelian populations. Genetics 16:97-159, 1931.

[18] James D. Watson, Francis H.C. Crick. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. Nature 171:737-738, 1953.

[19] Manfred Eigen. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465-523, 1971.

[20] James F. Crow, Motoo Kimura. An introduction to population genetics theory. Harper & Row, New York, 1970, p.265, eq.6.4.1.

[21] Manfred Eigen, Peter Schuster. The Hypercycle – A principle of natural self-organization. Springer-Verlag, Berlin, 1979.

[22] John Maynard Smith, Eörs Szathmáry. The major transitions in evolution. Oxford University Press, Oxford, UK, 1995.

[23] Peter Schuster. How does complexity arise in evolution. Nature's recipe for mastering scarcity, abundance, and unpredictability. Complexity 2(1):22-30, 1996.

[24] Eörs Szathmáry, John Maynard Smith. The major evolutionary transitions. Nature 374:227-232, 1995.

[25] Britta Wlotzka, John S. McCaskill. A molecular predator and its prey: Coupled isothermal amplification of nucleic acids. Chemistry & Biology 4:25-33, 1987

[26] Tracey A. Lincoln, Gerald F. Joyce. Self-Sustained Replication of an RNA Enzyme. Science 323:2778-2782, 2009.

[27] Nilesh Vaidya, Michael L. Manapat, Irene A. Chen, Ramon Xulvi-Brunet, Eric J. Hayden, Niles Lehman. Spontaneous network formation among cooperative RNA replicators. Nature 491:72-77, 2012.

[28] Peter Schuster. Increase in complexity and information through molecular evolution. Entropy 18:e397, 2016a.

[29] Peter Schuster. A mathematical model of evolution. MATCH Communications in Mathematical and in Computer Chemistry 80:547-585, 2018.

[30] Lanny D. Schmidt. The Engineering of chemical reactions. Second Ed. Oxford University Press, Oxford, UK, 2004.

[31] Gerald F. Joyce. Forty years of *in vitro* evolution. Angew.Chem.Internat.Ed. 73:791-836, 2004.

[32] Andre Koltermann, Ulrich Kettling. Principles and methods of evolutionary biotechnology. Biophys.Chem. 66: 159-177, 1997.

[33] Günther Strunk, Tobias Ederhof. Machines for automated evolution experiments *in vitro* based on the serial-transfer concept. Biophys.Chem. 66: 193-202, 1997.

[34] Peter Schuster, Karl Sigmund. Dynamics of evolutionary optimization. Ber.Bunsenges.Phys.Chem. 89:668-682, 1985.

[35] Manfred Eigen, Peter Schuster. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften 64:541-565, 1977.

[36] Jörg Swetina, Peter Schuster. Self-replication with errors – A model for polynucleotide replication. Biophys.Chem. 16:329-345, 1982.

[37] Esteban Domingo, Peter Schuster. What is a quasispecies? Historical origins and current scope. In: Esteban Domingo, Peter Schuster, eds. Quasispecies: From theory to experimental systems. Current Topics in Microbiology and Immunology. Volume 392, pp.1-22. Springer International, Cham, CH, 2016b.

[38] Thomas Wiehe. Model dependency of error thresholds: The role of fitness functions and contrasts between the finite and infinite sites models. Genet.Res.Camb. 69:127-136, 1997.

[39] Esteban Domingo, ed. Virus entry into error catastrophe as a new antiviral strategy. Virus Research 107(2):115-228, 2005.

[40] James J. Bull, Rafael Sanjuán, Claus O. Wilke. Theory of lethal mutagenesis for viruses. J.Virology 81:2930-2939, 2007.

[41] Héctor Tejero, Arturo Marín, Francisco Montero. Effect of lethality on the extinction and on the error threshold of quasispecies. J.Theor.Biol. 262:733-741, 2010.

[42] Peter Schuster. Quasispecies on fitness landscapes. In: Esteban Domingo, Peter Schuster, eds. Quasispecies: From theory to experimental systems. Current Topics in Microbiology and Immunology. Volume 392, pp.61-120. Springer International, Cham, CH, 2016c.

[43] Peter Schuster. Some mechanistic requirements for major transitions. Phil.Trans.Roy.Soc.B 371:e2015439, 2016d.

[44] Nancy A. Moran. Symbiosis. Current Biology 16:R866-R871, 2006.

[45] Lynn Margulis. Origin of eukaryotic cells. Yale University Press, 1970.

[46] Steven Strogatz. Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering. Westview Press, Boulder, CO, 2001.

[47] Hermann Joseph Muller. The relation of recombination to mutational advance. Mutation Research 106:2-9, 1964.

[48] Joseph Felsenstein. The evolutionary advantage of recombination. Genetics 78:737-756, 1974.

[49] Narendra S. Goel, Nira Richter-Dyn. Stochastic models in biology. Academic Press, New York, 1974.

[50] Peter Schuster. Stochasticity in processes. Fundamentals and applications in chemistry and biology. Springer Series in Synergetics. Springer-Verlag, Berlin, 2016e.

[51] B. L. Jones, H. K. Leung. Stochastic analysis of a non-linear model for selection of biological macromolecules. Bull.Math.Biol. 43:665-680, 1981.

[52] Motoo Kimura. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, UK, 1983.

[53] Martijn A. Huynen, Peter F. Stadler, Walter Fontana. Smoothness within ruggedness. The role of neutrality in evolution. Proc.Natl.Acad.Sci.USA 93:397-401, 1996.

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

[55] Anne Kupczok, Peter Dittrich. Determinants of simulated RNA evolution. J.Theor.Biol. 238:726-735, 2006.

[56] Barrett Steinberg, Marc Ostermeier. Environmental changes bridge evolutionary valleys. Sci.Adv. 2:e1500921, 2016

[57] Sewall Wright. The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution. In: D. F. Jones, ed. Int. Proceedings of the Sixth International Congress on Genetics. Volume 1, pp.356-366. Brooklyn Botanic Garden, Ithaca, NY, 1932.

[58] Peter Schuster, Walter Fontana, Peter F. Stadler, Ivo L. Hofacker. From sequences to shapes and back. A case study in RNA secondary structures. Proc.Roy.Soc.Lond.B 255:279-284, 1994.

[59] Peter Schuster. Prediction of RNA secondary structures: From theory to models and real molecules. Rep.Prog.Phys. 69:1419-1477, 2006.

[60] Peter F. Stadler. Fitness landscapes. In: Michael Lässig, Angelo Valeriani, eds. Biological evolution and statistical physics. Lecture Notes in Physics 585. Springer-Nature, Berlin, 2002, pp.183-204.

[61] José I. Jiménez, Ramon Xulvi-Brunet, Gregory W. Campbell, Rebecca Turk-MacLeod, Irene A. Chen. Comprehensive experimental fitness landscape and evolutionary network for small RNA. Proc.Natl.Acad.Sci.USA 110:14984-14989, 2013.

[62] Ramon Xulvi-Brunet, Gregory W. Campbell, Sudha Rajamani, José I. Jiménez, Irene A. Chen. Computational analysis of fitness networks from *in vitro* evolution experiments. Methods 106;86-96, 2016.

[63] Adam S. Lauring, Raul Andino. Exploring the fitness landscape of an RNA virus by using a universal barcode microarray. J.Virology 85:3780-3791, 2011.

[64] Roger D. Kouyos, Gebriel E. Leventhal, Trevor Hinkley, Mojgan Haddad, Janette M. Whitcomb, Christos J. Petropoulos, Sebastian Bonhoeffer. Exploring the complexity of the HIV-1 fitness landscape. PLoS Genetics 8:e1002551, 2012.

[65] Héctor Cervera, Jasna Lalić, Santiago F. Elena. Effect of host species on topography of the fitness landscape for a plant RNA virus. J.Virology 90:10160-10169, 2016.

[66] José Aguilar-Rodriguez, Joshua L. Payne, Andreas Wagner. A thousand empirical adaptive landscapes and their navigability. Nature Ecology & Evolution 1:e0045, 2017.

[67] Michael S. Waterman, Temple F. Smith. RNA secondary structure: A complete mathematical analysis. Math.Biosciences 42:257-266, 1978.

[68] Michael S. Waterman, Temple F. Smith. Rapid dynamic programming algorithms for RNA secondary. Adv.Appl.Math. 7:167-212, 1986.

[69] Michael Zuker, Patrick Stiegler. Optimal Computer Folding of Larger RNA Sequences Using Thermodynamic and Auxiliary Information. Nucleic Acids Research 9:133-148, 1981.

[70] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 31:3406-3415, 2003.

[71] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, Peter Schuster. Fast folding and comparison of RNA secondary structures. Mh.Chem. 125:167-188, 1994.

[72] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, Ivo L. Hofacker. ViennaRNA Package 2.0. Algorithms for Molecular Biology 6:e26, 2011.

[73] Walter Grüner, Robert Giegerich, Dirk Strothmann, Christian Reidys, Jaqueline Weber, Ivo L. Hofacker, Peter F. Stadler, Peter Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. Mh.Chem. 127:355-374, 1996.

[74] Walter Grüner, Robert Giegerich, Dirk Strothmann, Christian Reidys, Jaqueline Weber, Ivo L. Hofacker, Peter F. Stadler, Peter Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering. Mh.Chem. 127:375-389, 1996.

[75] Walter Fontana, Wolfgang Schnabl, Peter Schuster. Physical aspects of evolutionary optimization and adaptation. Phys.Rev.A 40:3301-3321, 1989.

[76] Walter Fontana, Peter Schuster. Shaping Space: The possible and the attainable in RNA genotype-phenotype mapping. J.theor.Biol. 194:491-515, 1998.

[77] Bärbel M. R. Stadler, Peter F. Stadler, Günter P. Wagner, Walter Fontana. The topology of the possible: Formal spaces underlying patterns of evolutionary change. J.theor.Biol. 213:241-274, 2001.

[78] Karen Hopkin. The evolving definition of a gene. BioScience 59:928-931, 2009.

[79] Mark B. Gerstein, Can Bruce, Joel S. Rozosky, Deyou Zheng, Jiang Du, Jan O. Korbel, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, Michael Snyder. What is a gene, post-ENCODE? History and updated definition. Genome Res. 17:669-681, 2007.

