Peter Schuster

Evolution and Quasispecies

A collection of three recent articles

Affiliations: Universität Wien, Institut für Theoretische Chemie,
Währingerstraße 17, 1090 Wien, Austria

The Santa Fe Institute, 1399 Hyde Park Road,
Santa Fe, NM 87501, USA

Three recent contributions:

[1] P. Schuster. Mathematical modeling of evolution. Solved and open problems. *Theory Biosci.*, 130:71–89, 2011.

[2] E. Domingo and P. Schuster. What is a quasispecies? Historical origins and current scope. In E. Domingo and P. Schuster, editors, *Quasispecies: From Theory to Experimental Systems*, volume 392 of *Current topics in Microbiology and Immunology*, chapter 1, pages 1–22. Springer-Verlag, Berlin, 2016.

[3] P. Schuster. Quasispecies on fitness landscapes. In E. Domingo and P. Schuster, editors, *Quasispecies: From Theory to Experimental Systems*, volume 392 of *Current Topics in Microbiology and Immunology*, chapter 4, pages 61–120. Springer-Verlag, Berlin, 2016.

Theory Biosci. (2011) 130:71–89
DOI 10.1007/s12064-010-0110-z

ORIGINAL PAPER

Mathematical modeling of evolution. Solved and open problems

Peter Schuster

Received: 8 December 2009 / Accepted: 4 July 2010 / Published online: 1 September 2010

Abstract Evolution is a highly complex multilevel process and mathematical modeling of evolutionary phenomenon requires proper abstraction and radical reduction to essential features. Examples are natural selection, Mendel's laws of inheritance, optimization by mutation and selection, and neutral evolution. An attempt is made to describe the roots of evolutionary theory in mathematical terms. Evolution can be studied in vitro outside cells with polynucleotide molecules. Replication and mutation are visualized as chemical reactions that can be resolved, analyzed, and modeled at the molecular level, and straightforward extension eventually results in a theory of evolution based upon biochemical kinetics. Error propagation in replication commonly results in an error threshold that provides an upper bound for mutation rates. Appearance and sharpness of the error threshold depend on the *fitness landscape*, being the distribution of fitness values in genotype or *sequence space*. In molecular terms, fitness landscapes are the results of two consecutive mappings from sequences into structures and from structures into the (nonnegative) real numbers. Some properties of genotype–phenotype maps are illustrated well by means of sequence–structure relations of RNA molecules. Neutrality in the sense that many RNA sequences form the same (coarse grained) structure is one of these properties, and characteristic for such mappings. Evolution cannot be fully understood without considering fluctuations—each mutant originates form a single copy, after all. The existence of neutral sets of genotypes called *neutral networks*, in particular, necessitates stochastic modeling, which is introduced here by simulation of molecular evolution in a kind of flowreactor.

Keywords Error threshold · Molecular evolution · Natural selection · Neutral evolution · Quasi-species · RNA structures

P. Schuster (✉)
Institut für Theoretische Chemmie, Universität Wien,
Währingerstraße 17, 1090 Wien, Austria
e-mail: pks@tbi.univie.ac.at

Introduction

Although most of individual ideas concerning biological evolution were raised already in the eighteenth century and earlier, the concept of population-level evolution based on variation and natural selection is due to the great naturalist Charles Darwin who derived it from a wealth of observations. Almost at the same time, Greogor Mendel uncovered the laws of inheritance by performing carefully designed breeding experiments with plants and statistical evaluation of the results. About 60 years later the path-breaking discoveries of both scholars were united by the work of the famous mathematician and population geneticists Ronald Fisher: Early population genetics describes the interplay of genetics and selection by means of differential equations. Modeling in population genetics has been an enormous abstraction since differential equations can encapsulate only certain features of population dynamics. Stochasticity, for example, is missing and mutation, the driving force of innovation is not part of the model but operates rather like a *deus ex machina* injecting new genotypes into the system. Deviations from Mendel's laws were detected and described by quantitative phenomenology of genetic recombination but no satisfactory mechanistic explanation was available.

Molecular biology originating from the determination of biopolymer structures (Judson 1979) provided a new and solid foundation of biology rooted in physics and chemistry. Reproduction could be reduced to replication of nucleic acid molecules, recombination and mutation fell out as biochemical reactions just as correct copying of molecules. Since molecules replicate readily in proper assays outside cells, evolution can be studied in cell-free system allowing for analysis by the full repertoire of methods from physics and chemistry: modeling evolution in vitro became a case study in chemical kinetics. The development of novel and highly efficient sequencing techniques for DNA (Maxam and Gilbert 1977; Sanger et al. 1977) changed molecular genetics entirely. The whole cell or the complete organism rather than individual biomolecules became the object of investigations and new disciplines, now aiming at a true exploration of the chemistry of life, originated. Genomics, for example, determines the genetic information of organisms through DNA sequencing, proteomics explores the full set of cellular proteins and their interactions, metabolomics is dealing with cellular metabolism as a gigantic network of biochemical reactions, functional genomics and systems biology, eventually, head for describing all functions of biomolecules and modeling the dynamics of whole cells. Needless to say, present day molecular biology is not yet there, but new experimental and computational techniques are making fast progress and this highly ambitious goal is not completely out of reach.

This review starts out from an attempt to implement evolutionary thinking from Darwin and Mendel to Fisher in mathematical language ("Darwinian selection in mathematical language" section). Then, we focus on evolution in simple systems seen from a molecular perspective. In particular, the focus is laid on the interplay of mutation and selection ("Mutation driven evolution of molecules" section), and we shall make an attempt to include phenotypic properties in the model of evolution. The role of stochasticity in evolution of molecules, in particular neutrality with respect to selection, is investigated by means of computer simulation ("Modeling evolution shape in silico" section). The contribution is finished by "Concluding remarks" section).

Darwinian selection in mathematical language

In Charles Darwin's centennial work on the *Origin of Species* (Darwin 1859), we do not find a single mathematical equation. Accordingly, we can only speculate how Darwin might have formulated his theory of natural selection in case he had used mathematical language. Charles Darwin according to his own records had read Robert Malthus' (1798) *Essay on the Principle of Population* and was deeply impressed by the effects of population increase in the form of a geometric progression or exponential growth. Animal or human populations—according to Malthus—grow exponentially like every system capable of reproduction and the increase in the production of nutrition is at best linear as expressed by an arithmetic progression when we assume that the gain in land exploitable for agriculture is constant in time, i.e., the increase in the area of fields is the same every year. An inevitable result of the Malthusian vision of the world is the pessimistic view that populations will grow until the majority of individuals will die premature of malnutrition and hunger. Charles Darwin and also his younger contemporary Alfred Russell Wallace took from Malthus' population theory that in the wild, where birth control does not exist and individuals fight for food, the major fraction of progeny will die before they reach the age of reproduction and only the strongest will have a chance to multiply. Natural selection, ultimately, appears as a result of exponential growth and finite carrying capacity of ecosystems. Presumably not known to Darwin the Belgium mathematician Jean François Verhulst complemented the theory of exponential growth by the introduction of finite resources (Verhulst 1838). The Verhulst or logistic equation is of the form

$$\frac{\mathrm{d}N}{\mathrm{d}t} = rN\left(1 - \frac{N}{K}\right), \tag{1}$$

where N denotes the number of individuals of species X. It can be solved exactly,

$$N(t) = N(0)\frac{K}{N(0) + (K - N(0))\exp(-rt)}. \tag{2}$$

Apart from the initial number of individuals of X, $N(0)$, the Verhulst equation has two parameters: (i) the Malthusian parameter or the growth rate r and (ii) the carrying capacity K of the ecological niche or the ecosystem. A population of size $N(0)$ grows exponentially at short times: $N(t) \approx N(0)\exp(rt)$ for $N(0) \ll K$ at t sufficiently small. As shown in Fig. 1, the population size approaches the carrying capacity asymptotically for long times: $\lim_{t\to\infty} N(t) = K$.

The two parameters are taken as criteria to distinguish different evolutionary strategies: Species that are r-selected exploit ecological niches with low density, produce a large number of offspring each of which has a low probability to survive to adulthood, whereas K-selected species are strongly competing in crowded niches and invest heavily in few offspring that have a high probability to survive to adulthood. The two cases, r- and K-selection, are the extreme situations of a continuum of mixed selection strategies. In the real world, the r-selection strategy is an appropriate adaptation to fast changing environments, whereas K-selection pays in slowly varying or constant environments.

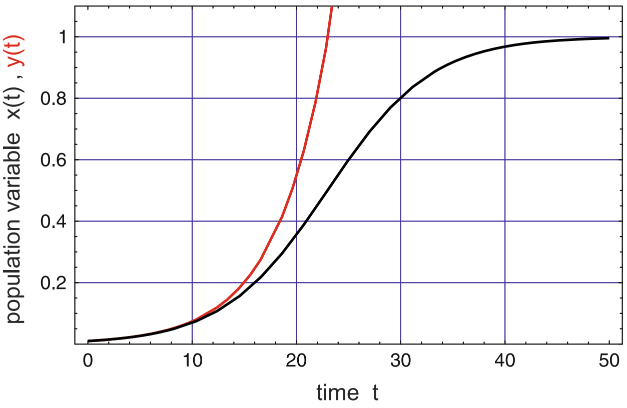


Fig. 1 Exponential growth and finite carrying capacity. The figure compares solution curves for the Verhulst equation (2) and exponential growth, $y(t) = y(0)\,\exp(rt)$ (*black* and *red curve*, respectively). The variable of the Verhulst equation is normalized: $x(t) = N(t)/K$ with $\lim_{t\to\infty} x(t) = 1$ and $0 < x(t) \le 1$ for $0 < x(0) \le 1$. Choice of parameters: $r = 0.2$, $K = 1$, and $x(0) = y(0) = 0.02$. Unrestricted exponential growth outgrows any constant or linearly growing resource. In the print version red appears as gray

The Verhulst equation can be used to derive a selection equation in the spirit of Darwin's theory. The single species X is replaced by several species or variants forming a population: $\Pi = \{X_1, X_2, \ldots, X_n\}$, the numbers of individuals are represented by a vector $\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))$ with $\sum_{i=1}^{n} N_i(t) = C(t)$. The carrying capacity is defined for all n species together: $\lim_{t\to\infty} \sum_{i=1}^{n} N_i(t) = K$. The Malthus parameters or fitness values are denoted by $f_1, f_2, \ldots, f_n$, respectively. The differential equations for individual species are now of the form

$$\frac{dN_j}{dt} = N_j\left(f_j - \frac{C}{K}\phi(t)\right) \quad \text{with } \phi(t) = \frac{1}{C}\sum_{i=1}^{n} f_i N_i(t) \tag{3}$$

being the mean fitness of the population. Summation of over all species yields a differential equation for the total population size

$$\frac{dC}{dt} = C\left(1 - \frac{C}{K}\right)\phi(t), \tag{4}$$

that can be solved analytically

$$C(t) = C(0)\frac{K}{C(0) + (K - C(0))e^{-\Phi}} \quad \text{with } \Phi = \int_0^t \phi(\tau)d\tau,$$

where $C(0)$ is the population size at time $t = 0$. The function $\Phi(t)$ depends on the distribution of fitness values within the population and its time course. For $f_1 = f_2 = \cdots = f_n = r$ the integral yields $\Phi = rt$, and we retain Eq. 2. In the long time limit, Φ grows to infinity and $C(t)$ converges to the carrying capacity K.

As an exercise, we perform stability analysis: From $dC/dt = 0$ follow two stationary states of Eq. 4: (i) $\bar{C} = 0$ and (ii) $\bar{C} = K$.[1] For conventional stability analysis, we calculate the (1×1) Jacobian and obtain for the eigenvalue

$$\lambda = \frac{\partial(dC/dt)}{\partial C} = \phi(t) - \frac{C}{K}\left(2\phi(t) - K\frac{\partial\phi}{\partial C}\right) - \frac{C^2\partial\phi}{K\,\partial C}.$$

Insertion of the stationary values yields $\lambda^{(i)} = +\phi > 0$ and $\lambda^{(ii)} = -\phi < 0$, state (i) is unstable and state (ii) is asymptotically stable. The total population size converges to the value of the carrying capacity, $\lim_{t\to\infty} C(t) = K$ as, of course, derived already from the exact solution.

The primary issue in the multi-species case is to describe the time course of the distribution of species within the population. For this goal, we introduce normalized variables: $x_j(t) = N_j(t)/C(t)$ with $\sum_{i=1}^{n} x_i(t) = 1$. The ODE in normalized variables,

$$\frac{dx_j}{dt} = x_j(f_j - \phi(t)), \quad j = 1, 2, \ldots, n \quad \text{with}$$
$$\phi(t) = \frac{1}{C}\sum_{i=1}^{n} f_i N_i = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} x_i} = \sum_{i=1}^{n} f_i x_i, \tag{5}$$

Equation 5 can be solved exactly by means of integrating factor transformation (Zwillinger 1998, p. 322ff):

$$x_i(t) = z_i(t) \cdot \exp\left(-\int_0^t \phi(\tau)d\tau\right),$$

which after insertion into (3) and solution for $z_j(t)$ yields

$$x_j(t) = \frac{x_j(0) \cdot \exp(f_j t)}{\sum_{i=1}^{n} x_i(0) \cdot \exp(f_i t)}. \tag{6}$$

Two properties of the selection process that are relevant for evolution follow straightforwardly (Fig. 2): (i) The mean fitness, $\phi(t)$ is a non-decreasing function of time, and (ii) a population variable $x_j(t)$ increases if and only if the differential fitness of the corresponding species is positive, $\delta\phi_j(t) = f_j - \phi(t) > 0$, and decreases if and only if the differential fitness is negative, $\delta\phi_j(t) = f_j - \phi(t) < 0$.

First, we present a proof for the first statement (non-decreasing ϕ): The time dependence of the mean fitness or flux ϕ is given by

[1] There is also a third stationary state defined by $\phi = 0$. For strictly positive fitness values, $f_i > 0 \;\forall\; i = 1, 2, \ldots, n$, this condition can only be fulfilled by $x_i = 0 \;\forall\; i = 1, 2, \ldots, n$, which is identical to state (i). If some f_i values are zero—corresponding to lethal variants—the respective variables vanish in the infinite time limit because of $dx_i/dt = -\phi(t)\,x_i$ with $\phi(t) > 0$.

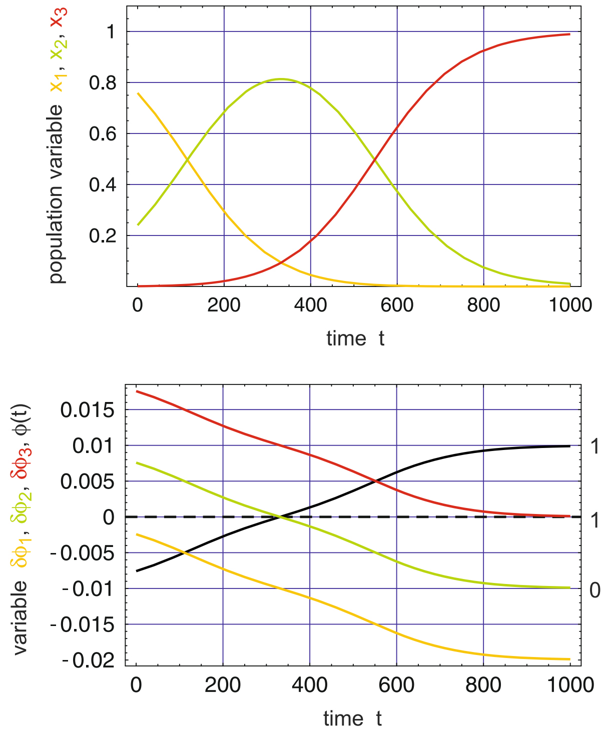


Fig. 2 Differential fitness and selection. In the *upper part* of the figure, we show the time development of a population at constant population size, $C = K = 1$. The three species differ in initial presence and fitness values: X_1, $x_1(0) = 0.759$, $f_1 = 0.99$ (*yellow*); X_2, $x_2(0) = 0.240$, $f_2 = 1.00$ (*green*); X_3, $x_3(0) = 0.001$, $f_3 = 1.01$ (*red*). The *lower part* of the figure compares differential fitness of individual species, $\delta\phi_j; j = 1, 2, 3$ (*yellow*, *green*, and *red*; left ordinate scale), and the mean fitness of the population, $\phi(t)$ (*black*; right ordinate scale). The population variable of a species increases if the differential fitness is positive and decreases for negative differential fitness (as follows directly from a comparison of the two plots). The mean fitness is a non-decreasing function of time. In the print version red appears as the darkest gray and other colors differ in their gray-value

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \sum_{i=1}^{n} f_i \dot{x}_i = \sum_{i=1}^{n} f_i \left(f_i x_i - x_i \sum_{j=1}^{n} f_j x_j \right)$$
$$= \sum_{i=1}^{n} f_i^2 x_i - \sum_{i=1}^{n} f_i x_i \sum_{j=1}^{n} f_j x_j = \overline{f^2} - \left(\bar{f}\right)^2 = \mathrm{var}\{f\} \geq 0. \qquad (7)$$

Since a variance is always nonnegative, Eq. 7 implies that $\phi(t)$ is a non-decreasing function of time, and hence it is optimized during selection. □

The second statement (differential fitness $\delta\phi_j$) is trivial but provides insight into the selection mechanism. At $t = 0$ all population variables with a fitness below average, i.e., with a negative differential fitness, $\delta\phi < 0$, will decrease, all variables with $\delta\phi > 0$ will increase. The result is an increase in $\phi(t)$ in agreement with Eq. 7. As time progresses and $\phi(t)$ increases, more and more species fall under the $\delta\phi < 0$-criterion, will decrease and finally disappear. Ultimately, only the species with the largest fitness value, $X_m : f_m = \max\{f_1, f_2, \ldots, f_n\}$, will remain and the mean fitness has reached its maximal value: $\phi(t) = f_m$. Selection of the fittest has occurred!

The Augustinian monk Gregor Mendel was a contemporary of Charles Darwin and had the missing piece of Darwin's theory, a mechanism of inheritance (Mendel 1866, 1870) in hand, but his works were ignored by evolutionary biologists until the turn of the century. The English statistician and geneticist Ronald Fisher succeeded in uniting natural selection with Mendelian genetics (Fisher 1930). His selection equation describes the evolution of the distribution of alleles at a single gene locus:

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \sum_{i=1}^{n} a_{ji} x_i x_j - x_j \phi(t) = x_j \left(\sum_{i=1}^{n} a_{ji} x_i - \phi(t) \right)$$
$$= x_j \left(\bar{f}_j - \phi(t) \right), \quad \text{with } \bar{f}_j = \sum_{i=1}^{n} a_{ji} f_i, j = 1, 2, \ldots, n$$
$$\text{and } \phi(t) = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ji} x_i x_j. \qquad (8)$$

The variables denote the frequencies of the alleles in the population $x_j = [X_j]$, normalization yields $\sum_{j=1}^{n} x_j = 1$, and a_{ij} is the fitness of the (diploid) genotype $X_i X_j$. A diploid organisms carries two alleles of each gene on a autosome[2]—one being transferred from the father and one coming from the mother—and the contribution to the change of the frequency of allele X_j in time is proportional to the fitness of the genotype, a_{ij}, and the frequencies of the two alleles, x_i and x_j. In conventional genetics the properties of a phenotype are assumed to be independent of the origin of alleles—it does not matter whether the alleles comes form the father or from the mother—and therefore, we have $a_{ji} = a_{ij}$ (Fig. 3): The matrix of fitness values $A = \{a_{ij}\}$ is symmetric. In this case, it is straightforward to prove that $\phi(t)$, the mean fitness of the alleles is a non-decreasing function of time as shown for the simple selection case analyzed in Eq. 7.

In contrast to the simple selection case (3), Fisher's selection equation may have several asymptotically stable stationary states and therefore the outcome of selection depends on initial conditions. A straightforward example is provided by higher fitness of the homozygote genotypes compared to the heterozygote: the states corresponding to the homozygotes X_1X_1 and X_2X_2 ($x_1 = 1$, $x_2 = 0$ and $x_1 = 0$, $x_2 = 1$, respectively) are asymptotically stable whereas the heterozygous states X_1X_2 and X_2X_1 ($x_1 = x_2 = 0.5$) is unstable.[3] Unfortunately—but fortunately for

[2] All chromosomes are autosomes except the sexual chromosomes X and Y.

[3] In case matrix A is not symmetric, the dynamical system (10) may show more complex dynamics like oscillations, deterministic chaos, etc.

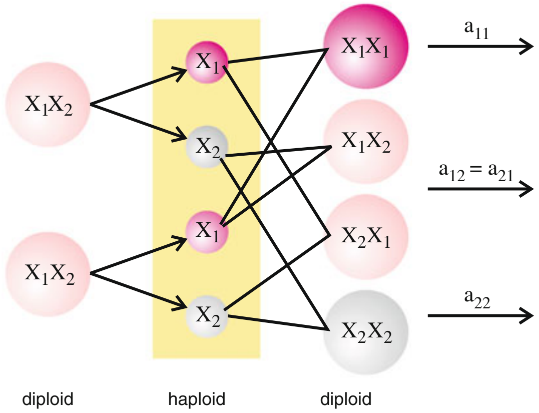


Fig. 3 Mendelian genetics and sexual reproduction. In sexual reproduction the two copies of a gene X in a diploid organism are separated to yield two haploid gametes. In the offspring the two genes in the two gametes are combined at random resulting in recombination. The figure sketches the progeny of two heterozygous organisms—these are organisms carrying two different alleles of the gene. The fitness of the diploid phenotype, X_i X_j, is denoted by a_{ij}

population geneticists and theoretical biologists because it provided and provides a whole plethora of problems to solve—Fisher's selection equation holds only for independent genes. Two and more locus models with gene interaction turned out to be much more complicated and no generally valid optimization principle has been found so far: Natural selection in the sense of Charles Darwin is an extremely powerful optimization heuristic but no theorem. Nevertheless, Fisher's fundamental theorem is much deeper than the toy version that has been presented here. The interested reader is referred to a few, more or less arbitrarily chosen references from the enormous literature on this issue (Price 1972; Edwards 1994; Okasha 2008).

Mutation driven evolution of molecules

Molecular biology was born when Watson and Crick published their centennial paper on the structure of DNA. Further development provided information on the chemistry of life at a breathtaking pace (Judson 1979). A closer look on the structure of DNA revealed the discrete nature of base pairing—two nucleotides make a base pair that fits into the double helix or they do not. With this restriction, the natural nucleobases allow only for four combinations: AU, UA, GC, and CG. This fact is sufficient for an understanding of the molecular basis of genetics: genetic information is of digital nature and multiplication of information is tantamount to copying. Mutation, the process that leads to innovation in evolution, was disclosed as imperfect reproduction or an error in the copying process. Correct reproduction and mutation at the molecular level are seen as parallel chemical reactions (figure 4). In order to guarantee inheritance, correct copying must occur more frequently than mutation (as indicated in the figure caption). In "The kinetic model of replication and mutation" section, we shall cast this intuitive statement into a quantitative expression, whereby for the sake of simplicity only point mutations will be considered. This, however, should not mean that other changes in genomes like insertions, deletions, duplications, and other genome rearrangements are unimportant.

The more general a model is, the wider is its range of applicability. The enormous success of Darwin's natural selection is its almost universal applicability and this results from the lack of specific assumptions on the process of multiplication and variation. On the other hand, specificity is required for working out mathematical models, which can provide explanation for observations and which are suitable for experimental test. DNA replication is an extremely complicated process involving some twenty proteins,[4] and has not yet been studied thoroughly by biochemical kinetics. Compared to DNA replication, replication of RNA viruses, in particular bacteriophages, is rather simple in the sense that it usually requires only a single enzyme. Since the mechanism of replication has been resolved down to molecular details in few systems only, we describe here replication by the specific replicase from the bacteriophage Qβ (Biebricher and Eigen 1988; Eigen and Biebricher 1988; Nakaishi et al 2002; Hosoda et al 2007) as an illustration of complete bottom-up understanding of evolution in vitro.

Virus-specific RNA replication

Qβ-replicase is a virus-specific, RNA-dependent RNA polymerase and amplifies suitable RNA molecules in a medium containing the activated nucleotides, ATP, UTP, GTP, and CTP, in excess (Mills et al 1967). in early experiments Qβ-replicase was isolated from *Escherichia coli* bacteria infected by Qβ bacteriophage, at present production of the enzyme makes use of genetic engineering.[5] Replication of Qβ-RNA is initiated by a single strand RNA molecule that binds to the enzyme Qβ-replicase at sequence specific recognition sites (Brown and Gold 1996; Küppers and Sumper 1975). Through enzyme action, the

[4] Protein synthesis in vivo is regulated by a complex network controlling gene activity called *gene expression*. The network involves regulation of transcription (DNA → RNA), post-transcriptional modification and maturation of the messenger-RNA, its translation into protein, and post-translational modification before the protein unfolds its function.

