1

Physical principles of evolution

Peter Schuster

Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, 1090 Wien, Austria, pks@tbi.univie.ac.at

Abstract. Theoretical biology without a comprehensive theory of evolution is incomplete, since evolution is in the core of biological thought. Evolution is visualized as a migration process in genotype or sequence space that is either an adaptive walk driven by some fitness gradient or a random walk in absence of (sufficiently large) fitness differences. The Darwinian concept of natural selection consisting in the interplay of variation and selection is based on a dichotomy: All variations occur on genotypes whereas selection is operating on phenotypes and relations between genotypes and phenotypes as encapsulated in a mapping form genotype space into phenotype space are central for an understanding of evolution. Fitness is conceived as a function of the phenotype represented by a second map from phenotype space into nonnegative real numbers. In the biology of organisms, genotype-phenotype maps are enormously complex and relevant information on them is exceedingly scarce. The situation is better in the case of viruses but so far only one example of a genotype-phenotype map, the mapping of RNA sequences into RNA secondary structures, has been investigated in sufficient detail. It provides direct information of RNA selection *in vitro* and test-tube evolution, and it is a basis for testing *in silico* evolution on a realistic fitness landscape. Most of the modeling efforts in theoretical and mathematical biology of today are done by means of differential equations but stochastic effects are of undeniably great importance for evolution. Population sizes are much smaller than the numbers of genotypes constituting sequence space. Every mutant, after all, has to begin with a single copy. Evolution can be modeled by a chemical master equation, which (in principle) can be approximated by a stochastic differential equation. In addition, simulation tools are available that compute trajectories for master equations. The accessible population sizes in the range of $10^7 \leq N \leq 10^8$ molecules are commonly too small for problems in chemistry but sufficient for biology.

1.1 Mathematics and biology

The beginning of modern science in the sixteenth century has been initiated by the extremely fruitful marriage between physics and mathematics. Nobody has expressed the close relation between mathematics and physics

clearer than Galileo Galilei in his famous statement [1]: *Philosophy (science) is written in this grand book, the universe, ... It is written in the language of mathematics, and its characters are triangles, circles and other geometric features. ... .* Indeed, physics and mathematics have cross-fertilized each other from the beginnings of modern science until present day. Theoretical physics and mathematical physics are highly respected disciplines and no physics journal will accept empirical observations without an attempt to bring it into a context that allows for quantification and interpretation by theory. General concepts and successful abstractions have a high reputation in physics and the reductionists' program[1] is the accepted scientific approach towards complex systems. This view is common in almost all subdisciplines of contemporary physics and, in essence, is shared with chemistry and molecular biology.

Conventional biology, in this context, is very different: Great works of biology like Charles Darwin's *Origin of Species* [2] or in recent years Ernst Mayr's *Growth of Biological Thought* [3] do not contain a single mathematical expression, theoretical and mathematical biology had and still have a bad reputation among macroscopic biologists, special cases are preferred over generalizations, which are looked upon with scepticism, and holistic views are commonly more appreciated than reductionists' explanations no matter whether they are in a position to provide insight into problems or not. A famous and unique exception among others is Charles Darwin's theory of *natural selection* by reproduction and variation in finite populations. Although not cast into mathematical equations, the theory is based on a general concept whose plausibility is erected upon a wealth of collected and carefully interpreted empirical observations. Darwin's strategy has something in common with the conventional mathematical approach based on observation, abstraction, conjecture, and proof: On different islands of the Galapagos archipelago Darwin observed similar looking species in different habitats and concluded correctly that these different species are closely related and owe their existence to histories of adaptation to different environments on the individual islands. The occurrence of adaptations has been attributed to natural selection as a common mechanism through abstraction from specific cases. Darwin's conjecture combines three facts known at his times:

(i) *Multiplication*: All organisms multiply by cell division, parthenogenesis or sexual reproduction, multiplication is accompanied by inheritance –

[1] The reductionist program, also called methodological reductionism, aims at an exploration of complex objects through breaking them up into modular, preferentially molecular parts and studying the parts in isolation before reassembling the object. Emergent properties are assumed to be describable in terms of the phenomena from and the processes by which they emerge. The reductionist program is different from ontological reductionism, which denies the idea of ontological emergence by the claim that emergence is merely a result of the system's description and does not exist on a fundamental level.

'progeny resembles parents', and under the condition of unlimited resources multiplication results in exponential growth of population size.

(ii) *Variation*: All natural populations show variance in phenotypic properties either continuously varying features like body size or discontinuously varying features like the number of limbs, the number of digits, color of flowers, skin patterns or seeds shapes, and it is straightforward to relate variation to inheritance.[2]

(iii) *Selection*: Exponential growth results in overpopulation of habitats,[3] only a small fraction of offspring can survive and have progeny of their own, and this stringent competition prevents less efficient variants from reproduction.

Taking together the three items and introducing the notion of fitness for the number of offspring, which reach the age of fertility, the conjecture could be formulated in the following way:

> **Natural selection**: In nonhomogeneous populations the frequencies of variants with fitness values below the population average are decreasing, those with fitness values above average are increasing and consequently the population average itself will increase until it reaches the maximum value corresponding to a homogeneous population of the best adapted or fittest variant.

Darwin's *Origin of Species* is an overwhelming collection of observations from nature, from animal breeders and from nursery gardens that provide strong evidence for the correctness of Darwin's conjecture. This enormous collection in a way is the empirical substitute for a mathematical proof.

Although Gregor Mendel analyzed his experiments on inheritance in peas by mathematical statistics and found thereby the explanatory regularities, mathematics did not become popular in biology. In contrary, Mendel's work has been largely ignored by the biological community for more than thirty years. Then, Mendel has been rediscovered and genetics became an important discipline of biology. Population genetics has been founded by the three scholars Ronald Fisher [8], J.B.S. Haldane [9] and Sewall Wright [10]. In the nineteen thirties they succeeded to unite Mendelian genetics and Darwin's natural selection, and to cast evolution into a rigorous mathematical frame but conventional geneticists and evolutionary biologists continued to fight until the completion of the synthetic theory almost twenty years later [3].

[2] Gregor Mendel was the first to investigate such relations experimentally [4–6] and discovered the transmittance of properties in discrete packages from the parents to offspring. His research objects were the pea (*pisum*) from where he derived his rules of inheritance and the hawkweed (*hieracium*), which was rather confusing for him, because it is apomictic, i.e. it reproduces asexually. Charles Darwin, on the other hand, had a mechanism of inheritance in mind, which was entirely wrong. It was based on the idea of blending of the parents' properties,

[3] According to his own records Charles Darwin has been influenced strongly by Robert Malthus and his demographic theory of [7].

Modeling in biology became an important tool for understanding complex dynamical phenomena. Representative for many other approaches we mention here only three: (i) Modeling of coevolution in a predator-prey system was introduced by Alfred Lotka [11] and Vito Volterra [12] by means of differential equations that were borrowed from chemical kinetics. In a way, they were the pioneers of theoretical ecology, which has been developed by the brothers Howard and Eugene Odum [13] and became a respectable field of applied mathematics later [14]. (ii) A model for pattern formation based on the reaction-diffusion (partial differential) equation with a special chemical mechanism has been suggested and analyzed by Alan Turing [15]. Twenty years later the Turing model was applied to biological morphogenesis [16, 17] and provided explanations for patterns formed during development [18, 19]. (iii) Based on experimental studies of nerve pulse propagation in the squid giant axon Alan Hodgkin and Andrew Huxley formulated a mathematical model for nerve excitation and pulse propagation [20] that became the standard model for single nerve dynamics in neurobiology. They both were awarded the Nobel Prize in Medicine in 1963. A second breakthrough in understanding neural systems came from modeling networks of neurons. John Hopfield conceived an exceedingly simple model of neurons in networks [21] that initiated a whole new area of scientific computing: computation with *neural networks*, in particular modeling and optimization of complex systems. Despite these undeniable and apparent successes, the scepticism of biologists with respect to theory and mathematics, nevertheless, continued for almost the entire rest of the twentieth century.

The advent of molecular biology in the nineteen fifties brought biology closer to chemistry and physics, and changed the general understanding of nature in a dramatic way [22]. Inheritance got a profound basis in molecular genetics and reconstruction of phylogenies became possible through comparison of biopolymer sequences from present day organisms. Structures of biomolecules at atomic resolution were determined by refined techniques from physical chemistry and they gave deep insights into biomolecular functions. Spectroscopic techniques, in particular nuclear magnetic resonance, require a solid background in mathematics and physics for conceiving and analyzing conclusive experiments. A novel era of biology was initiated in the nineteen seventieth when the highly efficient new methods for DNA sequencing developed by Walter Gilbert and Frederick Sanger became available [23, 24]. Sequencing whole genomes became technically within reach and financially affordable. The first two complete bacterial genomes were published in 1995 [25] and the following years saw a true explosion of sequencing data. High-throughput techniques using chip technology for genome wide analysis of translation and transcription products known as proteomics and transcriptomics followed and an amount of data was created that has never been seen before. In this context it is worth to cite the Nobel laureate Sydney Brenner [26] who made the following statement in 2002 to characterize the situation in molecular biology:

> *"I was taught in the pre-genomic era to be a hunter. I learnt how to identify the wild beasts and how to go out, hunt them down and kill them. We are now, however, being urged to be gatherer. To collect everything lying about and and put it into storehouses. Someday, it is assumed someone will come and sort through the storehouses, discard the junk and keep the rare finds. The only difficulty is how to recognize them."*

Who else but a theorist should be this *"someone"*? The current development seems to indicate that *"someday"* is not too far away. The flood of data and the urgent need for a comprehensive theory have driven back the aversion for computer science and mathematics of the biologists. Modern genetics and genome analysis without bioinformatics are unthinkable and understanding network dynamics without mathematics and computer modeling is impossible.

The new discipline of systems biology has the ambitious goal to find holistic descriptions for cells and organisms without giving up the roots in chemistry and physics. Although still in its infancy and falling into one trap after another, modeling in systems biology progresses slowly towards larger and more detailed models for regulatory modules in cell biology. New techniques are developed and applied, examples are flux-balance analysis [27] and application of inverse methods [28], whereby the primary challenge is up-scaling to larger systems like whole organisms. Recent advances in experimental evolution allow for an extension of detailed models to questions of evolution, which is of central importance of biology as Theodosius Dobzhansky has encapsulated in his famous sentence: *"Nothing in biology makes sense except in the light of evolution"* [29]. From a conceptional point of view, theoretical biology is in a better position than theoretical physics where the attempts of unification of the two fundamental theories, quantum mechanics and relativity theory, have not been successful so far. Biology has one comprehensive theory, the theory of evolution, and present day molecular biology is building the bridge to chemistry and physics. Missing are a proper language and efficient techniques to handle the enormous complexity and to build proper models.

1.2 Darwin's theory in mathematical language

If Charles Darwin would have been a mathematician, how might he have formulated his theory of natural selection? Application of mathematics to problems in biology has a long history. The first example that is relevant for evolution dates back to medieval times. In the famous *Liber Abaci* written in the year 1202 by Leonardo Pisano also known as Fibonacci (*filius Bonacci*) we find a counting example of the numbers of pairs of rabbits in subsequent time spans. Every adult pair is assumed to give birth to another pair, new born rabbits have to wait one time interval before they become fertile adults. Starting from a single couple yields the following series:

$$(0)\quad 1\quad 1\quad 2\quad 3\quad 5\quad 8\quad 13\quad 21\quad 34\quad 55\quad 89\quad \ldots\ .$$

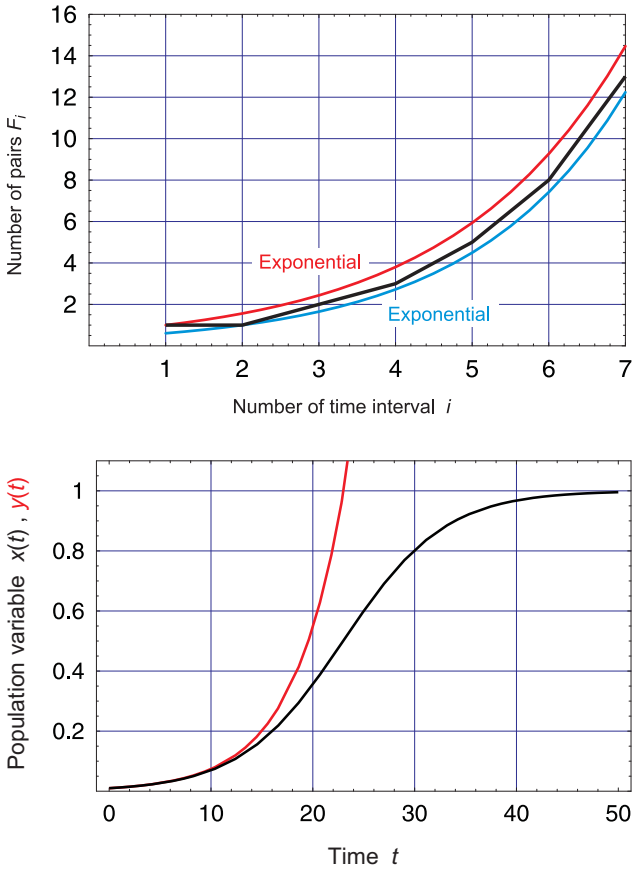


Fig. 1.1. Fibonacci series, exponential functions, and limited resources. The Fibonacci series (black; upper plot) is embedded between two exponential functions in the range $0 < i \leq 10$: $n_{\text{upper}}(t) = \exp\bigl(0.4453 \cdot (t-1)\bigr)$ (red) and $n_{\text{lower}}(t) = \exp\bigl(0.5009 \cdot (t-2)\bigr)$ (blue), wherein the time t is the continuous equivalent to the discrete (generation) index i. The lower plot compares the exponential function, $y(t) = y_0 \exp(r\,t)$ for unlimited growth (red; $y_0 = 0.02$, $r = 0.1$) with the normalized solution of the Verhulst equation ($x(t)$, black; $x_0 = 0.02$, $r = 0.1$, and $C = 1$ by definition)

Every number is the sum of its two precursors and the Fibonacci series is defined by the recursion

$$F_{i+1} = F_i + F_{i-1} \quad \text{with} \quad F_0 = 0 \ \text{ and } \ F_1 = 1\,. \tag{1.1}$$

It is straight forward to show that the Fibonacci series can be approximated well by exponential functions as upper and lower limits (Fig. 1.1). The exponential function, however, was not known before the middle of the eighteenth

century, it was introduced in the fundamental work of the Swiss mathematician Leonhard Euler [30]. Robert Malthus – although living fifty years later – still uses a geometric progression, $2, 4, 8, 16, \ldots$, for the unlimited growth of populations [7]. The consequences of unlimited growth for demography are disastrous and, as said, Malthus' work was influential on Darwin's thoughts.