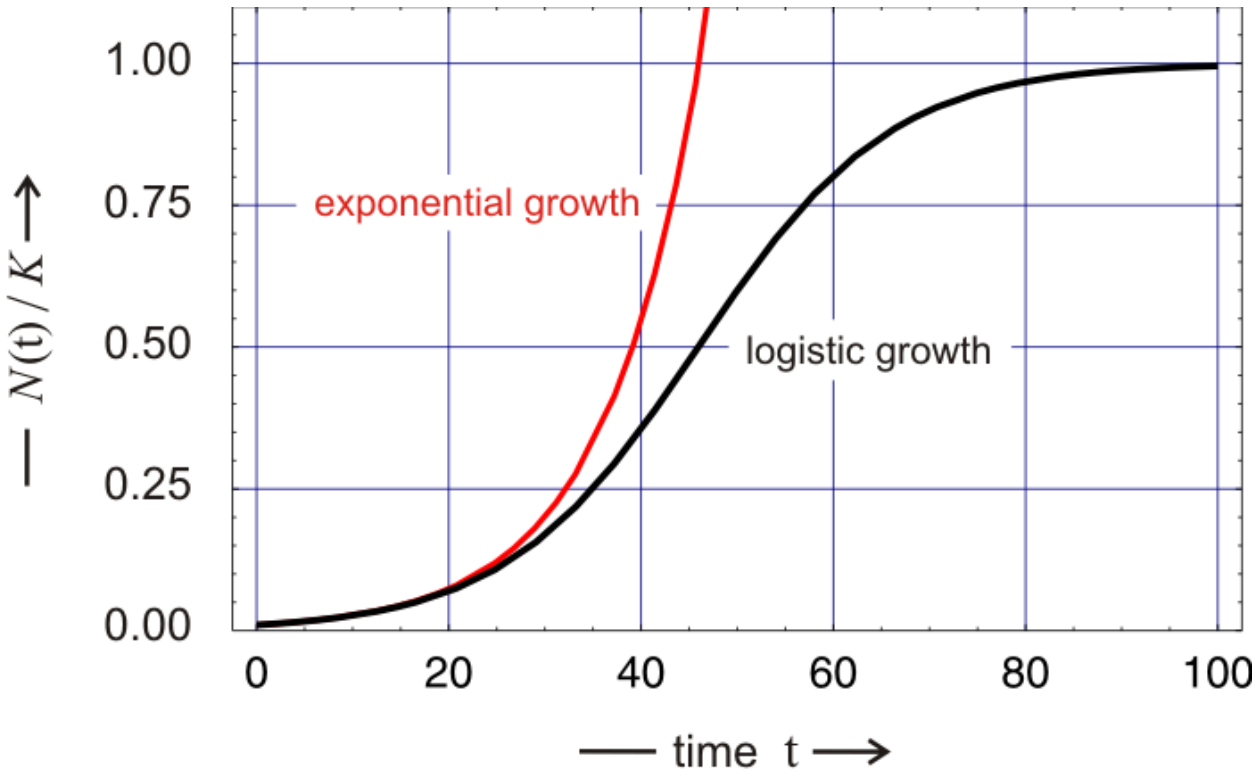
[80] Petter Portin, Adam Wilkins. The evolving definition of the term gene. Genetics 205-1353, 2017.

[81] Shelley L. Berger, Tony Kouzarides, Ramin Shiekhattar, Ali Shilatifard. An operational definition of epigenetics. Genes & Development 23:781-783, 2009.

[82] Eva Jablonka, Gal Raz. Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. Quart.Rev.Biol. 84:131-176, 2009.

[83] Edith Heard, Robert A. Martienssen. Transgenerational epigenetic inheritance: Myths and mechanisms. Cell 157:95-105, 2014.

[84] Manfred Eigen. Preface to the second edition, in: Esteban Domingo, Colin R. Parrish, John J. Holland, eds. Origin and evolution of viruses, 2nd edition. Elsevier-Academic Press, Amsterdam, NL, 2008.


1.00
0.75
0.50
0.25
0.00
N(t) / K
exponential growth
logistic growth
0
20
40
60
80
100
time t

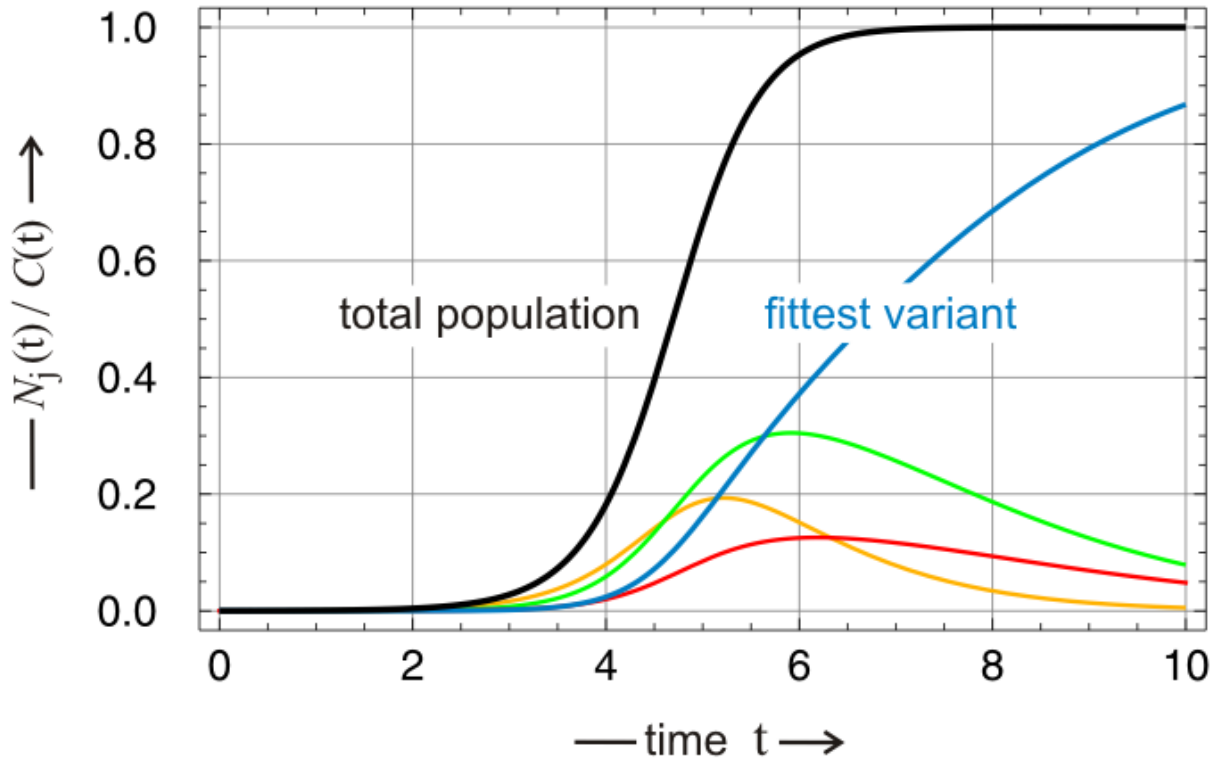

1.0
0.8
0.6
0.4
0.2
0.0
Nj(t) / C(t)
total population
fittest variant
0
2
4
6
8
10
time t