[5] Qβ-replicase is an enzyme consisting of four subunits. Three subunits are host proteins involved in translation, the ribosomal protein S_1 and the elongation factors Ef-Tu and Ef-Ts. The fourth subunit is a virus-specific protein encoded by the viral RNA.

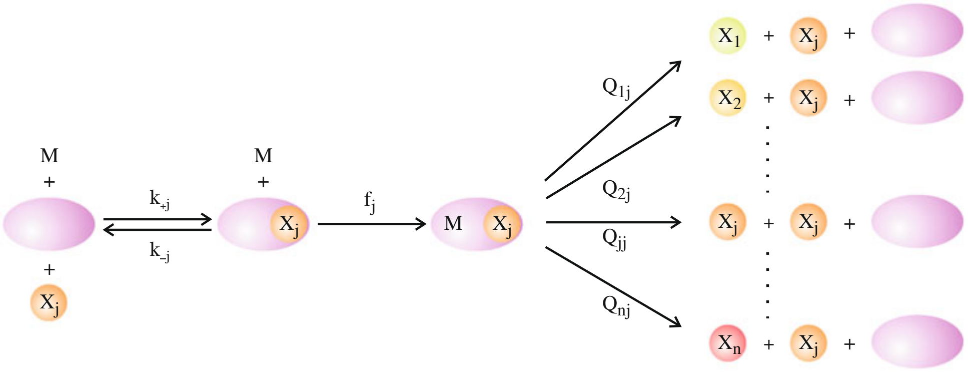


Fig. 4 A molecular view of replication and mutation. The replicase molecule (*violet*) binds the template RNA molecule (X_j, *orange*) with a binding constant $K_j = k_{+j}/k_{-j}$ and replicates with a rate parameter f_j. The reaction leads to a correct copy with frequency Q_{jj} and to a mutant X_k with frequency Q_{kj} with $Q_{jj} \gg Q_{kj}\ \forall\ k \neq j$. Stoichiometry of replication requires $\sum_{i=1}^{n} Q_{ij} = 1$, since the product has to be either correct or incorrect. The sum of all activated monomers is denoted by M. In the print version the mutant spectrum differs in the gray-value from lightest for X_1 to darkest for X_n

template strand is completed—nucleotide after nucleotide in the direction from the 3′- to the 5′-end—and forms locally a double-helical RNA duplex. The process of replication follows a simple principle making use of strand complementarity and is often denoted as *complementary replication*: Like in the historical silver-based photography, the plus strand acts as template for the synthesis of the minus strand, and vice versa, the minus strand is the template for plus strand synthesis. In vivo and in vitro, Qβ-replicase plays a twofold role: (i) It increases the accuracy of replication by reinforcing correct base pairing (A=U and G≡C) and (ii) it assists separation of the two complementary strands—template and newly synthesized RNA molecule—in the RNA duplex into individual strands during replication (Mills et al 1967; Weissmann 1974). Strand separation is essential for successful replication, because dissociation of the complete RNA duplex is thermodynamically so unfavorable that it does not occur at the temperature applied for replication.[6] In Qβ RNA replication, the whole length RNA duplex helix is never formed since the double helical stretch needed for template polymerization is separated into a plus and a mins strand on the fly (Fig. 5), both strands form their energetically favored specific single strand structures and prevent duplex formation. In this context it is worth mentioning that an enzyme-free experiment of cross-catalytic reproduction of RNA molecules with rich single strand structure has been successful (Lincoln and Joyce 2009).

Complementary replication is as efficient for population growth as direct replication is: after internal stationarity has been achieved, the plus–minus ensemble grows like a single unit of reproduction. Under the conditions of a closed system (no exchange of materials with the environment, see Fig. 6) RNA replication passes three phases of growth: exponential growth, linear growth, and saturation (Biebricher et al. 1983, 1984, 1985). Selection and Darwinian evolution require exponential growth and accordingly, an open system is indispensable in order to maintain replication within the exponential phase (Phillipson and Schuster 2009, chaps. 2, 3) by means of a flow. The easiest way to achieve this goal is to supply all materials consumed in RNA synthesis by a constant influx and to remove RNA in excess by an outflux that, in addition, compensates also for the increase in volume caused by the influx. The relatively low accuracy of viral RNA replication (see section: "The error threshold of reproduction") produces a sufficiently rich variety of variants that provide the basis for in vitro evolution (Joyce 2007).

The kinetic model of replication and mutation

The kinetic reaction mechanism of RNA replication in vitro has been studied in great detail (Biebricher et al. 1983, 1984, 1985): Under suitable conditions, excess replicase and nucleotide triphosphates (ATP, UTP, GTP, and CTP), the concentration of the RNA plus–minus ensemble grows exponentially (Fig. 6). The population maintains exponential growth when the consumed material is replenished either by a suitable flow device or serial transfer of small quantities of the reaction mixture into fresh medium (Spiegelman 1971). Under these conditions, replication kinetics can be simplified and properly described by the differential equation:

[6] Standard amplification of single stranded DNA by means of the polymerase chain reaction (PCR) is a frequently used technique for replication that circumvents isothermal duplex dissociation by means of a temperature program: Single stranded DNA is completed to a double helical duplex by means of a polymerase from *Thermophilus aquaticus* (Taq), the duplex is dissociated into single stands at higher temperature, and cooling of single strands completes the cycle (see also Cahill et al. 1991).

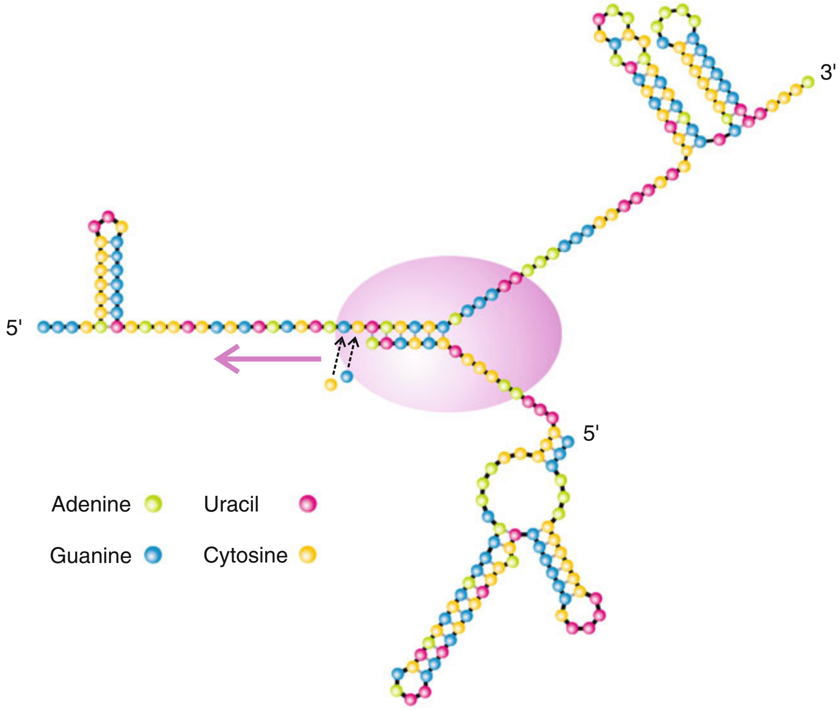


Fig. 5 Sketch of RNA replication by Qβ-replicase. An RNA template—here the plus-strand of the SV11 variant of Qβ-RNA (Biebicher and Luc, 1992)—is bound to the replicase and replication proceeds by adding single activated nucleotides one after the other to the growing product, the minus strand. The replicase operates on single stranded stretches. Double helical structural elements on the template strand are opened when they are encountered by the enzyme. Still on the enzyme, the duplex formed during replication is separated in order to allow for independent structure formation of both strands

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \sum_{i=1}^{n} Q_{ji} f_i x_i - \phi(t) x_j, \quad j = 1,2,\ldots,n \quad \text{with } \phi(t) = \sum_{i=1}^{n} f_i x_i \quad \text{or in vector notation } \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = (\mathrm{Q}\cdot\mathrm{F} - \phi(t))\boldsymbol{x}, \tag{9}$$

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 (Fig. 4). Equation 9 can be transformed into a linear ODE by means of integrating factor transformation and than solved by means of an eigenvalue problem (Thompson and McBride 1974; Jones et al. 1976):

$$z(t) = \boldsymbol{x}(t)\cdot\exp\left(\int_0^t \phi(\tau)d\tau\right),$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \mathrm{Q}\cdot\mathrm{F}z = \mathrm{W}z \text{ and } \mathrm{W} = \mathrm{B}\cdot\Lambda\cdot\mathrm{B}^{-1} \text{ or}$$
$$\Lambda = \mathrm{B}^{-1}\cdot\mathrm{W}\cdot\mathrm{B},$$

with Λ being a diagonal matrix containing the eigenvalues of W, $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$. Whenever a path of consecutive single point mutations can be found from every X_i to every X_j the matrix W is primitive[7] and fulfils Perron–Frobenius theorem (Seneta 1981, pp. 3, 22). Accordingly, the largest eigenvalue, λ_0, is strictly positive and non-degenerate and the corresponding right hand eigenvector ζ_0 has only positive entries. The calculation of the solutions x_j is somewhat lengthy but straightforward:

$$x_j(t) = \frac{\sum_{k=0}^{n-1} b_{jk} \sum_{i=1}^{n} h_{ki} x_i(0)\exp(\lambda_k t)}{\sum_{l=1}^{n}\sum_{k=0}^{n-1} b_{lk} \sum_{i=1}^{n} h_{ki} x_i(0)\exp(\lambda_k t)}, \quad j = 1,2,\ldots,n. \tag{10}$$

The new quantities in this equation are the elements of the two transformation matrices:

$$\mathrm{B} = \{b_{jk}; j = 1,2,\ldots,n; k = 0,1,\ldots,n-1\} \text{ and}$$
$$\mathrm{B}^{-1} = \{h_{kj}; k = 0,1,\ldots,n-1; j = 1,2,\ldots,n\}$$

The columns of B and the rows of B^{-1} represent the right hand and left hand eigenvectors of the matrix W. For example, we have

$$\zeta_0 = \begin{pmatrix} b_{10} \\ b_{20} \\ \vdots \\ b_{n0} \end{pmatrix}.$$

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

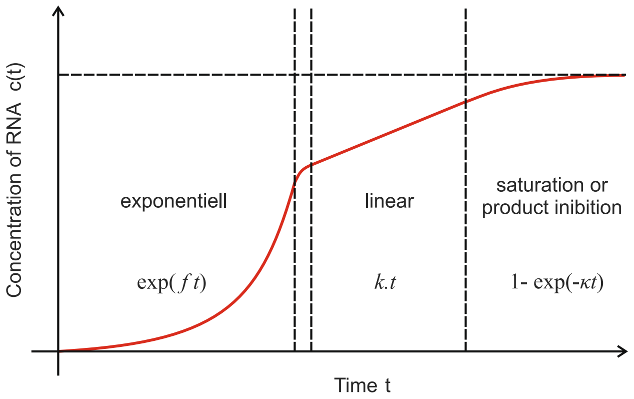


Fig. 6 Kinetics of RNA replication in closed systems. The time course of RNA replication by Qβ-replicase shows three distinct growth phases: (i) an exponential phase, (ii) a linear phase, and (iii) a phase characterized by saturation through product inhibition (Biebricher et al. 1983, 1984, 1985). The experiment is initiated by transfer of a very small sample of RNA suitable for replication into a medium containing Qβ-replicase and the activated monomers, ATP, UTP, GTP, and CTP in excess (consumed materials are not replenished in this experiment). In the phase of exponential growth, there is shortage of RNA templates, every free RNA molecule is instantaneously bound to an enzyme molecule and replicated, and the corresponding over-all kinetics follows $dx/dt = f{\cdot}x$ resulting in $x(t) = x_0 \cdot \exp(ft)$. In the linear phase, the concentration of template is exceeding that of enzyme, every enzyme molecule in engaged in replication, and over-all kinetics is described by $dx/dt = k' \cdot e_0^{(E)} = k$, wherein $e_0^{(E)}$ is the total enzyme concentration, and this yields after integration $x(t) = x_0 + kt$. Further increase in RNA concentration slows down the dissociation of product (and template) RNA from the enzyme–RNA complex and leads to a phenomenon known as product inhibition of the reaction. At the end, all enzyme molecules are blocked by RNA in complexes and no more RNA synthesis is possible, $c(t) \to c_\infty$

Since $\lambda_0 > \lambda_1 \geq \lambda_2 \cdots \geq \lambda_{n\text{-}1}$, the stationary solution contains only the contributions of the largest eigenvector, ζ_0 :

$$\lim_{t\to\infty} x_j(t) = \bar{x}_j = \frac{b_{j0}\sum_{i=1}^{n} h_{0i}x_i(0)}{\sum_{l=1}^{n} b_{l0}\sum_{i=1}^{n} h_{0i}x_i(0)}, \quad j = 1,2,\ldots,n. \qquad (11)$$

In other words, ζ_0 describes the stationary distribution of mutants and represents the genetic reservoir of an asexually reproducing species similarly to the gene pool of a sexual species. For this reason, ζ_0 has been called *quasi-species*.

The error threshold of reproduction

The dependence of quasi-species on the frequency of mutation is considered in this subsection. In general, the mutation rate is not tunable, but it can be varied within certain limits in suitable experimental assays. In order to illustrate, mutation rate dependence and to subject it to mathematical analysis, a simplifying model assumption called *uniform error rate* model is made (Eigen 1971). The error rate per nucleotide and replication, p, is assumed to be independent of the position and the nature of the nucleotide exchange (for example, A → U, A → G or A → C occur with the same frequency p and the total error rate at a given position is $3p$). Then the elements of the mutation matrix Q depend only on three quantities: the chain length of the sequence to be replicated, ℓ, the error frequency p, and the Hamming distance between the template, X_i, and the newly synthesized sequence, X_j, denoted by d_{ij}^{H},[8]

$$Q_{ji} = (1-(\kappa-1)p)^{\ell - d_{ij}^{\mathrm{H}}} \cdot p^{d_{ij}^{\mathrm{H}}} = (1-(\kappa-1)p)^{\ell}\varepsilon^{d_{ij}^{\mathrm{H}}} \quad \text{with } \varepsilon = \frac{p}{1-(\kappa-1)p}. \qquad (12)$$

The size of the nucleotide alphabet is denoted by κ—for natural polynucleotides we have $\kappa = 4$ corresponding to {A, U(T),G,C}. The explanation of the two terms in Eq. 12 is straightforward: The two sequences differ in d_{ij}^{H} positions and hence $\ell - d_{ij}^{\mathrm{H}}$ nucleotides have to be copied correctly, each contributing a factor $1-(\kappa-1)p$, and d_{ij}^{H} errors with frequency p have to be made at certain positions. Since the Hamming distance is a metric, we have $d_{ij}^{\mathrm{H}} = d_{ji}^{\mathrm{H}}$, and within the approximation of the uniform error rate model, the mutation matrix Q is symmetric.

For $p = 0$, we encounter the selection case (6): in absence of degeneracy—all fitness values f_j are different—the species of highest fitness, the master sequence X_m, is selected and all other variants disappear in the long time limit. The other extreme is random replication, a condition under which all single nucleotide incorporations, correct or incorrect, namely A → A, A → U, A → G, and A → C, are equally probable and occur with frequency $\tilde{p} = 0.25$. Generalization from four to κ letters is straightforward: Then, for $\tilde{p} = \kappa^{-1}$ all elements of matrix Q are equal to $\kappa^{-\ell}$ where ℓ is again the sequence length. If all sequences are considered in the model the matrix W contains $n = \kappa^{\ell}$ identical rows and takes on the following form at $p = \tilde{p}$

$$\tilde{\mathrm{W}} = \kappa^{-\ell}\begin{pmatrix} f_1 & f_2 & \ldots & f_n \\ f_1 & f_2 & \ldots & f_n \\ \vdots & \vdots & \ddots & \vdots \\ f_1 & f_2 & \ldots & f_n \end{pmatrix}.$$

The uniform distribution $\Pi = \{\bar{x}_j = n^{-1} \forall \;\; j = 1,2,\ldots, n \;\; \text{with } n = \kappa^{\ell}\}$ is the eigenvector corresponding to the largest eigenvalue $\lambda_0 = \kappa^{-\ell}\sum_{i=1}^{n} f_i$, whereas all all other eigenvalues of W vanish.[9] In the whole range $0 \leq p \leq \kappa^{-1}$, the stationary distribution changes from the homogeneous

[8] The Hamming distance d_{ij}^{H} between two strings, X_i and X_j of equal length counts the number of positions in which the two end-to-end aligned strings differ (Hamming 1986).

[9] It can be proven by means of a recursion that the eigenvalues of the matrix $\tilde{W}$ fulfill the relation $\lambda^{n-1}\left(\lambda - \kappa^{-\ell}\sum_{i=1}^{n} f_i\right) = 0$.

population, $\Xi_m = \{\bar{x}_m = 1, \bar{x}_j = 0 \forall j \neq m\}$ to the uniform distribution Π. A remark concerning the uniform distribution is required: the number of possible polynucleotide sequences $-\kappa^\ell = 4^\ell$ for natural molecules—exceeds by far any accessible population size already for small RNAs with $\ell \approx 30$. Although Eq. 10 predicts the uniform distribution in theory, no stationary population is possible in practice, and we expect populations to drift randomly through sequence space (Derrida and Peliti 1991; Huynen et al. 1996; and "Modeling evolution in silico" section). A limitation of modeling by differential equations is encountered [see also the localization threshold of mutant distributions (McCaskill 1984; Eigen et al. 1989)].

Between the two extremes, the function $\bar{x}_m(p)$ was approximated by Manfred Eigen through neglect of back-flow from mutants to the master sequence. He obtained for $\mathrm{d}x_m/\mathrm{d}t = 0$ (Eigen 1971):

$$\bar{x}_m = Q_{mm} - \frac{\bar{f}_{-m}}{f_m} = Q_{mm} - \sigma_m^{-1} \quad \text{with } \bar{f}_{-m} = \frac{\sum_{i=1, i \neq m}^{n} f_i \bar{x}_i}{1 - \bar{x}_m}. \tag{13}$$

The quantity $\sigma_m = f_m / \bar{f}_{-m}$ is denoted as the *superiority* of the master sequence. In this rough, zeroth order approximation, the frequency of the master sequence becomes zero at a critical value of the mutation rate parameter, $p_{\max}$, for constant chain length ℓ or at a maximal chain length $\ell_{\max}$ for constant replication accuracy p,

$$p_{\max} \approx \frac{\ln \sigma_m}{(\kappa - 1)\ell} \quad \text{or } \ell_{\max} \approx \frac{\ln \sigma_m}{(\kappa - 1)p},$$

respectively. The critical replication accuracy has been characterized as *error threshold* of replication. As we shall see in the "Fitness landscapes and error thresholds" section, the error threshold reminds of a phase transition in which the quasi-species changes from a mutant distribution centered around a master sequence to some other distribution that is only weakly dependent on p or independent at all, for example the uniform distribution.[10] In other words, the solution that becomes exact at $p = \tilde{p}$ is closely approached at $p = p_{\max}$ already. For the purpose of illustration for a superiority of $\sigma_m = 1.1$ and a chain length of $\ell = 100$, we obtain $p_{\max} = 0.00032$ compared to $\tilde{p} = 0.25$.

Both relations for the error threshold, maximum replication accuracy and maximum chain length, were found to have practical implications: (i) RNA viruses replicate at mutation rates close to the maximal value (Drake 1993). A novel concept for the development of antiviral drugs makes use of this fact and aims at driving the virus population to mutation rates above the error threshold (Domingo 2005). (ii) There is a limit in chain length for faithful replication that depends on the replication machinery: the accuracy limit of enzyme-free replication is around one error in one hundred nucleotides, RNA viruses with a single enzyme and no proof reading can hardly exceed accuracies of one error in 10,000 nucleotides, and DNA replication with repair on the fly reaches one error in 10^8 nucleotides. For prokaryotic DNA replication, post-replication repair increases the accuracy to 10^{-9}–10^{-10}, which is roughly one mutation in 300 duplications of bacterial cells (Drake et al. 1998).

Fitness landscapes and error thresholds

The approximation of the error threshold through neglect of mutational back-flow (13) caused the results to be independent of the distribution of replication parameters of mutants, since only the mean replication rate, $\bar{f}_{-m}$, enters the expression. As a matter of fact, the appearance of an error threshold and its shape depend on the fitness landscape (Wiehe 1997; Phillipson and Schuster 2009, pp. 51–60). In this subsection we shall now consider the influence of the distribution of fitness values in two steps: (i) different fitness values are applied for sequences with different Hamming distances from the master sequence, and (ii) different fitness values are assigned to individual sequences. In the first case, all sequences X_j with Hamming distance $d^{\mathrm{H}}_{m,j} = k$ fall into the *error class* k. Although the assumption that all sequences in a given error class have identical fitness is not well justified on the basis of molecular data, it turns out to be useful for an understanding of the threshold phenomenon.

The following five model landscapes or fitness matrices $\mathrm{F} = \{F_{ij} = f_i \cdot \delta_{ij}\}$ were applied (Fig. 7): (i) the single-peak landscape corresponding to a mean field approximation, (ii) the hyperbolic landscape, (iii) the step-linear landscape, (iv) the multiplicative landscape, and (v) the additive or linear landscape. Examples for the dependence of the quasi-species distribution on the error rate are shown in Fig. 8.

For analyzing error thresholds, it is useful to consider three separable features: (i) the decay in the frequency of the master sequence—$x_m(p) \to 0$ in the zeroth order approximation (13), (ii) the phase transition-like sharp change in the mutant distribution, and (iii) the transition from the quasi-species to the uniform distribution. All the three phenomena coincide on the single-peak landscape (Fig. 8; upper part). Characteristic for most hyperbolic landscapes is an abrupt transition in the distribution of sequences according to (ii) but—in contrast to the single-peak landscape—the transition does not lead to the uniform

[10] A sharp transition from the structured quasi-species to the uniform distribution is found for the single-peak landscape and some related landscapes only (see "Fitness landscapes and error thresholds" section.