A contemporary of Charles Darwin, the mathematician Pierre-François Verhulst [31], formulated a model based on differential equations combining exponential growth and limited resources (Fig. 1.1):

$$\frac{dN}{dt} \;=\; \dot{N} \;=\; r\,N\,\left(1-\frac{N}{C}\right) \tag{1.2}$$

with $N(t)$ describing the number of individuals at time t, r being the Malthusian parameter and C the carrying capacity of the ecosystem. Equ. (1.2) consists of two terms: (i) the exponential growth term, rN, and (ii) the constraint to finite population size expressed by the term $-rN^2/C$. In other words, the ecosystem can only support $N = C$ individuals and $\lim_{t\to\infty} N(t) = C$. The solution of the differential equation (1.2) is of the form

$$N(t) \;=\; \frac{N_0\;C}{N_0\;+\;(C-N_0)\exp(-rt)}. \tag{1.3}$$

Herein $N_0 = N(0)$ is the initial number of individuals. It its straightforward to normalize the variable to the carrying capacity, $x(t) = N(t)/C$ yielding

$$x(t) \;=\; \frac{x_0}{x_0\;+\;(1-x_0)\exp(-rt)} \tag{1.3'}$$

with $x_0 = N_0/C$. It will turn out to be useful to cast the term representing the constraint into the form $N\,\phi(t)/C = x\,\phi(t)$. Then, we obtain for the Verhulst equation

$$\frac{dx}{dt} \;=\; \dot{x} \;=\; x\left(r\;-\;\phi(t)\right) \quad \text{with} \quad \phi(t) = x\,r \tag{1.2'}$$

being the (mean) reproduction rate of the population.

Eventually, we generalize to the evolution of n species or variants[4] in the population $\boldsymbol{\Xi} = \{X_1, X_2, \ldots, X_n\}$. The numbers of individuals are now denoted by $[X_i] = N_i$ with $\sum_{i=1}^{n} N_i = N$ and the normalized variables $x_i = N_i/N$ with $\sum_{i=1}^{n} x_i = 1$. Each variant has its individual Malthus parameter of fitness value f_i, and for the selection constraint leading to constant population size we find now $\phi(t) = \sum_{i=1}^{n} x_i\,f_i$, which is the mean reproduction rate of the entire population. The selection constraint $\phi(t)$ can be used for modeling much more general situations than constant population size by

[4] In this chapter we shall not consider sexual reproduction or other forms of recombination. In asexual reproduction a strict distinction between variants and species is neither required nor possible. We shall briefly come back to the problem of bacterial or viral species in section 1.7.

means of the mean reproduction rate. As we shall see in section 1.5, the proof for the occurrence of selection can be extended to very general selection constraints $\phi(t)$ as long as the population size does not become zero, $N > 0$.

The kinetic differential equation in the multi-species case, denoted as selection equation,

$$\dot{x}_j \;=\; x_j\left(f_j \,-\, x_j\sum_{i=1}^{n} x_i\,f_i\right) \;=\; x_j\left(f_j \,-\, x_j\,\phi(t)\right),\; j=1,2,\ldots,n\;, \quad (1.4)$$

can be solved exactly by the integrating factors transform ([32], p.322ff.)

$$z_j(t) \;=\; x_j(t)\cdot\exp\left(\int_0^t \phi(\tau)d\tau\right)\;. \quad (1.5)$$

Insertion into (1.4) yields

$$\begin{aligned}
\dot{z}_j \;&=\; f_j z_j \;\text{ and }\; z_j(t) \;=\; z_j(0)\cdot\exp(f_j t)\;,\\
x_j(t) \;&=\; x_j(0)\cdot\exp(f_j t)\cdot\exp\left(-\int_0^t \phi(\tau)d\tau\right) \;\text{ with}\\
\exp\left(\int_0^t \phi(\tau)d\tau\right) \;&=\; \sum_{i=1}^{n} x_i(0)\cdot\exp(f_i t)\;,
\end{aligned}$$

where we have used $z_j(0) = x_j(0)$ and the condition $\sum_{i=1}^{n} x_i = 1$. The solution finally is of the form

$$x_j(t) \;=\; \frac{x_j(0)\cdot\exp(f_j t)}{\sum_{i=1}^{n} x_i(0)\cdot\exp(f_i t)}\,;\; j=1,2,\ldots,n\;. \quad (1.6)$$

The interpretation is straightforward. The term with the largest fitness value, $f_m = \max\{f_1, f_2, \ldots, f_n\}$, dominates the sum in the denominator after sufficiently long time:[5]

$$\sum_{i=1}^{n} x_i(0)\cdot\exp(f_i t) \;\rightarrow\; x_m(0)\cdot\exp(f_m t) \;\text{ for large } t \;\text{ and }\; x_m(t) \;\rightarrow\; 1\;.$$

Optimization in the sense of Charles Darwin's principle of selection of the fittest variant, X_m, takes place.

The occurrence of selection in equ.(1.4) can be verified also without knowing the solution (1.6). For this goal we consider the time dependence of the constraint ϕ, which is given by

[5] We assume here that the largest fitness value f_m is non-degenerate, i.e. there is no second species having the same (largest) fitness value. In section 1.5 we shall drop this restriction.

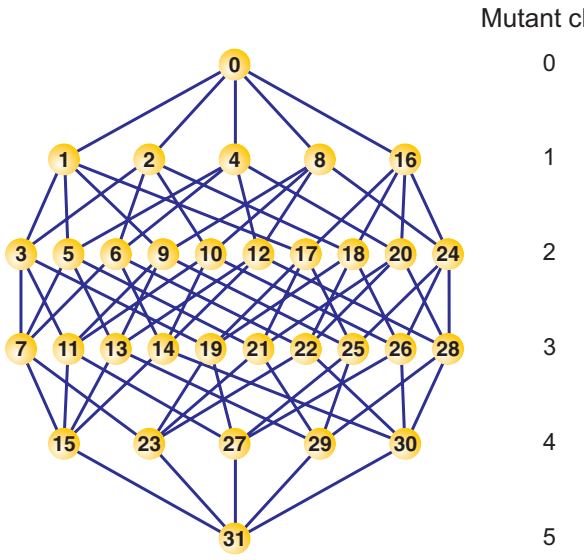


Fig. 1.2. Sequence space of binary sequences of chain length $\ell = 5$. The sequence space $\mathcal{Q}_5^{\{0,1\}}$ comprises 32 sequences. Every sequence is represented by a point. The numbers in the yellow balls are the decimal equivalents of the binary sequences and can be interpreted as sequences of two nucleotides, "**0**" ≡ "**C**" and "**1**" ≡ "**G**". Examples are $0 \equiv$ **00000** ≡ **CCCCC**, $14 \equiv$ **01110** ≡ **CGGGC** or $29 \equiv$ **11101** ≡ **GGGCG**. All positions of a (binary) sequence space are equivalent in the sense that each sequence has ℓ nearest neighbors, $\ell(\ell-1)/2$ next nearest neighbors, etc. Accordingly, sequences are properly grouped in mutant classes around the reference sequence, here 0.

$$\begin{aligned}
\frac{d\phi}{dt} &= \sum_{i=1}^{n} f_i \dot{x}_i = \sum_{i=1}^{n} f_i \left(f_i x_i - x_i \sum_{j=1}^{n} f_j x_j \right) = \\
&= \sum_{i=1}^{n} f_i^2 x_i - \sum_{i=1}^{n} f_i x_i \sum_{j=1}^{n} f_j x_j = \\
&= \overline{f^2} - \left(\overline{f}\right)^2 = \operatorname{var}\{f\} \geq 0 \,. \qquad (1.7)
\end{aligned}$$

Since a variance is always nonnegative, equ.(1.7) implies that $\phi(t)$ is a nondecreasing function of time. The value $\operatorname{var}\{f\} = 0$ implies a (local) maximum of ϕ and hence, ϕ is optimized during selection. Zero variance is tantamount to a homogeneous population containing only one variant. Since ϕ is at a maximum, this is the fittest variant X_m.

1.3 Evolution in genotype space

Evolution can be visualized as a process in an abstract genotype or sequence space, $\mathcal{Q}$. At constant chain lengths ℓ of polynucleotides the sequence space is specified as $\mathcal{Q}_\ell^{\mathcal{A}}$ were $\mathcal{A}$ is the alphabet, for example $\mathcal{A} = \{\mathbf{0}, \mathbf{1}\}$ or $\mathcal{A} = \{\mathbf{G}, \mathbf{C}\}$ is the binary alphabet and $\mathcal{A} = \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$ the natural nucleotide alphabet. The gain of such a comprehensive view of genotypes is generality and the frame for reduction to the essential features, the shortcomings, obviously, are lack of detail. Building a model for evolution upon a space that fulfils all requirements required for the molecular view of biology and which may, eventually, bridge microscopic and macroscopic views, is precisely what we are aiming for here. The genotypes are DNA or RNA sequences and the proper genotype space is sequence space. The concept of a static sequence space [33, 34] has been invented in the early nineteen seventieth in order to bring some ordering criteria into the enormous diversity of possible biopolymer sequences. Sequence space $\mathcal{Q}_\ell^{\mathcal{A}}$ as long we are only dealing with reproduction and mutation is a metric space with the Hamming distance[6] serving as the most useful metric for all practical purposes. Every possible sequence is a point in the discrete sequence space and in order to illustrate the space by a graph, sequences are represented by nodes and all pairs of sequences with Hamming distance one by edges (Fig. 1.2 shows a space of binary sequences as an example. Binary sequence spaces are hypercubes of dimension ℓ being the length of the sequences).

Two properties of sequence spaces are important: (i) All nodes in a sequence space are equivalent in the sense that every sequence has the same number of nearest neighbors with Hamming distance $d_\mathrm{H} = 1$, next nearest neighbors with Hamming distance $d_\mathrm{H} = 2$, and so on, which are grouped properly in mutant classes. (ii) All nodes of a sequence space are at the boundary of the space or, in other words, there is no interior. Both features are visualized easily by means of hypercubes:[7] All points are positioned at equal distances from the origin of the (Cartesian) coordinate system. What makes sequence spaces difficult to handle are neither internal structures nor construction prin-

[6] The Hamming distance $d_\mathrm{H}(X_i, X_j)$ [35] counts the number of positions at which two aligned sequences X_i and X_j differ. It fulfils the four criteria for a metric in sequence space: (i) $d_\mathrm{H}(X_i, X_j) \geq 0$ (non-negativity), (ii) $d_\mathrm{H}(X_i, X_j i) = 0$ if an only if $X_i = X_j$ (identity of indiscernibles), (iii) $d_\mathrm{H}(X_i, X_j) = d_\mathrm{H}(X_i, X_j)$ (symmetry), and (iv) $d_\mathrm{H}(X_i, X_j) \leq d_\mathrm{H}(X_i, X_k) + d_\mathrm{H}(X_i, X_k)$ (triangle inequality). For sequences of equal chain length ℓ end-to-end alignment is the most straightforward alignment, although it may miss close relatedness that is a consequence of deletions and insertions, which are mutations that alter sequence length.

[7] An ℓ-dimensional hypercube in the Cartesian space of dimension ℓ is the analogue of a (three-dimensional) cube. The ℓ-dimensional hypercube is constructed by drawing 2ℓ (hyper)planes of dimension $(\ell - 1)$ perpendicular to the coordinate axes at the positions $\pm a$. The corners of the hypercubes are the 2ℓ points where ℓ planes cross.

ciples but the hyper-astronomically large numbers of points: $|\mathcal{Q}_\ell^{\mathcal{A}}| = \kappa^\ell$ for the sequences of length ℓ over an alphabet of size κ with $\kappa = |\mathcal{A}|$.