Figure 1

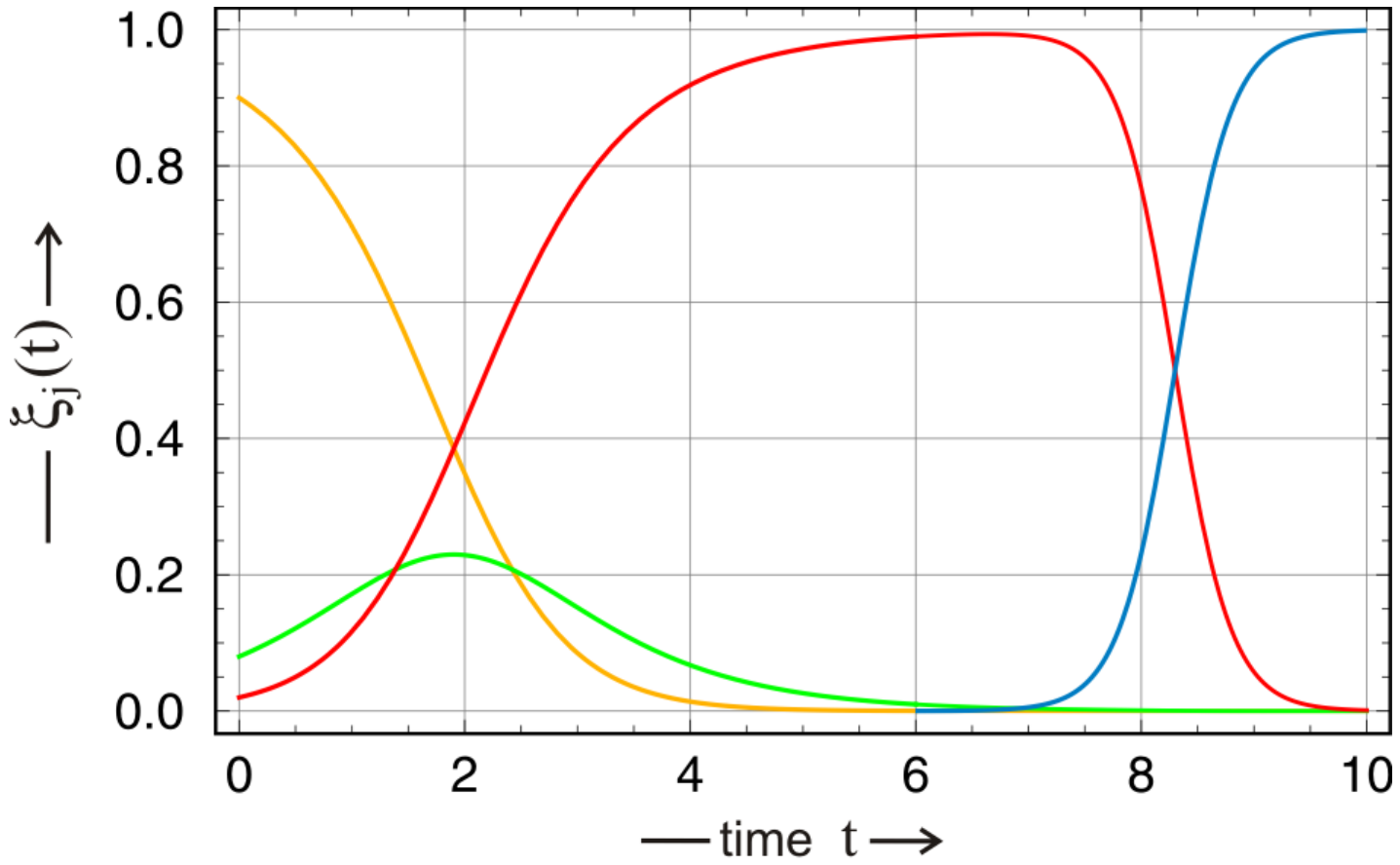


Figure 2

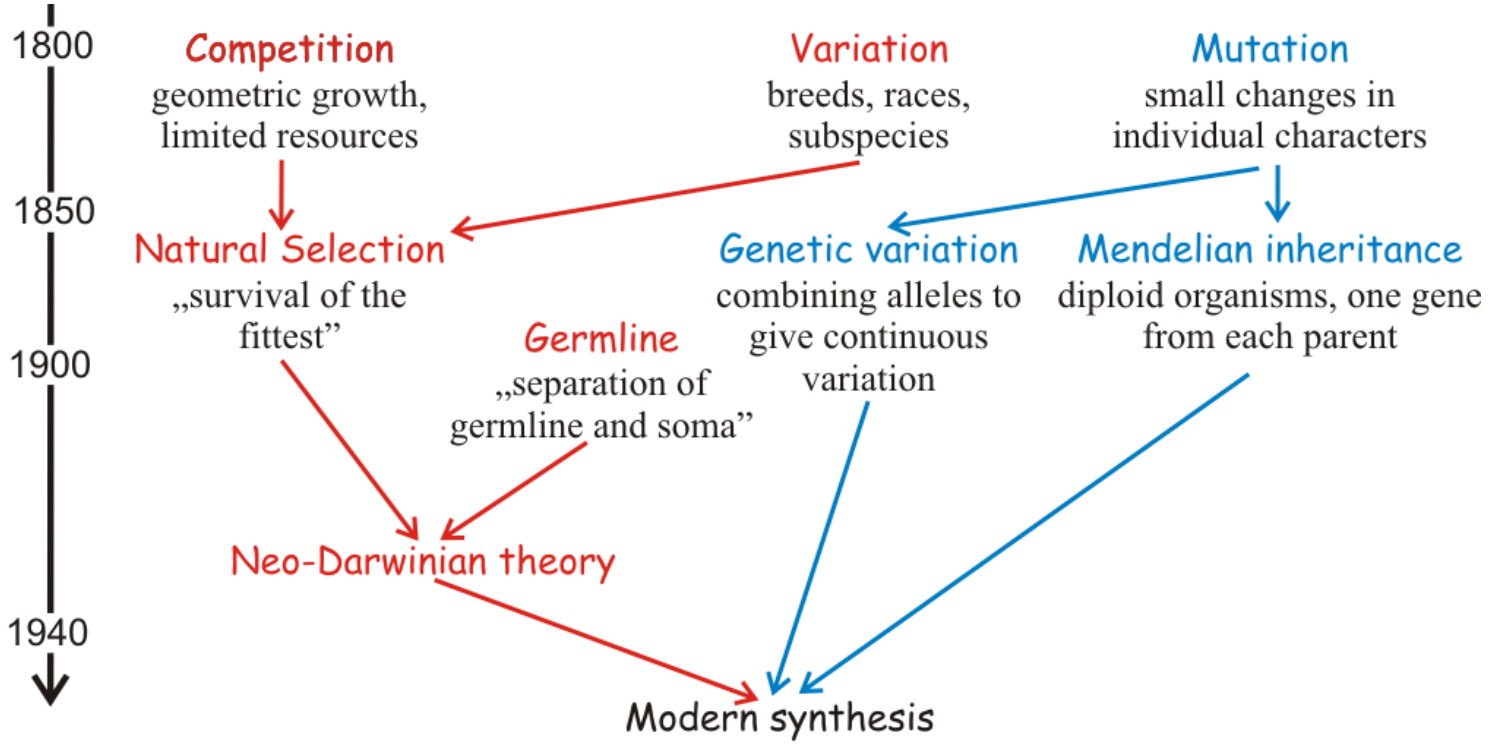
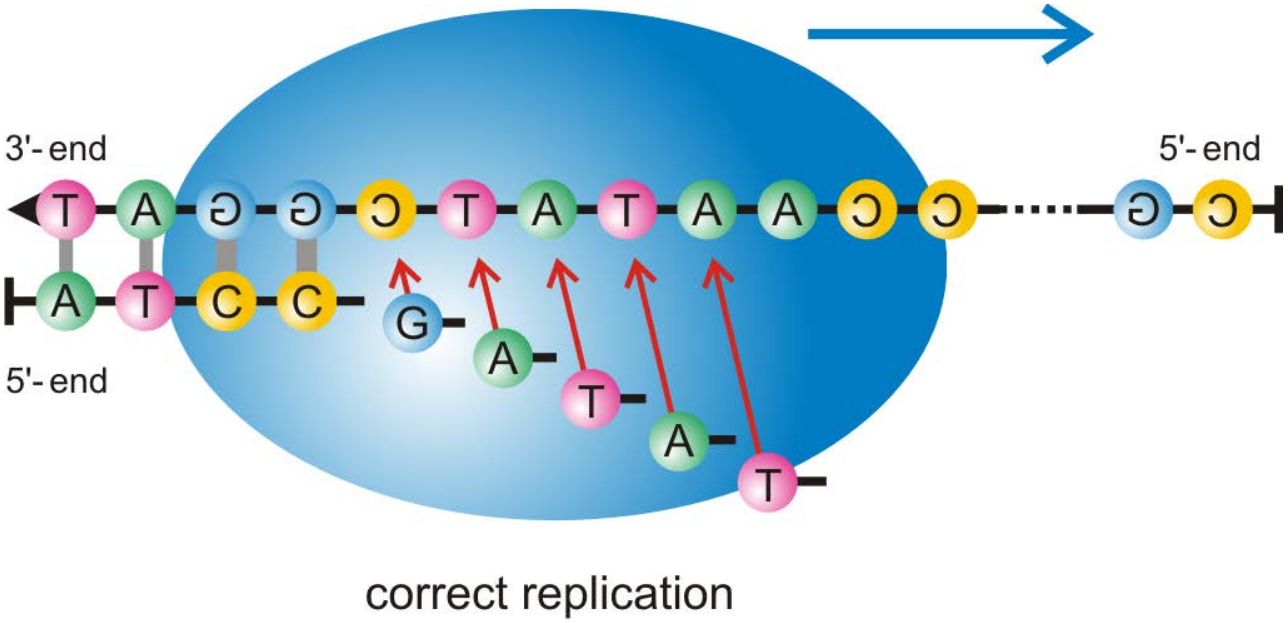
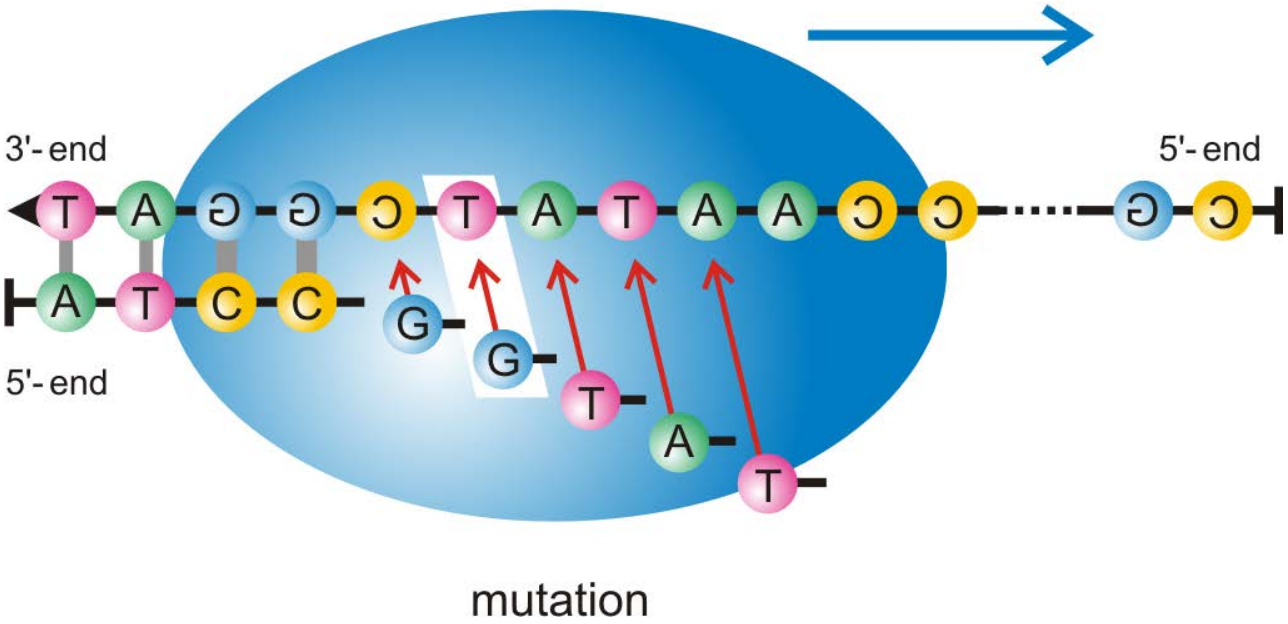

1800
1850
1900
1940
Competition
geometric growth, limited resources
Variation
breeds, races, subspecies
Mutation
small changes in individual characters
Natural Selection
„survival of the fittest"
Germline
„separation of germline and soma"
Genetic variation
combining alleles to give continuous variation
Mendelian inheritance
diploid organisms, one gene from each parent
Neo-Darwinian theory
Modern synthesis