Mathematical Modeling of Evolution – Erratum

The denominator in equation (13) in Peter Schuster, *Theory in Biosciences* 130:71-89, 2011, page 79 is missing. The correct equation is of the form

$$\bar{x}_m \;=\; \frac{Q_{mm} - \sigma_m^{-1}}{1 - \sigma_m^{-1}} \;\text{ with }\; \sigma_m \;=\; \frac{f_m}{f_{-m}} \;\text{ and }\; f_{-m} \;=\; \frac{\sum_{i=1, i \neq m}^{n} f_i x_i}{1 - \bar{x}_m}\;. \qquad (13)$$

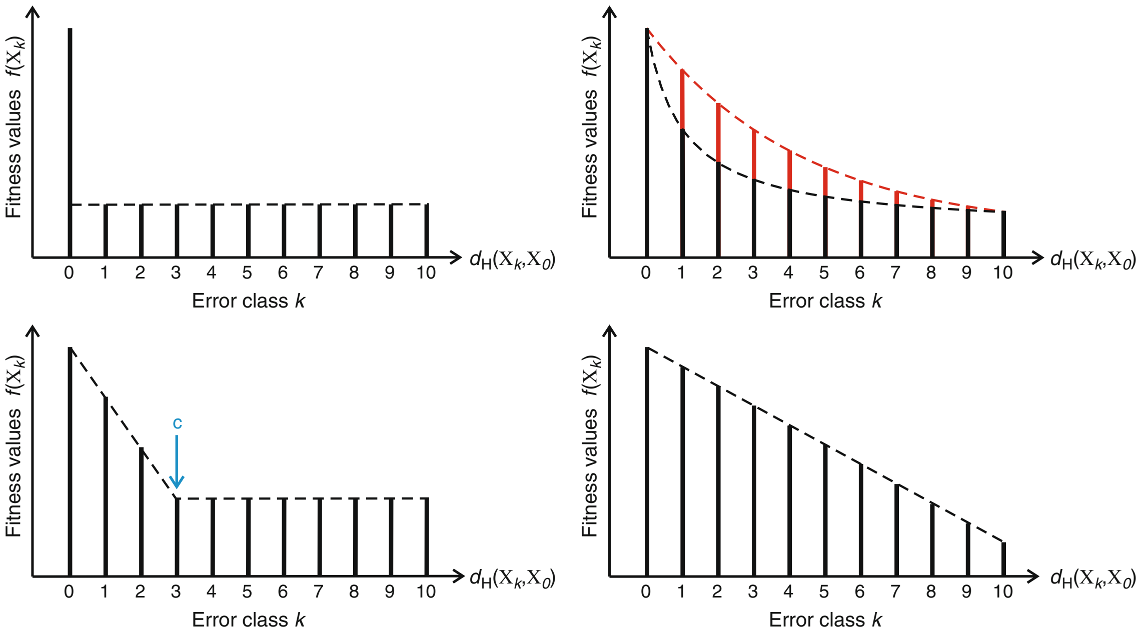


Fig. 7 Some examples of model fitness landscapes. The figure shows five model landscapes with identical fitness values for all sequences in a given error class: (i) the single peak landscape ($f(X_0) = f_0$ and $f(X_j) = f_n\ \forall\ j = 1, \ldots, n$; *upper left drawing*), (ii) the hyperbolic landscape ($f(X_j) = f_0 - (f_0 - f_n)(n+1)j/(n(j+1))\ \forall j = 0, \ldots, n$; *upper right drawing*, *black curve*), (iii) the step-linear landscape ($f(X_j) = f_0 - (f_0 - f_n)j/k\ \forall\ j = 0, \ldots, k$ and $f(X_j) = f_n\ \forall\ j = k+1, \ldots, n$; *lower left drawing*), (iv) the multiplicative landscape ($f(X_j) = f_0\ (f_n/f_0)^{j/n}\ \forall j = 0, \ldots, n$; *upper right drawing*, *red curve*), and (v) the additive or linear landscape ($f(X_j) = f_0 - (f_0 - f_n)j/n\ \forall\ j = 0, \ldots, n$; *lower right drawing*). In the print version red appears as gray

distribution, instead another distribution is formed that changes gradually into the uniform distribution, which becomes the exact solution at the point $p = \tilde{p}$. The step-linear landscape illustrates the separation of the decay range (i) and the phase transition to the uniform distribution (ii and iii). In particular, variation in the position of the step (‘c’ in Fig. 7) that the phase transition point $p_{\max}$ shifts towards higher values of p when the position of the step moves towards higher error-classes, whereas the decrease in the decay of the master sequence moves in opposite direction. The additive and the multiplicative landscape, the two landscapes that are often used in population genetics, do not sustain threshold-behavior. On these two landscapes, the quasi-species is transformed smoothly with increasing p into the uniform distribution.

Error thresholds on realistic fitness landscapes can be modeled straightforwardly by the assumption of a scattered distribution of fitness values within a given band of width d for all sequences except the master sequence[11]:

$$f(X_j) = \bar{f}_{-m} + d(\eta_{\mathrm{rnd}}(j) - 0.5) - 1,\quad j = 1, 2, \ldots, \kappa^{\ell},\ j \neq m. \tag{14}$$

In this expression ’$\eta_{\mathrm{rnd}}(j)$’ is a random number drawn from some random number generator with a uniform distribution of numbers in the range $0 \leq \eta_{\mathrm{rnd}}(j) \leq 1$ with j being the index of the consecutive calls of the random function and d is the band width of fitness values. Similarly the uniform error rate model (12) is only a rough approximation to the distribution of mutation frequencies. In order to relax the stringent constraint here, we define a local mutation rate p_k for each position k ($k = 1, 2, \ldots, \ell$) along the sequence and assume again that the individual p_k values vary within a given band width. The computational capacities of today allow for studies of error thresholds at the resolution of individual sequences up to chain lengths $n = 10$. Further increase in computational power raises expectation to be able to reach $n = 20$, which in case of binary sequences is tantamount to the diagonalization of $10^6 \times 10^6$ matrices.

Three questions are important in the context of resolution of fitness values down to individual sequences: (i) How does the dispersion of fitness values expressed in terms of the band width d change the characteristics of the error threshold, (ii) how does variation in local mutation rates influence error threshold and (iii) what happens if two more sequences have the same maximal fitness value f_m. The answers to question (i) and (ii) follow readily from the

[11] The data obtained from biomolecules suggest a high degree of ruggedness for the landscapes derived for structures and functions: nearby sequences may lead to identical or very different structures. By the same token functions like fitness values may be the same or very different for close by lying genotypes. Ruggedness is an intrinsic property of mapping from biopolymer sequences into structures or functions.

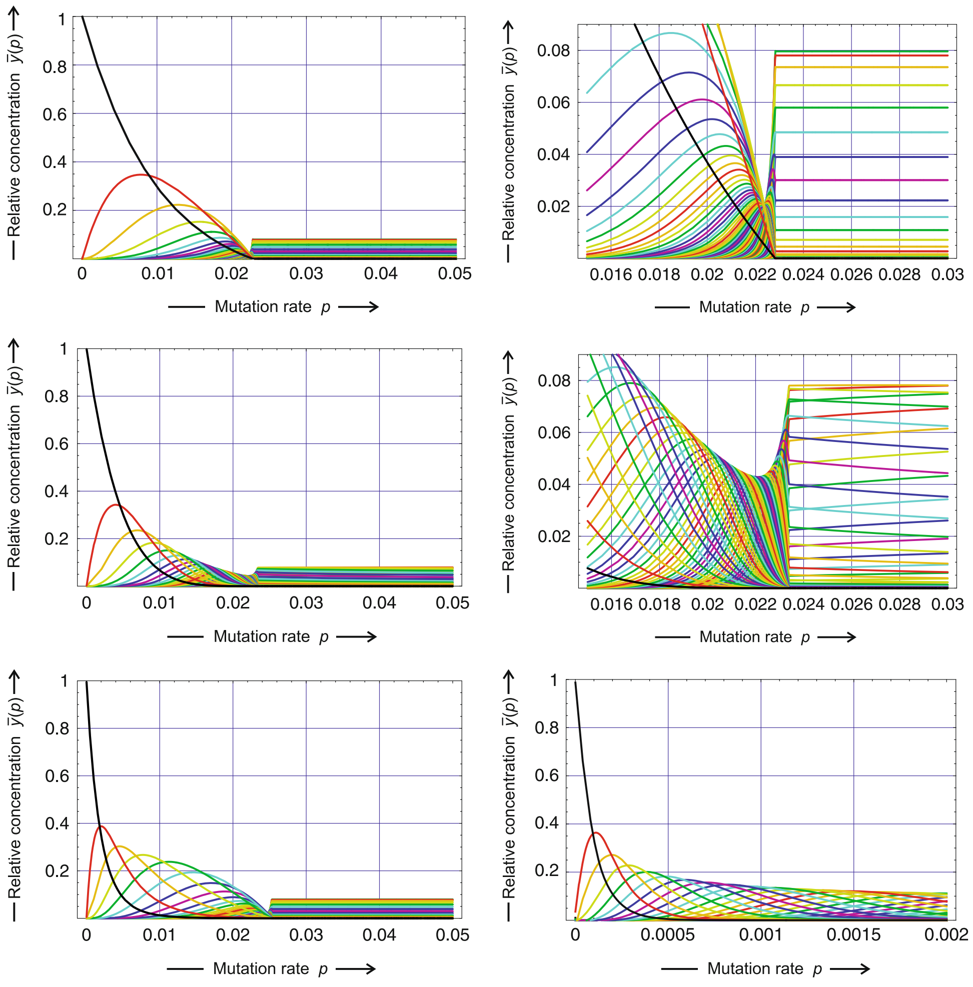


Fig. 8 Error thresholds on different model fitness landscapes. Relative stationary concentrations of entire error classes $\bar{y}_k(p)$ $(k = 0, 1, \ldots, \ell; \bar{y}_k = \sum_{i=1, d_H(X_i,X_m)=k}^{n} \bar{x}_i)$ are plotted as functions of the mutation rate p (the different error classes are color coded, $d_{\mathrm{H}}(X_i,X_m) = 0$, *black*; $d_{\mathrm{H}}(X_i,X_m) = 1$, *red*; $d_{\mathrm{H}}(X_i,X_m) = 2$, *yellow*, $d_{\mathrm{H}}(X_i,X_m) = 3$, *chartreuse*; $d_{\mathrm{H}}(X_i,X_m) = 4$, *green*, etc). The pictures at the top show the threshold behavior on the single-peak fitness landscape (enlarged on the *right hand side*) where the three conditions (i), (ii), and (iii)—decay of master, phase transition, and transition to uniform distribution—coincide. The *two pictures* in the *middle* were computed for the hyperbolic landscape (enlarged on the *right hand side*) where the phase transition leads to a distribution that changes gradually into the uniform distribution and (i) has a slight offset to the *left* of (ii). The *left-hand figure* at the *bottom* corresponds to the step-linear landscape and fulfils (ii) and (iii) whereas (i) has a large offset to the left, and eventually the additive landscape (*bottom*, *right-hand side*) does not sustain an error threshold at all. Parameters used in the calculations: $\ell = 100$, $f_m = f_0 = 10$, $f_n = 1$ (except the hyperbolic landscape where we used $f_n = 0.9091$ in order to have $\bar{f}_{-\mathrm{m}} = 1$ as for the single peak landscape), and $c = 5$ for the step-linear landscape. In the print version red appears as the darkest gray and other colors differ in their gray-value

calculated results: the position at which the frequency of the master sequence in the population reaches a given small value migrates towards smaller f-values with increasing band width d. This observation agrees fully with expectation because the fitness value closest to f_m becomes larger for broader bands of fitness values. The scatter of fitness values at the same time broadens the transition. Relaxation of the uniform error rate assumption causes smoothing of the error threshold and a shift of $p_{\max}$ towards higher values of p.

Degeneracy of fitness values implies that two or more genotypes have the same fitness and this is commonly denoted as *neutrality* in biology. An investigation of the role of neutrality requires an extension of Eq. 14. A certain fraction of sequences, expressed by the degree of neutrality λ, is assumed to have the highest fitness value f_0, and the fitness values of the remaining fraction $1 - \lambda$ are assigned as in the non-neutral case (14). This random choice of neutral sequences together with a random dispersion of the other fitness values yields an interesting result: random selection in the sense of Motoo Kimura's neutral theory of evolution (Kimura 1983) occurs only for sufficiently distant fittest sequences. In full agreement with the exact result derived for the limit $p \to 0$ (Schuster and Swetina

1988) we find that two fittest sequences of Hamming distance $d_H = 1$, 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 $p = p_{\max}$. For clusters of more than two Hamming distance one sequences, the frequencies of the individual members of the cluster are obtained from the largest eigenvector of the adjacency matrix. Pairs of fittest sequences with Hamming distance $d_H = 2$, 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 up to the error threshold $p = p_{\max}$. Strong coupling of fittest sequences manifests itself in virology as systematic deviations from consensus sequences of populations as is indeed observed in nature. For fittest sequences with $d_H \geq 3$ random selection chooses one sequence arbitrarily and eliminates all others as predicted by the Kimura's neutral theory of evolution.

Mapping sequences into structures

Modeling evolution of molecules by means of chemical kinetics solves one vital problem of the theory of evolution: fitness can be determined independently of the evolutionary process by measuring the rate parameters of replication and the sometimes raised argument that *survival of the fittest* is nothing but a tautology, because there is no other way to measure fitness except running evolution, is obsolete. A full understanding of evolution, however, is confronted with enormous complexity even in the simple case of nucleic acid molecules in the test tube. How does the fitness of a molecule change in response to mutation? This question is tantamount to asking for the prediction of molecular function from known biopolymer sequences, which is a notoriously hard problem. Commonly prediction of function is addressed in two steps: (i) prediction of structure from known sequence and (ii) prediction of function from known structure. Both tasks are hard in general and useful solutions are available for special cases only. An exception are RNA structures on the level of so-called secondary structures: Structure prediction is accessible by mathematical and computational methods (Schuster 2006). The discreteness of nucleotide interactions—either two nucleotides form a base pair or they do not—facilitates the analysis of RNA structures and allows for the application of efficient dynamic programming algorithms to structure prediction (Hofacker et al. 1994a; Zuker and Stiegler, 1981; Zuker, 1989a, b). The relation between structure and function can be modeled straightforwardly for a number of special cases. One example, is binding between RNA molecules called RNA hybridization (Hofacker et al. 1994b; Dimitrov and Zuker 2004).

The basic principle of folding RNA sequences into secondary structures is double helix formation, in essence the same as used in nucleic acid replication: the single stranded molecule folds back onto itself when the sequence allows for (partial) duplex formation (Fig. 9) whereby the driving force is lowering Gibbs free energy. Since base pairing logic applies as well to structure formation as to replication, secondary structures are objects that can be analyzed by means of combinatorics. Simple logic on one hand side is counteracted by complexity originating from nonlocal interactions. As illustrated in the example of Fig. 9 distant nucleotides as well close by lying ones may form base pairs. The full three-dimensional structure of RNA molecules is built through forming additional nucleotide interactions called tertiary interactions, which are often stabilized by divalent cations, especially by $\mathrm{Mg}^{2\oplus}$. Tertiary interactions are either sequence specific and can be catalogued therefore (Leontis et al. 2006) or they follow a general principle like, for example, 'end-on-end' stacking of helices from secondary structure (Moore 1999).

Because of the discreteness of RNA structure space, mappings from RNA sequence space into structure space can be addressed by combinatorics and have been studied extensively (Fontana et al 1993; Schuster et al. 1994; Reidys et al. 1997; Fontana and Schuster 1998b; Stadler et al. 2001). Six properties of these mappings appear to be relevant for evolution (Schuster 2006):

(i) The numbers of RNA sequences exceed by far the numbers of RNA secondary structures and neutrality with respect to structures is inevitable.

(ii) Sequences folding into the same structure form neutral networks that are the pre-images of structures in sequence space.

(iii) Depending on the degree of neutrality, neutral networks are either connected or split into components. The critical connectivity threshold depends only on the number of letters in the nucleotide alphabet.

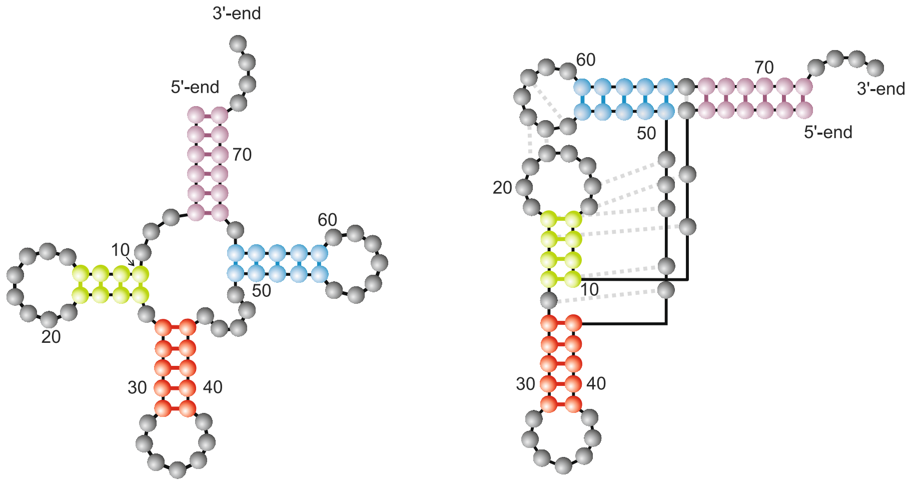
(iv) Neutral networks in the conventional {A,U,G,C} -space are larger and more likely to be connected than neutral networks in the binary or {G,C}-space.

(v) Neutral networks are embedded in sets of sequences that are compatible with the structure.[12]

(vi) The intersection of the compatible sets of two structures is always non-empty. In other words, for

[12] Compatibility means that the sequence can form the structure but not necessarily as the minimum free energy structure.

5’end-GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGG-
-AGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA-3’end



((((((...((((........)))).(((((.......))))).....(((((.......))))).))))))....

Fig. 9 The secondary structure of a typical transfer RNA. The nucleotide sequence folds back on itself through forming double helical stacks whenever sequence complementarity allows for it (*left hand sketch*; individual stacks are color coded). In the *symbolic notation*, a secondary structure is represented by an equivalent string with *parentheses* and dots whereby each single nucleotide is represented by a *dot*, each base pair by a parenthesis, and mathematical notation applies (*string* at the *bottom* of the figure). The secondary structure is converted into the full three-dimensional structure by forming additional stabilizing interactions between nucleotides (*right hand sketch*). One general principle in tertiary structure formation is extension of helices through ’end-on-end’ stacking: The *green* helix extends the *red* one, and the *violet* helix extends the *blue* one in the figure above. The molecular example shown is the phenylalanyl-transfer RNA ($tRNA_{phe}$) from the yeast *Saccharomyces cerevisiae*. The letters D, M, Y, T,and P denote modified nucleotides

two given structures it is always possible to find a sequence that can form both.

Evidence for the existence and strong hints on the properties of neutral networks come from RNA selection experiments (Schultes and Bartel 2000; Held et al. 2003; Huang and Szostak 2003). For a more complete understanding of neutrality in evolution of molecules further development of the theory and appropriate experiments are needed.

Modeling evolution in silico

Stochasticity is essential for evolution—each mutant after all starts out from a single copy and random drift in the sense of Motoo Kimura is a pure stochastic phenomenon. A large number of studies have been conducted on stochastic effects in population genetics (Blythe and McKane 2007). Not too much work, however, has been done so far on the development of a general stochastic theory of molecular evolution. We mention two examples representative for others (Jones and Leung 1981; Demetrius et al. 1985). 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 solutions of the deterministic Eq. 9 in the limit of large populations.

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 (Fontana and Schuster 1987; Fontana et al. 1989; Fontana and Schuster 1998b): The sequence is the genotype and the RNA secondary structure represents the phenotype. The simulation tool starts from a population of RNA molecules and simulates chemical reactions corresponding to replication and mutation in a continuous stirred flow reactor (CSTR) by using Gillespie’s algorithm (Gillespie 1976, 1977, 2007). Fitness parameters are predefined functions of RNA structures—Eq. 15 presents an example. Molecules replicate in the reactor and produce correct copies and mutants according to a stochastic version of the mechanism shown in Fig. 4, the material consumed is supplied by a continuous influx of stock solution into the reactor, and excess material is removed by means of an outflux compensating the increase in volume. Whenever a new sequence is produced by mutation, the corresponding structure and its fitness are calculated. The stochastic process in the reactor is constructed to have two absorbing states: (i) extinction—all RNA molecules are diluted out of the reaction vessel, and (ii) success—the reactor has produced the predefined target structure. The population size determines the outcome of the computer experiment: Below $N = 18$ the reactor goes into extinction with a probability greater 0.5 and it reaches the target with a high probability close to one for population sizes $N > 20$. For sufficiently large populations the probability of extinction is very small, for population sizes reported here, $N \geq 1000$, extinction has been never observed.

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 between the symbolic notations of the structure formed by the sequence, $S_k = f(X_k)$ and the target structure S_T,

$$f_k(S_k, S_T) = \frac{1}{\alpha + d_H(S_k, S_T)/\ell}. \tag{15}$$

An adjustable parameter α is introduced in order to avoid infinite fitness when the target is reached (here it was chosen to be 0.1). The fitness increases when S_k approaches the target, a trajectory is completed when the population reaches a sequence that folds into the target structure. A typical trajectory is shown in Fig. 10. In this simulation a homogenous population consisting of N molecules with the same random sequence and the corresponding structure is chosen as initial condition. The target structure was chosen to be the well-known secondary structure of phenylalanyl-transfer RNA (tRNA phe) shown in Fig. 9. The mean distance to target of the population decreases in steps until the target is reached (Fontana et al. 1989; Fontana and Schuster 1998a, b; Schuster 2003). Individual (short) adaptive phases are interrupted by long quasi-stationary epochs.

Optimization dynamics in phenotype space is reconstructed in terms of a time ordered series of structures that leads from an initial structure S_I to the target structure S_T. This series, called the *relay series*, is a uniquely defined and uninterrupted sequence of structures in the flow reactor. It is retrieved through backtracking, that is in opposite direction from the final structure to the initial structure: 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 structure from which it descended by mutation. Now, we record time and structure, highlight the parent structure, and repeat the procedure. Recording further backwards yields a series of structures and times of first appearance, which ultimately ends in the initial population.[13] Usage of the relay series and its theoretical background allows for classification of transitions (Fontana and Schuster 1998a; Stadler et al. 2001): Minor or frequent transitions occur almost instantaneously, they are manifested by small changes in the structures commonly involving one or two base pairs, and major or rare transitions, which require random drift in neutral subspaces in order to find an appropriate starting point for the successful mutation. Major transition are accompanied by larger structural changes (Fontana and Schuster 1998a; and Fig. 11).

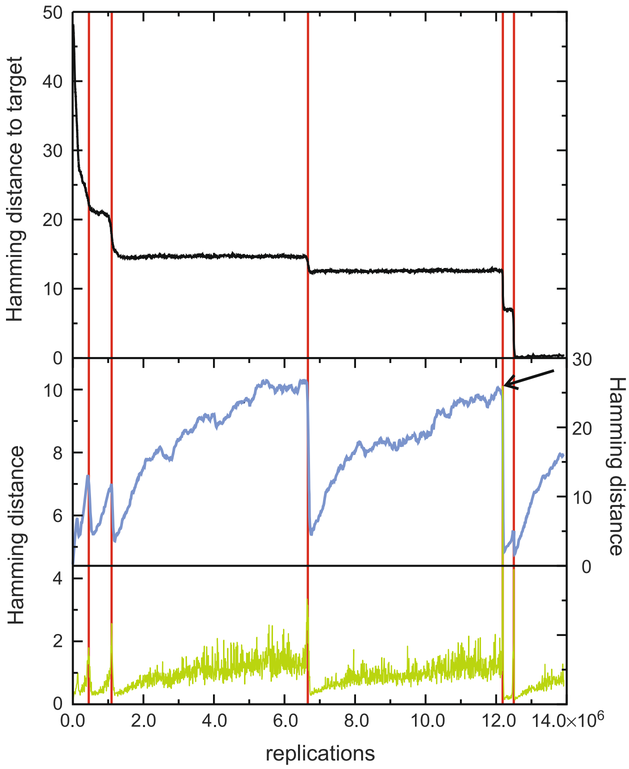


Fig. 10 A trajectory of evolutionary optimization. The topmost plot presents the mean distance to the target structure of a population of 1000 molecules. The plot in the middle shows the width of the population in Hamming distance between sequences and the plot at the bottom is a measure of the velocity 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 Huynen et al (1996). A remarkable synchronization is observed: At the end of each quasi-stationary 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 Eq. 15, and initial as well as target structure are shown in Table 1

Inspection of the relay series together with the sequence record on the quasi-stationary plateaus provides an explanation for the stepwise approach towards the target and allows for a distinction of two scenarios:

(i) The structure is constant and we observe neutral evolution in the sense of Kimura's theory of neutral

[13] 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 a time-ordered series of genotypes.