The population $\boldsymbol{\Xi} = \{X_1, X_2, \ldots, X_n\}$ is represented by a vector with the numbers of species as elements $\mathbf{N} = (N_1, N_2, \ldots, N_n)$, the population size is the L_1-norm:

$$N \;=\; \|\mathbf{N}\|_1 \;=\; \sum_{i=1}^{n} |N_i| \;=\; \sum_{i=1}^{n} N_i \;,$$

where absolute values are dispensable since particle numbers are real and non-negative by definition. Normalization of the variables yields $\mathbf{x} = \mathbf{N}/\|\mathbf{N}\|$ or $x_i = N_i/N$ and $\sum_{i=1}^{n} x_i = 1$, respectively. A population is thus represented by an L_1-normalized vector $\mathbf{x}$ and the population size N. An important property of a population is its *consensus sequence*, $\bar{X}$, consisting of a nucleotide distribution at each position of the sequence. This consensus sequence can be visualized as the center of the population in sequence space.

A sequence is conventionally understood as a string of ℓ symbols chosen from some predefined alphabet with κ letters, which can be written as

$$X_j \;=\; \left(b_1^{(j)}, b_2^{(j)}, \ldots, b_\ell^{(j)}\right) \quad \text{with} \quad b_i^{(j)} \in \mathcal{A} = \{\alpha_1, \ldots, \alpha_\kappa\} \;.$$

The natural nucleotide alphabet contains four letters: $\mathcal{A} = \{\mathsf{A}, \mathsf{U}, \mathsf{G}, \mathsf{C}\}$, but RNA molecules with catalytic functions have been derived also from three- and two-letter alphabets [36,37]. For the forthcoming considerations it is straightforward to adopt slightly different definitions: A sequence X_j results from the multiplication of the alphabet vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\kappa)$ with a $\kappa \times \ell$ matrix $\mathcal{X}_j$ having only 0 and 1 as entries:

$$\begin{aligned} X_j \;&= \boldsymbol{\alpha} \cdot \mathcal{X}_j \;=\; \boldsymbol{\alpha} \cdot \left(\boldsymbol{\beta}_1^{(j)}, \boldsymbol{\beta}_2^{(j)}, \ldots, \boldsymbol{\beta}_\ell^{(j)}\right) \quad \text{with} \\ \boldsymbol{\beta}_i^{(j)} \;&\in \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\} . \end{aligned} \tag{1.8}$$

In other words, the individual nucleotides in the sequence X_j are replaced by products of two vectors, $b_i^{(j)} = \boldsymbol{\alpha} \cdot \boldsymbol{\beta}_i^{(j)}$.

With the definition (1.8) it is straightforward to compute the consensus sequence of a population $\boldsymbol{\Xi}_k$:

$$\boldsymbol{\Xi}_k \;=\; \boldsymbol{\alpha} \cdot \sum_{j=1}^{n} x_j^{(k)} \, \mathcal{X}_j \;, \tag{1.9}$$

and the distribution of nucleotides at position "i" is given by

$$\boldsymbol{b}_i^{(k)} \;=\; \boldsymbol{\alpha} \cdot \sum_{j=1}^{n} x_j^{(k)} \, \boldsymbol{\beta}_i^{(j)} \;. \tag{1.9'}$$

It is important to note the difference between $b_i^{(j)}$ and $\boldsymbol{b}_i^{(k)}$: The former refers to the nucleotide at position "i" in a given sequence whereas the latter describes the nucleotide distribution at position "i" in the population. In case one nucleotide is dominating at every position the distribution can be collapsed to a single sequence, the consensus sequence.

The internal structure of every sequence space $\mathcal{Q}_\ell^{\mathcal{A}}$ is induced by point mutation and this is essential for inheritance because it creates a hierarchy in the accessability of genotypes. Suppose we have a probability p to make one error in the reproduction of a sequence then, provided mutation at different positions is assumed to be independent, the probability to make two errors is p^2, to make three errors is p^3, etc. Inheritance requires sufficient accuracy of reproduction – otherwise children would not resemble their parents – and this implies p has to be sufficiently small. Then, p^2 is smaller and the power series $p^{d_{\mathrm{H}}}$ decreases further with increasing distance from the reference sequence. This ordering of sequences according to a probability criterion that is intimately related to the Hamming metric (see section 1.5). As a matter of fact mutation is indeed a fairly rare event in evolution and populations are commonly dominated by a well-defined single consensus sequence since single nucleotide exchanges that occur at many different positions do not contribute significantly to the average.

Evolutionary dynamics is understood as change of the population vectors in time: $\mathbf{N}(t)$ or $\mathbf{x}(t)$. This change can be modeled by means of differential equations (section 1.5) or stochastic processes (section 1.6). A practical problem concerns the representation of genotype space. Complete sequence space, $\mathcal{Q}_\ell^{\mathcal{A}}$ has the advantage to cover all possible genotypes but its extension is huge and, since the numbers of possible genotypes exceed all even the largest populations by far we are confronted with the problem that most degrees of freedom are empty and very likely will never be populated during the evolutionary process described. Alternatively the description could be restricted to those genotypes, which are actually present in the population and which constitute the *population support* $\boldsymbol{\Phi}(t)$ that is defined by

$$\boldsymbol{\Phi}(t) \doteq \{X_j | N_j(t) \geq 1\} \,. \tag{1.10}$$

The obvious advantage is a drastic reduction in the degrees of freedom to a tractable size but one has to pay the price for the simplification: The population support is time dependent and changes whenever a new genotype is produced by mutation or an existing one goes extinct [38]. Depending on population size population dynamics on the support can be described either by differential equations or modeled as a stochastic process. Support dynamics, on the other hand, is intrinsically stochastic since every mutant starts from a single copy.

Finally, it is important to mention that recombination without mutation can be modeled successfully as a process in an abstract recombination space [39–41] and plays a major role in the theory of genetic algorithms [42, 43]. A great challenge for theorists is the development of a genotype space for

both, mutation and recombination. Similarly, convenient sequence spaces for genotypes with variable chain lengths are not at hand.

1.4 Modeling genotype-phenotype mappings

Unfolding genotypes to yield phenotypes is studied in developmental biology and provides the key to understanding evolution and, in particular, the origin of species. For a long time it has been common knowledge already that the same genotype can develop into different phenotypes depending on differences in the environmental conditions and epigenetic effects.[8] Current molecular biology provides explanations for several epigenetic observations and reveals mechanisms for the inheritance of properties that are not encoded by the DNA of the individual. Still, genetics is shaping the phenotypes – otherwise progeny would not resemble parents – but epigenetics and environmental influences provide additional effects that are indispensable for understanding and modeling the relations between genotypes and phenotypes. Here we shall adopt the conventional strategy of physicists and consider simple cases in which the genotypes unfolds unambiguously into a unique phenotype. This condition is fulfilled, for example, in evolution *in vitro* when biopolymer sequences form (the uniquely defined) minimum free energy structures as phenotypes. Bacteria in constant environments provide other cases of simple genotype-phenotype mappings (the long-term experiments of Richard Lenski [44–46] may serve as examples; see section 1.6). Under this simplifying assumption genotype-phenotype relations can be modeled as mappings from an abstract genotype space into a space of phenotypes or *shapes*. A counter example in a simple system is provided by biopolymers with metastable suboptimal conformations, which can serve a models where a single genotype – a sequence – can give rise to several phenotypes being molecular structures [47].

Since only point mutations shall be considered here, the choice of an appropriate genotype space is straightforward. It is the sequence space $\mathcal{Q}_\ell^{\mathcal{A}}$ with the Hamming distance d_{H} as metric. The phenotype space or shape space $\mathcal{S}_\ell$ is the space of all phenotypes formed by all genotypes of chain length ℓ. Although the definition of a physically or biologically meaningful distance between phenotypes is not at all straightforward, some kind of metric can always be found. Accordingly the genotype-phenotype mapping ψ can be characterized by

$$\psi : \left\{\mathcal{Q}_\ell^{(\mathcal{A})};\, d_{\mathrm{H}}(X_i, X_j)\right\} \stackrel{\text{mfe}}{\Longrightarrow} \left\{\mathcal{S}_\ell; d_S(S_i, S_j)\right\} \quad \text{or} \quad S_k = \psi(X_k)\,. \tag{1.11}$$

The map ψ need not be invertible. In other words, several genotypes can be mapped onto the same phenotype when we are dealing with a case of neutrality.

[8] *Epigenetics* was used as a term subsuming phenomena that could not be explained by conventional genetics.

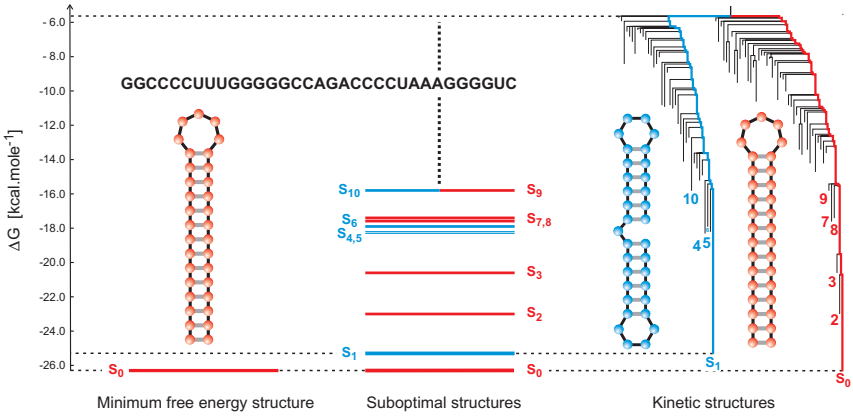


Fig. 1.3. Secondary structures of ribonucleic acid molecules (RNAs). Conventional RNA folding algorithms compute the minimum free energy (mfe) structure for a given sequence [48, 49]. Hairpin formation is shown as an example on the l.h.s. of the figure. In addition, the sequence can fold also into a large number of suboptimal conformations (diagram in the middle of the figure), which are readily computed by efficient computer programs [50, 51]. In case a suboptimal structure is separated from the mfe-structure by a sufficiently high activation barrier, the structure is metastable. The metastable structure in the example shown here is a double hairpin (r.h.s. of the figure). The activation energy of more than 20 kcal/mole does not allow for interconversion of the two structures at room temperature (For the calculation of kinetic structures see, for example, [52, 53]).

An example of a genotype-phenotype mapping that can be handled straightforwardly by analytical tools is provided by *in vitro* evolution of RNA molecules [54–56]. RNA molecules are transferred to a solution containing activated monomers as well as a virus-specific RNA replicase. The material consumed by the replication reaction is replenished by serial transfer of a small sample into fresh solution. The replicating ensemble of RNA molecules optimizes the mean RNA replication rate of the population in the sense of Darwinian evolution (see equ. 1.6). The interpretation of RNA evolution *in vitro* identifies the RNA sequence with the genotype. The RNA structure, the phenotype, is responsible for binding to the enzyme and for the progress of reproduction, since the structure of the template molecules has to opened in order to allow for replication [57–59]. In case of RNA aptamer selection[9] the binding affinity is a function of molecular structure and sequence-structure mapping is an excellent model for the relation between genotype and phenotype.

[9] An aptamer is a molecule that binds to a predefined target molecule. Aptamers are commonly produced by an evolutionary selection process [60].

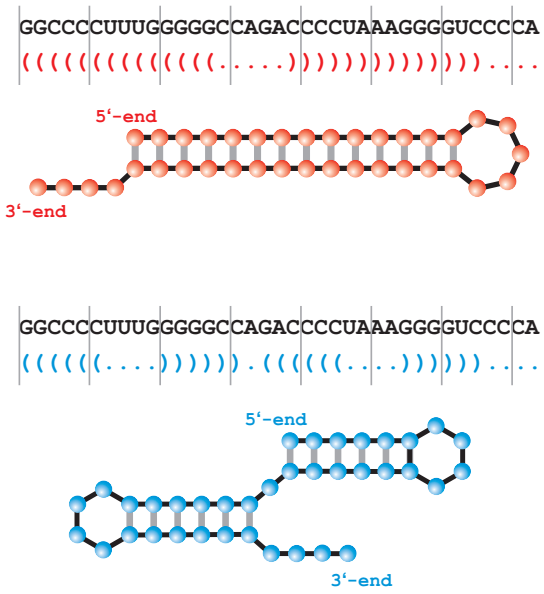


Fig. 1.4. Symbolic notation of RNA secondary structures. RNA molecules have two chemically different ends, the 5'- and the 3'-end. A general convention determines that all strings corresponding to RNA molecules (sequences, symbolic notation, etc) start from the 5'-end and have the 3'-end at the r.h.s. The symbolic notation is equivalent to graphical representation of secondary structures. Base pairs are denoted by parentheses where the opening parenthesis corresponds to the nucleotide closer to the 5'-end and the closing parenthesis to the nucleotide closer to the 3'-end of the sequence. In the figure we compare the symbolic notations with the conventional graphic representations for two structures formed by the same sequence.