Figure 3



correct replication



mutation

Figure 4

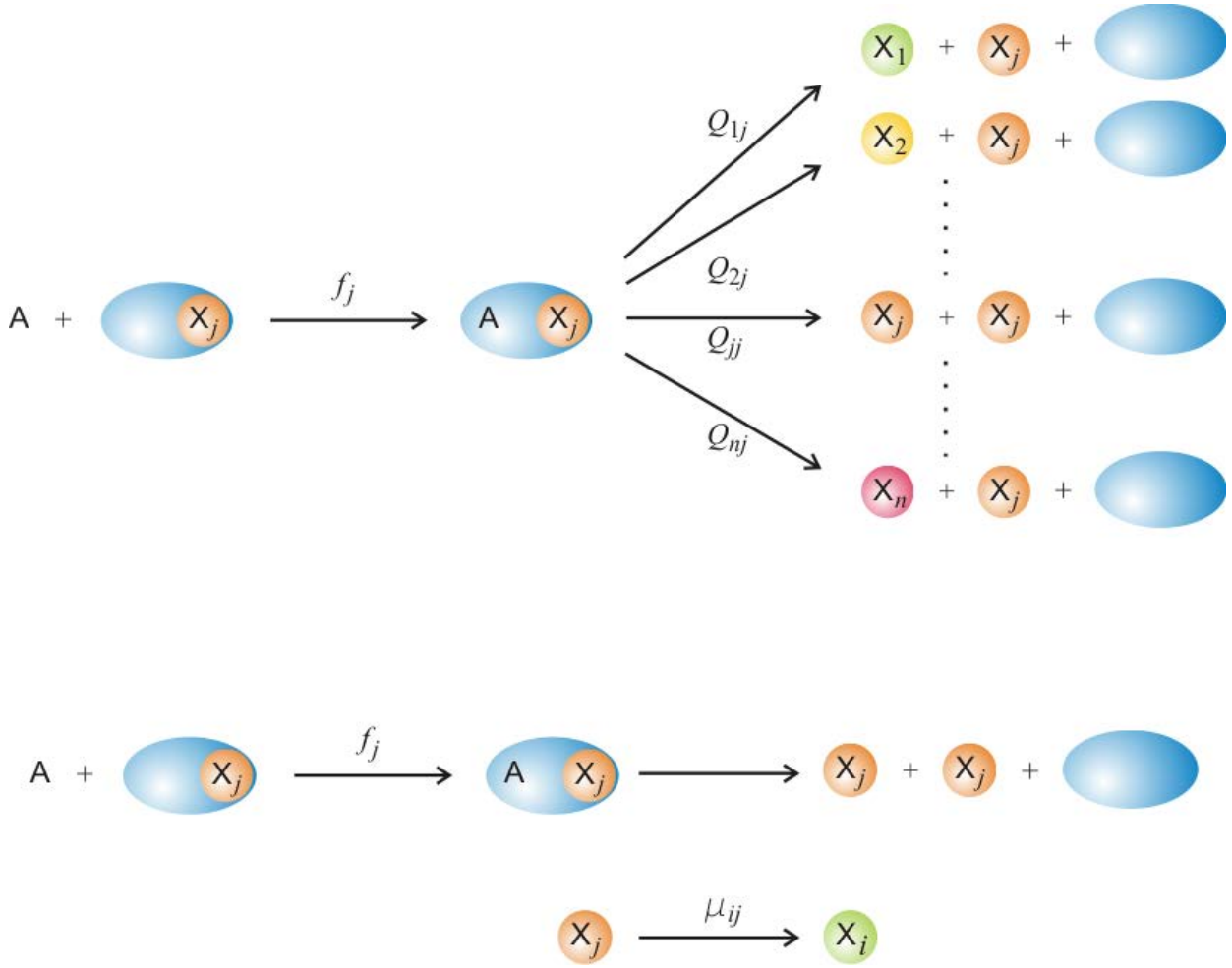

A + X_j
f_j
A X_j
Q_1j
Q_2j
Q_jj
Q_nj
X_1 + X_j +
X_2 + X_j +
X_j + X_j +
X_n + X_j +
A + X_j
f_j
A X_j
X_j + X_j +
X_j
μ_ij
X_i

Figure 5

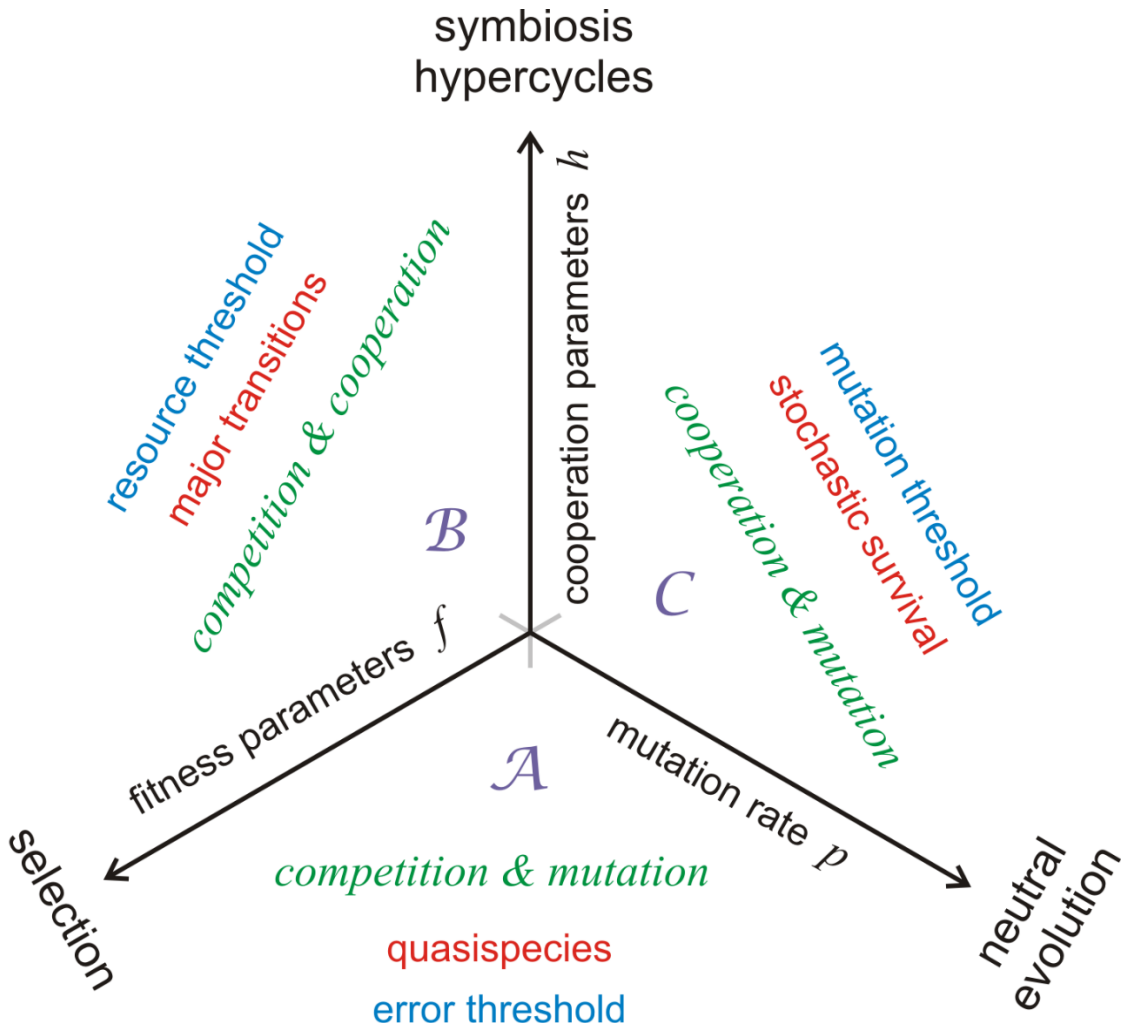

symbiosis
hypercycles
cooperation parameters h
resource threshold
major transitions
competition & cooperation
$\mathcal{B}$
f
fitness parameters
selection
$\mathcal{C}$
cooperation & mutation
mutation threshold
stochastic survival
$\mathcal{A}$
mutation rate p
neutral
evolution
competition & mutation
quasispecies
error threshold

Figure 6

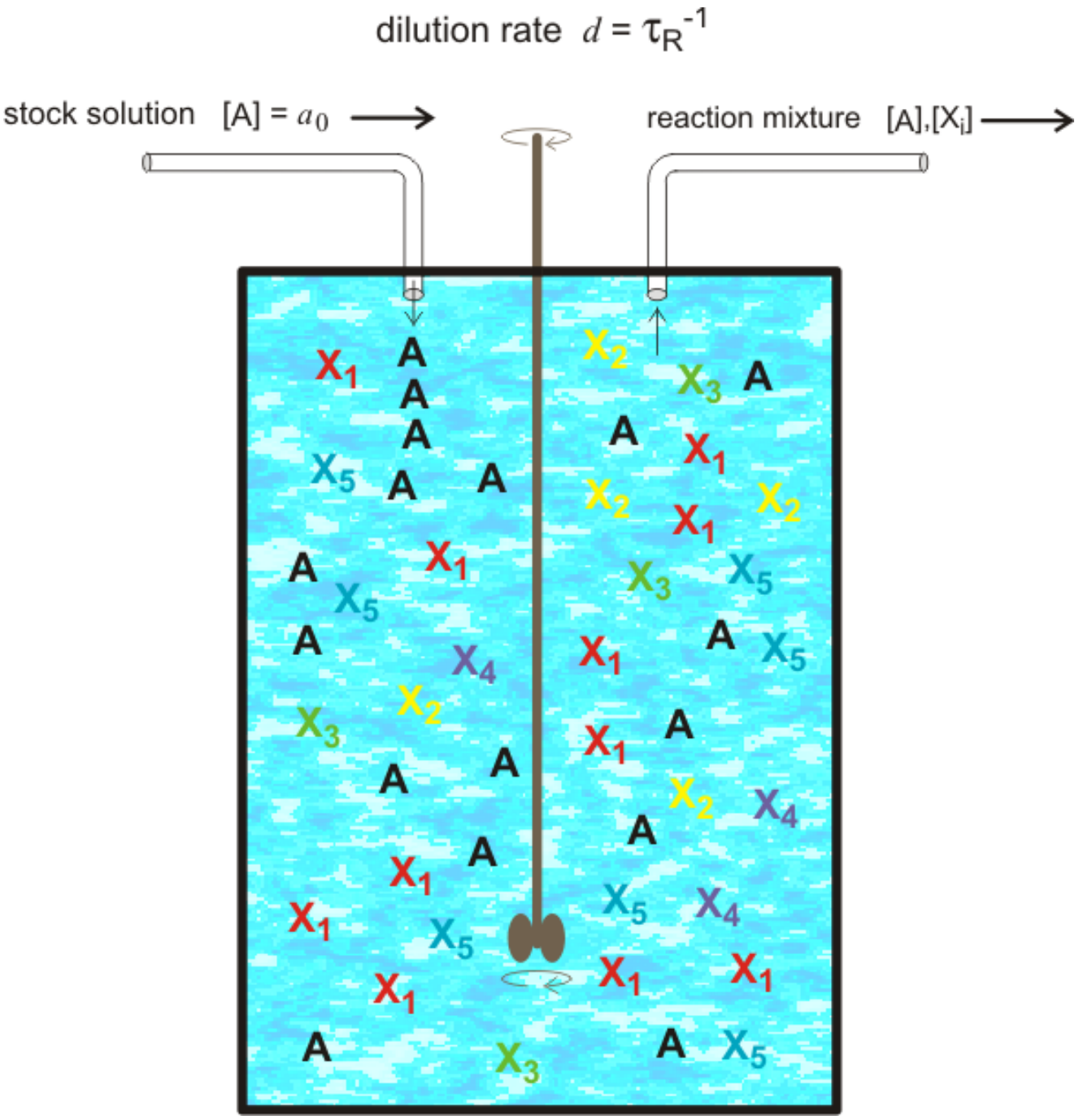
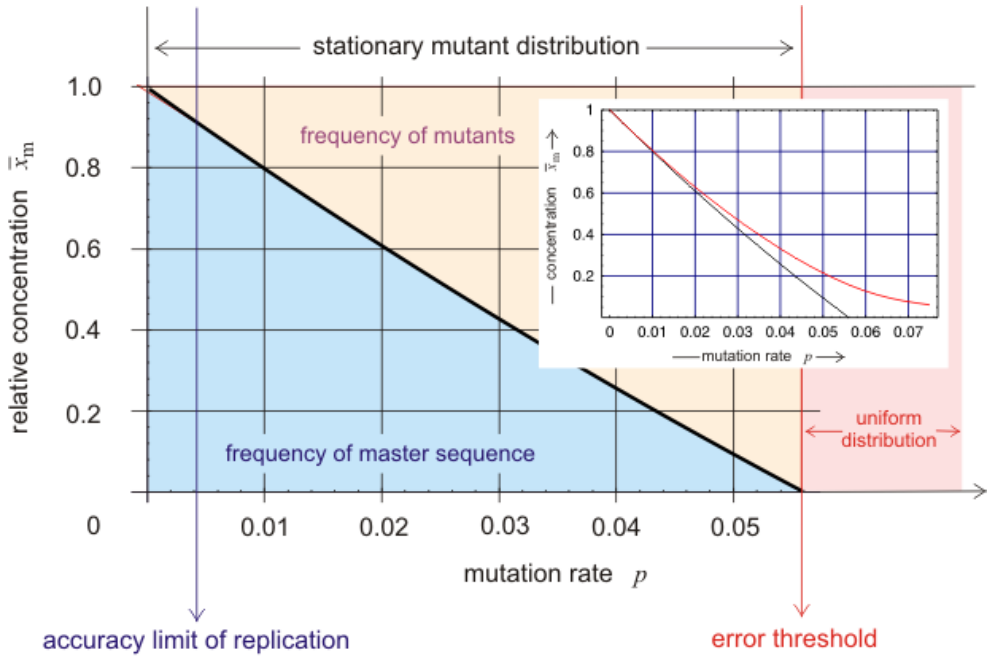