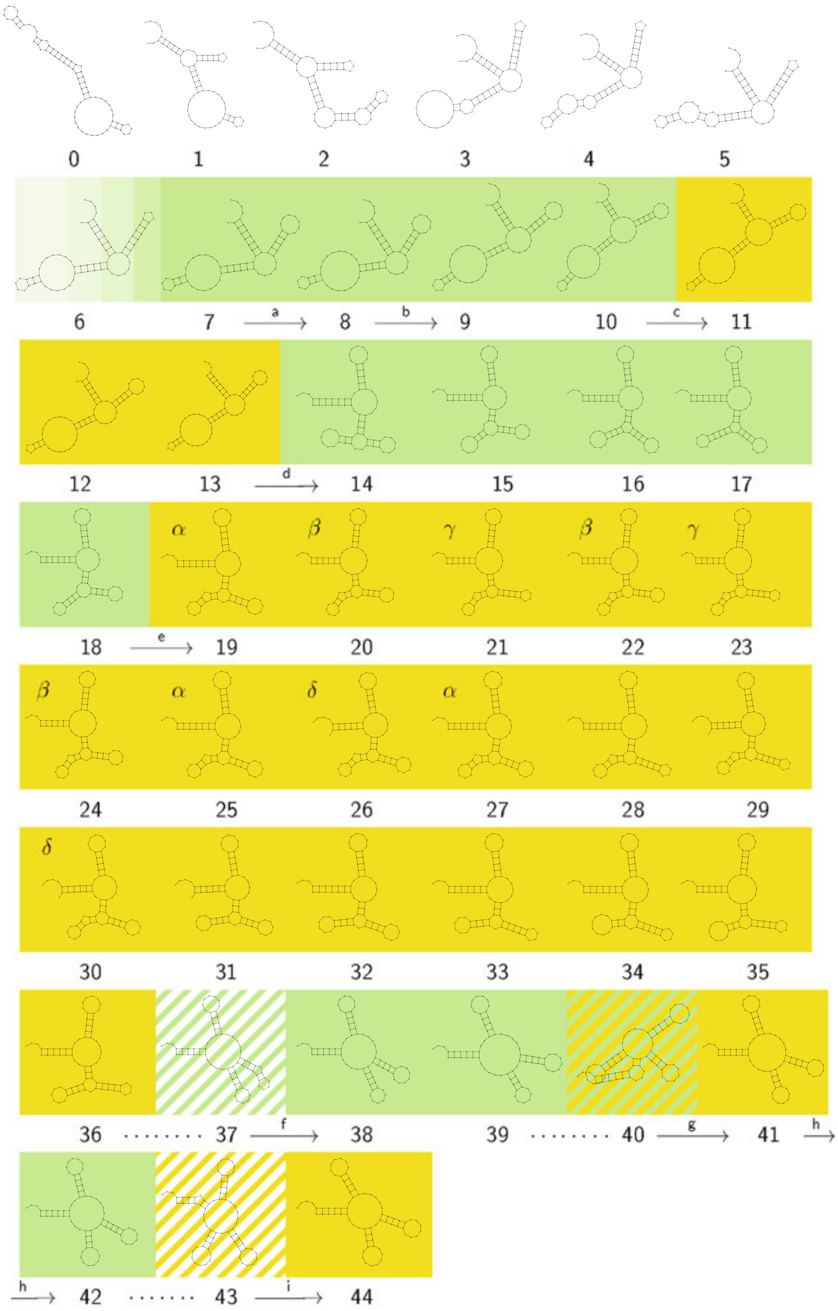


Fig. 11 The relay series of an in silico evolution experiment. The relay series consists of 44 steps leading from the initial structure $S_I = S_0$ to the target structure $S_T = S_{44}$. *Lower case roman letters* (a, b, c,…) indicate major transitions and lower case Greek letters (α, β, γ …) identify closely related structures. The *background color* indicates stretches of closely related structures in the relay series. It is worth noticing that the same structures can appear several times in the relay series (e.g., shapes *19–30*). The construction of relay series is described in the text. The figure is taken from Fontana and Schuster (1998a, suppl. 1)

evolution (Kimura 1983). 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 diffusion process on the corresponding neutral network (Huynen et al. 1996, see also Fig. 10).

(ii) The process during the stationary epoch involves several structures with identical replication rates and the relay series reveal 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. 10 that shows the width of the population as a function of time (Schuster 2003). The population width increases during the quasi-stationary epoch and sharpens almost instantaneously after a sequence had been produced 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: On the plateaus the drift is very slow but becomes fast at the end of the plateau when the population center moves quickly or 'jumps' in sequence space from one point to another point from where a new adaptive phase can be initiated (as manifested by the peaks in Fig. 10). A closer look at the figure reveals the coincidence of the 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* (Lenski et al. 1991) 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 20 years, a great adaptive innovation in one colony (Blount et al. 2008). 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 food source that led to a substantial increase in colonial growth. The mutation leading to citrate import into the cell is reproducible with earlier isolates of this particular colony that has apparently traveled on the neutral network to a position from where the adaptive mutation is within reach. All other eleven colonies did not give rise to mutations with 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 (1977) famous quote: "Nothing makes sense in biology except in the light of evolution".

Table 1 collects some numerical data harvested by sampling of evolutionary trajectories under identical conditions.[14] Individual trajectories show enormous scatter in the time or in 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 times as long as the corresponding process in AUGC sequence space in agreement with the difference in neutral network structure (Schuster, 2003, 2006).

(ii) The time to target decreases with increasing population size.

(iii) The number of replications required to reach target increases with population size.

Combining items (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 sufficiently large to avoid extinction.

A simulation study on the parameter dependence of in silico RNA evolution has been reported recently (Kupczok and Dittrich 2006). Increase in mutation rate leads to an error threshold phenomenon that is closely related to the one observed with quasi-species on a single-peak landscape as described above (Eigen et al. 1989). Evolutionary optimization becomes more efficient[15] with increasing error rates until the error threshold is reached. Further increase in the error rate leads to an abrupt breakdown of the optimization process. As expected, the distribution of replication rates or fitness values f_j in sequence space is highly relevant too: steep decrease of fitness with the distance to the master structure—represented by the target that has the highest fitness value—leads to sharp threshold behavior reminding of a single-peak landscape, whereas flat landscapes show broad maxima of optimization efficiency without an indication of a threshold-like behavior.

Concluding remarks

The exceedingly complex phenomenon of evolution takes place on multiple organizational levels, which range from cell organelles and cells to organs, organisms and populations. All these levels are different manifestations of the phenotype. A comprehensive description is not yet at hand and mathematical modeling as well as experimental studies inevitably have to concentrate on individual aspects or modules of the system. Nevertheless, the reductionists' program to partition the whole into tractable subsystems and to reconstitute it with the detailed knowledge of a lower level of description turned out to be impressively successful. In case of the cell, for example, molecular biology has first reduced the highly complex entity to individual biomolecules—nucleic acids, proteins, carbohydrates, lipids and others—and subjected the parts to biochemical and biophysical analysis. Next followed the study of the supramolecular complexes, molecular machines, and organelles within the cell. Starting with genomics and proteomics in the nineteen nineties and

[14] Identical means here that everything was kept unchanged in the computer experiments except the seeds for the random number generator.

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

Table 1 Statistics of the optimization trajectories

Alphabet	Population size	Number of runs	Real time from start to target		Number of replications [10^7]	
	N	n_R	Mean value	σ	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	–	–

The table shows the results of sampled evolutionary trajectories leading from a random initial structure S_I to the structure of tRNA$^{\text{phe}}$, S_T as target. Simulations were performed with an algorithm introduced by Gillespie (1976, 1977, 2007). The time unit is here undefined and *real time* implies proportionality to real measurements of time. 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 the data of the simulations. For less than 50 runs, no statistics is given for the number of replication because the uncertainty is too large

The structures S_I and S_T were used in the optimization:

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

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

continuing with systems biology at the turn of the century the object of investigation at the molecular level has been shifted from single molecules to higher units, cells, organs, and eventually organisms. The goal of the new biology is to complete the bottom-up approach from chemistry and physics and to provide a novel access to the understanding of the complexity of life as well as to develop new tools and techniques for the exploration of biology specific phenomena. Still there is a long way to go before this goal will be reached and the unsolved problems exceed by far the available solutions, but the contours of a new and comprehensive theoretical biology that is rooted in mathematics, physics, and chemistry are already apparent.

Evolution of molecules, viruses, and bacteria is studied under simplified conditions in vitro and in silico, and, in principle, allows for the incorporation of molecular mechanisms of reproduction, mutation, and recombination into the equations of the evolutionary process. Chemical kinetics of virus specific RNA replication is well understood and even in this very simple case the process or reproduction is quite complicated. Modeling DNA replication kinetics at full molecular resolution is still a great challenge but solvable by means of the current experimental techniques. If replication is already so complicated, how can the Darwinian theory of variation and selection be fairly simple and work? The answer is straightforward: Only the numbers of individuals, parents and progeny, are counted and the internal structure of the replicating entities, molecules, cells, organisms or societies, plays no role. Moreover replication in nature does never operate under conditions of excess of nucleic acids, because cellular division controls the number of genetic information carriers. Virus reproduction in the host cell is the only well known counterexample: Replication continues until the available resources are exhausted. Genetics like natural selection was discovered on a completely empirical basis. No explanations were at hand, neither for the observations nor for the deviations from the idealized ratios. In the second half of the twentieth century, an understanding of the Mendelian rules was provided by the molecular approach to heredity and at the same time a natural explanation was given for the deviations from them. Epigenetics was invoked as a *deus ex machina* in order to explain phenomena that escaped explanations by genetics. Recently, most of these previously strange observations found a straightforward explanation on the molecular level.

Modeling evolution by differential equations is well established and—although not yet available—a comprehensive stochastic approach will complement conventional modeling. Computer simulation of chemical kinetics of evolution is still in an early state and the simulations are lacking a more systematic approach. One still unsolved problem concerns the parameter space for molecular properties. A true wealth of data is currently obtained in genomics and related disciplines but only some of them will be useful for modeling evolution. The problem is how to find the pearls in a mess of rubbish. A second major limitation comes from the hyper-astronomical size of sequence space containing κ^{ℓ} different genotypes. At present, all genotypes can be considered routinely for short chain lengths up to $\ell = 10$. Extension to $\ell = 20$ provides already a challenge to numerical methods even for binary sequences, since the number of genotype interactions reaches a magnitude of $10^6 \times 10^6$. Nature itself provides the solution: Whole sequence spaces are never covered, clones are confined to tiny parts of the space and drift slowly by the mechanisms of evolution. The enormous size of sequence space, on the other hand, provides a convincing explanation for the existence of bacterial species: Their genomes are so far apart in sequence space that they do not require reproductive isolation. Despite steady migration they won't meet and merge.

The real complexity in biology arises from genotype–phenotype relations. Simple model landscapes are used in population genetics but they are far from being realistic and lead to wrong predictions. Even in the simplest case of in vitro evolution of molecules, the genotype–phenotype map requires understanding of the folding of biopolymer sequences into structures and the derivation of function from structure. This understanding as well as the predictive power of theories and algorithms in this field is still poor. Nevertheless, a few hints can be derived already from these simple model systems. Mappings involving biopolymers are commonly rugged, and optimization on rugged landscapes is an especially hard problem, because search strategies including the evolutionary approach involving populations are trapped very likely in some minor local optimum. The solution is obvious from the properties of biopolymers: nearby sequences may give rise to entirely different structures and functions but very often they lead to phenotypes that are indistinguishable by selection or *neutral*. Biological landscapes are not only rugged they are also characterized by a fairly high degree of neutrality. Sequence space is high-dimensional and this implies the existence of a large number of independent directions—commonly denoted as *orthogonal*. If the population during an adaptive walk in one direction is caught in a local optimum, there is a good chance that an escape from the trap is possible in some other direction where neutral sequences are found. A landscape structured in this way—ruggedness accompanied by a sufficiently high degree of neutrality—enables stepwise optimization: Short adaptive phases are supplemented by long quasi-stationary periods of random drift on neutral networks. At the same time the optimization process receives a memory on its past: Because of the high dimensionality of sequence space, individual trajectories are unique in the long run. A population migrating along a path collects series of stochastic events, which are not repeatable. Short term migration of populations, however, can be reproduced and this leads to an interesting kind of biological contingency: the recent past is repeatable whereas developments taking long time and collecting long series of mutations are not. Although the underlying dynamics is very different, the physicist will be reminded a little bit of irreversibility in thermodynamics.

References

Biebicher CK, Luce R (1992) In vitro recombination and terminal elongation of RNA by Q_β replicase. EMBO J 11(13):5129–5135

Biebricher C, Eigen M (1988) Kinetics of RNA replication by $Q\beta$ replicase. In: Domingo E, Holland J, Ahlquist P (eds) RNA genetics, vol I. CRC Press, Boca Raton, pp. 1–21

Biebricher CK, Eigen M, William C, Gardiner J (1983) Kinetics of RNA replication. Biochemistry 22:2544–2559

Biebricher CK, Eigen M, William C, Gardiner J (1984) Kinetics of RNA replication: plus-minus asymmetry and double-strand formation. Biochemistry 23:3186–3194

Biebricher CK, Eigen M, William C, Gardiner J (1985) Kinetics of RNA replication:competition and selection among self-replicating RNA species. Biochemistry 24:6550–6560

Blount ZD, Z C, Lenski RE (2008) Historical contingency an the evolution of a key innovation in an experimental population of *Escherichia coli*. Proc Natl Acad Sci USA 105:7898–7906

Blythe RA, McKane A (2007) Stochastic models of evolution in genetice, ecology and linguistics. J Stat Mech P07018:1–58

Brown D, Gold L (1996) RNA replication by $Q\beta$ replicase: a working model. Proc Natl Acad Sci USA 93:11558–11562

Cahill P, Foster K, Mahan DE (1991) Polymerase chain reaction and $Q\beta$ replicase amplification. Clin Chem 37:1482–1485

Darwin C (1859) On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life. John Murray, London

Demetrius L, Schuster P, Sigmund K (1985) Polynucleotide evolution and branching processes. Bull Math Biol 47:239–262

Derrida B, Peliti L (1991) Evolution in a flat fittness landscape. Bull Math Biol 53:355–382

Dimitrov RA, Zuker M (2004) Prediction of hybridization and melting for double-stranded nucleic acids. Biophys J 87:215–226

Dobzhansky T, Ayala FJ, Stebbins GL, Valentine JW (1977) Evolution. W. H. Freeman & Co., San Francisco

Domingo E (ed) (2005) Virus entry into error catastrophe as a new antiviral strategy. Virus Res 107(2):115–228

Drake JW (1993) Rates of spontaneous mutation among RNA viruses. Proc Natl Acad Sci USA 90:4171–4175

Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148:1667–1686

Edwards AW (1994) The fundamental theroem of natural selection. Biol Rev 69:443–474

Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523

Eigen M, Biebricher C (1988) Sequence space and quasispecies distribution. In: Domingo E, Holland J, Ahlquist P (eds) RNA Genetics, vol III, CRC Press, Boca Raton, pp 211–245

Eigen M, McCaskill J, Schuster P (1989) The molecular quasispecies. Adv Chem Phys 75:149–263

Fisher RA (1930) The genetical theory of natural selection. Oxford University Press, Oxford

Fontana W, Schuster P (1987) A computer model of evolutionary optimization. Biophys Chem 26:123–147

Fontana W, Schuster P (1998a) Continuity in evolution: on the nature of transitions. Science 280:1451–1455

Fontana W, Schuster P (1998b) Shaping space. The possible and the attainable in RNA genotype-phenotype mapping. J Theor Biol 194:491–515

Fontana W, Schnabl W, Schuster P (1989) Physical aspects of evolutionary optimization and adaptation. Phys Rev A 40: 3301–3321

Fontana W, Konings DAM, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. Biopolymers 33:1389–1404

Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comp Phys 22:403–434

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340–2361

Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58:35–55

Hamming RW (1986) Coding and information theory, 2nd edn. Prentice-Hall, Englewood Cliffs

Held DM, Greathouse ST, Agrawal A, Burke DH (2003) Evolutionary landscapes for the acquisition of new ligand recognition by RNA aptamers. J Mol Evol 57:299–308

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994a) Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie 125:167–188

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994b) Vienna RNA package. Current Version 1.8.4. Institute for Theoretical Chemistry, University of Vienna, Vienna.

Hosoda K, Matsuura T, Kita H, Ichihashi N, Tsukada K, Yomo T (2007) Kinetic analysis of the entire RNA amplification process by Qβ replicase. J Biol Chem 282:15516–15527

Huang Z, Szostak JW (2003) Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer. RNA 9:1456–1463

Huynen MA, Stadler PF, Fontana W (1996) Smoothness within ruggedness. The role of neutrality in adaptation. Proc Natl Acad Sci USA 93:397–401

Jones BL, Leung HK (1981) Stochastic analysis of a non-linear model for selection of biological macromolecules. Bull Math Biol 43:665–680

Jones BL, Enns RH, Rangnekar SS (1976) On the theory of selection of coupled macromolecular systems. Bull Math Biol 38:15–28

Joyce GF (2007) Forty years of in vitro evolution. Angew Chem Int Ed 46:6420–6436

Judson H (1979) The eighth day of creation. The makers of the revolution in biology. Jonathan Cape, London

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kupczok A, Dittrich P (2006) Determinants of simulated RNA evolution. J Theor Biol 238:726–735

Küppers B, Sumper M (1975) Minimal requirements for template recognition by bacteriophage Qβ replicase: approach to general RNA-dependent RNA synthesis. Proc Natl Acad Sci USA 72:2640–2643

Lenski RE, Rose MR, Simpson SC, Tadler SC (1991) Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. Am Nat 38:1315–1341

Leontis NB, Lescoute A, Westhof E (2006) The building blocks and motifs of RNA architecture. Curr Opin Struct Biol 16:279–287

Lincoln TA, Joyce GF (2009) Self-sustained replication of an RNA enzyme. Science 323:1229–1232

Malthus TR (1798) An Essay of the principle of population as it affects the future improvement of society. J. Johnson, London

Maxam A, Gilbert W (1977) A new method of sequencing DNA. Proc Natl Acad Sci USA 74:560–564

McCaskill JS (1984) A localization threshold for macromolecular quasispecies from continuously distributed replication rates. J Chem Phys 80:5194–5202

Mendel G (1866) Versuche über Pflanzen-Hybriden. Verhandlungen des naturforschenden Vereins in Brünn 4:3–47

Mendel G (1870) Über einige aus künstlicher Befruchtung gewonnenen Hieracium-Bastarde. Verhandlungen des naturforschenden Vereins in Brünn 8:26–31

Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci USA 58:217–224

Moore PB (1999) Structural motifs in RNA. Annu Rev Biochem 68:287–300

Nakaishi T, Iio K, Yamamoto K, Urabe I, Yomo T (2002) Kinetic properties of Qβ replicase, an RNA dependent RNA polymerase. J Biosci Bioeng 93:322–327

Okasha S (2008) Fisher's fundamental theorem of natural selection—a philosophical analysis. Br J Phil Sci 59:319–351

Phillipson PE, Schuster P (2009) Modeling by nonlinear differential equations. Dissipative and conservative processes. World Scientific Series on Nonlinear Science A, vol 69. World Scientific, Singapore

Price GR (1972) Fisher's "fundamental theorem" made clear. Annals of Human Genetics 36:129–140

Reidys C, Stadler PF, Schuster P (1997) Generic properties of combinatory maps. Neutral networks of RNA secondary structure. Bull Math Biol 59:339–397

Sanger F, Nicklen S, Coulson A (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schultes EA, Bartel DP (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. Science 289:448–452

Schuster P (2003) Molecular insight into the evolution of phenotypes. In: Crutchfield JP, Schuster P (eds) Evolutionary dynamics—exploring the interplay of accident, selection, neutrality, and function, Oxford University Press, New York, pp 163–215

Schuster P (2006) Prediction of RNA secondary structures: From theory to models and real molecules. Reports on Progress in Physics 69:1419–1477

Schuster P, Swetina J (1988) Stationary mutant distribution and evolutionary optimization. Bull Math Biol 50:635–660

Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc Lond B 255:279–284

Seneta E (1981) Non-negative matrices and Markov chains, 2nd edn. Springer-Verlag, New York

Spiegelman S (1971) An approach to the experimental analysis of precellular evolution. Q Rev Biophys 4:213–253

Stadler BRM, Stadler PF, Wagner GP, Fontana W (2001) The topology of the possible: formal spaces underlying patterns of evolutionary change. J Theor Biol 213:241–274

Thompson CJ, McBride JL (1974) On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules. Math Biosci 21:127–142

Verhulst P (1838) Notice sur la loi que la population pursuit dans son accroisement. Corresp Math Phys 10:113–121

Weissmann C (1974) The making of a phage. FEBS Lett 40:S10–S18

Wiehe T (1997) Model dependency of error thresholds: the role of fitness functions and contrasts between the finite and infinite sites models. Genet Res Camb 69:127–136

Zuker M (1989a) On finding all suboptimal foldings of an RNA molecule. Science 244:48–52

Zuker M (1989b) The use of dynamic programming algorithms in RNA secondary structure prediction. In: Waterman MS (ed) Mathematical methods for DNA sequences, CRC Press, Boca Raton, pp 159–184

Zuker M, Stiegler P (1981) Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. Nucl Acids Res 9:133–148

Zwillinger D (1998) Handbook of differential equations, 3rd edn. Academic Press, San Diego

What Is a Quasispecies? Historical Origins and Current Scope

Esteban Domingo and Peter Schuster

Abstract The quasispecies concept is introduced by means of a simple theoretical model that uses as little chemical kinetics and mathematics as possible but fully in the spirit of Albert Einstein who said: "Things should be made as simple as possible but not simpler." More elaborate treatments follow in the forthcoming chapters. It is shown that the most important results of the theory, in particular the existence of error thresholds, are not dependent on simplifying assumptions concerning the distribution of fitness values. Error thresholds are regularly found on landscapes with large and irregular scatter of fitness. After the introduction to theory, it will be shown how experimental data on the evolution of molecules or viruses may be fit to the theoretical model.

Contents

E. Domingo (✉)
Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Consejo Superior de Investigaciones Científicas (CSIC), Campus de Cantoblanco, 28049 Madrid, Spain
e-mail: edomingo@cbm.csic.es

E. Domingo
Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Barcelona, Spain

P. Schuster (✉)
Institut für Theoretische Chemie der Universität Wien, Währingerstraβe 17,
1090 Vienna, Austria
e-mail: pks@tbi.univie.ac.at

P. Schuster
The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Current Topics in Microbiology and Immunology (2016) 392: 1–22
DOI 10.1007/82_2015_453

Published Online: 30 August 2015

1 Evolution on the Cross-Road of Chemistry and Biology

A theory of evolution at the molecular level was conceived by Manfred Eigen (Eigen 1971; Eigen and Schuster 1977, 1978a, 1978b) through merging population dynamics with the knowledge of molecular biology. In this way, evolution could be integrated into chemical kinetics without losing the characteristic features of biology, in particular the role of genetic information stored in nucleic acid molecules and the nature of mutations were fully preserved. The key to evolution is reproduction, and at the level of DNA or RNA, reproduction is replication, which can be simply understood as copying of genetic information, which is error prone in general but can be error free or correct in a particular replication event. Modeling the basic property of molecular copying mechanisms, correct replication and mutation are parallel chemical reaction channels (Fig. 1) and accordingly, the same model assumptions hold for low and high mutation rates. The assumption that mutation is a byproduct of replication is straightforward for virus populations. One important consequence of this assumption is the factorization of fitness and mutation effects that is indispensable for the fitness landscape concept, which turned out to be very useful in understanding viral infection (see, e.g., Kouyos et al. 2012). In population genetics, for particular in the Crow–Kimura model of asexual evolution (Crow and Kimura 1970, p. 265), replication and mutation are considered as independent events, but there an entirely different mechanism is assumed: Mutation is not related to reproduction and occurs by external action during the whole lifetime of the organism. In order to be able to study evolution of molecules, environmental conditions may be kept constant in the model, but the extension to changing condition is straightforward.

General results derived from the theory of molecular evolution in constant environment are as follows:

(i) In error-free replication,

$$\mathbf{A} + \mathbf{X}_k \rightarrow 2\mathbf{X}_k; \quad k = 1, 2, \ldots \mathrm{n}, \tag{1}$$

selection in the sense of Charles Darwin's survival of the fittest results from chemical reactions approaching a stable stationary state, and is a straightforward consequence of the reaction mechanism. The approach toward stationarity is accompanied by optimization of the mean fitness of the population. Accordingly, the mean fitness of the population $\bar{f}$ is steadily increasing during the selection process, the selected molecular species $\mathbf{X}_m$ is the one with the highest fitness value: $f_m = \max(f_1, f_2, \ldots, f_n)$, and survival of the fittest is tantamount to optimization of the fitness of the entire population. The final result of selection is unique, a stationary homogeneous population containing only the fittest molecular species $\mathbf{X}_m$, no matter what the initial sequence distribution in the population was (provided it contained $\mathbf{X}_m$ at some, maybe very small amount).