RNA sequences fold spontaneously into secondary structures consisting of double helical stacks and single stranded stretches. Within a stack nucleotides form base pairs that are elements of a pairing logic $\mathcal{B}$, which consists of six allowed base pairs in case of RNA structures: $\mathcal{B} = \{\mathbf{AU}, \mathbf{UA}, \mathbf{GC}, \mathbf{CG}, \mathbf{GU}, \mathbf{UG}\}$. Further structure formation, very often initiated by the addition of two-valent cations mostly $\mathbf{Mg}^{2+}$, folds secondary structure into three-dimensional structures by means of sequence specific *tertiary* interactions of nucleotide bases called motifs [61, 62]. Secondary structures have the advantage of computational and conceptional simplicity allowing for the application of combinatorics to global analysis of sequence-structure mappings [47, 63]. A conventional RNA secondary structure consists exclusively of base pairs and unpaired nucleotides and can be represented in a formal three-letter alphabet with the symbols '·', '(', ')' for unpaired nucleotides, downstream bound and upstream bound nucleotides, respectively (Fig.1.4). A straightforward way to annotate

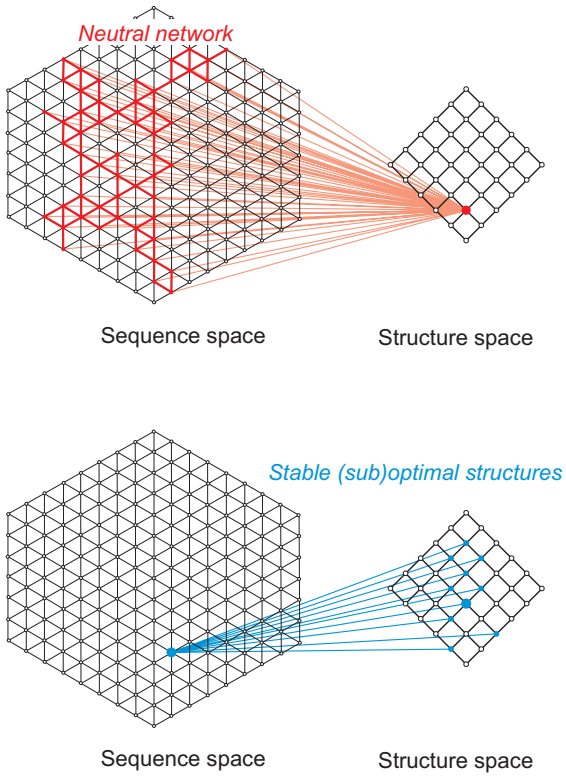


Fig. 1.5. Mappings from sequence space onto shape space and back. In the upper part of the figure we sketch a mapping from sequence space onto structure or shape space.[a] One structure is uniquely assigned to every sequence. The drawing shows the case of a mapping, which is many-to-one and non-invertible: Many sequences fold into the same secondary structure and build a *neutral network*. The lower part of the figure sketches the set of stable (sub)optimal structures that are formed by a single sequence. The mfe structure is indicated by a larger circle.

[a] Both sequence space and shape space are high-dimensional. The two-dimensional representation is used for the purpose of illustration only.

pairs in structures is given by the *base pair count* $S_i = [\gamma_1^{(i)}, \ldots, \gamma_\ell^{(i)}]$, which we illustrate here by means of the lower (blue) structure in the figure as an example:[10]

[10] The base pair count is another equivalent representation of RNA secondary structures. In case of conventional secondary structures the symbolic notation is converted into the base pair count by an exceedingly simple algorithm: Starting with zero at the 5'-end and proceeding from left to right a positive integer counting the

$$S_i \;=\; [1,2,3,4,5,6,0,0,0,0,6,5,4,3,2,1,0,7,8,9,10,11,12,0,0,0,0,12,11,10,9,8,7,0,0,0,0]$$

Consecutive numbers are assigned to first nucleotides of base pairs corresponding to an opening parenthesis in the sequence, in which they appear in the structure, and the same number is assigned to the corresponding closing parenthesis lying downstream. Unpaired nucleotides are denoted by '0'. In total the structure contains n_{p} base pairs and n_{s} single nucleotides with $2n_{\mathrm{p}} + n_{\mathrm{s}} = \ell$.

Molecular physics provides an excellent tool for modeling folding of molecules into structures, the concept of *conformation space*: A free energy is assigned to or calculated for each conformation of the molecule. Commonly, the variables of conformation space are continuous, bond lengths, valence angles or torsion angles may serve as examples. The free energy (hyper)surface or free energy landscape of a molecule presents the free energy as a function of the conformational variables. The mfe structure corresponds to the global minimum of the landscape, metastable states to local minima. In the case of RNA secondary structures conformation space and shape space are identical, and they are discrete spaces, since a nucleotide is either paired or unpaired. Whether a given conformation, a given base pairing pattern, is a local minimum or not depends also on the set of allowed moves in shape space $\mathcal{S}$. The move set defines the distance between structures, the metric $d_S(S_i, S_j)$ in equation (1.11). An appropriate move set for RNA secondary structures comprises three moves: (i) base pair closure, (ii) base pair opening, and (iii) base pair shift [47, 52]. The first two moves need no further explanation, the shift move combines base pair opening and base pair formation with neighboring unpaired nucleotides. This set of three moves corresponds to a metric $d_S(S_i, S_j)$, which is the Hamming distance between the symbolic notations of the two structures S_i and S_j.

Conventional structure prediction is dealing with single structures derived from single sequence inputs. Structure formation depends on external conditions like temperature, pH-value, ionic strength or the nature of the counterions and in order to obtain a unique solution these conditions have to be specified. Commonly the search goes for the most stable structure, the minimum free energy (mfe) structure, which corresponds to the global minimum of the conformational free energy landscape of the RNA molecule. In fig.1.3 the mfe structure $S_0 = \psi(X)$ is a single long hairpin shown (in red) at the l.h.s. of the picture. A sequence that forms a stable mfe structure S_0 (free energy of folding:[11] $\Delta G_{\mathrm{fold}}(S_0) < 0$) commonly forms almost always a set of suboptimal conformations $\{S_1, S_2, \ldots, S_m\}$ with higher free energies of formation, $\Delta G_{\mathrm{fold}}(S_i) > \Delta G_{\mathrm{fold}}(S_0)$ for $i \neq 0$. In fig.1.3 (middle) the ten lowest suboptimal structures are listed; together with S_0 they represent the

number of open parenthesis is assigned to every position along the sequence. The base pair count is not only more convenient for base pair assignments but also more general. It is, for example, applicable to RNA structures with pseudoknots.

[11] The free energy of folding is the difference in free energy between the structure S_i and the unfolded (open) chain $\mathcal{O}$: $\Delta G_{\mathrm{fold}}(S_i) = G(S_i) - G(\mathcal{O})$.

eleven lowest states of the spectrum of structures associated with the sequence X. Low-lying suboptimal conformations may have influence on the molecular properties in particular when conformational changes are involved. The Boltzmann-weighted contributions of all suboptimal structures at temperature T are readily calculated by means of the partition function of RNA secondary structures [49, 64]. Instead of base pairs the analysis of the partition function yields base pairing probabilities that tell how likely it is to find two specific nucleotides forming a base pair in the ensemble of structures at thermal equilibrium.

Although folding RNA sequences into secondary structures is, presumably, the simplest conceivable case of a genotype-phenotype map, it is at the same time an example for the origin of complexity at the molecular level. The base pairing interaction is essentially non-local since a nucleotide can pair with another nucleotide from almost any position of the sequence.[12] The strongest stabilizing contributions to the free energy of structure formation come from neighboring base pairs and are therfore local. The combination of local and non-local effects is one of the most common sources of complex relations in mappings.

The relation of an RNA sequences and its suboptimal structures is sketched in fig.1.5 (lower part). A single sequence X gives rise to a whole set of structures spread all over shape space. In principle, all structures that are *compatible* with the sequence appear in the spectrum of suboptimals but only a subset is stable in the sense that the structure S_i $(i = 1, \ldots)$ corresponds to a local minimum of the conformational energy surface and the free energy of folding is negative ($\Delta G_{\text{fold}}(S_i) < 0$). Using the base pair count the set of all structures that are compatible with the sequence X_h can be defined straightforwardly:

$$S_i \in \mathcal{C}(X_h) \text{ iff } \{\gamma_j^{(i)} = \gamma_k^{(i)} \implies b_j^{(h)} b_k^{(h)} \in \mathcal{B} \; \forall \; \gamma_j \neq 0,\, j = 1, \ldots, \ell\} \quad (1.12)$$

In other words, a structure S_i is compatible with a sequence X_h if, and only if, two nucleotides that can form a base pair, appear in the sequence at all pairs of positions, which are joined by a base pair in the structure. For an arbitrary sequence the number of compatible structures is extremely large but the majority of them has either positive free energies of folding ($\Delta G_{\text{fold}}(S_i) > 0$) and/or represent saddle points rather than local minima of the conformational energy surface. Fig.1.5 indicates the relation between an RNA sequence, its mfe structure, and its stable suboptimal conformations.

Studies of mfe structures or suboptimal structures refer to a certain set of conditions – for example, temperature T, pH, ionic strength – but time is missing since free energy differences (ΔG) or partition functions are equilibrium properties. The structures that are determined and investigated experimentally, however, refer always to some time window – we are not dealing with

[12] Pairing with nearest neighbors is excluded for geometrical reasons. In other words, base pairs of two adjacent nucleotides have such a high positive free energy of formation that they are never observed.

equilibrium ensembles but with metastable states. The finite time structures of RNA are obtained by kinetic folding (see, e.g., [52,53]). The RNA example shown in fig.1.3 represents the case of a bistable molecule: The most stable suboptimal structure S_1, a double hairpin conformation (blue), is the most stable representative of a whole family of double hairpin structures forming a broad basin of the free energy landscape of the molecule. This basin is separated from the basin of the single hairpin structure S_0 by a high energy barrier of about 20 kcal/mole and this implies that practically no interconversion of the two structures will take place at room temperature. We are dealing with an RNA molecule with one stable and one metastable conformation, a so called RNA switch. RNA switches are frequent regulatory elements in procaryotic regulation of translation [65].

1.5 Chemical kinetics of evolution

Provided population sizes N are sufficiently large, mutation rates are high enough, and stochastic effects are reduced by statistical compensation, evolution can be described properly by means of differential equations. In essence, we proceed as described in section 1.2 and find for replication and mutation as an extension of the selection equation (1.4)

$$\begin{aligned} \frac{dx_j}{dt} &= \sum_{i=1}^{n} Q_{ji}\, f_i\, x_i \;-\; \phi(t)\, x_j\,,\; j = 1,\ldots,n \text{ with } \phi(t) = \sum_{i=1}^{n} f_i x_i \\ \text{or}\quad \frac{d\boldsymbol{x}}{dt} &= \Big(\mathrm{Q}\cdot\mathrm{F} \;-\; \phi(t)\Big)\,\boldsymbol{x} \;=\; \Big(\mathrm{W} \;-\; \phi(t)\Big)\,\boldsymbol{x}\,, \end{aligned} \tag{1.13}$$

where $\boldsymbol{x}$ is an n-dimensional column vector; Q and F are $n \times n$ matrices. The matrix Q contains the mutation probabilities – Q_{ji} referring to the production of X_j as an error copy of template X_i – and F is a diagonal matrix whose elements are the replication rate parameters or fitness values f_i.

Solutions of the mutation-selection equation (1.13) can be obtained in two steps: (i) integrating factor transformation allows for an elimination of the nonlinear term $\phi(t)$ and (ii) the remaining linear equation is solved in terms of an eigenvalue problem [66–69].

$$x_j(t) \;=\; \frac{\sum_{k=1}^{n} b_{jk} \sum_{i=1}^{n} h_{ki}\, x_i(0)\, \exp(\lambda_k\, t)}{\sum_{l=1}^{n} \sum_{k=1}^{n} b_{lk} \sum_{i=1}^{n} h_{ki}\, x_i(0)\, \exp(\lambda_k\, t)}\,,\; j = 1,\ldots,n\,. \tag{1.14}$$

The new quantities in this equation, b_{jk} and h_{kj}, are the elements of two transformation matrices:

$$\mathrm{B} = \{b_{jk}; j = 1,\ldots,n; k = 1,\ldots,n\} \text{ and}$$
$$\mathrm{B}^{-1} = \{h_{kj}; k = 1,\ldots,n; j = 1,\ldots,n\}$$

The columns of B and the rows of B^{-1} represent the right hand and left hand eigenvectors of the matrix $\mathrm{W} = \mathrm{Q} \cdot \mathrm{F}$ with $\mathrm{B}^{-1} \cdot \mathrm{WB} = \boldsymbol{\Lambda}$ being a diagonal matrix containing the eigenvalues of W. The elements of the matrix W are non-negative by definition since they are the product of a fitness value or replication rate parameter f_i and a mutation probability Q_{ji}, which both are non-negative. If, in addition, W is a non-negative primitive matrix[13] – implying that every sequence can be reached from every sequence by a finite chain of consecutive mutations – the conditions for the validity of Perron-Frobenius theorem [70] are fulfilled. Two (out of six) properties of the eigenvalues and eigenvectors of W are important for replication-mutation dynamics:
(i) The largest eigenvalue λ_1 is non-degenerate, $\lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$, and
(ii) the unique eigenvector belonging to λ_1 denoted by $\boldsymbol{\xi}_1$ has only positive elements, $\xi_j^{(1)} > 0 \,\forall\, j = 1, \ldots, n$.