dilution rate $d = \tau_R^{-1}$
stock solution [A] = a_0
reaction mixture [A],[X_i]

Figure 7


stationary mutant distribution
frequency of mutants
frequency of master sequence
uniform distribution
relative concentration $\bar{x}_m$
1.0
0.8
0.6
0.4
0.2
0
0.01
0.02
0.03
0.04
0.05
mutation rate p
accuracy limit of replication
error threshold
concentration $\bar{x}_m$
0.07
0.06

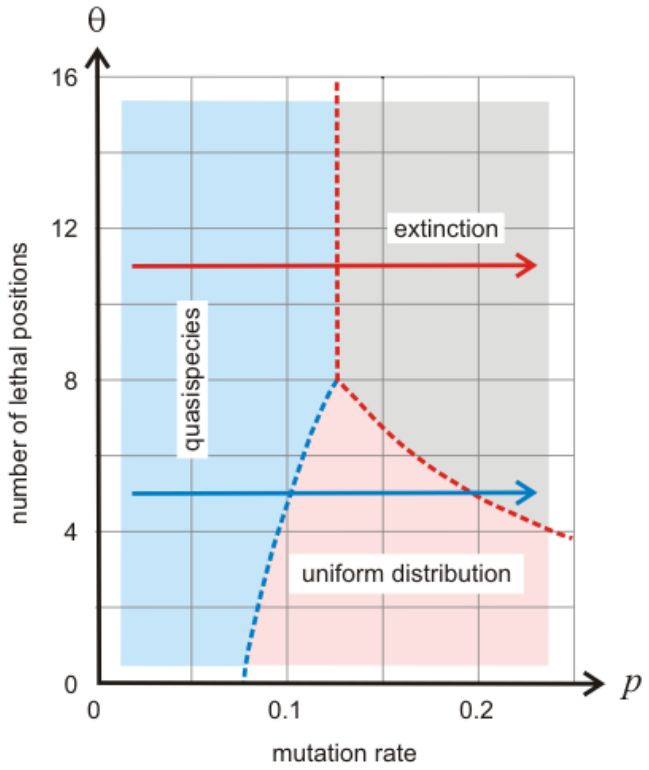

θ
16
12
8
4
0
number of lethal positions
extinction
quasispecies
uniform distribution
0.1
0.2
p
mutation rate

Figure 8

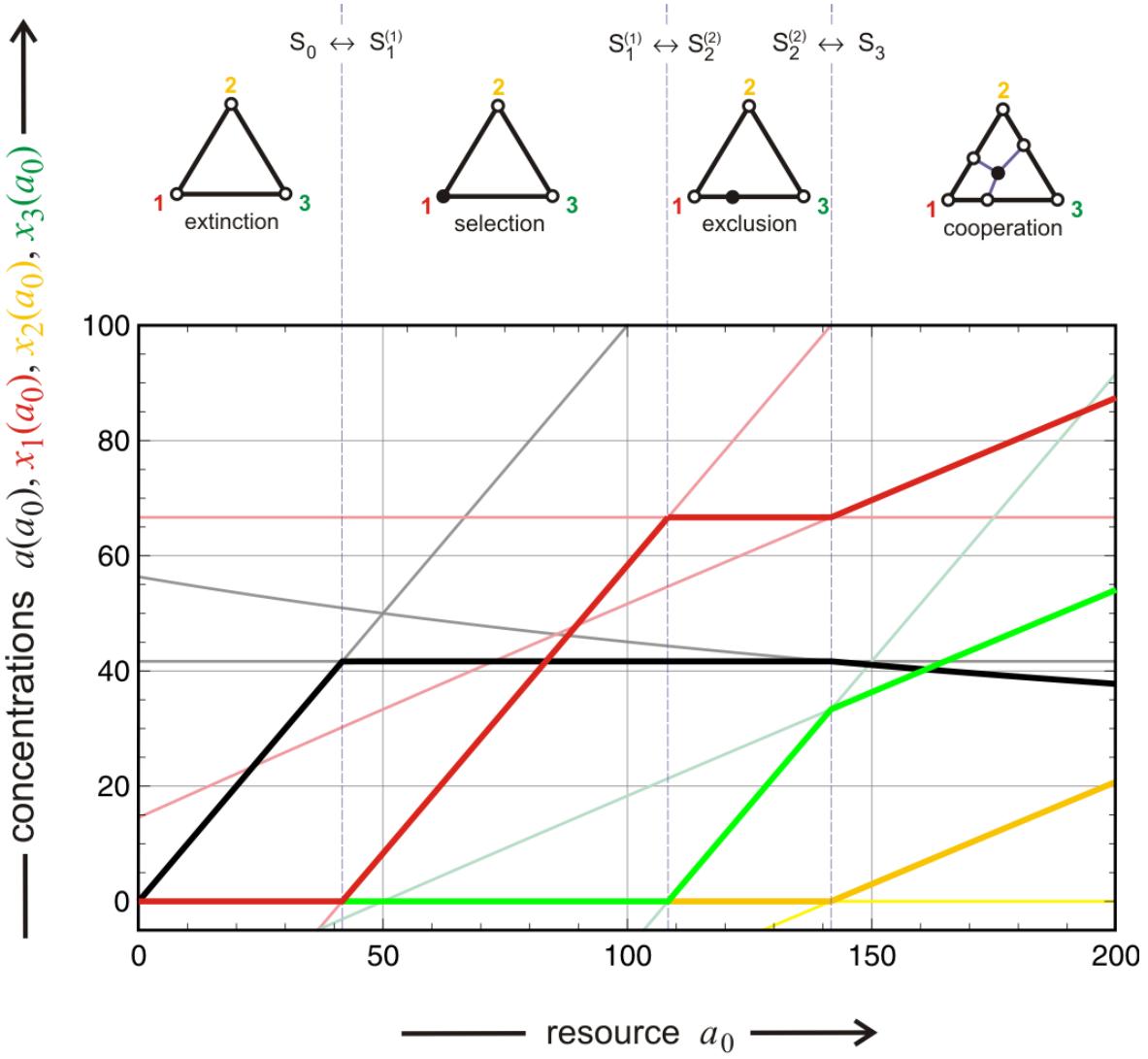
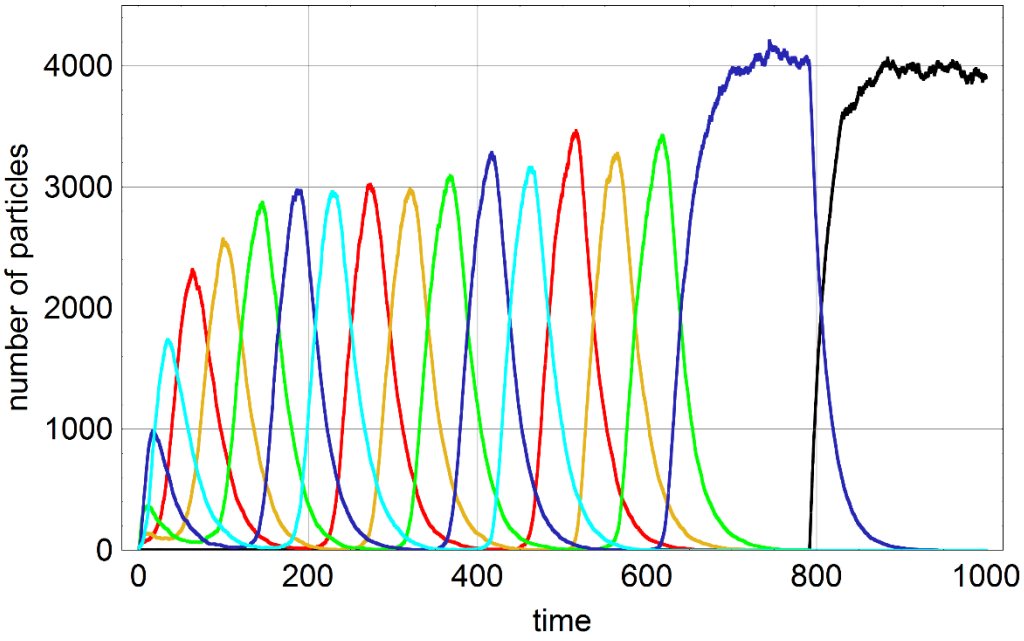

$S_0 \leftrightarrow S_1^{(1)}$
$S_1^{(1)} \leftrightarrow S_2^{(2)}$
$S_2^{(2)} \leftrightarrow S_3$
extinction
selection
exclusion
cooperation
1
2
3
concentrations $a(a_0), x_1(a_0), x_2(a_0), x_3(a_0)$
resource a_0
0
20
40
60
80
100
0
50
100
150
200