(ii) Errors occurring during the replication process,

$$\mathbf{A} + \mathbf{X}_k \rightarrow \mathbf{X}_j + \mathbf{X}_k; \quad j, k = 1, 2, \ldots n;\ j \neq k, \tag{2}$$

produce mutations (Fig. 1) and change the features of correct replication kinetics discussed in (i). After sufficiently long time, the replication–mutation process approaches a stationary state, which does not consist of not a single fittest species $\mathbf{X}_m$ only but is a collective of replicating variants, symbolized by γ. The name "quasispecies" has been coined for this longtime sequence distribution in order to point at the fact that asexual reproduction like sexual reproduction forms genetic reservoirs, which provide pools of variants for future evolution. For a given parameter set, the quasispecies is unique: No matter what the population looked like initially the same longtime sequence distribution will result. The question of fitness optimization is more subtle than in the previous case (i): For most initial conditions, fitness will increase during the replication–mutation process and selection of the quasispecies γ is

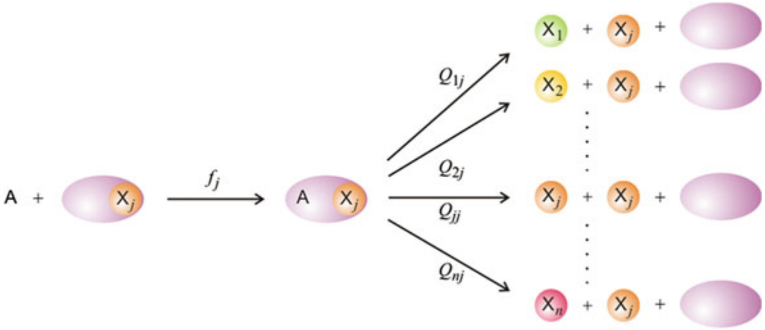


Fig. 1 A molecular view of replication and mutation. The replication device **E** (*violet*), commonly a single replicase molecule—as in polymerase chain reaction (PCR) or in many examples of simple viruses—or a multienzyme complex binds the template DNA or RNA molecule ($\mathbf{X}_j$, *orange*) in order to form a replication complex $\mathbf{E} \cdot \mathbf{X}_j$ and replicates with a rate parameter f_j. During the template-copying process, reaction channels leading to mutations are opened through replication errors. The reaction leads to a correct copy with frequency Q_{jj} and to a mutant $\mathbf{X}_k$ with frequency Q_{kj}. Commonly, we have $Q_{jj} \gg Q_{kj}$ for all $k \neq j$. In other words, correct replication dominates mutant formation. Stoichiometry of replication requires $\sum_{i=1}^{n} Q_{ij} = 1$, since the product has to be either correct or incorrect. The reaction is terminated by full dissociation of the replication complex. The sum of all activated monomers is denoted by **A**. A consequence of the model is factorization of the contributions from fitness and mutation: $w_{kj} = Q_{kj} \cdot f_j$

accompanied by an increase of the mean fitness $\bar{f}$ like in the mutation-free selection process. Nevertheless, situations are possible where this is not the case. Consider, for example, a homogeneous initial population consisting of the fittest genotype $\mathbf{X}_m$ only; then replication and mutation will create the quasispecies that contains other sequences too; these sequences have lower fitness; and mean fitness of the population is doomed to decrease during quasispecies formation: The paradigm of fitness optimization is invalidated. More complex cases can be constructed by choosing appropriate initial conditions, and then, the mean fitness may even change non-monotonously and go through a maximum or minimum during the approach toward the quasispecies γ. Optimization of mean fitness is not a global property in replication–mutation systems, but it is confined to a certain region of initial conditions. This region, the optimization region, is by far the largest part of the space of all possible initial conditions, and accordingly, optimization is observed in the majority of all replication–mutation experiments, artificial or in nature. Consequently, the Darwinian principle is a very powerful heuristic in the replication–mutation system, despite not being a universal law.

(iii) The population structure of quasispecies shows some regularities that turn out to be important for applications. The distribution of mutants is centered around a most frequent sequence called the "master sequence" $\mathbf{X}_m$, which commonly but not always is the sequence with largest fitness. The frequency of a mutant $\mathbf{X}_k$ in the stationary distribution is determined by two quantities: (a) the minimum number of point mutations or the Hamming distance d_{km} separating $\mathbf{X}_k$ from the master $\mathbf{X}_m$, and (b) the difference in the fitness values of the two sequences, $f_m - f_k$. Quasispecies theory explains also an empirical fact: The mean sequence of a population called the consensus sequence coincides with the master sequence unless the mutation rate is very high.

(iv) Considering the quasispecies as a function of the mutation rate p, a threshold phenomenon is predicted by the theory: Error-free replication leads to a quasispecies that contains the master sequence exclusively; the relative amount of the master sequence decreases gradually with increasing mutation rate p until a maximum mutation rate $p_{\max}$ is reached above which no stationary population exists; and no stable transfer of genetic information from generation to generation is possible (Fig. 2).

Chemical kinetics of RNA replication by means of virus-specific replicases is a rather complicated multistep process, but as Christof Biebricher (Biebricher 1983; Biebricher et al. 1983) has shown, conditions can be found under which the mechanism follows to a good approximation the simple autocatalytic overall kinetics mentioned above (Eq. 2). The conditions for the occurrence of the simple kinetics are excess material for RNA synthesis, here denoted by $\mathbf{A}$, and replicating enzyme an RNA-specific RNA polymerase in excess to the template polynucleotide $\mathbf{X}_k$. Few enzymes can support synthesis of infectious viral genomic RNA in the test

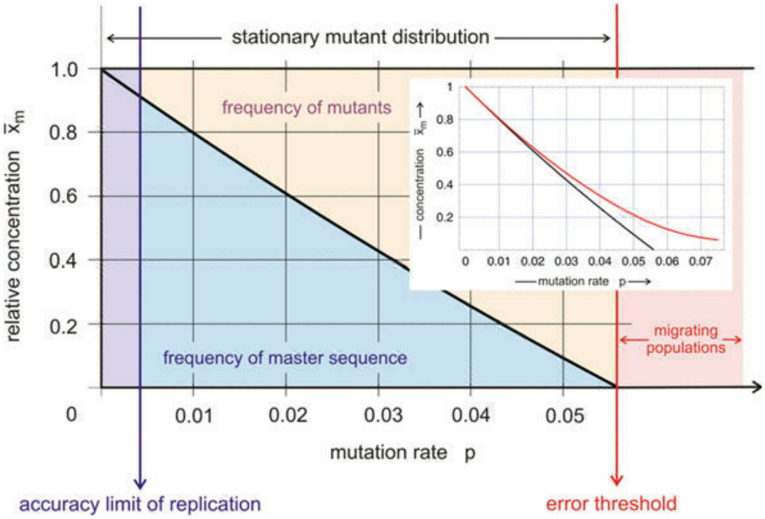


Fig. 2 The error threshold. The stationary frequency of the master sequence $\mathbf{X}_m$ is shown as a function of the local mutation rate p. In the approximation neglecting mutational backflow, the function $\overline{x_m}(p)$ is almost linear in the particular example shown here. In the insert, the approximation (*black*) is shown together with the exact solution (*red*). The 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, $l = 6$, $\sigma = 1.4131$

tube in the way Qβ replicase does (Biebricher 1983). Important advances in the enzymology of viral RNA replication have been made by Craig Cameron and his colleagues working with poliovirus polymerase. They devised simplified template-primer molecules that have allowed calculation of basic kinetic parameters for nucleotide incorporation, and the quantification of polymerase fidelity, an extremely important development, that will be discussed in several chapters of this book [(Castro et al. 2005) and references therein].

DNA replication involving some twenty or more proteins and enzymes is much more complex than RNA replication, but again suitable conditions can be found where the process can be approximated by simple autocatalysis (Eq. 2). Life cycles of viruses follow a complex multistep process with the overall stoichiometry $\mathbf{A} + \mathbf{X} \rightarrow n\,\mathbf{X}$ with n being the number of virions produced in one life cycle through the infection of a single cell. This process obeys the same laws as simple autocatalysis with the only difference that n reaction channels corresponding to n virions are chosen simultaneously rather than a single one (Fig. 1). Bacteria and more complex organisms adopt highly complex and perfectly controlled mechanisms of cell division that allow for modeling by simple autocatalysis since reproduction occurs at the level of cells rather than at the level of molecules. Simple autocatalysis (Eq. 2), direct or as overall kinetics, is the basis for the applicability of replication–mutation kinetics (Fig. 1) to the analysis of evolutionary phenomena sketched in the next section. It is not unimportant to verify the overall mechanism of reproduction in order to make sure that quasispecies theory in the form presented here is applicable, particularly in the case of cell-free evolution of molecules.

2 A Few Quantitative Relations

The replication–mutation mechanism sketched in Fig. 1 can be cast into ordinary differential equations as used in conventional chemical kinetics:

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \sum_{k=1}^{n} Q_{jk} f_k x_k - x_j \bar{f};\; j = 1, \ldots, n \quad \text{with} \quad \bar{f} = \sum_{i=1}^{n} f_i x_i \Big/ \sum_{i=1}^{n} x_i \tag{3}$$

The symbols used in this equation are as follows: The extensive variable $x_j = [\mathbf{X}_j]$ is the amount of species $\mathbf{X}_j$ present in the system expressed as concentration, but sometimes the usage of particle numbers $X_j = [\mathbf{X}_j]$ is of advantage, f_k is the fitness of species $\mathbf{X}_{k,}\,\bar{f}$ is the mean fitness of the population, and Q_{jk} finally is the frequency with which species $\mathbf{X}_j$ is produced as an error copy of $\mathbf{X}_k$ (Fig. 1). Considering the population as a whole, we introduce a total concentration or population size $c = \sum_{i=1}^{n} x_i$ and $\frac{\mathrm{d}c}{\mathrm{d}t} = \sum_{i=1}^{n} \frac{\mathrm{d}x_i}{\mathrm{d}t}$. Since all replication products are either correct or incorrect, we have a conservation relation $\sum_{j=1}^{n} Q_{jk} = 1$, the term $-\sum x_j \bar{f} = -c\bar{f}$ compensates exactly the population growth $\sum_{j=1}^{n} \sum_{k=1}^{n} Q_{jk} f_k x_k$, and the total concentration or population size is constant.

Exact solutions of Eq. (3) can be derived via an eigenvalue problem, and this implies that they are available in numerical form only (Jones et al. 1976; Eigen et al. 1989; Thompson and McBride 1974). For many purposes, however, approximations are perfect when they can be obtained in analytical form. We shall present here simple and illustrative expressions that are accurate enough for almost all practical purposes.

The most efficient approximation in this context is based on the assumption of "zero mutational backflow" (Eigen 1971): If the mutational flow from species $\mathbf{X}_k$ to species $\mathbf{X}_j$ is denoted by $\Phi_{k\rightarrow \mathrm{j}}$, we can symbolize the flow from the master to the mutation cloud by $\Phi_{m\rightarrow(\mathrm{j})}$ where (j) stands for $j = 1, \ldots, n; j \neq m$. Then, mutational backflow from the cloud to the master is written $\Phi_{m\leftarrow(\mathrm{j})}$. In other words, only mutations from the master sequence to the various mutants are allowed and all back mutations, $\mathbf{X}_j \rightarrow \mathbf{X}_m$, are forbidden. To be consistent, all mutations within the mutant cloud are neglected too. Equation (3) implies constant population size, and the modification of the mutation term requires a compensation in the flow term $-x_j\bar{f}$, which when introduced leads to a useful approximation for small mutation rates (see Chap. 4). Eigen leaves the flow term unchanged, and accordingly, the population size is no longer a constant. We shall denote this procedure that will turn out a posteriori as extremely successful as "phenomenological approach."

The mutation rate for single nucleotides, denoted by p, is assumed to be independent of the position on the sequence or, in other words, mutations are assumed to occur with the same frequency at each site. This approximation has been characterized as “uniform error rate” assumption, and it simplifies substantially the calculation of the mutation rates Q_{jk}. The probability to be copied correctly is the same for all sequences $\mathbf{X}_k$ and has the form

$$Q_{kk} = Q = (1-p)^l \quad \text{for all } k = 1,\ldots,n, \tag{4a}$$

where l is the chain length of the polynucleotide. This equation comprises the trivial fact that for error-free replication, $p = 0$, all copies are correct. Then, the fraction of correct replicas decreases monotonously with increasing mutation rate p (Fig. 2). The error containing copies $\mathbf{X}_j$ are obtained with the frequency

$$Q_{jk} = (1-p)^{l-d_{jk}} p^{d_{jk}} = Q\varepsilon^{d_{jk}} \quad \text{with} \quad \varepsilon = p/1-p \tag{4b}$$

The exponent d_{jk} is the Hamming distance between the two sequences $\mathbf{X}_j$ and $\mathbf{X}_k$. The Hamming distance is the (minimal) number of positions in which the two sequences differ. With these approximations and notations, it is straightforward to calculate the stationary concentration of the master sequence $\mathbf{X}_m$ that we denote by $\hat{x}_m^{(0)}$:

$$\hat{x}_m^{(0)} = \frac{Q-\sigma_m^{-1}}{1-\sigma_m^{-1}}\hat{c} \quad \text{with } \sigma_m = \frac{f_m}{\bar{f}_{-m}} \quad \text{and} \quad \bar{f}_{-m} = \frac{\sum_{i=1,i\neq m}^{n} f_i \hat{x}_i^{(0)}}{\hat{c}-\hat{x}_m^{(0)}}, \tag{5a}$$

where $\hat{c} = \sum_{i=1}^{n} \hat{x}_i^{(0)}$ and where we indicate stationary concentrations by a hat and the zero mutational backflow approximation by the superscript “(0)”. The mean fitness of all sequences except the master or, in other words, the mean fitness of the mutants is denoted by $\bar{f}_{-m}$, and finally, σ_m is the superiority of the master expressing the ratio of the fitness of the master and the mean fitness of the rest of the population. For the mutants $\mathbf{X}_j$, we obtain by the same token

$$\hat{x}_j^{(0)} = \varepsilon^{d_{jm}} \frac{f_m}{f_m - f_j}\hat{x}_m^{(0)} = \varepsilon^{d_{jm}} \frac{f_m^2(Q-\sigma_m^{-1})}{(f_m - f_j)(f_m - \bar{f}_{-m})}\hat{c}. \tag{5b}$$

In essence, the frequency at which a mutant is present in the quasispecies depends on two quantities: (i) the Hamming distance d_{jm} between sequence $\mathbf{X}_j$ and the master $\mathbf{X}_m$—the closer related to the master a sequence is the higher is its share in the stationary distribution—and (ii) the fitness difference between $\mathbf{X}_m$ and $\mathbf{X}_j$—the higher the fitness of the mutant, the higher is its frequency in the quasispecies. Accordingly, a quasispecies is not some arbitrary collective of variants but a highly ordered distribution with a master sequence in the center and mutant cloud surrounding it in sequence space.

Within the phenomenological approach, the stationary concentration of the master sequence as well as the concentrations of the mutations contains a factor $(Q - \sigma_m^{-1})$, which expresses the dependence of the concentrations on the mutation rate p, and vanishes if the condition $Q = \sigma_m^{-1}$ is fulfilled. The mutation rate $p_{\max}$ at which this happens is easily calculated:

$$Q = (1 - p_{\max})^l = \sigma_m^{-1} \quad \text{leading to} \quad p_{\max} \approx \frac{\ln \sigma}{l} \quad \text{or } l_{\max} \approx \frac{\ln \sigma}{p}. \tag{6}$$

The notation $p_{\max}$ points already at the fact that a conventional quasispecies exists only in the range $0 \leq p < p_{\max}$. As discussed in the next paragraph, at mutation rates higher than the threshold value, we get no information on the nature of the longtime solution of the replication–mutation system from the phenomenological approach. The phenomenon of a maximal mutation rate as described by Eq. (6) has been called the "error threshold": In order to guarantee evolutionary stability of the genetic information stored in nucleic acid sequences, the inaccuracy of replication must not exceed some critical value, which is defined by the sequence length l and the superiority of the master sequence σ_m. Alternatively, when the replication accuracy is given by some replication machinery, the error threshold defines some polynucleotide chain length $l_{\max}$ that cannot be exceeded without jeopardizing inheritance of genetic information. A comparison of the genome lengths of organisms from viroids to higher eukaryotes with replication machineries of largely different copying fidelity (Gago et al. 2009) yields a clear cut inverse relation between mutation rates and genome sizes. The error threshold concept has inspired an active field of antiviral research termed "virus transition into error catastrophe" or "lethal mutagenesis", consisting in virus extinction through defective viral gene products induced by excess of mutations (Chaps. 7 and 14).

It is important to stress that the existence of error thresholds is not restricted to a few model landscapes of fitness values. On the contrary, and as outlined in Chap. 5, all fitness distributions with strong scatter of the individual values show the threshold phenomenon and the width of the transition depends on the broadness of the dispersion of fitness values.

3 What Happens Beyond Error Threshold?

Since almost all analytical expressions of the quasispecies theory that are used in applications were derived within the frame provided by the phenomenological approach, we shall have a closer look on this assumption here. In particular, the error threshold in Eq. (6) has been derived from the application of the zero mutational backflow approximation to the mutation term, and therefore, it is legitimate to ask whether this result is a real phenomenon or an artifact of the approximation. First, we compare with the exact solution of Eq. (3) and consider the correct stationary concentration of the master sequence, $\bar{x}_m$, and its

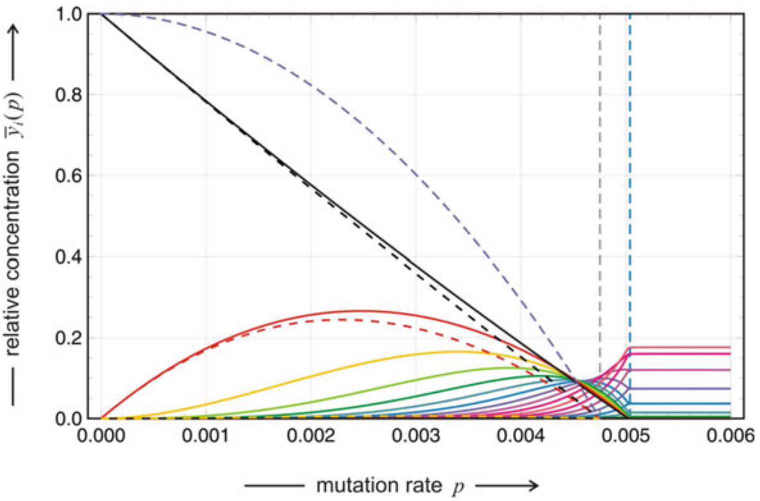


Fig. 3 The phenomenological approach. In the plot, the (*exact*) stationary solutions (*full lines*) are compared with the results derived from the phenomenological equations (*dashed lines*). The violet dashed line is the total relative concentration or population size in the phenomenological approach. The relative concentration of the master sequence (*black*) and the one-error class ($\bar{y}_1(p)$; *red*) agree well with the exact curves, whereas in case of all other error classes, the agreement is very poor (see, for example, $\bar{y}_2(p)$, *yellow*). The error threshold derived from the (*exact*) computation is $p_{\mathrm{cr}} = 0.00507$ (*dashed blue line* obtained from level-crossing and from class-merging as outlined in Chap. 5) and compares well to the value $p_{\mathrm{cr}} = 0.00475$ (*gray dashed line*) of the phenomenological approach. Choice of parameters: $l = 20$, single peak landscape with $f_m = 1.1$ and $f = 1.0$

approximation, $\hat{x}_m^{(0)}$. In order to facilitate the comparison, we assume that all sequences in a given error class have the same fitness. As illustrated in Fig. 3, there is little difference in the curves for master sequence and for the class of single point mutations. We remark that for sequences with two or more errors, this is not the case and the reasons for agreement and disagreement can be readily analyzed (Chap. 4 and Schuster 2012). Here, we use a plausibility argument: The dominant contribution to the mutational flow to one-error mutants comes from the master sequence, and it is taken into account correctly by the zero mutational backflow approximation. For mutants with two and more errors, the largest mutation flow comes from the mutants with one error less and mutations coming from other mutants are neglected. The take home lesson is that the phenomenological approach (5a) provides an excellent approximation for the master (5a) and the class of mutants carrying a single point mutation but only for them.

Whereas the stationary solution of the phenomenological approach is zero at the error threshold and becomes negative for $p > p_{\max}$ and thus no acceptable solution is available above threshold, the exact stationary solution becomes almost indistinguishable from the uniform distribution, which contains all variants—master sequence and mutants—at the same concentrations: $\bar{x}_m = \bar{x}_1 = \cdots = \bar{x}_n = \bar{c}/\kappa^l$, where κ denotes the size of the nucleotide alphabet. Numbers κ^l are "hyperastronomical," for example, there are 3 × 10^{120} different sequences for the smallest viroid genome with the size $l \approx 200$ in the natural alphabet ($\kappa = 4$). Population sizes in evolution experiments with replicating molecules are the largest that can be achieved, they may be as large as 10^{15}, and this implies that in a sample with uniform concentrations, we would be dealing with values in the range $\bar{x} \approx 10^{-105}$. This result simply tells that uniform concentrations cannot exist, the only conceivable alternative are clonal populations migrating in the huge space of all sequences (Derrida and Peliti 1991; Huynen et al. 1996), and indeed no stationary population can exist for $p > p_{\max}$. The enormous size of sequence space has another consequence that will be important in several other contributions in this book: In practice, all realistic populations of viruses, viroids, or polynucleotide molecules are local in sequence space and no global solutions exist in reality. Under favorable circumstances like in case of the quasispecies, the global solutions coincide with some local solutions for all practical purposes. In other words, we can use the results of the quasispecies concepts up to a certain Hamming distance, and for mutants being further away from the master sequence, the results have no practical meaning. As described in several chapters of this book, many important phenotypic changes in viruses depend on one or few mutations. That is, the transitions between different phenotypes depend on short distance migrations (small Hamming distances) within the locally occupied sequence space.

4 Origin of the Experimental Quasispecies Concept

When the development of quasispecies theory and hypercyclic organization was well advanced (Eigen 1971; Eigen and Schuster 1979), Charles Weissmann and his colleagues were in the process of founding "reverse genetics", a term first proposed by Weissmann four decades ago (Weissmann et al. 1977). The principles of reverse genetics were established using the RNA *Escherichia coli* bacteriophage Qβ as model system. Sol Spiegelman had achieved replication of Qβ RNA in the test tube (in the absence of cells or cells extracts) by using Qβ RNA as template, purified Qβ replicase and a host factor protein as the replicative machinery, and nucleoside triphosphates under adequate ionic conditions (particulary the presence of Mg^{2+}). This reaction mixture supported efficient synthesis and many-fold amplification of genomic Qβ RNA (Mills et al. 1967). The experimental system allowed in vitro RNA evolution experiments by introducing selective pressures (intercalating dyes, ribonuclease, etc.) during RNA synthesis. The in vitro evolution experiments of Spiegelman were one of the incentives of Eigen to develop his first mathematical treatment of error-prone replication (Eigen 1971).

Weissmann and colleagues applied the in vitro Qβ RNA synthesis with the purpose of producing specific, site-directed mutants in bacteriophage Qβ and to analyze the biological consequences of the mutation introduced. Until then, the standard procedure in genetics was to generate mutations at random by chemical mutagenesis and then select and study viruses (or cells) with a desired phenotypic trait. Hence, the term “reverse genetics” refers to the opposite strategy, to generate a precisely known mutation at a genomic site and to study its consequences.

One of the Qβ RNA genomic regions of interest at the time was the 3′-extracistronic (untranslated) region (3′-UTR) because of its conservation among related bacteriophages. Weissmann’s team developed the procedure to generate specific 3′-UTR mutations taking advantage of a unique property of Qβ replicase that allowed a stepwise Qβ RNA synthesis (by limiting the types of nucleoside triphosphates made available to the replicase at each synthesis step) until a desired position at the growing minus (complementary) strand was reached. Then, a mutagenic nucleotide analogue was incorporated at the selected position, and the complementary strand synthesis completed by allowing the reaction to proceed with the four standard nucleotides (Flavell et al. 1974). While a first mutation introduced at 3′-UTR position 16 was not viable despite the RNA being efficiently replicated by Qβ replicase (Flavell et al. 1974), a mutant RNA with an A → G transition at position 40 (termed mutant 40) from the 3′-end yielded progeny virus upon transfection on *E. coli* spheroplasts (Domingo et al. 1976). However, the mutant reverted to the wild type with a kinetics sufficiently slow to permit quantifying the proportion of mutant 40 and true wild-type revertants as a function of passage number. Competition between wild type and mutant 40 Qβ phages in *E. coli*, and reversion of the mutant to wild type in the course of serial passages, allowed Eduard Batschelet to calculate a mutation rate for the G → A reversion of 10^{-4} substitutions per genome doubling (Batschelet et al. 1976). Despite the calculation referring to a single genomic site, the value obtained is quite representative of mutation rates for RNA viruses that were subsequently estimated by other procedures (Steinhauer and Holland 1986; Eigen and Biebricher 1988; Ward and Flanegan 1992; Drake 1993; Drake and Holland 1999; Sanjuan et al. 2010).