After sufficiently long time the population converges to the largest eigenvector $\boldsymbol{\xi}_1$ which is, therefore, the stationary state of equ. (1.13). Since $\boldsymbol{\xi}_1$ represents the genetic reservoir of an asexually replicating species it was called the *quasispecies* [68]. A quasispecies commonly consists of a fittest genotype, the *master sequence*, and a mutant distribution surrounding the master sequence in sequence space. Although the solution of the mutation-selection is straightforward, an experimental proof for the existence of a stationary mutant distribution and its analysis are quite involved [71]. The work has been conducted with relatively short RNA molecules (chain length: $\ell = 87$). Genotypic heterogeneity in virus populations was detected already in the nineteen seventies [72]. Later, the existence of quasispecies in nature has been demonstrated for virus populations (For an overview and a collection of reviews see [73, 74]). Since it is very hard if not impossible to prove that a natural population is in a steady state, the notion *virus quasispecies* was coined for virus populations observed *in vitro* and *in vivo*.

In order to explore quasispecies as a function of the mutation rate p a crude or zeroth order approximation consisting of neglect of backward mutations has been adopted [33]. The differential equation for the master sequence is then of the form

$$\frac{dx_m^{(0)}}{dt} \;=\; Q_{mm} f_m \, x_m^{(0)} - x_m^{(0)} \, \phi(t) \;=\; x_m^{(0)} \Big(Q_{mm} f_m - \bar{f}_{-m} - x_m^{(0)} (f_m - \bar{f}_{-m}) \Big),$$

with $\bar{f}_{-m} = \big(\sum_{j=1, j \neq m}^{n} f_j x_j \big) / (1 - x_m)$. We apply the uniform error approximation and assume that the mutation rate per nucleotide and replication event, p, is independent of the nature of the nucleotide (**A**, **U**, **G** or **C**) and the position along the sequence and find for the elements of the mutation matrix Q

[13] A square non-negative matrix $\mathrm{W} = \{w_{ij}; i, j = 1, \ldots, n;\, w_{ij} \geq 0\}$ is called *primitive* if there exists a positive integer m such that W^m is strictly positive: $\mathrm{W}^m > 0$ which implies $\mathrm{W}^m = \{w_{ij}^{(m)}; i, j = 1, \ldots, n; w_{ij}^{(m)} > 0\}$.

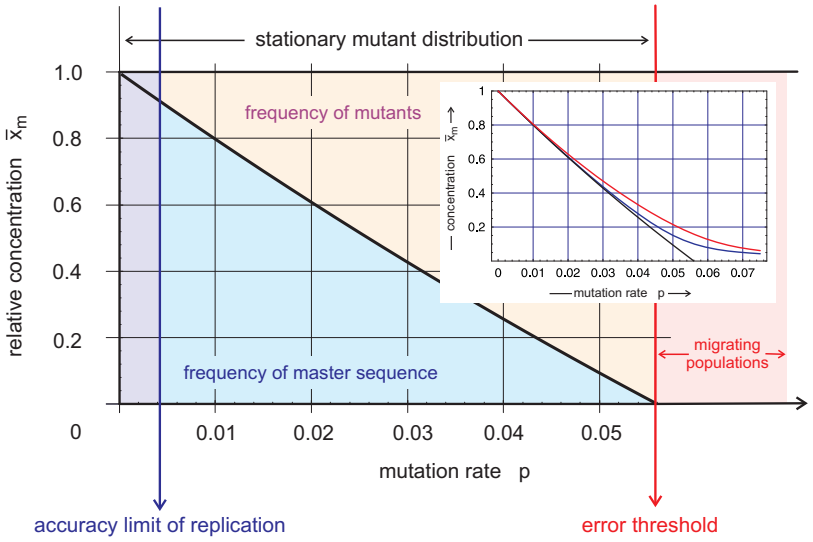


Fig. 1.6. The error threshold in RNA replication. The stationary frequency of the master sequence X_m is shown as a function of the mutation rate p. In the zeroth order approximation neglecting mutational backflow the function $\bar{x}_m^{(0)}(p)$ is almost linear in the particular example shown here. In the insert the zeroth order approximation (black) is shown together with the exact function (red) and an approximation applying the uniform distribution to the mutational cloud ($\bar{x}_j = (1-\bar{x}_m)/(n-1)\,\forall\, j \neq m$; blue), which is exact at the mutation rate $p = 0.5$ for binary sequences. The error rate p has two natural limitations: (i) the physical accuracy limit of the replication process provides a lower bound for the mutation rate and (ii) the error threshold defines a minimum accuracy of replication that is required to sustain inheritance and sets an upper bound for the mutation rate. Parameters used in the calculations: binary sequences, $\ell = 6$, $\sigma = 1.4131$.

$$Q_{jj} \;=\; (1-p)^{\ell} \;\text{ and }\; Q_{ji} \;=\; (1-p)^{\ell}\left(\frac{p}{1-p}\right)^{d_{\mathrm{H}}(X_i,X_j)}\;, \qquad (1.15)$$

and obtain for the stationary concentration of the master sequence

$$\bar{x}_m^{(0)} \;=\; \frac{Q_{mm}-\sigma_m^{-1}}{1-\sigma_m^{-1}} \;=\; \frac{1}{\sigma_m - 1}\Big(\sigma_m\,(1-p)^{\ell} \;-\; 1\Big)\,,$$

where $\sigma_m = f_m/\bar{f}_{-m} > 1$ is the *superiority* of the master sequence and $\bar{f}_{-m}$ is defined by

$$\bar{f}_{-m} \;=\; \frac{1}{1-x_m}\sum_{i=1,i\neq m}^{n} x_i\, f_i\;.$$

In this zeroth order approximation the stationary concentration $\bar{x}_m^{(0)}(p)$ vanishes at the critical value

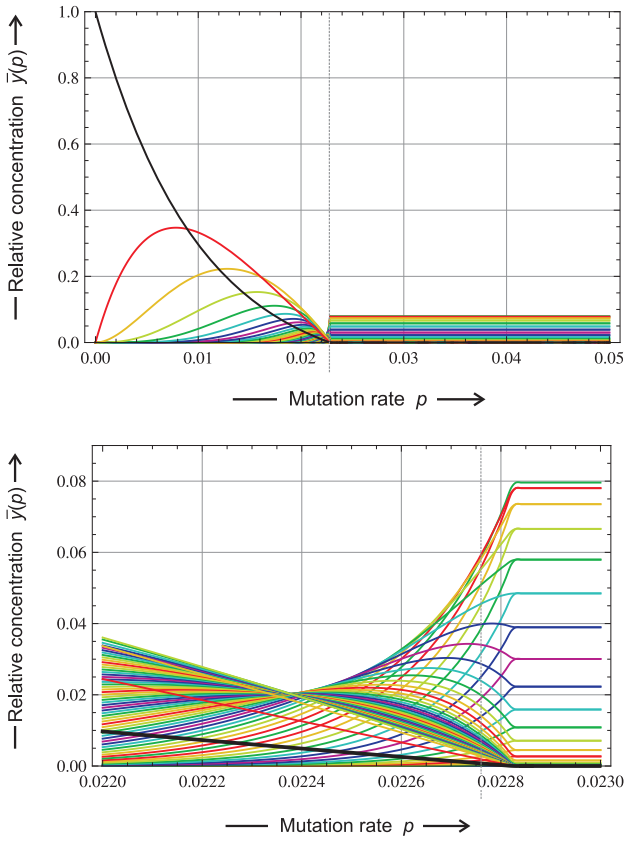


Fig. 1.7. The error threshold on single peak fitness landscapes. The upper part of the figure shows the quasispecies as a function of the mutation rate p. The variables $\bar{y}_k(p)$ $(k = 0, 1, \ldots, \ell)$ represent the total concentrations of all sequences with Hamming distance $d_{\mathrm{H}} = k$: $\bar{y}_0 = \bar{x}_m$ (black) is the concentration of the master sequence, $\bar{y}_1 = \sum_{i=1,d_{\mathrm{H}}(X_i,X_m)=1}^{n} \bar{x}_i$ (red) is the concentration of the one-error class, $\bar{y}_2 = \sum_{i=1,d_{\mathrm{H}}(X_i,X_m)=2}^{n} \bar{x}_i$ (yellow) that of the two-error class and, accordingly, we have $\bar{y}_k = \sum_{i=1,d_{\mathrm{H}}(X_i,X_m)=k}^{n} \bar{x}_i$ for the k-error class. The lower part shows an enlargement. The position of the error threshold computed from the zeroth order approximation (1.16) is shown as by a dotted line (grey). Choice of parameters: $\kappa = 2$, $\ell = 100$, $f_m = 10$, $f_0 = 1$ and hence $\sigma_m = 10$ and $p_{\mathrm{cr}} = 0.02276$.

$$p_{\mathrm{cr}} \approx 1 - (\sigma_m)^{-1/\ell} . \tag{1.16}$$

Needless to say, zero concentration of the master sequence is an artifact of the approximation, because the exact concentration of the master sequence cannot vanish by Perron-Frobenius theorem as long as the population size is nonzero. In order to find out what really happens at the critical mutation rate p_{cr} computer solutions of the complete equation (1.13) were calculated for the single

peak fitness landscape.[14] These calculations [75] show a sharp transition from the ordered quasispecies to the uniform distribution, $\bar{x}_j = \kappa^{-\ell}\,\forall\, j = 1, \ldots, \kappa^{\ell}$. At the critical mutation rate p_{cr} replication errors accumulate and (independently of initial conditions) all sequences are present at the same frequency in the long time limit as is reflected by the uniform distribution. The uniform distribution is the exact solution of the eigenvalue problem at equal probabilities for all nucleotide incorporations (**A**→**A**, **A**→**U**, **A**→**G**, and **A**→**C**) occurring at $\tilde{p} = \kappa^{-1}$. The interesting aspect of the *error threshold* phenomenon consists in the fact that the quasispecies approaches the uniform distribution at a the critical mutation rate p_{cr}, which is far below the random mutation value $\tilde{p}$. As a matter of fact the appearance of an error threshold and its shape depend on details of the fitness landscape [76, pp.51-60]. Some landscapes show no error threshold at all but a smooth transition to the uniform distribution [77]. More realistic fitness landscapes with a distribution of fitness values reveal a much more complex situation: For constant superiority the value of p_{cr} becomes smaller with increasing variance of fitness values. The error threshold phenomenon can be split into three different observations that coincide on the single peak landscape: (i) vanishing of the master sequence x_m, (ii) phase transition like behavior, and (iii) transition to the uniform distribution. On suitable model landscapes the three observations do not coincide and thus can be separated [78, 79].

How do populations behave at mutation rates above the error threshold? In reality a uniform distribution of variants as requested for the stationary state can't be realized. In RNA selection experiments population sizes hardly exceed 10^{15} molecules, the smallest aptamers have chain lengths of $\ell = 27$ nucleotides [80] and this implies $4^{27} \approx 18 \times 10^{15}$ different sequences. Even in this most favorable case we are dealing with more sequences than molecules in the population: a uniform distribution cannot exist. Although the origin of the lack of selective power is completely different – high mutation rates wiping out the differences in fitness values versus fitness differences being zero or too small for selection, the most likely scenarios to occur are migrating populations similar to evolution on a flat landscape [81]. Bernard Derrida and Luca Peliti find that the populations break up into clones, which migrate into different directions in sequence space. Migrating populations are unable to conserve a genotype over generations, and unless a large degree of neutrality allows for maintenance of a phenotype despite changing genotypes, evolution becomes impossible because inheritance breaks down.

Because of high selection pressure resulting from the hosts' defense systems virus population operate at mutations rates as high as possible in order

[14] The single peak fitness landscape is a kind of mean field approximation: A fitness value f_m is assigned to the master sequence, whereas all other variants have the same fitness f_0. For this particular landscape the position $\bar{x}_m^{(0)} = 0$ calculated within the zeroth-order approximation almost coincides with the position of the critical change in the population structure (Fig.1.7).

to allow for fast evolution, and this is just below the error threshold [82]. Increasing the mutation rate should drive the virus population beyond threshold where sufficiently accurate replication is no more possible. Therefore virus populations are doomed to die out at mutation rates above threshold and this suggested a novel antiviral strategy that has led to the development of new drugs [83]. A more recent discussion of the error threshold phenomenon tries to separate the error accumulation phenomenon from mutation caused fitness effects leading to virus extinction called *lethal mutagenesis* [84,85]. As a matter of fact lethal mutagenesis describes the error threshold phenomenon for variable population size N as required for $\lim N \to 0$, but an analysis of population dynamics without and with stochastic effects at the onset of migration of populations is still missing. In addition, more detailed kinetic studies on replication *in vitro* near the error threshold are required before the mechanism of virus extinction at high mutation rates will be understood.