Figure 9


number of particles
4000
3000
2000
1000
0
0
200
400
600
800
1000
time

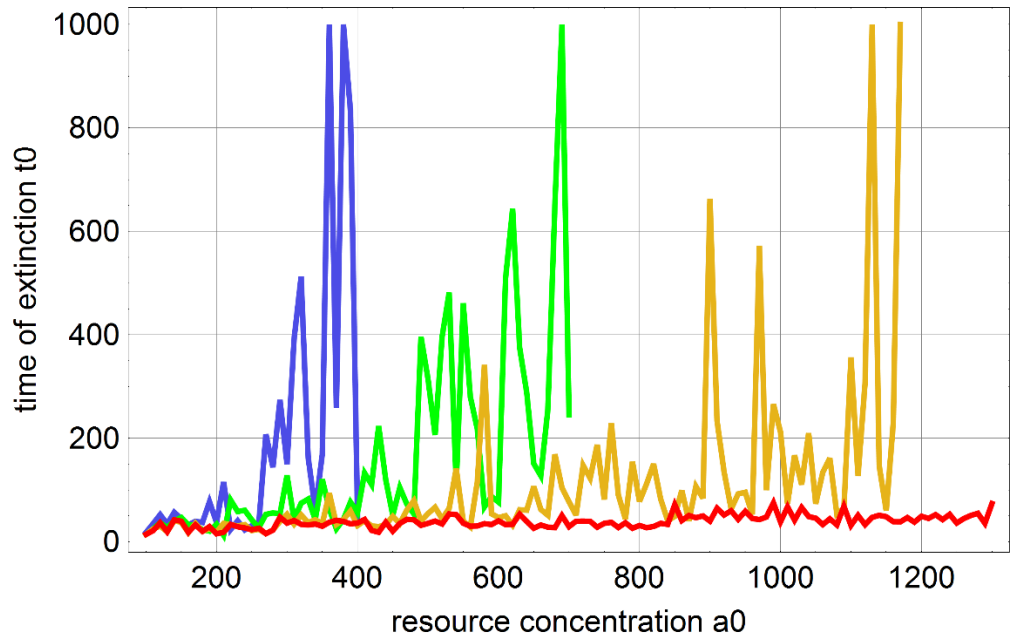

time of extinction t0
1000
800
600
400
200
0
200
400
600
800
1000
1200
resource concentration a0

Figure 10

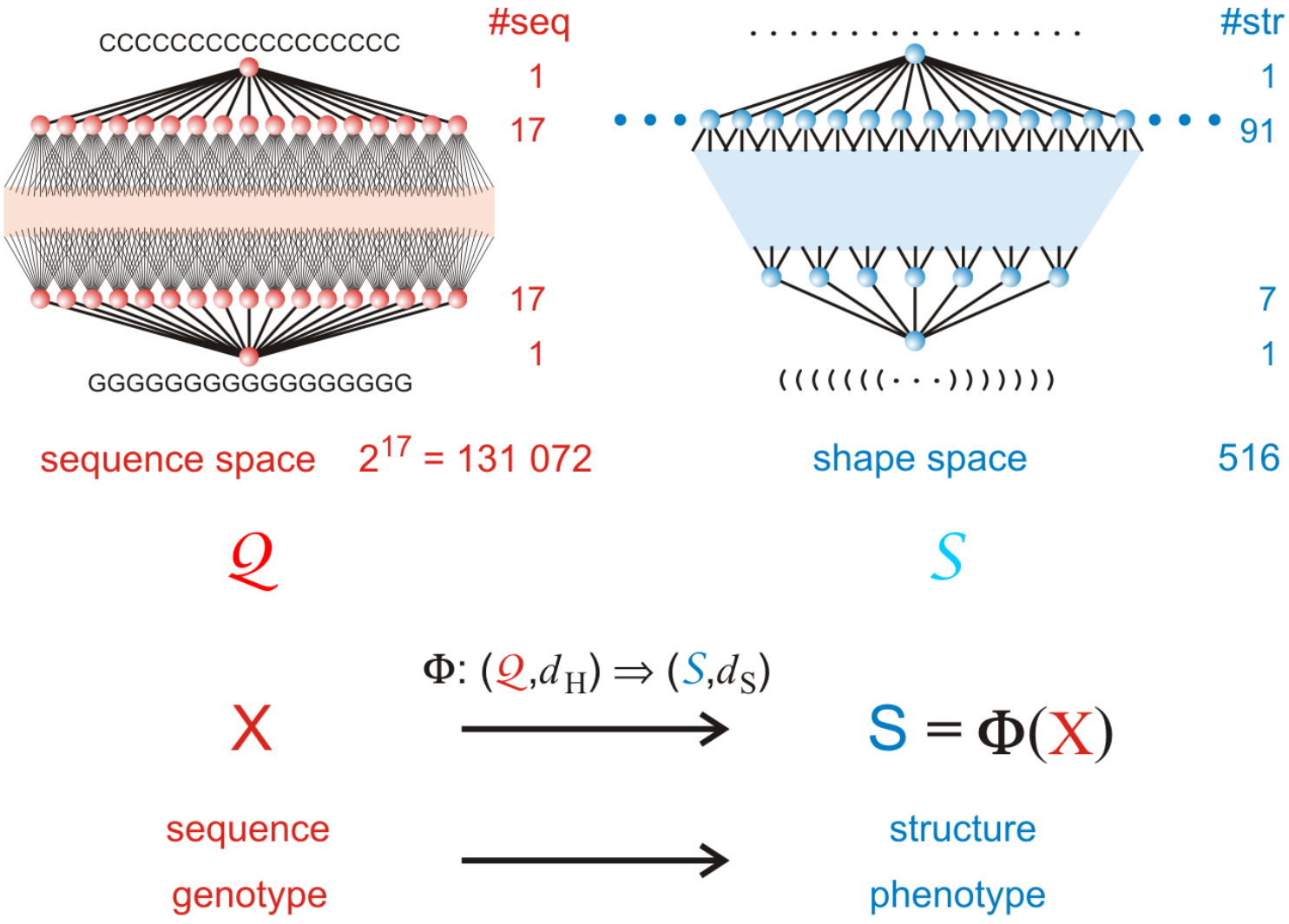
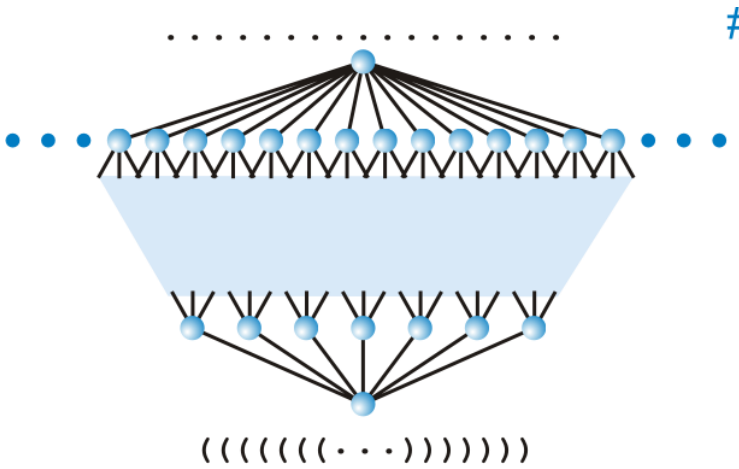
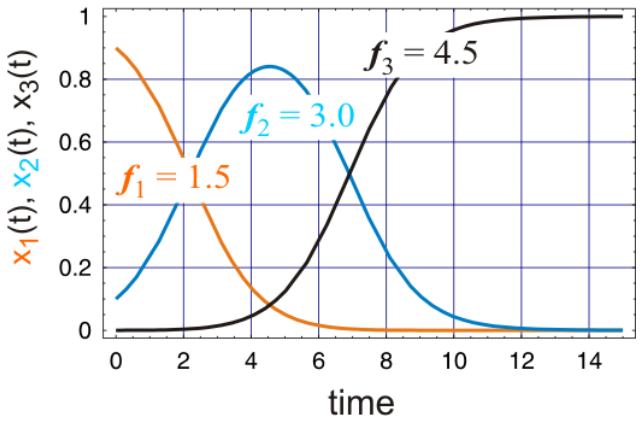

CCCCCCCCCCCCCCCCC
GGGGGGGGGGGGGGGGG
#seq
1
17
17
1
. . . . . .
(((((((· · ·)))))))
#str
1
91
7
1
sequence space
2^{17} = 131 072
shape space
516
$\mathcal{Q}$
S
Φ: $(\mathcal{Q},d_{\mathrm{H}}) \Rightarrow (S,d_{\mathrm{S}})$
X
S = Φ(X)
sequence
genotype
structure
phenotype