In the course of the experiments on site-directed mutagenesis, the RNA of many biological clones of bacteriophage Qβ was analyzed by T_1-oligonucleotide fingerprinting, a method of nucleotide sequence sampling used at the time because cDNA synthesis, molecular cloning, and rapid sequencing were not available. Martin Billeter had sequenced and mapped in the Qβ genome the large T_1-oligonucleotides so that changes in the fingerprints could be interpreted by the occurrence of point mutations in the RNA (Billeter 1978). The result of the survey of biological clones was astonishing: Virtually, the RNA of each biological clone from a multiply passaged phage population differed in one to two nucleotide positions from the average sequence of the corresponding parental population. Experiments in which individual biological clones were passaged to generate heterogeneous populations led to the following conclusion “A Qβ phage population is in a dynamic equilibrium, with viable mutants arising at a high rate on the one hand, and being strongly selected against on the other. The genome of Qβ phage cannot be described as a defined unique structure, but rather as a weighted average of a large number of individual sequences ” (Domingo et al. 1978).

We know now that this statement applies to RNA viruses in general, as evidenced by work by many authors, initiated with foot-and-mouth disease virus (FMDV) by

Esteban Domingo and colleagues (Domingo et al. 1980; Sobrino et al. 1983) and vesicular stomatitis virus (VSV) by John Holland and colleagues (Holland et al. 1979, 1982). The early work on experimental quasispecies with bacterial, animal, and plant RNA viruses, as well as its impact for RNA genetics, was reviewed during the decade that followed the initial Qβ work (Domingo et al. 1985, 1988).

5 Quasispecies Theory and Experimental RNA Virus Quasispecies

During the 1970s, transdisciplinarity in science was less intense than today probably because of limited means of information exchange among practitioners of different scientific disciplines. Also, while theoretical physicists often asked general and fundamental issues of broad significance, experimental biologists focused on more detailed questions. Molecular biologists approached basic (but specific) problems of genome organization and expression, while virologists aimed at understanding viruses as disease agents. In the prevalent view at the time, disease mechanisms were unrelated to evolutionary concepts, a situation which is no longer tenable at present. Despite science compartmentalization, Manfred Eigen held a highly multidisciplinary Max Planck Winter Seminar at the Swiss village of Klosters, a stimulating scientific forum that continues until nowadays. In Winter 1978, Weissmann presented the experimental results on Qβ genome heterogeneity, and Eigen was thrilled to see the principles of quasispecies theory at work with a real virus. This key Klosters encounter and its impact have been described (Domingo et al. 1995, 2001; Holland 2006; Domingo et al. 2012), and it was the beginning of a stimulating collaboration between theoreticians and experimentalists that is partly responsible for the writing of the present book.

There is general agreement among theoretical biologists and experimental virologists that the Qβ population dynamics is directly connected with the generation of mutant distributions inherent to quasispecies theory. Nevertheless, a few population geneticists questioned the relevance of quasispecies theory for RNA viruses and some are still questioning it today. The main point in their argument goes as follows: RNA viruses are steadily evolving and cannot form stationary mutant distributions as required by quasispecies theory because there is not enough time for reaching a mutational equilibrium, and therefore, RNA virus populations cannot be seriously approached within the framework of quasispecies theory. Instead, the claim is raised that RNA virus heterogeneity is an extension of the classical concept of genetic polymorphism known in population biology since the 1960s, the only distinctive feature being that mutation rates of RNA viruses are orders of magnitude higher than standard mutation rates of cells. It is worth noting that classical population genetics is based on the assumption of small mutation rates, and in the classical concept of polymorphism, alleles that were not present in at least 10 % of individuals of a biological species were not counted as polymorphic sites (Spiess 1977). Deep sequencing applied to analyses of mutant spectra is presently revealing the presence of many minority genomes at much lower levels (for example, at the 0.1–1 % level that is the current range of cutoff values for reliable mutant frequencies), which are relevant players in the continuous dynamics of replacement of some minority subpopulations by others. This dynamics can certainly not be identified with genetic polymorphism in the classical sense.

Two points must be added here regarding the suitability of quasispecies as a theoretical framework for viral population dynamics. First, quasispecies theory is an extension of populations dynamics in order to make it fit for the incorporations of molecular data, and therefore, it is not incompatible with the classical Wright–Fisher models of mutation–selection balance (Wilke 2005). In fact, quasispecies theory enables going one step further due to the internal interactions within a mutant spectrum that converts the entire viral quasispecies into a unit of selection. Intra-mutant spectrum interactions can be of complementation (individuals display lower replicative fitness than the ensemble) or interference (infectivity of fully infectious individuals can be suppressed by the mutant ensemble). Several chapters of this book deal with intra-mutant spectrum interactions that frequently occur through *trans*-acting proteins expressed from viral genomes (see Chaps. 10 and 14). In the mutant distributions of quasispecies theory, the critical element that permits the quasispecies to behave as a unit of selection is the connectivity among closely related genomes established through frequent mutation. Cross talk among genomes is very intense when genomes are close neighbors in sequence space although more distant interactions may be also established thanks to the high connectivity of sequence space (Eigen and Biebricher 1988) (Chap. 4). Selection does not pull an individual but a connected set of individuals.

The equilibrium argument is worth being considered in more detail. It is commonplace stating that nothing on the Earth is at thermodynamic equilibrium because our planet as such is exposed to a steady flow of energy and entropy that goes from sun into outer space but nevertheless, there exists a plentitude of phenomena that are perfectly described by quasi-equilibrium theories. The notion quasi-equilibrium expresses the fact that a system looks as if it were in an equilibrium state within a certain time span but is changing on longer time scales. What matters here is not the fact of change in the long run, but the validity of timescale separation. In rigorous mathematical terms, the problem is characterized as center manifold reduction (Carr et al. 1981). In a nutshell, it says that the final state is approached in two phases (i) a fast process leading from the initial conditions to the center manifold, and (ii) a slow process during which the population moves along the center manifold to some final state. The question whether or not a quasi-equilibrium hypothesis can be justified, boils down to the existence or non existence of a center manifold (see Chap. 4). Here, we illustrate the concept of center manifold reduction by addressing its meaning for viruses and virus evolution: (i) The fast process is the formation of a mutation-balanced clan of sequences consisting of the master sequence and its most frequent mutants that are, in essence, derived from single or at maximum double nucleotide exchange mutations, and (ii) the selection-based and neutral drift of the population through the appearance of rare mutations and the occurrence of environmental changes. A necessary condition for the existence of center manifold and the meaningfulness of the quasispecies concept as well as any other quasi-equilibrium model is the formation of frequent mutants that occurs faster than the environment changes or, in other words, the environment is essentially constant during the formation of the mutation-balanced clans. In case of viroids and almost all RNA viruses, the postulation of a quasi-equilibrium seems to be on the safe side because the mutation rate is in the order of one per replication event (Gago et al. 2009). One remark about the frequency of mutations is important here: According to Eq. (4b), this frequency is proportional to the mutation rate raised to a power being the Hamming distance between the mutant sequence $\mathbf{X}_j$ and the master sequence $\mathbf{X}_m$, $\varepsilon^{d_{jm}}$. Since the mutation rate p—or $\varepsilon = p/(1-p)$—is small and the Hamming distance between to virus genomes can vary enormously, we shall be always dealing with a core of frequent mutants being at quasi-equilibrium with the master sequence and a plethora of rare variants whose appearance are a stochastic events. Neutrality with respect to fitness is another biological phenomenon that requires notions of stationarity, which are more sophisticated than simple quasi-equilibria (see Chap. 4).

Extensions of quasispecies theory to finite populations and variable fitness landscapes have been developed by many authors, including Eigen himself (Nowak and Schuster 1989; Alves and Fontanari 1998; Eigen 2000; Wilke et al. 2001; Nowak 2006; Ochoa 2006; Saakian and Hu 2006; Saakian et al. 2006; Takeuchi and Hogeweg 2007; Saakian et al. 2009; Park et al. 2010; Schuster 2010a, 2010b). Finite quasispecies populations in variable fitness landscapes are further treated in Chaps. 3 and 4 of this book. In theoretical biology, it is quite frequent to develop a deterministic model in mathematically solvable terms and then to extend it to real situations by introducing stochastic components in the model formulation. The same schools that initially opposed quasispecies suggested also that the heterogeneity of mutant spectra had been overestimated due to misincorporations during the enzymatic procedures involved in the preparation of molecular clones for nucleotide sequencing. As discussed elsewhere (Arias et al. 2001; Domingo et al. 2004), these arguments have proven incorrect since the impact of artifactual mutations can be controlled, and they have not affected significantly the heterogeneity measurements. Application of deep sequencing methodologies has amply confirmed the extensive genomic heterogeneity of RNA virus populations (Chap. 8), in agreement with the results obtained by classic biological or molecular cloning and Sanger sequencing.

Thus, quasispecies theory (despite its limitations, see last section) has provided the theoretical framework to interpret key characteristics of RNA viral populations: extreme genetic heterogeneity, mutant ensembles acting as a unit of selection, evolution (both short-term or intra-host and long-term or inter-host) understood fundamentally as replacement of genome subpopulations by others, and movements in sequence space as the basis to generate new phenotypes which are extremely relevant to virus biology. These aspects are amply discussed in different chapters of the present volume.

Despite the overwhelming evidence of quasispecies dynamics for RNA viruses in their natural environment, a few geneticists still advocate using undefined terms such as “variation” (or similar) rather than quasispecies. Avoidance of the term quasispecies may be acceptable provided scientists are aware of the nature of viral populations. However, unexpected side effects can derive from ambiguous terms. Millions of dollars and euros have gone into projects on antiviral and vaccine strategies doomed to failure because quasispecies dynamics was not incorporated as a relevant feature prior to the designs. Thus, there are pressing scientific (and even economic!) arguments to incorporate the term quasispecies in the fields of experimental and clinical virology. Several chapters of this book cover relevant aspects.

Different definitions of quasispecies have been used in physics, chemistry, and biology. In physics, quasispecies has been defined as a cloud in sequence space. To chemists, quasispecies are mutant swarms composed of related, nonidentical genomic sequences, the definition most familiar to virologists. To biologists, quasispecies is the target of selection, without the term implying a modification of the species concept in biology. In connection with the present volume, the most widely used quasispecies definition in virology is as follows: “a collection of related viral genomes subjected to a continuous process of genetic variation, competition, and selection that act as a unit of selection” (Domingo et al. 2012). Interesting new developments outside virology may require some more general definition of quasispecies that render it applicable to non-replicative systems. Some such developments are summarized next.

6 Extensions of Quasispecies to Non-viral Systems

Replication with a regular production of error copies is not privative of viruses, but it is a feature shared by cellular and subcellular systems endowed with replicative machineries that display limited template-copying fidelity. Connections have been made between viral quasispecies and cellular collectivities in two aspects: (i) error-prone replication with its ensuing competition dynamics among cells and (ii) collective behavior arising from interacting cell ensembles [for review see (Mas et al. 2010; Ojosnegros et al. 2011; Domingo et al. 2012; Solé et al. 2014)].

Concerning the first aspect, error-prone replication is prominent in mutator bacteria (which are characterized by mutation rates which are 10^2- to 10^3-fold larger than standard bacterial mutation rates) as well as in cancer cells. In both cases, enhanced mutation rates provide a selective advantage to the cells, either to expand the range of phenotypes for increased adaptability or to enhance cellular proliferation. A difference with viral quasispecies is in place here. The capacity of exploration of the sequence space available to viruses is far greater than the capacity exhibited by cells. The main reason is the difference in genome size between cells and viruses in relation to the usual population size of viruses and cellular organisms in nature. As an example, a viral genome of 10,000 nucleotides has a maximum of 3×10^4 single mutants, a number which is lower than the population size of many viruses, that can attain 10^{10} to 10^{12} particles per infected individual. All single mutants and many multiple mutants are potentially present (excluding fitness effects) in a viral population infecting a single host. In contrast, the potential number of single mutants in a mammalian genome will approach 10^{10}, a far larger value than the population size of mammalian species. These and other parameters [population heterogeneity, number of mutations needed for a biological change, and fecundity or the capacity to generate progeny; see further discussion in (Domingo et al. 2012)] render quasispecies a far more effective adaptive strategy for viruses than for cells, even if their population dynamics follow similar principles.

Cancer cell dynamics has been extensively studied both theoretically and from a clinical perspective. Martin Nowak reviewed the conceptual origins of cancer viewed as a genetic disease, the types of genetic lesions that render cancer cells an error-prone system that favors tumor progression, and the basic mathematics of tumor cell proliferation (Nowak 2006). Very early work emphasized the relevance of cancer cell heterogeneity, clonal evolution, and the consideration of tumor metastasis as an adaptive process (Nowell 1976; Nicolson 1987). Recent models view cancer as cell collectivities that have restricted their functional genetic information to that required for cell integrity and proliferation, but free of the constraints inherent to cellular differentiation (Gatenby and Frieden 2002; Solé et al. 2014). This is reminiscent of the result of evolution of Qβ RNA in the test tube (the classic Spiegelman–Weissmann passage experiments discussed earlier) in which maintenance of RNA infectivity was no longer needed, and the only remaining requirement to the RNA was to replicate. In the words of the authors: “What will happen to the RNA molecules if the only demand made is the Biblical injunction, multiply, with the biological proviso that they do so as rapidly as possible?” (Mills et al. 1967). The result was selection of RNAs with extensive deletions than were adapted to bind efficiently to the replicase and to undergo rapid replication; infectivity was rapidly lost.

The search for the minimum requirements for cancer cell proliferation may help providing the basis to produce an error catastrophe in cancer (Solé and Deisboeck 2004; Fox and Loeb 2010), following the strategy under investigation for viruses (Chaps. 7 and 14). Tumor cell heterogeneity is a determinant of adaptability and limits the efficacy of anticancer drugs, because of the ease of selection of drug-escape mutant cells through several molecular mechanisms. The problem of treatment failure due to selection of drug resistance within a tumor cell population is very similar to that faced in the case of viral infections (Chaps. 12 and 14), and strategies alternative to the standard anticancer chemotherapeutic protocols have been suggested (Gatenby et al. 2009; Luo et al. 2009). In the course of adaptive RNA virus evolution in natural environments, in particular during intra-host expansions of viral populations, mutation rates are expected to remain constant, except in rare cases in which a specific fidelity mutation may be incorporated in the viral polymerase gene and become dominant. In contrast, the cascade of molecular events during cancer progression, mainly mutations that increase the cell division rate and mutations that increase the cellular mutation rate (that include tumor suppressor genes, oncogenes and genetic instability genes), is more complex. As a consequence, and interestingly, mutation rates are unlikely to remain constant through tumor progression. Evolutionary dynamics under constant versus increasing mutation rates deserves further theoretical and experimental investigation.

Concerning collective behavior due to cell to cell interactions, they have been also recognized within tumors, in particular regarding competition between fitter chemosensitive cells and less fit, drug-resistant cells during therapy (Gatenby et al. 2009). A parallel with the internal interactions among components of mutant spectra in viruses (Chaps. 10 and 14) has been also found in the behavior of bacterial collectivities [(Ojosnegros et al. 2011) and references therein]. In particular, quorum sensing in bacteria has been proposed as a factor to modulate virulence, so that an important biological trait is the result of cooperative interactions among individuals.

Recently, a striking conceptual parallelism has been established between the conformational heterogeneity of prions and viral quasispecies (Li et al. 2010; Weissmann et al. 2011; Weissmann 2012). Prions are infectious agents composed only of protein, without a nucleic acid. They are propagated through transmission of a misfolded form of a cell-coded protein (Castilla et al. 2008; Barria et al. 2009). Despite having the same amino acid sequence, distinguishable prion “strains” are characterized by different conformations. A “mutation” in a prion represents a change in conformation that may occur through environmental changes and confer altered pathogenic potential and drug sensitivity (Ghaemmaghami et al. 2009; Mahal et al. 2010). As in the case of viruses, both drug-resistant and drug-dependent prions can be selected (Oelschlegel and Weissmann 2013). Prion populations are heterogeneous in the sense that they include subsets of protein molecules with minority conformations, a parallel with the minority components of mutant spectra of viral quasispecies (Weissmann et al. 2011; Bateman and Wickner 2013; Vanni et al. 2014). Conformational variants can be either positively selected or remain in equilibrium with other variants (conformomers). In remarkable parallelism with viral quasispecies, the population size of a prion subjected to amplification can be a determinant of its evolution, and bottleneck transfers lead to reduced “replicative fitness” of prions (Vanni et al. 2014). How can such a parallel Darwinian behavior of a replicative and a non-replicative system originate? Mutations in genetic systems are the result of elementary molecular fluctuations events that determine base mispairings. Similar fluctuations may influence amino acid–amino acid interactions that determine protein conformation. A specific conformation may act as a nucleation point for the conversion of neighbor proteins into a similar conformation (Bernacki and Murphy 2009). Certainly, it would be extremely interesting to develop a theory for Darwinian evolution in non-genetic systems, search for protein transitional states and Darwinian behavior in proteins other than prions, and define the molecular basis of collective conformational transitions in protein ensembles. Such research may open new avenues for the control of neurological disease. Thus, the basic concepts emanating from quasispecies are permeating many domains of biological sciences, a demonstration of the experiment-provoking power of quasispecies theory.

7 Limitations and Strengths of the Quasispecies Concept

The concept of quasispecies refers to the level of populations in a homogeneous or mostly homogeneous environment, and this need not be realistic in case of real virus infections in heterogeneous host populations. In a sufficiently diverse population, for example, the master sequence in one host need not coincide with the master sequence in another host. Heterogeneity of environments may be important for many other reasons, but these are not quasispecies specific problems. Theoretical epidemiology is struggling with the effects of complex environments as well, and this for rather long time already.

In its current form, quasispecies theory does not account for stochastic effects. Small particle numbers up to several hundred infectious units can be important because of the autocatalytic nature of the replication process, and special stochastic effects such as incomplete packaging of genome segments in viruses with a segmented genome such as influenza A or early extinction due to replication accidents may need to be taken into account by virus-specific modeling. The major problem with stochastic modeling is not of principal nature. It concerns the numerical simulation techniques that are extremely time consuming even for medium-size systems and the unavailability of analytical methods for many component systems. The current best way to overcome this problem is to sacrifice generality and to construct virus group-specific stochastic models.

Although conceptually rooted in the same grounds as population genetics, the theory of the quasispecies has several advantages and can be more easily extended:

(i) The model is constructed at the molecular level, and this provides a frame that can be readily adapted to the desired level of details. The replication–mutation reaction (2) comprises the simplest conceivable mechanism. Provided one does not spare the effort, a detailed viral mechanism, for example, the RNA bacteriophage replication kinetics (Biebricher et al. 1983), could be introduced into the kinetic differential equations, and numerical analysis based on kinetic differential equations would be possible. By the same token, entirely different forms of reproduction can be incorporated, for example, the proliferation of prions (Weissmann et al. 2011) or mitosis of cancer cells (Gatenby and Frieden 2002; Gatenby et al. 2009). In the future, it will be desirable and possible to integrate complex regulation of gene expression into molecular models. Important examples are RNA-based epigenetic mechanisms.

(ii) Inherent in the molecular replication–mutation mechanism that understands mutation as a parallel reaction channel to correct copying is the possibility to factorize the selective value into one factor coming from the spectrum of fitness values and a second factor containing mutation frequencies. This handle on separability is not only an important tool for theoretical work but it also suggests to adopt two different strategies in the development of antiviral agents: reduction of fitness through interfering, for example, with the binding of the virus to the replication machinery or increase in mutation rate in order to drive the replicating virus beyond the error threshold.

(iii) The conventional quasispecies concept is based on the assumption that populations have reached a stationary or a quasi-stationary state. Although the validity of this assumption may be questionable, replication–mutation dynamics provides an appropriate tool for rigorous tests based on the center manifold theorem. The time a population system requires for a close approach to quasi-stationarity is well defined as a first passage time in the stochastic model and has been studied in the past [for an example with more references on this topic see the publication by Marin et al. (2012)]. Nevertheless, more detailed investigations are required to adapt the quasispecies theory questions concerning appropriate times, for example, the optimal duration of patient treatments.

(iv) Virus evolution is determined by the fitness landscape, which may be dynamic in a changing environment. Given a high degree of ruggedness as follows form empirical data, e.g., Kouyos et al. (2012) or the experience with biopolymer landscapes (Schuster 2006) quasispecies will commonly be unstable against changes in mutation rates. Quasispecies theory makes the prediction that migration into other regions in sequence space where "strong quasispecies" can be formed makes the population evolutionary stable (Chap. 4).

The application of quasispecies theory to the understanding of virus dynamics in infected organisms has opened the way to a rational design of antiviral interventions which until now have been basically an empirical endeavor. The increasing applicability of next generation, deep sequencing of viral populations as they replicate in their hosts, has unveiled the complexity of natural mutant spectra and estimates of relative fitness levels of minority genomes (Chap. 8). These analyses should permit personalized treatments with selected standard inhibitors and virus-specific mutagenic agents, used sequentially or in combination (Chap. 14). These are important practical consequences derived from the new understanding of viral populations that became clear when populations were examined under the focus of quasispecies theory.

Acknowledgments Peter Schuster wishes to acknowledge support by the University of Vienna, Wien, Austria and the Santa Fe Institute, Santa Fe, USA. Esteban Domingo acknowledges the support of grants BFU 2011-23604 and SAF2014-52400-R from Ministerio de Economía y Competitividad, grant S2013/ABI-2906 (PLATESA) from Comunidad Autónoma de Madrid, CIBERehd (Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas) which is funded by Instituto de Salud Carlos III, and Fundación Ramón Areces.