Sequence-structure mappings of nucleic acid molecules (1.4) and proteins provide ample evidence for neutrality in the sense that many genotypes give rise to the same phenotype and identical or almost identical fitness values that cannot be discriminated by natural selection. The possible occurrence of neutral variants has been discussed already by Charles Darwin [2, chapter iv]. Based on the results of the first sequence data from molecular biology Motoo Kimura formulated his neutral theory of evolution [86,87]. In absence of fitness differences between variants random selection occurs because of stochastic enhancement through autocatalytic processes: More frequent variants are more likely to be replicated than less frequent ones. Ultimately a single genotype becomes *fixated* in the population. The average time of replacement for a dominant genotype is the reciprocal mutation rate, $\nu^{-1} = (\ell p)^{-1}$, which, interestingly, is independent of the population size. Are Kimura's results valid also for large population sizes and high mutation rates as they occur, for example, with viruses? Mathematical analysis [88] together with recent computer studies [78] yields the answer: Random selection in the sense of Kimura occurs only for sufficiently distant (master) sequences. In full agreement with the exact result in the limit $p \to 0$ we find that two fittest sequences of Hamming distance $d_{\mathrm{H}} = 1$, two nearest neighbors in sequence space, are selected as a strongly coupled pair with equal frequency of both members. Numerical results demonstrate that this strong coupling occurs not only for small mutation rates but extends over the whole range of p-values from $p = 0$ to the error threshold $p = p_{\mathrm{cr}}$. For clusters of more than two Hamming distance one sequences the frequencies of the individual members of the cluster is determined by the largest eigenvector of the adjacency matrix. Pairs of fittest sequences with Hamming distance $d_{\mathrm{H}} = 2$, i.e. two next nearest neighbors with two sequences in between, are also selected together but the ratio of the two frequencies is different from one. Again coupling extends from zero mutation rates to the error threshold. Strong coupling of fittest sequences manifests itself in virology as systematic deviations from consensus sequences of populations as indeed observed in nature. For two fittest sequences with

$d_{\mathrm{H}} \geq 3$ random selection chooses arbitrarily one of both and eliminates the other one as predicted by the neutral theory.

The function $\phi(t)$ was introduced as the mean fitness of a population in order to allow for straightforward normalization of the population variables. A more general interpretation considers $\phi(t)$ as a flux out of the system. Then the equation describing evolution of the column vector of particle numbers $\boldsymbol{N} = (N_1, \ldots, N_n)$ is of the form [89]

$$\frac{dN_j}{dt} \;=\; F_j(\boldsymbol{N}) \;-\; \frac{N_j}{C(t)}\,\phi(t)\,,\; i = 1, \ldots, n\;,$$

where $F_j(\boldsymbol{N})$ is the function of unconstrained reproduction. An example is provided by equation (1.13): $F_j(\boldsymbol{N}) = \sum_{i=1}^{n} Q_{ji} f_i N_i$. Explicit insertion of the total concentration $C(t) = \sum_{i=1}^{n} N_i(t)$ yields

$$\phi(t) \;=\; \sum_{i=1}^{n} F_i(\boldsymbol{N}) \;-\; \frac{dC}{dt} \quad \text{or} \quad C(t) \;=\; C_0 \;+\; \int_0^t \left(\sum_{i=1}^{n} F_i(\boldsymbol{N}) - \phi(\tau) \right) d\tau\;.$$

Either $C(t)$ or $\phi(t)$ can be chosen freely, the second function is then determined by the equation given above. For normalized variables we find

$$\frac{dx_j}{dt} \;=\; \frac{1}{C(t)} \left(F_j(\boldsymbol{N}) \;-\; x_j \sum_{i=1}^{n} F_j(\boldsymbol{N}) \right)\;.$$

For a large number of examples and for the most cases important in evolution the functions $F_j(\boldsymbol{N})$ are homogeneous functions in N. For homogeneity of degree γ we have $F_j(\boldsymbol{N}) = F_j(C \cdot \boldsymbol{N}) = C^{\gamma} F_j(\mathbf{x})$ and find

$$\frac{dx_j}{dt} \;=\; C^{\gamma-1} \left(F_j(\mathbf{x}) \;-\; x_j \sum_{i=1}^{n} F_j(\mathbf{x}) \right)\;,\; j = 1, \ldots, n\;. \tag{1.17}$$

Two conclusions can be drawn from this equation: (i) For $\gamma = 1$, e.g. the selection equation (1.4) or the replication-mutation equation (1.13), the dependence on the total concentration C vanishes and the solution curves in normalized variables $x_j(t)$ are the same in stationary ($C = \text{const}$) and non-stationary systems as long as $C(t)$ remains finite and does not vanish, and (ii) if $\gamma \neq 1$ the long term behavior determined by $\dot{\mathbf{x}} = 0$ is identical for stationary and non-stationary systems unless the population dies out $C(t) \to 0$ or explodes $C(t) \to \infty$.

1.6 Evolution as a stochastic process

Stochastic phenomena are essential for evolution – each mutant after all starts out from a single copy and a large number of studies have been conducted on stochastic effects in population genetics [90]. Not too much work, however, has been devoted so far to the development of a general stochastic theory of molecular evolution. We mention two examples representative for others [91, 92]. In the latter case the reaction network for replication and mutation was analyzed as a multi-type branching process and it was proven that the stochastic process converges to the deterministic equation (1.13) in the limit of large populations. What is still missing is a comprehensive treatment, for example by means of chemical master equations [93]. Then the deterministic population variables $x_j(t)$ are replaced by stochastic variables $\mathcal{X}_j(t)$ and the corresponding probabilities

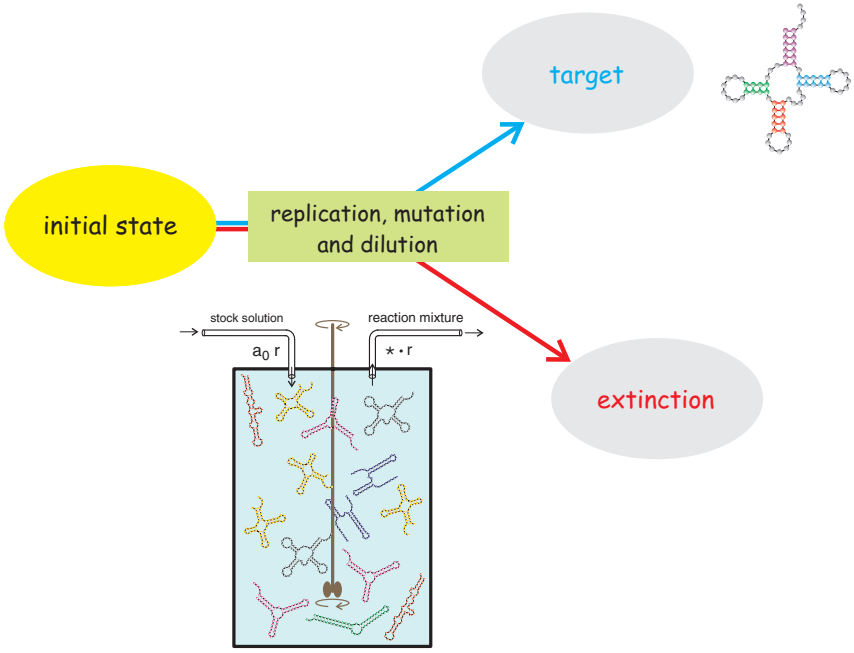
$$P_k^{(j)}(t) \;=\; Prob\{\mathcal{X}_j \;=\; k\}\,,\; k = 0, 1, \ldots, N\,;\; j = 1, \ldots, n\,. \tag{1.18}$$

The chemical master equation translates a mechanism into a set of differential equation for the probabilities. The pendant of equation (1.13), for example, is the master equation

$$\begin{aligned}\frac{dP_k^{(j)}}{dt} \;=\; &\left(\sum_{i=1}^{n} Q_{ji} f_i \sum_{s=1}^{n} s\, P_s^{(i)}\right) P_{k-1}^{(j)} \;-\; \phi(t)\, P_k^{(j)} \;- \\ &- \left(\sum_{i=1}^{n} Q_{ji} f_i \sum_{s=1}^{n} s\, P_s^{(i)}\right) P_k^{(j)} \;+\; \phi(t)\, P_{k+1}^{(j)}\;.\end{aligned} \tag{1.19}$$

The only quantity that has to be specified further in this equation is the flux term $\phi(t)$. For the stochastic description it is not sufficient to have a term that is just compensating the increase in population size due to replication, a detailed model of the process is required. Examples are (i) the Moran process [94–96] with strictly constant population size and (ii) the flow reactor (CSTR) with a population size fluctuating within the limits of a $\sqrt{N}$-law [97, 98].[15] The Moran process assumes that for every newborn molecule one molecule is instantaneously eliminated. Strong coupling of otherwise completely independent processes has the advantage of mathematical simplicity but it is lacking a physical background. The flowreactor, on the other hand, is harder to treat in the mathematical analysis but it is based on solid physical grounds and can be easily implemented experimentally. In computer simulation both models require comparable efforts and for molecular systems preference is given therefore to the flowreactor.

[15] All thermodynamically admissible processes obey a so-called $\sqrt{N}$ law: For a mean population size of N the actual population size fluctuates with a standard deviation proportional to $\sqrt{N}$.



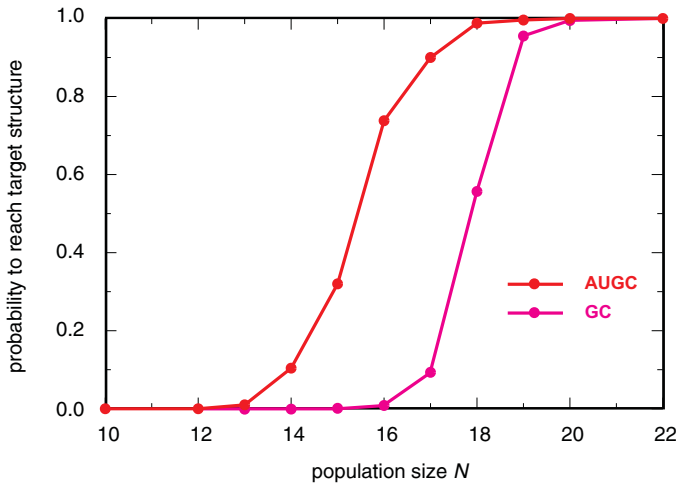


Fig. 1.8. Survival in the flowreactor. Replication and mutation in the flowreactor is implemented according to the mechanism (1.20). The stochastic process has two absorbing states: (i) extinction, $\mathcal{X}_j = 0 \forall\, j = 1, \ldots, n$, and (ii) a predefined target state – here the structure of tRNA^{phe}. A rather sharp transition in the long time behavior of the population is shown in the lower plot: Populations of natural sequences (**AUGC**) switch from almost certain extinction to almost certain survival in the range $13 \leq N \leq 18$ and for binary sequences (**GC**) the transition is even sharper but requires slightly larger population sizes.

For evolution of RNA molecules through replication and mutation in the flowreactor, the following reaction mechanism has been implemented:

$$\begin{aligned} * &\xrightarrow{a_0\, r} A\ , \\ A + X_i &\xrightarrow{Q_{ji} f_i} X_i + X_j\,;\ i,j = 1,\ldots,n\ , \\ A &\xrightarrow{r} \varnothing\ , \quad \text{and} \\ X_j &\xrightarrow{r} \varnothing\,;\ j = 1,\ldots,n\ . \end{aligned} \tag{1.20}$$

Stock solution is flowing into the reactor with a flow rate r and it feeds the reactor with the material required for polynucleotide synthesis – schematically denoted by A and consisting, for example, of activated nucleotides, **ATP**, **UTP**, **GTP** and **CTP** as well as a replicating enzyme – into the system. The concentration of A in the stock solution is denoted by a_0. The molecules X_j are produced by the second reaction either by correct copying or by mutation. The third and the fourth reaction describe the outflux of material and compensate the increase in volume caused by the influx of stock solution. The reactor is assumed to be perfectly mixed at every instant (*continuous stirred-tank reactor* = CSTR). For target search the stochastic process in the reactor is constructed to have two absorbing states (Fig.1.8): (i) extinction – all RNA molecules are diluted out of the reaction vessel, and (ii) survival – the predefined target structure has been produced in the reactor. The population size determines the outcome of the computer experiment: Below population sizes of $N = 13$ the reaction in the CSTR goes almost certainly extinct but it reaches the target with a probability close to one for $N > 20$. The probability of extinction is very small for sufficiently large populations and for population sizes, $N \geq 1000$, as reported here, extinction has been never observed.