Figure 11





shape space 516 parameter space

S $\mathbb{R}_{\geq 0}$

$\Psi: (S, d_S) \Rightarrow \mathbb{R}_{\geq 0}$

$S = \Phi(X)$ $f = \Psi(S)$

structure function

phenotype selection

Figure 12

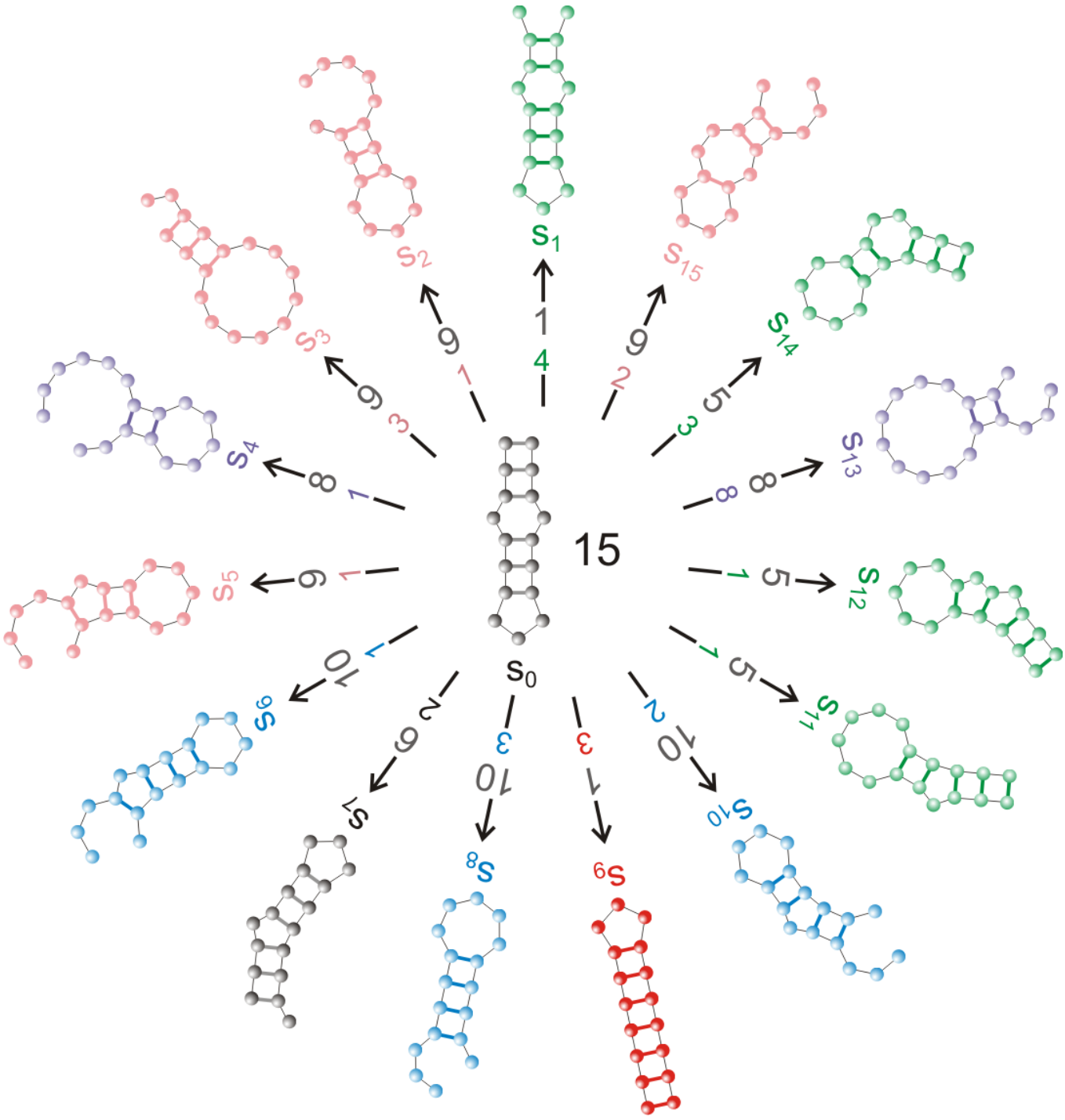

S0
15
S1
14
S2
6
7
S3
6
3
S4
8
1
S5
9
1
S6
10
1
S7
2
6
S8
10
3
S9
1
3
S10
10
2
S11
5
1
S12
5
1
S13
8
8
S14
5
3
S15
6
2

Figure 13

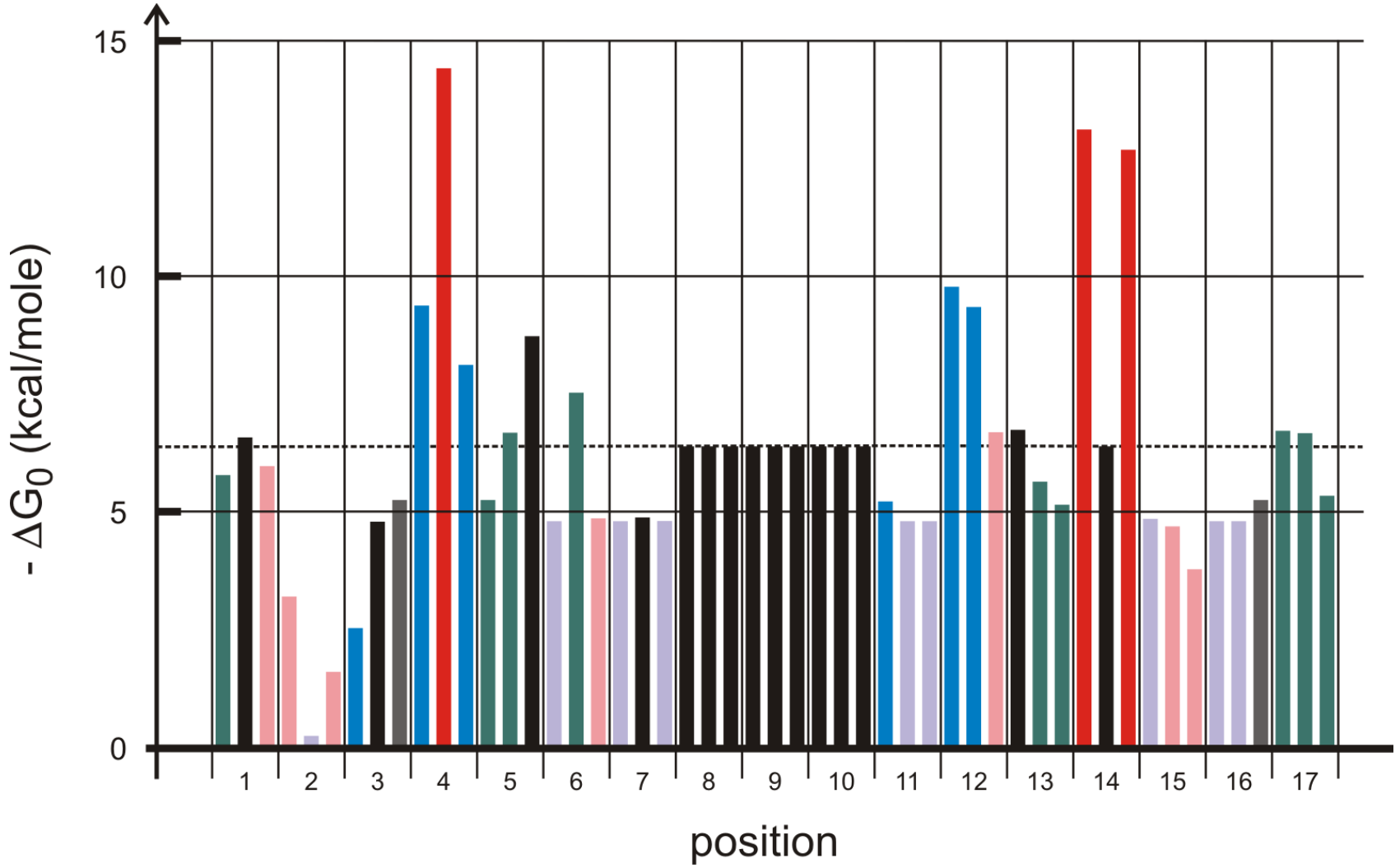

15
10
5
0
- ΔG0 (kcal/mole)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
position

Figure 14

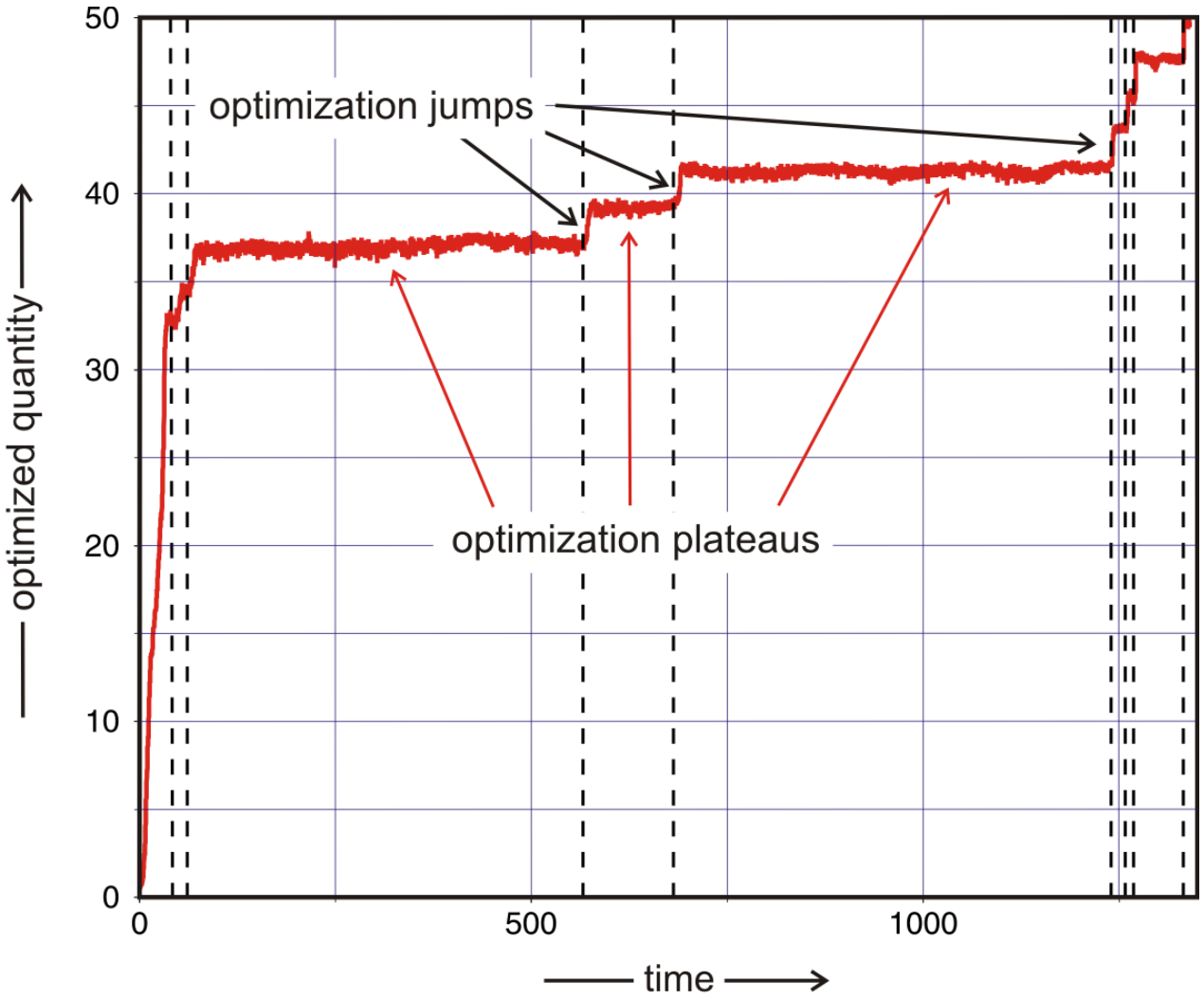

optimization jumps
optimization plateaus
optimized quantity
time
0
10
20
30
40
50
0
500
1000