References

Alves D, Fontanari JF (1998) Error threshold in finite populations. Phys Rev E 57:7008–7013

Arias A, Lazaro E, Escarmis C, Domingo E (2001) Molecular intermediates of fitness gain of an RNA virus: characterization of a mutant spectrum by biological and molecular cloning. J Gen Virol 82:1049–1060

Barria MA, Mukherjee A, Gonzalez-Romero D, Morales R, Soto C (2009) De novo generation of infectious prions in vitro produces a new disease phenotype. PLoS Pathog 5(5):e1000421

Bateman DA, Wickner RB (2013) The [PSI+] prion exists as a dynamic cloud of variants. PLoS Genet 9(1):e1003257

Batschelet E, Domingo E, Weissmann C (1976) The proportion of revertant and mutant phage in a growing population, as a function of mutation and growth rate. Gene 1:27–32

Bernacki JP, Murphy RM (2009) Model discrimination and mechanistic interpretation of kinetic data in protein aggregation studies. Biophys J 96(7):2871–2887

Biebricher CK (1983) Darwinian selection of self-replicating RNA molecules. Evol Biol 16:1–52

Biebricher CK, Eigen M, Gardiner WC Jr (1983) Kinetics of RNA replication. Biochemistry 22:2544–2559

Billeter M (1978) Sequence and location of large RNase T_1 oligonucleotides in bacteriophage Qβ RNA. J Biol Chem 253:8381–8389

Carr J (1981) Applications of centre manifold theory. Springer, Berlin

Castilla J, Morales R, Saa P, Barria M, Gambetti P, Soto C (2008) Cell-free propagation of prion strins. EMBO J 27(19):2557–2566

Castro C, Arnold JJ, Cameron CE (2005) Incorporation fidelity of the viral RNA-dependent RNA polymerase: a kinetic, thermodynamic and structural perspective. Virus Res 107:141–149

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York (Reprinted at The Blackburn Press, Caldwell, NJ, 2009)

Derrida B, Peliti L (1991) Evolution in a flat fitness landscape. Bull Math Biol 53:355–382

Domingo E, Flavell RA, Weissmann C (1976) In vitro site-directed mutagenesis: generation and properties of an infectious extracistronic mutant of bacteriophage Qβ. Gene 1:3–25

Domingo E, Sabo D, Taniguchi T, Weissmann C (1978) Nucleotide sequence heterogeneity of an RNA phage population. Cell 13:735–744

Domingo E, Davila M, Ortin J (1980) Nucleotide sequence heterogeneity of the RNA from a natural population of foot-and-mouth-disease virus. Gene 11:333–346

Domingo E, Martínez-Salas E, Sobrino F, de la Torre JC, Portela A, Ortín J, López-Galindez C, Pérez-Breña P, Villanueva N, Nájera R, VandePol S, Steinhauer D, DePolo N, Holland JJ (1985) The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance–a review. Gene 40:1–8

Domingo E, Holland JJ, Ahlquist P (1988) RNA genetics. CRC Press, Boca Raton

Domingo E, Holland JJ, Biebricher C, Eigen M (1995) Quasispecies: the concept and the word. In: Gibbs A, Calisher C, García-Arenal F (eds) Molecular evolution of the viruses. Cambridge University Press, Cambridge, pp 171–180

Domingo E, Biebricher C, Eigen M, Holland JJ (2001) Quasispecies and RNA virus evolution: principles and consequences. Landes Bioscience, Austin

Domingo E, Ruiz-Jarabo CM, Arias A, Garcia-Arriaza JF, Escarmís C (2004) Quasispecies dynamics and evolution of foot-and-mouth disease virus. In: Sobrino F, Domingo E (eds) Foot-and-mouth disease. Horizon Bioscience, Wymondham

Domingo E, Sheldon J, Perales C (2012) Viral quasispecies evolution. Microbiol Mol Biol Rev 76:159–216

Drake JW (1993) Rates of spontaneous mutation among RNA viruses. Proc Natl Acad Sci USA 90:4171–4175

Drake JW, Holland JJ (1999) Mutation rates among RNA viruses. Proc Natl Acad Sci USA 96:13910–13913

Eigen M (1971) Self-organization of matter and the evolution of biological macromolecules. Die Naturwissenschaften 58:465–523

Eigen M (2000) Natural selection: a phase transition? Biophys Chem 85:101–123

Eigen M, Schuster P (1977) The hypercycle—a principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften 64:541–565

Eigen M, Schuster P (1978a) The hypercycle—a principle of natural self-organization. Part B: The abstract hypercycle. Naturwissenschaften 65:7–41

Eigen M, Schuster P (1978b) The hypercycle—a principle of natural self-organization. Part C: The realistic hypercycle. Naturwissenschaften 65:341–369

Eigen M, Schuster P (1979) The hypercycle. A principle of natural self-organization, Springer, Berlin

Eigen M, Biebricher CK (1988) Sequence space and quasispecies distribution. In: Domingo E, Ahlquist P, Holland JJ (eds) RNA genetics. CRC Press Inc, Boca Raton, FL., pp 211–245

Eigen M, McCaskill J, Schuster P (1989) The molecular quasispecies. Adv Chem Phys 75:149–263

Flavell RA, Sabo DL, Bandle EF, Weissmann C (1974) Site-directed mutagenesis: generation of an extracistronic mutation in bacteriophage Q beta RNA. J Mol Biol 89:255–272

Fox EJ, Loeb LA (2010) Lethal mutagenesis: targeting the phenotype in cancer. Semin Cancer Biol 20(5):353–359

Gago S, Elena SF, Flores R, Sanjuan R (2009) Extremely high mutation rate of a hammerhead viroid. Science 323:1308

Gatenby RA, Frieden BR (2002) Application of information theory and extreme physical information to carcinogenesis. Cancer Res 62(13):3675–3684

Gatenby RA, Silva AS, Gilles RJ, Frieden BR (2009) Adaptive therapy. Cancer Res 69(11):4894–4903

Ghaemmaghami S, Ahn M, Lessard P, Giles K, Legname G, DeArmond SJ, Prusiner SB (2009) Continuous quinacrine treatment results in the formation of drug-resistant prions. PLoS Pathog 5(11):e1000673

Holland JJ (2006) Transitions in understanding of RNA viruses: an historical perspective. Curr Top Microbiol Immunol 299:371–401

Holland JJ, Grabau EA, Jones CL, Semler BL (1979) Evolution of multiple genome mutations during long-term persistent infection by vesicular stomatitis virus. Cell 16:495–504

Holland JJ, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S (1982) Rapid evolution of RNA genomes. Science 215:1577–1585

Huynen MA, Stadler PF, Fontana W (1996) Smoothness within ruggedness: the role of neutrality in adaptation. Proc Natl Acad Sci USA 93:397–401

Jones BL, Enns RH, Rangnekar SS (1976) On the theory of selection of coupled macromolecular systems. Bull Math Biol 38:15–28

Kouyos RD, Leventhal GE, Hinkley T, Haddad M, Whitcomb JM, Petropoulos CJ, Bonhoeffer S (2012) Exploring the complexity of the HIV-1 fitness landscape. PLoS Genet 8:e1002551

Li J, Browning S, Mahal SP, Oelschlegel AM, Weissmann C (2010) Darwinian evolution of prions in cell culture. Science 327:869–872

Luo J, Solimini NL, Elledge SJ (2009) Principles of cancer therapy: oncogene and non-oncogene addiction. Cell 136(5):823–837

Mahal SP, Browning S, Li J, Suponitsky-Kroyter I, Weissmann C (2010) Transfer of a prion strain to different hosts leads to emergence of strain varriants. Proc Natl Acad Sci USA 107 (52):22653–22658

Marin A, Tejero H, Nuño JC, Montero F (2012) Characteristic time in quasispecies evolution. J Theor Biol 303:25–32

Mas A, Lopez-Galíndez C, Cacho I, Gomez J, Martínez MA (2010) Unfinished stories on viral quasispecies and Darwinian views of evolution. J Mol Biol 397(4):865–877

Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci USA 58:217–224

Nicolson GL (1987) Tumor cell instability, diversification, and progression to the metastatic phenotype. From oncogene to oncophetal expression. Cancer Res 47(6):1473–1487

Nowak MA (2006) Evolutionary Dynamics. The Belknap Press of Harvard University Press, Cambridge

Nowak M, Schuster P (1989) Error thresholds of replication in finite populations mutation frequencies and the onset of Muller's ratchet. J Theor Biol 137:375–395

Nowell P (1976) The clonal evolution of tumor cell populations. Science 194:23–28

Ochoa G (2006) Error thresholds in genetic algorithms. Evol Comput 14:157–182

Oelschlegel AM, Weissmann C (2013) Acquisition of drug resistance and dependence by prions. PLoS Pathog 9:e1003158

Ojosnegros S, Perales C, Mas A, Domingo E (2011) Quasispecies as a matter of fact: viruses and beyond. Virus Res 162:203–215

Park JM, Munoz E, Deem MW (2010) Quasispecies theory for finite populations. Phys Rev 81:011902

Saakian DB, Hu CK (2006) Exact solution of the Eigen model with general fitness functions and degradation rates. Proc Natl Acad Sci USA 103:4935–4939

Saakian DB, Munoz E, Hu CK, Deem MW (2006) Quasispecies theory for multiple-peak fitness landscapes. Phys Rev E 73:041913

Saakian DB, Biebricher CK, Hu CK (2009) Phase diagram for the Eigen quasispecies theory with a truncated fitness landscape. Phys Rev 79:041905

Schuster P (2006) Prediction of RNA secondary structures: from theory to models and real molecules. Rep Prog Phys 69:1419–1477

Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R (2010) Viral mutation rates. J Virol 84:9733–9748

Schuster P (2010a) Genotypes and phenotypes in the evolution of molecules. In: Caetono-Anolles G (ed) Evolutionary genomics and systems biology. Wiley-Blackwell, New Jersey, pp 123–152

Schuster P (2010b) Mathematical modeling of evolution. Solved and open problems. Theory Biosci 130:71–89

Schuster P (2012) Evolution on ‘realistic’ fitness landscapes. Phase transitions, strong quasispecies, and neutrality. Santa Fe Institute working paper #12-06-006, Santa Fe Institute, Santa Fe

Sobrino F, Dávila M, Ortín J, Domingo E (1983) Multiple genetic variants arise in the course of replication of foot-and-mouth disease virus in cell culture. Virology 128:310–318

Solé RV, Deisboeck TS (2004) An error catastrophe in cancer? J Theor Biol 228(1):47–54

Solé RV, Valverde S, Rodriguez-Caso C, Sardanyés J (2014) Can a minimal replicating construct be identified as the embodiment of cancer? BioEssays 36:503–512

Spiess EB (1977) Genes in populations. Wiley, New York

Steinhauer DA, Holland JJ (1986) Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. J Virol 57:219–228

Takeuchi N, Hogeweg P (2007) Error-threshold exists in fitness landscapes with lethal mutants. BMC Evol Biol 7(15):author reply 15

Thompson CJ, McBride JL (1974) On Eigen’s theory of the self-organization of matter and the evolution of biological macromolecules. Math Biosci 21:127–142

Vanni I, Di Bari MA, Pirisinu L, D’Agostino C, Agrimi U, Nonno R (2014) In vitro replication highlights the mutability of prions. Prion 8:154–160

Ward CD, Flanegan JB (1992) Determination of the poliovirus RNA polymerase error frequency at eight sites in the viral genome. J Virol 66:3784–3793

Weissmann C (2012) Mutation and selection of prions. PLoS Pathog 8:e1002582

Weissmann C, Tanaguchi T, Domingo E, Sabo D, Flavell RA (1977) Site-directed mutagenesis as a tool in genetics. In: Schultz J, Brada Z (eds) Genetic manipulation as it affects the cancer problem. Academic Press, New York, pp 11–36

Weissmann C, Li J, Mahal SP, Browning S (2011) Prions on the move. EMBO Rep 12:1109–1117

Wilke CO (2005) Quasispecies theory in the context of population genetics. BMC Evol Biol 5:44

Wilke CO, Ronnewinkel C, Martinetz T (2001) Dynamic fitness landscapes in molecular evolution. Phys Rep 349:395–446

Quasispecies on Fitness Landscapes

Peter Schuster

Abstract Selection–mutation dynamics is studied as adaptation and neutral drift on abstract fitness landscapes. Various models of fitness landscapes are introduced and analyzed with respect to the stationary mutant distributions adopted by populations upon them. The concept of quasispecies is introduced, and the error threshold phenomenon is analyzed. Complex fitness landscapes with large scatter of fitness values are shown to sustain error thresholds. The phenomenological theory of the quasispecies introduced in 1971 by Eigen is compared to approximation-free numerical computations. The concept of strong quasispecies understood as mutant distributions, which are especially stable against changes in mutations rates, is presented. The role of fitness neutral genotypes in quasispecies is discussed.

Contents

P. Schuster
Institut für Theoretische Chemie der Universität Wien,
Währingerstraße 17, 1090 Vienna, Austria

P. Schuster (✉)
The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
e-mail: pks@tbi.univie.ac.at; pks@santafe.edu

Current Topics in Microbiology and Immunology (2016) 392: 61–120
DOI 10.1007/82_2015_469

Published Online: 24 November 2015

1 Fitness Landscapes

The idea of an *adaptive landscape* or *fitness landscape* is commonly attributed to Wright (1931, 1932, 1988) who introduced it as a metaphor underlying the illustration of evolution as hill-climbing on a multi-peak potential (hyper)surface.[1] According to McCoy (1979), the concept of evolution as an adaptive process on a fitness landscape has been used the first time much earlier by Janet (1895) in order to provide an explanation for the lack of intermediate forms of species in the fossil record. Wright's *shifting balance* model of evolution consists of three phases: (i) random genetic drift splitting the global population in subpopulations, (ii) selection within subpopulations, and (iii) selection between subpopulations. The mean fitness of the population is assumed to decrease during phase (i) and to increase during phases (ii) and (iii). Wright's illustration visualizes a genotype recombination space with several alleles per locus. Before Watson and Crick published their model of the molecular structure of DNA (Watson and Crick 1953), the process creating mutations was not an integral part of the theory of evolution, operated like a *deus ex machina* unseen in the background, and could not be systematically related to moves in genotype space. Wright's fitness landscape is mapped upon a two-dimensional sketch of genotype space and contains many local peaks upon which his model of evolution is approaching the highest fitness optimum. Wright's model and the metaphor have been heavily debated in the following years [see, e.g., Provine (1986), Ruse (1996), and for a more recent well-founded analysis of Wright's landscape concept, we recommend Skipper (2004)]. Here, we shall understand the notion of landscape in a rigorous way as a mapping of genotypes onto nonnegative real numbers representing the fitness parameters, which enter the deterministic or stochastic dynamical systems describing the evolution of populations.

Evolution as an adaptive walk. Two basic elements define an adaptive walk: (i) a potential surface built upon genotype space and (ii) a move set being the collection of allowed changes of genotypes. Although Wright himself stresses multi-dimensionality of fitness landscapes as they are built upon genotype or sequence space,[2] which is a support of very high dimension, his landscapes, however, are always sketched on a continuous two-dimensional caricature of sequence space (Wright 1988, p. 117). Fisher (1941) challenges the usefulness of the two-dimensional metaphor by remarking correctly that the number of local optima decreases when the dimensionality of the support is raised, and in view of the enormously high dimensionality of genotype space, a single-peak landscape

[1] The expression *hypersurface* points at the fact that fitness landscapes are surfaces in high-dimensional space. Since we shall be dealing here almost exclusively with such high-dimensional objects, we drop the prefix 'hyper.'

[2] The genotype space in Wright's seminal paper (Wright 1932) is a space of genes, whereas we use virus genomes as elements of genotypes space. Accordingly, genotype space is identical with the space of DNA or RNA sequences of the chain length of virus genomes.

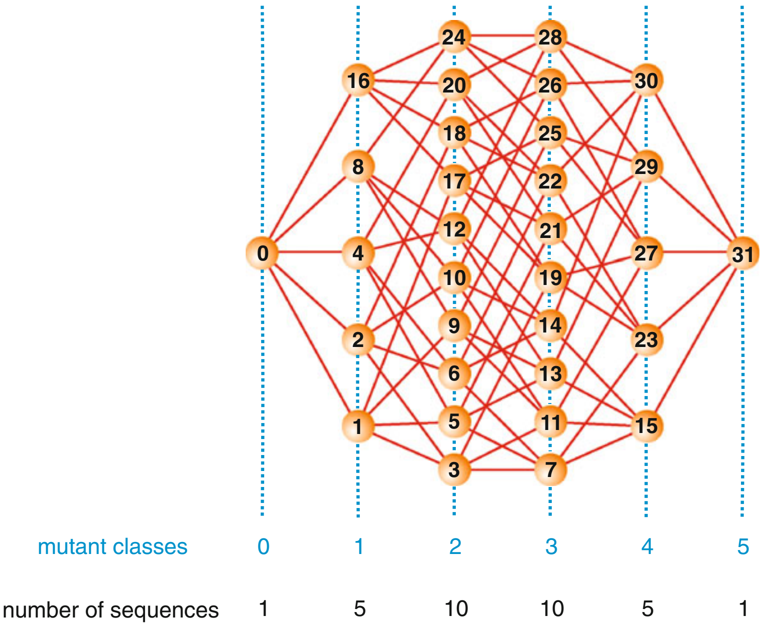


Fig. 1 Sketch of the binary sequence space with l = 5. The sequence space $\mathcal{Q}_5^{(2)}$ contains 32 sequences, which are indicated here by their equivalent decadic numbers and the assignments $0 \equiv \mathrm{C}$ and $1 \equiv \mathrm{G}$: $0 =$ `00000' $\equiv$ `CCCCC', $1 =$ `00001' $\equiv$ `CCCCG', $2 =$ `00010' $\equiv$ `CCCGC', ..., $31 =$ `11111' $\equiv$ `GGGGG'. Individual sequences are grouped in classes Γ_k that are defined by their Hamming distance to the reference sequence `CCCCC', $d^{\mathrm{H}}_{j0} = k$. The numbers of binary sequences in each Γ_k are given by the binomial distribution: $|\Gamma_k| = \binom{n}{k}$

will result that makes the sophisticated shifting balance process unnecessary since the summit can be reached by mutation and selection alone.

Sequence space (Fig. 1) is discrete, and local optima are simply defined by points that are higher in fitness than all their neighbors. Who the neighbors of a given genotype are is defined by the move set as the set of all sequences, which can be reached by a single move. Clearly, redefining the moves may turn local optima into saddle points or vice versa. An adaptive walk is a trajectory in sequence space that fulfills the condition of non-decreasing fitness $f_k = f(\mathsf{X}_k)$ in a time-ordered series of genotypes $\mathcal{X}(t) = \mathsf{X}_k$:

$$\begin{aligned}&(\mathcal{X}(t_1) = \mathsf{X}_1, \mathcal{X}(t_2) = \mathsf{X}_2, \ldots, \mathcal{X}(t_n) = \mathsf{X}_n) \quad \text{with} \\ &t_1 < t_2 < \ldots < t_n \quad \text{and} \quad f_1 \leq f_2 \leq \ldots < f_n,\end{aligned} \tag{1}$$

where we have implicitly assumed that the walk ends at a peak. The adaptive process is an illustration of evolutionary or natural selection in the sense of Darwin's *survival of the fittest*, although it is important to note that the adaptive walk refers to a single trajectory, whereas evolution deals with optimization of mean fitness in a population. Equation (1) has an immediate consequence for adaptive walks: The same sequence cannot be visited twice or more times unless the instances are separated exclusively by sequences of identical fitness or, in the other words, loops can occur only if the trajectory is confined to a neutral subset of sequences called a *neutral network* (Reidys et al. 1997). Adaptation on fitness landscapes is a frequently analyzed topic, and a large number of original papers, reviews, and books are available. Representative for others we mention here Gavrilets (1997), Jain and Krug (2007), McGhee (2007), Walsh and Blows (2009).

Point mutations and sequence space. Here, we are interested in population dynamics of viruses and other asexually reproducing species, and accordingly, genotype space will be represented by sequence space $\mathcal{Q}$, which is an abstract space where every different sequence of nucleotides is represented by a point and the distance between pairs of sequences X_i and X_j is given by the Hamming distance d_{ij}^{H} (Hamming 1950, 1986). The simplest and most straightforward move set in sequence space $\mathcal{Q}$ is point mutations, leading to single nucleotide exchanges $d_{ij}^{\mathrm{H}} = 1$. Figure 1 sketches the sequence space of binary sequences—sequences over an alphabet with $\kappa = 2$ letters—of chain length $l = 5$ denoted by $\mathcal{Q}_5^{(2)}$ and shows a natural grouping of sequences with respect to a given reference sequence into classes: A class Γ_k is the set of all sequences at Hamming distance $d_{ij}^{\mathrm{H}} = d_{\mathrm{H}} = k$ from a reference sequence X_0:

$$\Gamma_k = \{\mathsf{X}_i | d_{i0}^{\mathrm{H}} = k\}. \tag{2}$$

Accordingly, $d_{\mathrm{H}} = 0$ defines the reference sequence $\mathsf{X}_0 \equiv \Gamma_0$, which is a class by itself, the class Γ_1 with $d_{\mathrm{H}} = 1$ contains all one-error mutants, class Γ_2 with $d_{\mathrm{H}} = 2$ all two-error mutants, etc., and eventually Γ_l with $d_{\mathrm{H}} = l$ is the class whose members have different nucleotides from the reference at all positions. In the binary alphabet, this is the (unique) complementary sequence of the reference, $|\Gamma_l| = 1$ (where we denote the cardinality of a class by the *absolute value* symbol) and we have $\mathsf{X}_{2^l-1} \equiv \Gamma_l$. In the four-letter alphabet, this class contains $|\Gamma_l| = (\kappa - 1)^l = 3^l$ different sequences where κ as said is the number of different nucleotides in the alphabet.

Simple fitness landscapes. In the early days of population genetics and later on before extensive computer work became accessible rather, drastic simplifications were necessary for any modeling of adaptive walks on fitness landscapes. For example, the same fitness is assigned to all sequences within a given mutant class. The fitness of the genotype of largest fitness, the *master genotype* X_0, is the reference value f_0 and, in addition, at least one second fitness value f_n is required that still needs to be specified. For simple landscapes, the most straightforward definition chooses f_n as the lowest fitness value found in the population and assumes that all genotypes in a given class have the same fitness. Two typical assumptions are as follows: (i) *additive fitness* and (ii) *multiplicative fitness* (Fig. 2). In the first case,

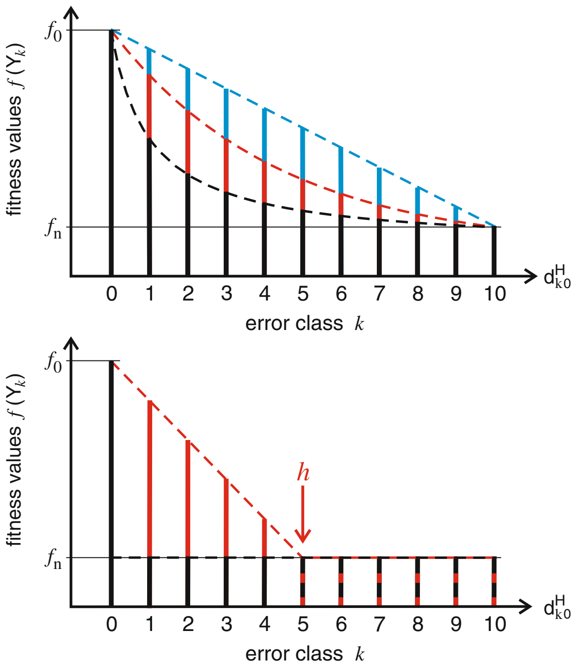


Fig. 2 Examples of simple fitness landscapes. The *upper sketch* shows three landscapes for which the fitness values of the different classes of sequences are given by continuous functions of the class index k: (i) the additive landscape (3a) in *blue*, (ii) the multiplicative landscape (3b) in *red*, and (iii) the hyperbolic fitness landscape (3c) in *black*. In the *lower drawing,* we present (iv) a single-peak landscape (3d) with a discontinuity in the derivative $\partial f/\partial k$ at $k = 0$ (*black*) and (v) the single-peak linear landscape (3e) where the discontinuity is located at $k = h$

every mutation decreases the fitness value of the master genotype by a constant amount $\Delta f/l$, and hence, the fitness of the genotypes in class Γ_k, $f_k = f(\Gamma_k)$ is

$$f_k = f_0 - \Delta f \frac{k}{l} \quad \text{with} \quad 0 < \Delta f = (f_0 - f_n) \leq f_0; k = 0, \ldots, l. \tag{3a}$$

The second case, multiplicative fitness, is characterized by

$$f_k = f_0 \cdot (\gamma_f)^{k/l} \quad \text{with} \quad 0 < \gamma_f = (f_n/f_0) < 1; k = 0, \ldots, l. \tag{3b}$$

Both cases are appropriate—if at all—for genes only and not for whole genotypes, since the basic argument for the usage of models (3a) or (3b) is the concept that species are located in local optima of fitness landscapes; hence, all mutations of reasonable probability are deleterious and reduce fitness. In addition, multiple mutations are assumed to have cumulative effects. As illustrated in Sect. 2, these requirements are not fulfilled by DNA of RNA sequences and point mutations as move set.

For the purpose of comparison, we mention a third landscape, the *hyperbolic fitness landscape*, which too has a continuous derivative $\partial f/\partial k$:

$$f_k = f_0 - \Delta f \frac{k}{l}\frac{l+1}{k+1} \text{ with } 0 < \Delta f = (f_0 - f_n) < f_0;\ k = 0, \ldots, l. \tag{3c}$$

The hyperbolic landscape is special, because it shares some features with landscapes that exhibit discontinuities in the derivative.

Eventually, we consider fitness landscapes that are modeled by functions with discontinuities in the derivatives. The most popular representative of this type of landscapes is the *single-peak landscape*, which reminds of the mean field approximation often used in physics: The highest fitness value, f_0, is assigned to the master genotype, and all other genotypes are assumed to have identical fitness, f_n (Fig. 2).

$$f_k = \begin{cases} f_0 & \text{for } k = 0, \\ f_n & \text{for } k = 1, \ldots, l. \end{cases} \tag{3d}$$

A generalization of the single-peak landscape characterized as *single-peak linear landscape* combines features of linear and single-peak landscapes: Fitness decreases linearly in the range $0 \le k \le h$ and is constant for the rest of the domain, $h \le k \le l$:

$$f_k = \begin{cases} f_0 - \Delta f \frac{k}{h} & \text{for } k = 0, 1, \ldots, h-1, \\ f_n & \text{for } k = h, \ldots, l \end{cases} \quad h = 1, \ldots, l. \tag{3e}$$

The landscapes (3a)–(3d) are completely described by the two parameters f_0 and f_n. Only the case (3e) requires a third parameter h defining the position of the discontinuity. We remark that single-peak linear landscapes with $h = 1$ are identical to single-peak landscapes and a landscape with $h = l$ is a linear landscape.

Fully resolved fitness landscapes. A fitness landscape is denoted as *fully resolved* when individual fitness values are determined for or assigned to different sequences and not only to classes as in case of simple fitness landscapes. The number of fitness values required is κ^l where κ denotes the number of different digits in the alphabet, e.g., $\kappa = 2$ for binary sequences and $\kappa = 4$ for natural nucleic acids. Within the last fifteen years, plenty of progress has been made in the determination of fitness values and trajectories of adaptive evolution of viruses (Betancourt and Bollback 2006; Elena and Sanjuán 2007) and successful attempts were made to measure distributions of fitness effects (Sanjuán et al. 2004). In general, exploration of fitness landscapes by site-directed mutagenesis is restricted to small neighborhoods—variants with Hamming distance $d_{\mathrm{H}} = 1, 2, 3$—from the master sequence X_0 or, in other words, to local areas in sequence space. Global information on fully resolved fitness landscapes of real systems is still far out of reach because of high dimensionality and hyper-astronomical numbers of sequences. It is also important to stress that fitness values and landscapes depend strongly on environmental effects and therefore, they can be determined efficiently only in approaches where a sufficiently large number of parameters can be kept constant, for example, in experimental evolution.