In order to simulate the interplay between mutation acting on the RNA sequence and selection operating on RNA structures, the sequence-structure map has to be turned into an integral part of the model [97, 98, 103]. The simulation tool starts from a population of RNA molecules and simulates chemical reactions corresponding to replication and mutation in a CSTR according to (1.20) by using Gillespie's algorithm [99–101]. Molecules replicate in the reactor and produce both correct copies and mutants, the materials to be consumed are supplied by the continuous influx of stock solution into the reactor, and excess volume is removed by means of the outflux of reactor solution. Two kinds of computer experiments were performed: Optimizations of properties on a landscape derived from the sequence-structure map and target searches in shape space where the target is some predefined structure.

Early simulations optimizing replication rates in populations of binary **GC**-sequences yielded two general results:
(i) The progress in evolution is stepwise rather than continuous as short adaptive phases are interrupted by long quasistationary epochs [97, 98].

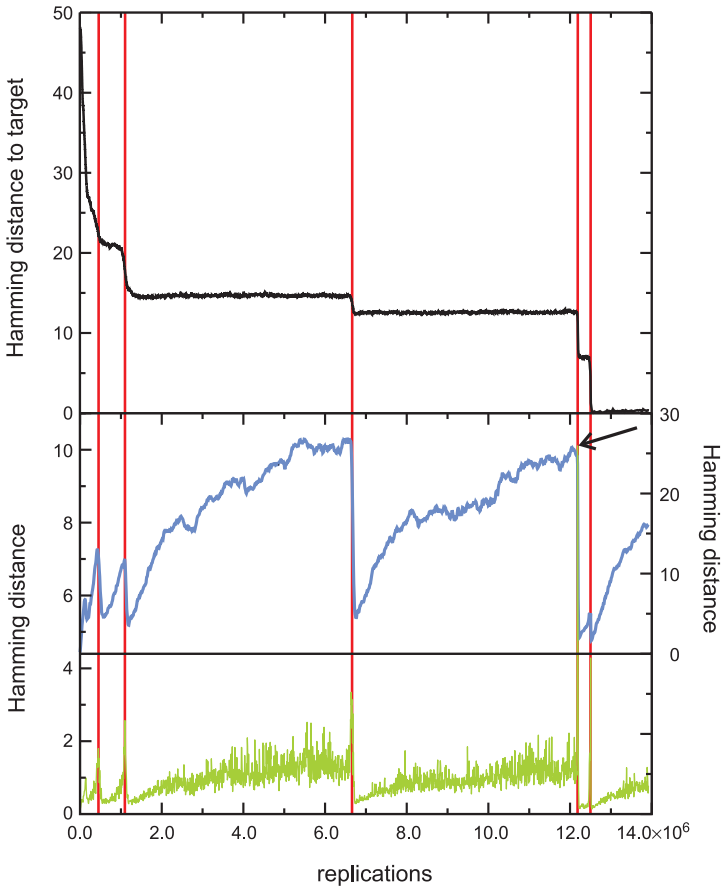


Fig. 1.9. A trajectory of evolutionary optimization. The topmost plot presents the mean distance to the target structure of a population of 1000 molecules. The plot in the middle shows the width of the population in Hamming distance between sequences and the plot at the bottom is a measure of the velocity with which the center of the population migrates through sequence space. Diffusion on neutral networks causes spreading on the population in the sense of neutral evolution [102]). A remarkable synchronization is observed: At the end of each quasistationary plateau a new adaptive phase in the approach towards the target is initiated, which is accompanied by a drastic reduction in the population width and a jump in the population center (The top of the peak at the end of the second long plateau is marked by a black arrow). A mutation rate of $p = 0.001$ was chosen, the replication rate parameter is defined in equation (1.21), and initial and target structures are shown in table 1.1.

Table 1.1. Statistics of the optimization trajectories. The table shows the results of sampled evolutionary trajectories leading from a random initial structure S_I to the structure of tRNA$^{\mathsf{phe}}$, S_T as target.[a] Simulations were performed with an algorithm introduced by Gillespie [99–101]. The time unit is here undefined. A mutation rate of $p = 0.001$ per site and replication was used. The mean and standard deviation were calculated under the assumption if a log-normal distribution that fits well the data of the simulations.

Alphabet	Population size N	Number of runs n_R	Real time from start to target Mean value	σ	Number of replications $[10^7]$ Mean value	σ
AUGC	1000	120	900	+1380 -542	1.2	+3.1 -0.9
	2000	120	530	+880 -330	1.4	+3.6 -1.0
	3000	1199	400	+670 -250	1.6	+4.4 -1.2
	10000	120	190	+230 -100	2.3	+5.3 -1.6
	30000	63	110	+97 -52	3.6	+6.7 -2.3
	100000	18	62	+50 -28	–	–
GC	1000	46	5160	+15700 -3890	–	–
	3000	278	1910	+5180 -1460	7.4	+35.8 -6.1
	10000	40	560	+1620 -420	–	–

[a] The structures S_I and S_T were used in the optimization:

S_I: ((.((((((((((((((.............(((....)))......))))))).)))))))..))...(((......)))

S_T: ((((((...((((........)))).(((((.......)))))).....(((((.......)))))..))))))....

(ii) Different computer runs with identical initial conditions[16] resulted in different structures with similar values for the optimized rate parameters. Despite identical initial conditions, the populations migrated in different – although

[16] *Identical* means here that everything in the computer runs was the same except the seeds for the random number generators and this implies different series of random events.

most orthogonal – directions in sequence space and gave rise to contingency in evolution thereby [98].

In target search problems the replication rate of a sequence X_k, representing its fitness f_k, is chosen to be a function of the Hamming distance[17] between the structure formed by the sequence, $S_k = f(X_k)$ and the target structure S_T,

$$f_k(S_k, S_T) \;=\; \frac{1}{\alpha \,+\, d_{\mathrm{H}}(S_k, S_T)/\ell}\,, \tag{1.21}$$

which increases when S_k approaches the target (α is an empirically adjustable parameter that was commonly chosen to be 0.1). A trajectory is completed when the population reaches a sequence that folds into the target structure – appearance of the target structure in the population is defined as an absorbing state of the stochastic process. A typical trajectory is shown in fig.1.9. In this simulation a homogenous population consisting on N molecules with the same random sequence and structure is chosen as initial condition. The target structure is the well-known secondary structure of phenylalanyl-transfer RNA (tRNA$^{\mathsf{phe}}$). The mean distance to target of the population decreases in steps until the target is reached [103–105] and, again the approach to target is stepwise rather than continuous: Short adaptive phases are interrupted by long quasistationary epochs. In order to reconstruct optimization dynamics, a time ordered series of structures was determined that leads from an initial structure S_I to the target structure S_T. This series, called *relay series*, is a uniquely defined and uninterrupted sequence of shapes. It is retrieved through backtracking, that is in opposite direction from the final structure to the initial shape. The procedure starts by highlighting the final structure and traces it back during its uninterrupted presence in the flow reactor until the time of its first appearance. At this point we search for the parent shape from which it descended by mutation. Now we record time and structure, highlight the parent shape, and repeat the procedure. Recording further backwards yields a series of shapes and times of first appearance which ultimately ends in the initial population.[18] Usage of the relay series and its theoretical background allows for classification of transitions [103, 104, 106]. Inspection of the relay series together with the sequence record on the quasistationary plateaus provides strong hints for the distinction of two scenarios:
(i) The structure is constant and we observe neutral evolution in the sense of Kimura's theory of neutral evolution [87]. In particular, the numbers of neutral mutations accumulated are proportional to the number of replications in the population, and the evolution of the population can be understood as a

[17] The distance between two structures is defined here as the Hamming distance between the two symbolic notations of the structures.

[18] It is important to stress two facts about relay series: (i) The same shape may appear two or more times in a given relay series series. Then, it was extinct between two consecutive appearances. (ii) A relay series is not a genealogy which is the full recording of parent-offspring relations in a time-ordered series of genotypes.

diffusion process on the corresponding neutral network [102].
(ii) The process during the quasistationary epoch involves several closely related structures with identical replication rates and the relay series reveals a kind of random walk in the space of these neutral structures.
The diffusion of the population on the neutral network is illustrated by the plot in the middle of fig.1.9 that shows the width of the population as a function of time [105]. The population width increases during the quasistationary epoch and sharpens almost instantaneously after a sequence had been created by mutation that allows for the start of a new adaptive phase in the optimization process. The scenario at the end of the plateau corresponds to a *bottle neck* of evolution. The lower part of the figure shows a plot of the migration rate or drift of the population center and confirms this interpretation: Migration of the population center is almost always very slow unless the center 'jumps' from one point in sequence space to a possibly distant point where the molecule initiating the new adaptive phase is located. A closer look at the three curves in fig.1.9 reveals coincidence of three events: (i) collapse-like narrowing of the population spread, (ii) jump-like migration of the population center, and (iii) beginning of a new adaptive phase.

It is worth mentioning that the optimization behavior observed in a long-term evolution experiment with *Escherichia coli* [46] can be readily interpreted in terms of random searches on a neutral network. Starting with twelve colonies in 1988, Lenski and his coworkers observed after 31 500 generation or twenty years a great adaptive innovation in one colony [45]: This colony developed a kind of membrane channel that allows for uptake of citrate, which is used as buffer in the medium. The colony thus conquered a new resource that led to a substantial increase in colonial growth. The mutation providing citrate import into the cell is reproducible when earlier isolates of this particular colony are used for a restart of the evolutionary process. Apparently this particular colony has traveled through sequence space to a position from where the adaptive mutation allowing for citrate uptake is within reach. All other eleven colonies did not give rise to mutations with a similar function. The experiment is a nice demonstration of contingency in evolution: The conquest of the citrate resource does not happen through a single highly improbable mutation but by means of a mutation with standard probability from a particular region of sequence space where the population had traveled in one case out of twelve – history matters, or repeating Theodosius Dobzhansky's famous quote: *"Nothing makes sense in biology except in the light of evolution"* [29].

Table 1.1 collects some numerical data sampled from evolutionary trajectories of simulations repeated under identical conditions. Individual trajectories show enormous scatter in the time or the number of replications required to reach the target. The mean values and the standard deviations were obtained from statistics of trajectories under the assumption of log-normal distributions. Despite the scatter three features are unambiguously detectable:
(i) The search in **GC** sequence space takes about five time as long as the corresponding process in **AUGC** sequence space in agreement with the difference

in neutral network structure.
(ii) The time to target decreases with increasing population size.
(iii) The number of replications required to reach target increases with population size.
Combination of the results (ii) and (iii) allows for a clear conclusion concerning time and material requirements of the optimization process: Fast optimization requires large populations whereas economic use of material suggests to work with small population sizes just sufficient to avoid extinction.

A study of parameter dependence of RNA evolution was reported in a recent simulation [107]. Increase in mutation rate leads to an error threshold phenomenon that is closely related to one observed with quasispecies on a single-peak landscape as described above [69, 75]. Evolutionary optimization becomes more efficient[19] with increasing error rate until the error threshold is reached. Further increase in error rates leads to a breakdown of the optimization process. As expected the distribution of replication rates or fitness values f_k in sequence space is highly relevant too: Steep decrease of fitness with the distance to the master structure represented by the target, which has the highest fitness value, leads to sharp threshold behavior as observed on single-peak landscapes, whereas flat landscapes show a broad maximum of optimization efficiency without an indication of threshold-like behavior.

1.7 Concluding remarks

Biology developed differently from physics because it refrained from using mathematics as a tool to analyze and unfold theoretical concepts. Application of mathematics enforces clear definitions and reduction of observations to problems that can be managed. Over the years physics became the science of abstractions and generalizations, biology the science of encyclopedias of special cases with all their beauties and peculiarities. Among others there is one exception of the rule: Charles Darwin presented a grand generalization derived from a wealth of personal and reported observations together with knowledge from economics concerning population dynamics. In the second half of the twentieth century the appearance of molecular biology on the stage changed the situation entirely. A bridge was built from physics and chemistry to biology and mathematical models from biochemical kinetics or population genetics became presentable in biology. Nevertheless, the vast majority of biologists still smiled at the works of theorists. Molecular genetics by the end of the twentieth century created such a wealth of data that almost everybody feels nowadays that progress cannot be made without a comprehensive theoretical foundation and a rich box of suitable computational tools. Nothing like this is at hand but indications for attempts in the right direction

[19] Efficiency of evolutionary optimization is measured by average and best fitness values obtained in populations after a predefined number of generations.

are already visible. Biology is going to enter the grand union of science that started with physics and chemistry and is progressing fast. Molecular biology started out with biomolecules in isolation and deals now with cells, organs, and organisms. Hopefully, this spectacular success will end the so far fruitless reductionism versus holism debate.

Insight into the mechanisms of evolution reduced to the conceivably most simple systems was provided here. These systems deal with evolvable molecules in cell-free assays and are accessible by rigorous mathematical analysis and physical experimentation. An extension to asexual species, in particular viruses and bacteria, is within reach. The molecular approach provides a simple explanation why we have species for this organisms despite the fact that there is neither restricted recombination nor reproductive isolation. The sequence spaces are so large that populations, colonies or clones can migrate for the age of the universe without coming close to another asexual species. We can give an answer to the question of the origin of complexity: Complexity in evolution results primarily from genotype-phenotype relations and from the influences of the environment. Evolutionary dynamics may be complicated in some cases but it is not complex at all. This has been reflected already by the sequence-structure map of our toy example. Conformation spaces depending on the internal folding kinetics as well as on environmental conditions and compatible sets are metaphors for more complex features in evolution proper.