Figure 15

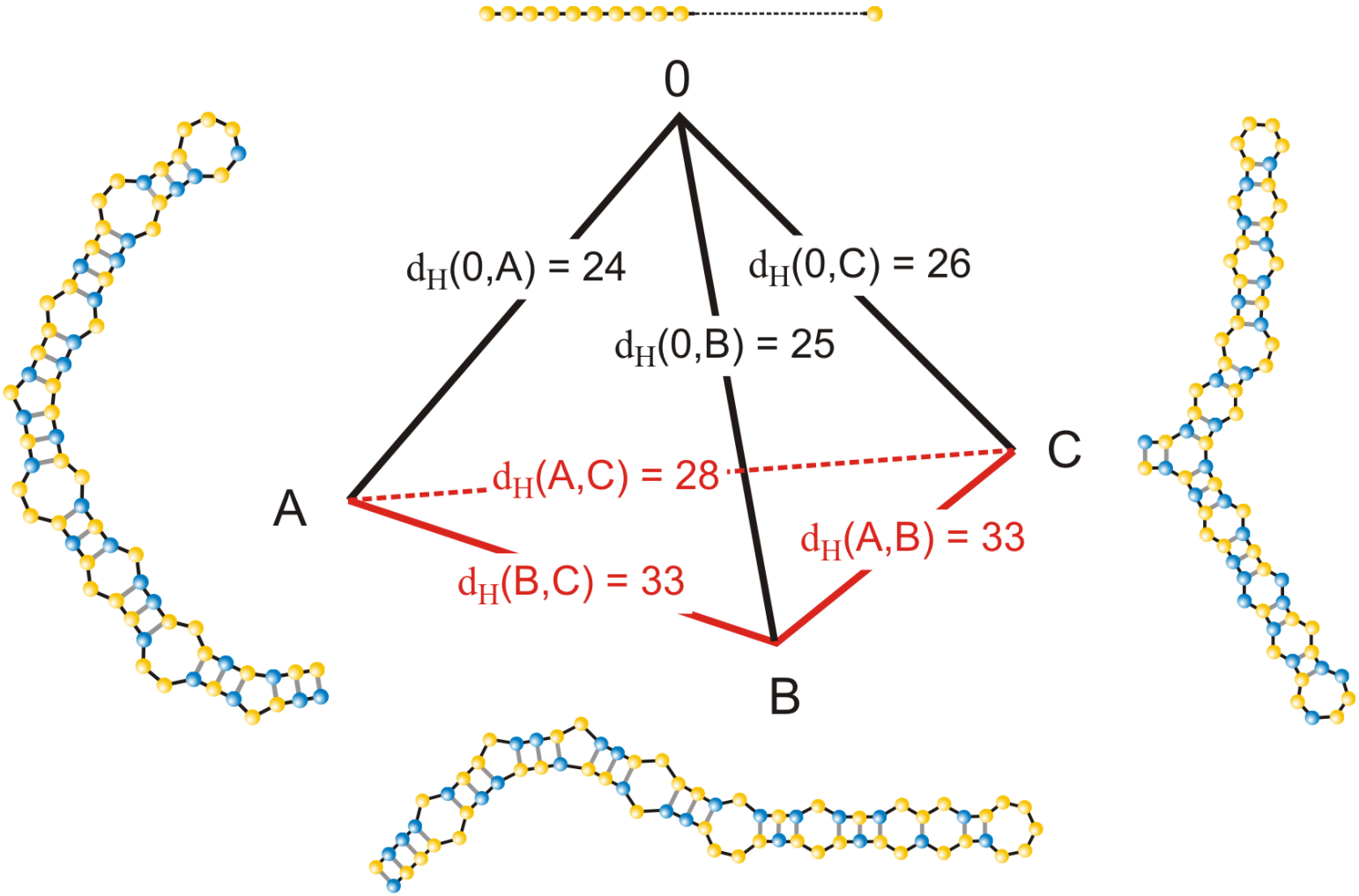


Figure 16

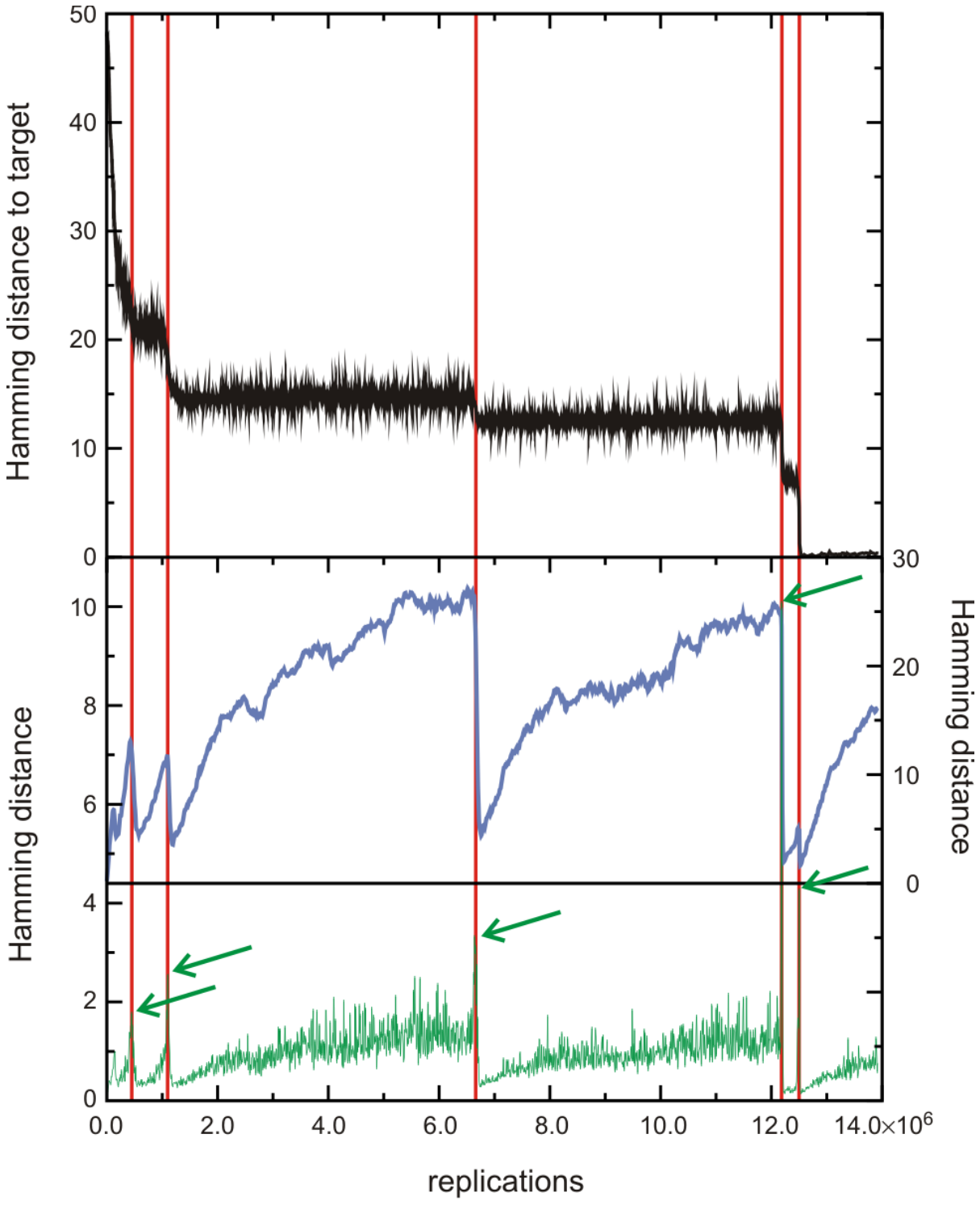

Hamming distance to target
50
40
30
20
10
0
Hamming distance
10
8
6
4
2
0
30
20
10
0
Hamming distance
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0×10^6
replications

Figure 17

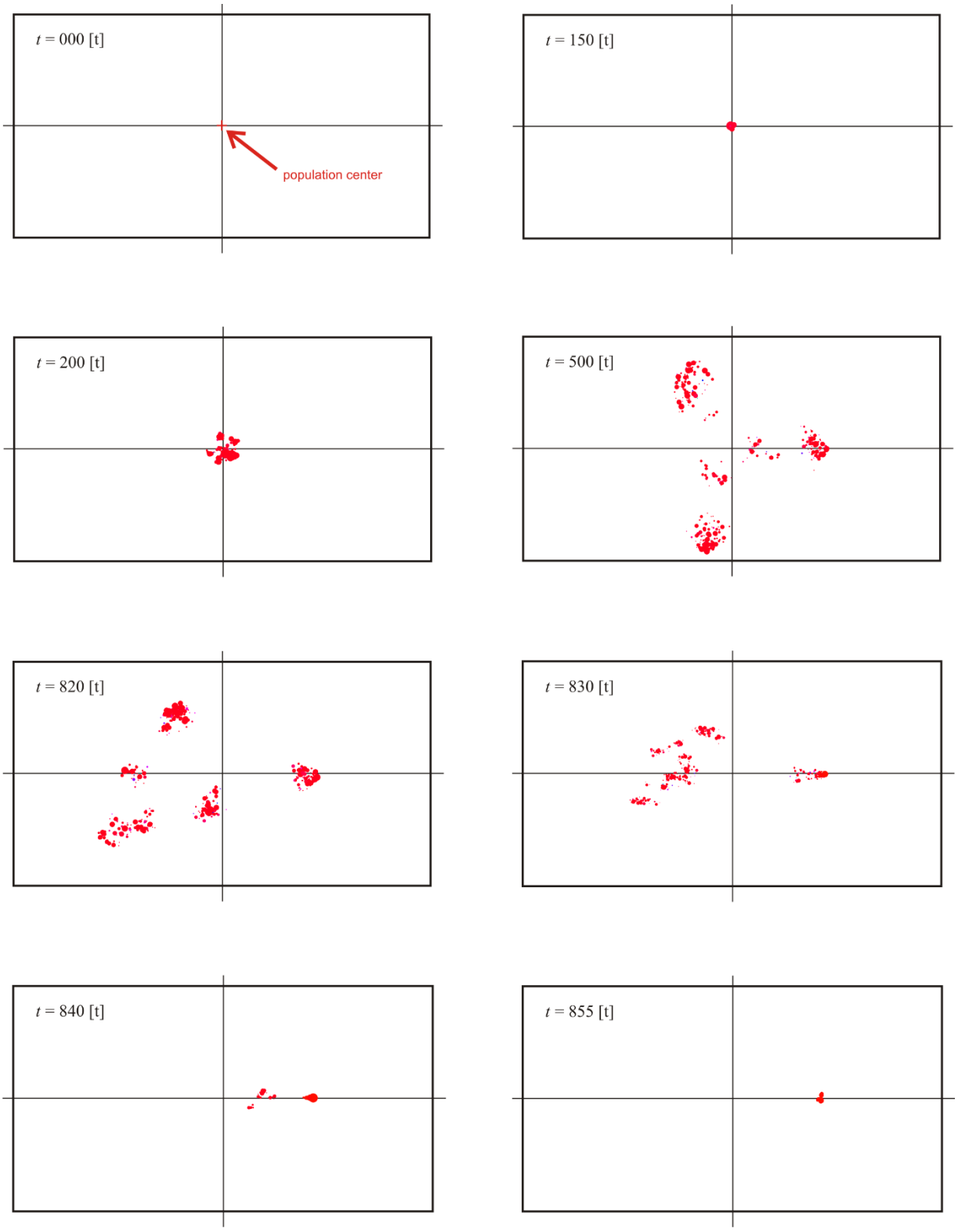

t = 000 [t]
population center
t = 150 [t]
t = 200 [t]
t = 500 [t]
t = 820 [t]
t = 830 [t]
t = 840 [t]
t = 855 [t]

Figure 18