Stochasticity is still an unsolved problem in molecular evolution. The mathematics of stochastic processes encounters difficulties in handling the equations of evolution in detail. A comprehensive stochastic theory is still not at hand and the simulations are lacking more systematic approaches since computer simulations of chemical kinetics of evolution are in an early state too. Another fundamental problem concerns the spatial dimensions: Almost all treatments are assuming spatial homogeneity but we have evidence for the solid particle like structure of the chemical factories of the cell. In the future, any comprehensive theory of the cell will have to deal with these structurally rich supramolecular structures too.

References

1. G. Galilei, *Il Saggiatore*, vol. 6, p.232 (Edition Nationale, Florence, IT, 1896). English Translation from Italian Original
2. C. Darwin, *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life* (John Murray, London, 1859)
3. E. Mayr, *The Growth of Biological Thought. Diversity, Evolution, and Inhertiance* (The Belknap Press of Harvard University Press, 1982)
4. G. Mendel, Verhandlungen des naturforschenden Vereins in Brünn **4**, 3 (1866)
5. G. Mendel, Verhandlungen des naturforschenden Vereins in Brünn **8**, 26 (1870)
6. G.A. Nogler, Genetics **172**, 1 (2006)

7. T.R. Malthus, *An Essay of the Principle of Population as it Affects the Future Improvement of Society* (J. Johnson, London, 1798)
8. R.A. Fisher, *The Genetical Theory of Natural Selection* (Oxford University Press, Oxford, UK, 1930)
9. J.B.S. Haldane, *The Causes of Evolution* (Longmans Green, London, 1932). Reprinted 1990 in the Princeton Science Library by Princeton University Press, Princeton, NJ
10. S. Wright, *Evolution and the Genetics of Populations*, vol. 1-4 (University of Chicago Press, Chicago, IL, 1968, 1969, 1977, 1978)
11. A.J. Lotka, *Elements of Physical Biology* (Williams & Wilkins, Baltimore, MD, 1925)
12. V. Volterra, Mem. R. Accad. Naz. dei Lincei **Ser.VI/2**, 31 (1926)
13. E.P. Odum, *Fundamentals of Ecology* (W. B. Saunders Company, Philadelphia, PA, 1953)
14. R.M. May (ed.), *Theoretical Ecology. Principles and Applications* (Blackwell Scientific Publications, Oxford, UK, 1976)
15. A.M. Turing, Phil. Trans. Roy. Soc. London Ser. B **237**, 37 (1952)
16. A.G. und Hans Meinhardt, Kybernetik **12**, 30 (1972)
17. H. Meinhardt, *Models of Biological Pattern Formation* (Academic Press, London, 1982)
18. J.D. Murray, Scientific American **258**(3), 62 (1988)
19. J.D. Murray, *Mathematical Biology II: Spatial Models and Biomedical Applications*, 3rd edn. (Springer-Verlag, New York, 2003)
20. A.L. Hodgkin, A.F. Huxley, J. Physiology **117**, 500 (1952)
21. J.J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982)
22. H. Judson, *The Eighth Day of Creation. The Makers of the Revolution in Biology* (Jonathan Cape, London, 1979)
23. A. Maxam, W. Gilbert, Proc. Natl. Acad. Sci. USA **74**, 560 (1977)
24. F. Sanger, S. Nicklen, A. Coulson, Proc. Natl. Acad. Sci. USA **74**, 5463 (1977)
25. J. Dale, M. von Schantz, *Form Genes to Genomes: Concepts and Applications of DNA Technology* (Wiley-VCh, New York, 2007)
26. S. Brenner, The Scientist **16**(4), 14 (2002)
27. J.S. Edwards, M. Covert, B.O. Palsson, Environ. Microbiol. **4**, 133 (2002)
28. H.W. Engl, C. Flamm, P. Kügler, J. Lu, S. Müller, P. Schuster, Inverse Problems **25**, 123014 (2009)
29. T. Dobzhansky, F.J. Ayala, G.L. Stebbins, J.W. Valentine, *Evolution* (W. H. Freeman & Co., San Francisco, CA, 1977)
30. L. Euler, *Introductio in Analysin Infinitorum, 1748. English Translation: John Blanton, Introduction to Analysis of the Infinite*, vol. I and II (Springer-Verlag, Berlin, 1988)
31. P. Verhulst, Corresp. Math. Phys. **10**, 113 (1838)
32. D. Zwillinger, *Handbook of Differential Equations*, 3rd edn. (Academic Press, San Diego, CA, 1998)
33. M. Eigen, Naturwissenschaften **58**, 465 (1971)
34. J. Maynard-Smith, Nature **225**, 563 (1970)
35. R.W. Hamming, *Coding and Information Theory*, 2nd edn. (Prentice-Hall, Englewood Cliffs, NJ, 1986)
36. J. Rogers, G. Joyce, Nature **402**, 323 (1999)
37. J.S. Reader, G.F. Joyce, Nature **420**, 841 (2002)

38. P. Schuster, Physica D **107**, 351 (1997)
39. P. Gitchoff, G.P. Wagner, Complexity **2**(1), 37 (1998)
40. P.F. Stadler, R. Seitz, G.P. Wagner, Bull. Math. Biol. **62**, 399 (2000)
41. B.R.M. Stadler, P.F. Stadler, M. Shpak, G.P. Wagner, Z. Phys. Chem. **216**, 217 (2002)
42. D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA, 1989)
43. L.M. Schmitt, Theoret. Computer Science **259**, 1 (2001)
44. J.E. Barrick, D.S. Yu, H. Jeong, T.K. Oh, D. Schneider, R.E. Lenski, J.F. Kim, Nature **441**, 1243 (2009)
45. Z.D. Blount, C. Z., R.E. Lenski, Proc. Natl. Acad. Sci. USA **105**, 7898 (2008)
46. R.E. Lenski, M.R. Rose, S.C. Simpson, S.C. Tadler, The American Naturalist **38**, 1315 (1991)
47. P. Schuster, Reports on Progress in Physics **69**, 1419 (2006)
48. M. Zuker, P. Stiegler, Nucleic Acids Research **9**, 133 (1981)
49. I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster, Mh. Chemie **125**, 167 (1994)
50. M. Zuker, Science **244**, 48 (1989)
51. S. Wuchty, W. Fontana, I.L. Hofacker, P. Schuster, Biopolymers **49**, 145 (1999)
52. C. Flamm, W. Fontana, I.L. Hofacker, P. Schuster, RNA **6**, 325 (1999)
53. M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, P.F. Stadler, J. Phys. A: Math. Gen. **37**, 4731 (2004)
54. D.R. Mills, R.L. Peterson, S. Spiegelman, Proc. Natl. Acad. Sci. USA **58**, 217 (1967)
55. S. Spiegelman, Quart. Rev. Biophys. **4**, 213 (1971)
56. G.F. Joyce, Angew. Chem. Internat. Ed. **46**, 6420 (2007)
57. C.K. Biebricher, M. Eigen, J. William C. Gardiner, Biochemistry **22**, 2544 (1983)
58. C.K. Biebricher, M. Eigen, J. William C. Gardiner, Biochemistry **23**, 3186 (1984)
59. C.K. Biebricher, M. Eigen, J. William C. Gardiner, Biochemistry **24**, 6550 (1985)
60. S. Klussmann (ed.), *The Aptamer Handbook. Functional Oligonucleotides and Their Applications* (Wiley-VCh Verlag, Weinheim, DE, 2006)
61. A. Lescoute, N.B. Leontis, C. Massire, E. Westhof, Nucleic Acids Res. **33**, 2395 (2005)
62. N.B. Leontis, A. Lescoute, E. Westhof, Curr. Opinion Struct. Biol. **16**, 279 (2006)
63. I.L. Hofacker, P. Schuster, P.F. Stadler, Discr. Appl. Math. **89**, 177 (1998)
64. J.S. McCaskill, Biopolymers **29**, 1105 (1990)
65. M. Mandal, B. Boese, J.E. Barrick, W.C. Winkler, R.R. Breaker, Cell **113**, 577 (2003)
66. C.J. Thompson, J.L. McBride, Math. Biosci. **21**, 127 (1974)
67. B.L. Jones, R.H. Enns, S.S. Rangnekar, Bull. Math. Biol. **38**, 15 (1976)
68. M. Eigen, P. Schuster, Naturwissenschaften **64**, 541 (1977)
69. M. Eigen, J. McCaskill, P. Schuster, Adv. Chem. Phys. **75**, 149 (1989)
70. E. Seneta, *Non-negative Matrices and Markov Chains*, 2nd edn. (Springer-Verlag, New York, 1981)
71. N. Rohde, H. Daum, C.K. Biebricher, J. Mol. Biol. **249**, 754 (1995)

72. E. Domingo, D. Szabo, T. Taniguchi, C. Weissmann, Cell **13**, 735 (1978)
73. E. Domingo, J. Holland, in *RNA Genetics. Vol.III: Variability of Virus Genomes*, vol. III, ed. by E. Domingo, J. Holland, P. Ahlquist (CRC Press, Boca Raton, FL, 1988), pp. 3–36
74. E. Domingo (ed.), *Quasispecies: Concepts and Implications for Virology* (Springer-Verlag, Berlin, 2006)
75. J. Swetina, P. Schuster, Biophys. Chem. **16**, 329 (1982)
76. P.E. Phillipson, P. Schuster, *Modeling by Nonlinear Differential Equations. Dissipative and Conservative Processes*, *World Scientific Series on Nonlinear Science A*, vol. 69 (World Scientific, Singapore, 2009)
77. T. Wiehe, Genet. Res. Camb. **69**, 127 (1997)
78. P. Schuster, Theory in Biosciences **129** (2010). Submitted
79. P. Schuster. Quasispecies and error thresholds on realistic fitness landscapes (2010). Preprint
80. L. Jiang, A.K. Suri, R. Fiala, D.J. Patel, Chemistry & Biology **4**, 35 (1997)
81. B. Derrida, L. Peliti, Bull. Math. Biol. **53**, 355 (1991)
82. J.W. Drake, Proc. Natl. Acad. Sci. USA **90**, 4171 (1993)
83. E. Domingo, ed., Virus Research **107**(2), 115 (2005)
84. J.J. Bull, L. Ancel Myers, M. Lachmann, PLoS Comp. Biol. **1**, 450 (2005)
85. J.J. Bull, R. Sanjuán, C.O. Wilke, J. Virology **81**, 2930 (2007)
86. M. Kimura, Nature **217**, 624 (1968)
87. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, UK, 1983)
88. P. Schuster, J. Swetina, Bull. Math. Biol. **50**, 635 (1988)
89. M. Eigen, P. Schuster, Naturwissenschaften **65**, 7 (1978)
90. R.A. Blythe, A. McKane, J. Stat. Mech.: Theor. Exp. p. P07018 (2007). DOI 10.1088/1742-5468/2007/07/P07018
91. B.L. Jones, H.K. Leung, Bull. Math. Biol. **43**, 665 (1981)
92. L. Demetrius, P. Schuster, K. Sigmund, Bull. Math. Biol. **47**, 239 (1985)
93. C.W. Gardiner, *Stochastic Methods. A Handbook for the Natural and Social Sciences*, 4th edn. Springer Series in Synergetics (Springer-Verlag, Berlin, 2009)
94. P. Moran, Proc. Cambridge Philos. Soc. **54**, 60 (1958)
95. P. Moran, *The Statistical Processes of Evolutionary Theory* (Clarendon Press, Oxford, UK, 1962)
96. M.A. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life* (The Belknap Press of Harvard University Press, Cambridge, MA, 2006)
97. W. Fontana, P. Schuster, Biophys. Chem. **26**, 123 (1987)
98. W. Fontana, W. Schnabl, P. Schuster, Phys. Rev. A **40**, 3301 (1989)
99. D.T. Gillespie, J. Comp. Phys. **22**, 403 (1976)
100. D.T. Gillespie, J. Phys. Chem. **81**, 2340 (1977)
101. D.T. Gillespie, Annu. Rev. Phys. Chem. **58**, 35 (2007)
102. M.A. Huynen, P.F. Stadler, W. Fontana, Proc. Natl. Acad. Sci. USA **93**, 397 (1996)
103. W. Fontana, P. Schuster, Science **280**, 1451 (1998)
104. W. Fontana, P. Schuster, J.Theor.Biol. **194**, 491 (1998)
105. P. Schuster, in *Evolutionary Dynamics – Exploring the Interplay of Accident, Selection, Neutrality, and Function*, ed. by J.P. Crutchfield, P. Schuster (Oxford University Press, New York, 2003), pp. 163–215
106. B.R.M. Stadler, P.F. Stadler, G.P. Wagner, W. Fontana, J. Theor. Biol. **213**, 241 (2001)
107. A. Kupczok, P. Dittrich, J. Theor. Biol. **238**, 726 (2006)

Index