One of the earliest assays for experimental evolution was developed by Spiegelman and coworkers (Mills et al. 1967; Spiegelman 1971): RNA molecules from the bacteriophage Qβ were transfected into a stock solution containing excess of all materials required for replication—the four RNA building blocks, **ATP**, **UTP**, **GTP**, **CTP**, and the enzyme Qβ-replicase—under suitably controlled conditions such as pH, ionic strength, and $\mathbf{Mg}^{2\oplus}$. Spiegelman's test tube experiment is an extreme example of adaptive evolution by loss of function since the RNA in the test tube needs little more than a suitable binding site for the enzyme and accordingly, the chain length of the viral RNA is reduced from $l = 4217$ to a few hundred through fitness increasing deletions. Detailed investigations of RNA replication kinetics revealed the molecular mechanism of in vitro evolution (Biebricher 1983; Biebricher et al. 1983, 1984, 1985). The most striking result of these very elegant and systematic studies is the observation that the selection mechanism changes with an increasing concentration ratio of RNA to replicating enzyme because of a change in the rate-limiting step of the multi-step kinetics, which consists of binding the RNA to the enzyme, initiation and propagation of complementary step synthesis, and product release. For the landscape concept, this finding has the immediate consequence that extracellular RNA evolution takes place upon different landscapes depending on whether or not enzyme is supplied in excess.

One of the most extensive construction and analysis of a viral fitness landscape has been performed in clinical studies with the human immunodeficiency RNA retrovirus (HIV-1) (Kouyos et al. 2012). Fitness is measured as the in vitro reproductive capacity of HIV-derived amplicons that were prepared and inserted into a constant resistance test vector (Kouyos et al. 2011). The empirical basis of the study is $\sim$70,000 clinical HIV-1 isolates taken in the absence of drug treatment or in the presence of a single drug chosen from a collection of fifteen. The fitness landscape is derived from this data set by means of a statistical model predicting fitness from the amino acid sequences of entire HIV protease (99 aa)[3] and parts of HIV reverse transcriptase (a heterodimer consisting of p66 with 560 aa and p52 with 440 aa) with a total chain length of 460. Landscapes are constructed by fitting of parameters to data from 65,000 isolates as training set, and the remaining 5000 are used as test for the predictive power of parameter set. The landscape fitted to the fitness values of the drug-free isolates is taken as reference. Two features, which will be analyzed and discussed in Sect. 2, were found to be characteristic for the HIV fitness landscape: (i) *ruggedness* in the sense of containing many local fitness maxima and (ii) *neutrality* expressed as an appreciable fraction of sample points share the same fitness. In addition, but nor surprisingly from the molecular point of view, the results confirm that epistasis is highly important since the effect of a given

[3] Here, 'aa' stands for 'amino acid residue.'

point mutation depends strongly on the presence or absence of other mutations in the isolate.[4] Although the HIV study (Kouyos et al. 2011) is very extensive indeed and reaches the upper limit that can be achieved straightforwardly at present, a commentary (Weinreich 2011) correctly says that much more work in theory and experiment is needed in order to allow for clinically valuable predictions. As examples of bacterial landscapes, we mention a study of fitness landscape defined by gene expression levels in the core metabolism of *Methylobacterium extorquens* (Chou et al. 2014) and an extensive analysis of epistatic interactions in the fitness landscape of *Escherichia coli* (Beerenwinkel et al. 2007).

Tunable resolved fitness landscapes with random assignments. Despite the enormous progress in the empirical determination of fitness landscapes reported in the previous paragraph, models for assigning fitness value to genotypes are required. There are, for example, 8×10^{2538} different RNA sequences of the chain length of the Qβ-bacteriophage, and even if the vast majority of sequences are functionless as genotypes, the remainder would be beyond all technical bounds. Accordingly, model landscapes that allow for fast calculation of a large number of fitness values were invented. We mention here two of them: (i) the random Nk landscape (RNkL) proposed by Kauffman (Kauffman and Levin 1987; Kauffman and Weinberger 1989) and (ii) the realistic rugged landscape (RRL) and its variant the realistic neutral landscape (RNL) introduced by the author (Schuster 2012, 2013).

Random Nk fitness landscapes. The RNkL (Altenberg 1997) is a stochastic model that generates fitness values f_j for binary sequences of chain length $l = N$. In other words, we are dealing with a genotype consisting of N loci and two alleles at each locus: $\mathsf{X}_j = (x_1^{(j)}, x_2^{(j)}, \ldots, x_N^{(j)})$ with $x_i^{(j)} \in \{0,1\} \forall i = 1, 2, \ldots, N$. The fitness of the genotype X_j is assumed to be the average of the fitness components $\phi_i^{(j)}$ contributed by the individual loci:

$$f_j = f(\mathsf{X}_j) = \frac{1}{N} \sum_{i=1}^{N} \phi_i^{(j)}(x_i^{(j)}; x_{i1}^{(j)}, x_{i2}^{(j)}, \ldots, x_{ik}^{(j)}) \qquad \text{with} \quad x_{il}^{(j)} \in \{x_1^{(j)}, x_2^{(j)}, \ldots, x_N^{(j)}\}; \quad l \neq i, \text{all } l \text{ different.} \tag{4}$$

The fitness component of position i in sequence X_j, $\phi_i^{(j)}$ clearly depends on the allele at this position, $x_i^{(j)}$, and through epistatic interactions, it depends also on the alleles at k other positions denoted by $x_{il}^{(j)}$ with $l = 1, \ldots, k; l \neq i$. Two possibilities were considered by Kauffman: (i) *adjacent neighborhoods* and (ii) *random*

[4] Considering single nucleotides as sites in structural RNA elements requires complementarity of the nucleobase at another locus for the formation of a base pair, and accordingly, the two sites are strongly coupled epistasis (see Sect. 2).

neighborhoods. In the first case, the k genes lying closest to position i on the chromosome are chosen, whereas in the second case, the genes are chosen at random. Epistatic contributions are calculated by assuming a *house of cards* model of fitness effects as proposed by Kingman (1978); see also (Kingman 1980, p. 15): When an allele at one locus is changed, the fitness components of all alleles, which interact with this locus, are changed without correlations to their previous values. The metaphor illustrates the situation as follows: If a single card is pulled out of a house of cards, the house collapses and must be rebuilt from scratch.

The parameter k is designed as a tunable parameter for the ruggedness of the landscape: $k = 0$ implies a smooth, single-peak linear landscape often called *Mount Fuji landscape,* and the maximal value $k = n - 1$ gives rise to fully developed randomness. The Nk landscape for $k = 2$ was shown to be closely related to a spin glass Hamiltonian in the sense that the Nk model describes a special class of spin glasses (Kauffman 1993, p. 43) [for more details, see Reidys and Stadler (2001, 2002)]. The Nk landscape with two adjacent neighbors ($k = 2$), for example, can be derived from a linear chain of genes by closing it to a loop, and in the random model, of course, no such assumption is required.

Realistic random fitness landscapes. In order to introduce a random distribution of fitness values in the single-peak fitness landscape, we consider a band of fitness values for all sequences except the master sequence. The lack of detailed empirical data is supplemented by a random input and a tunable parameter d that determines the width of this band, and neglecting neutrality, the fitness values are calculated from the expression (Schuster 2013, p. 608):

$$f(\mathsf{X}_j) = f_j = \begin{cases} f_0 & \text{if } j = 0, \\ f_n + 2d\Delta f(\eta_j^{(s)} - 0.5) & \text{if } j = 1, \ldots, \kappa^l - 1. \end{cases} \tag{5a}$$

The parameters f_0 and f_n are defined as before, and $\eta_j(s)$ is the jth output of a pseudorandom number generator that has been started by using s as seed. In order to make the procedure fully determined, the method used in the generation of pseudorandom numbers has to be specified. In addition, we need to predefine the distribution of the pseudorandom numbers. Here, we use a uniform distribution on the unit interval, $0 \leq \eta_j^{(s)} \leq 1$.

Neutrality can be readily incorporated into RRLs by means of a tunable degree of neutrality, λ: The fitness value f_0 is assigned to the master sequence and to all sequences X_j with pseudorandom numbers $1 \leq \eta_j^{(s)} \leq 1 - \lambda$, and random scatter in the sense of Eq. (5a) is chosen for all other sequences:

$$f(\mathsf{X}_j) = \begin{cases} f_0 & \text{if } \; j = 0, \\ f_0 & \text{if } \; \eta_j^{(s)} \geq 1 - \lambda, \\ f_n + \frac{2d}{1-\lambda}\Delta f(\eta_j^{(s)} - 0.5) & \text{if } \; \eta_j^{(s)} < 1 - \lambda, \\ & j = 1, \ldots, \kappa^l - 1; j \neq m. \end{cases} \tag{5b}$$

As shown in Eq. (5b), the interval $0 \leq \eta_j^{(s)} < 1 - \lambda$ is stretched to the full bandwidth of d for the determination of the remaining f_j-values. Clearly, Eq. (5a) results from (5b) through setting λ = 0 which is tantamount to *no neutrality*. Accordingly, an RRL or RNL is fully characterized by:

$$\begin{aligned} &\text{RNL} : \mathcal{L} = \mathcal{L}(l, \kappa, f_0, f_n; \lambda, d, s) \quad \text{and} \\ &\text{RRL} : \mathcal{L} = \mathcal{L}(l, \kappa, f_0, f_n; 0.0, d, s). \end{aligned} \tag{6}$$

It is important to stress that the definition of realistic random landscapes according to (5a, 5b) does not allow for landscape design since the relation between the random seed s and the calculated fitness values is too complicated in order to allow for a reconstruction by an inverse method. This concept of landscapes can be rather understood as a mean for performing a kind of experiment in computational biology in three steps: (i) choose seeds for the random number generator, e.g., $s \in \{000, \ldots, 999\}$, (ii) compute landscape $\mathcal{L}(l, \kappa, f_0, f_n; \lambda, d, s)$ in the form of the fitness values $f(\mathsf{X}_j), j = 1, \ldots, \kappa^l - 1$, and (iii) compute and analyze the mutant distribution $\Upsilon(d, \lambda; p, t)$.

The two major questions that will be studied in the rest of this chapter are: (i) How does the quasispecies distribution change as a function of the mutation rate p; and (ii) do abrupt transitions at critical mutation rates, p_{cr}, exist, and if they exist, how do they depend on the extend of random scatter d? The answer to the second question is of particular importance because doubts have been raised whether or not the scenarios derived from single-peak landscapes are specific for this simple landscape, and accordingly might not be relevant for more general rugged landscapes [(Baake and Wagner 2001; Charlesworth 1990; Wiehe 1997); see Sect. 4].

2 Sequence Structure Mappings

In the previous section, we introduced several classes of fitness landscapes and mentioned the available empirical support for two general features of real landscapes: (i) ruggedness and (ii) neutrality. Here, we present additional arguments for this conjecture by considering the properties of known genotype–phenotype mappings. In experimental evolution with molecular systems (Biebricher 1983), the genotype is considered as a polynucleotide sequence, DNA or RNA, and the phenotype is the molecular structure. Predicting biopolymer structures from known sequences is still kind of a scientific art, but in case of simplified structures of RNA molecules, so-called *secondary structures*, it is possible to derive shapes by simultaneous consideration of free energies of substructures and some principles from combinatorics. Secondary structures of polynucleotides are graphically illustrated listings of nucleotide pairs where the graphs of structures are equivalent to representations by strings over a three-letter alphabet: (i) '(' opening of a base pair, (ii) ')' closing of a base pair, and (iii) '•' an unpaired nucleotide. The assignment of opening parentheses to closing parentheses of base pairs follows mathematical rules, i.e., the first parenthesis opens a nucleotide pair and matches the parenthesis that encloses a complete set of closed parentheses. As an example, we show the string representation of the reference structure S_0 in Fig. 4

$$(((\cdot(((\cdot\cdot\cdot)))\cdot))),$$

where the three left opening parentheses match the three rightmost closing parentheses and the three inner parentheses form the hairpin loop.

Landscapes are built from sequences by two consecutive mappings: (i) the map of biopolymer sequences into molecular structures [for a review of the RNA model, see, e.g., Schuster (2003, 2006)] and (ii) the map from structures into molecular properties. In a given and constant environment, replication parameters tantamount to fitness values are functions of molecular structures. The current paradigm of structural biology is based on the conjecture that structures can be derived from sequences, and molecular properties in the form of parameters in functions are derivable from structures:

$$\text{sequence} \quad \Rightarrow \quad \text{structure} \quad \Rightarrow \quad \text{function.}$$

Sequences, structures, and parameters in functions are objects of metric spaces (Fig. 3), and relations between them are defined by mappings. The Hamming

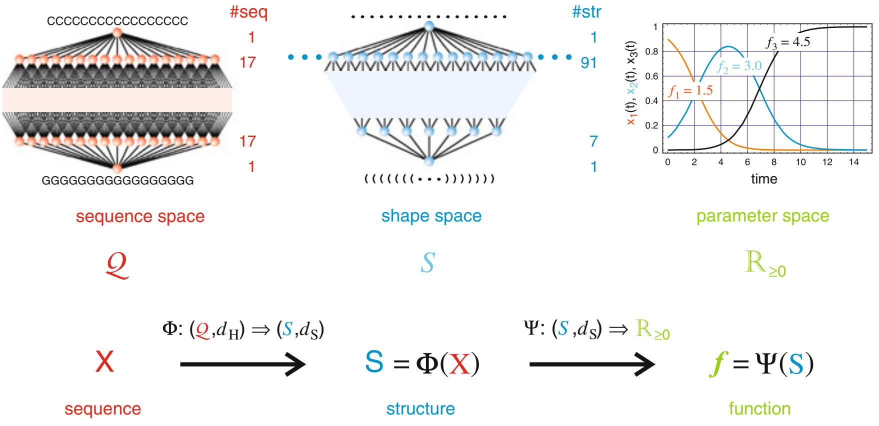


Fig. 3 The paradigm of structural biology. The relation between sequence, structure, and fitness is sketched as a sequence of two mappings from sequence space ($\mathcal{Q}_{17}^{(2)}$) into shape space ($\mathcal{S}_{17}^{(2)}$) and from shape space into nonnegative real numbers ($\mathbb{R}_{\geq 0}$). In order to facilitate drawing, sequences are assumed to chosen from a two-letter alphabet $(\mathbf{C}, \mathbf{G})$. For l = 17, sequence space contains 2^{17} = 131,072 sequences, which form 530 different *acceptable RNA secondary structures* (Schuster 2006). These structures determine the fitness values f. Sequence space and shape space are metric spaces with the Hamming distance d_H and some structure distance d_s representing the metric. Parameter space is based on real numbers $\mathbb{R}$, and the absolute value of the difference, $|f_i - f_j| = d_{ij}$, is the metric

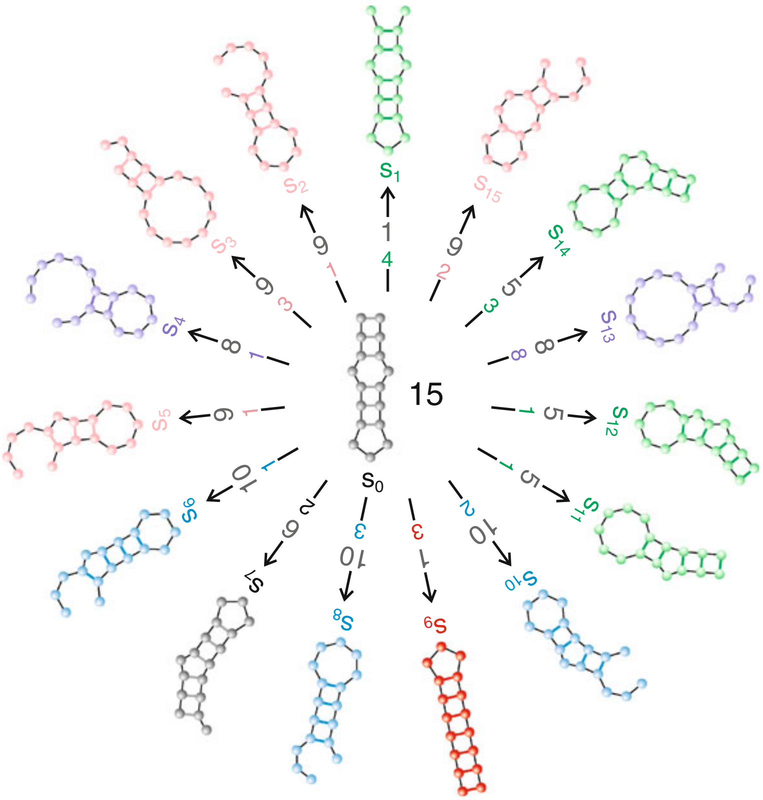


Fig. 4 Structures of the one-error mutant spectrum of a small RNA molecule. The figure presents the structures of all 51 single point mutations of sequence X_0. In total, 16 different structures S_k with $k = 0, 1, \ldots, 15$ were obtained. Structure S_0 in the *center* is the structure of the reference sequence X_0, and it is most frequent and occurs 15 times. The structures on the periphery are ordered according to their appearance in the series of consecutive mutations (Fig. 5). Inserted in the *arrows* pointing from S_0 to the individual structures S_k are (i) the numbers of occurrence (*color*) and (ii) the base pair distance $\mathrm{d}^{\mathrm{S}}_{0k}$ (larger numbers in *gray*). All drawings of structures begin at the 5′-end of the RNA, which is always the left end of the graph or string (in upright positioning), nucleotides are shown as beads, and base pairs are connected by a colored *thick line*. Colors encode number of base pairs: *red* 7, *black* 6, *green* 5, *blue* 4, *pink* 3, and *lavender* 2

distance d_{H} (Fig. 1) is the natural metric in sequence space for point mutations as the dominant changes in sequences. The base pair distance d_{H} can be chosen as a metric in *shape space* being the space of all RNA secondary structures that can be formed by all sequences of a given length l. It is defined as the number of base pairs in which two structures differ, for example, the base pair distance between the structures S_0 and S_7 (Fig. 4),

$$d^{\mathrm{S}}_{07} = \left\{ \begin{matrix} (((\cdot(((\cdot\cdot\cdot)))\cdot))) \\ \cdot((((((\cdot\cdot\cdot)))\cdot))) \end{matrix} \right\} = 6.$$

The three inner base pairs of the hairpin loop (Fig. 4) remain the same, but the three outer base pairs are replaced by three other base pairs, and this leads to a structure distance of $d^{\mathrm{S}}_{07} = 3 + 3 = 6$, since three base pairs have to be removed first and then three base pairs are added.

At the current state of the art, a determination of kinetic parameters from structures is not possible without largely simplifying assumptions. In a previous evolution model, we estimated replication parameters either by the free melting energies of structures, $-\Delta G_0^T$ or more elaborately by cooperative melting of stacking regions (Fontana et al. 1989; Fontana and Schuster 1987) and assumed the degradation rates to be determined by the unpaired nucleotides in the structure. This model introduces complex behavior since optimization of fitness leads to frustration (Toulouse 1977, 1980) in the sense of spin glass theory (Edwards and Anderson 1975; Sherrington and Kirkpatrick 1975). The RNA-based model has been used to analyze replication and mutation-based evolution in silico in population of up to 10,000 RNA molecules (Fontana and Schuster 1998a, b). In particular, mean fitness in the population shows a stepwise approach toward the optimum value and transitions can be classified as minor changes in structure occurring at almost constant mean fitness and major changes, which are commonly accompanied by fitness increases.

The bizarre nature of sequence to structure and structure to fitness mappings is illustrated by means of a simple but nevertheless representative example consisting of a very small RNA molecule with chain length $l = 17$ and the sequence $\mathsf{X} = ($**AGCUUACUUAGUGCGCU**$)$. This chain length is just enough to form a maximum of seven base pairs, and all properties can be either counted or calculated, or seen by inspection. Despite its simplicity, the example reflects the most important features of sequence structure mappings. The minimum free energy structure S_0 is calculated[5] for X_0, and a free energy of folding $\Delta G_0^{(0°\mathrm{C})} = \Delta G_0^0 = -6.39$ kcal/mole is obtained. Then, the same computations are performed for all 51 one-error mutants of X_0 and the numbers of occurrence for the individual structures are enumerated. The results are shown in Fig. 4: The most frequent structure is the structure of the reference sequence, S_0, and it is found 15 times, followed by structure S_{13}, which appears eight times, S_1 four times, S_3, S_8, S_9, and S_{14} three times each, S_7, S_{10} and S_{15} twice, and finally S_2, S_4, S_5, S_6, S_{11} and S_{12}, which occur only once. Thus, the local degree of neutrality in sequence space is $\lambda(\mathsf{X}_0) = 0.29$. Considering the free energy values of folding, ΔG_0^0 in Fig. 5 (upper plot), there is no relation between the frequency of occurrence and the folding energy: The most stable structure S_3 (red) appears three times, whereas the least stable structures S_2 and S_{13} occur

[5] Standard software packages are available for RNA secondary structure computation, for example, *mfold* (Zuker 1989) or the *Vienna RNA package* (Hofacker et al. 1994; Lorenz et al. 2011).

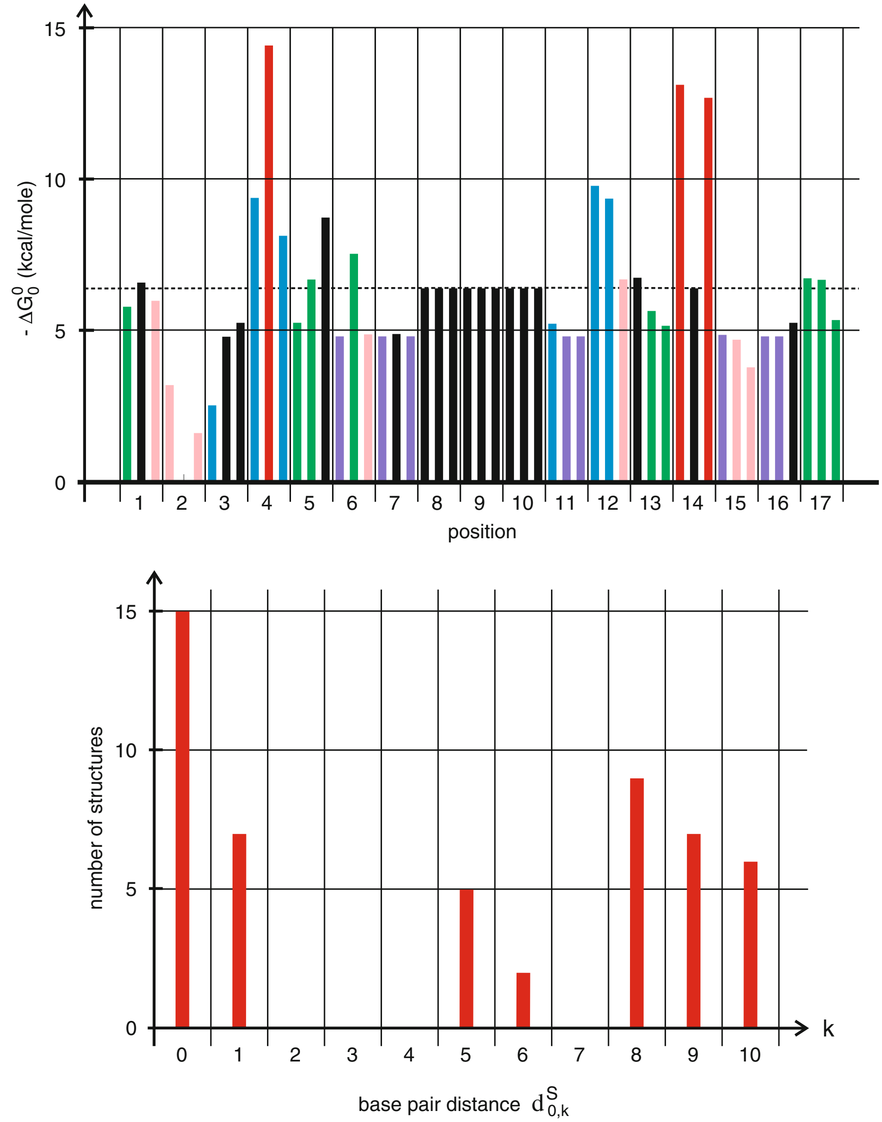


Fig. 5 Free energy of folding and base pair distances of one-error mutants. The *upper plot* shows the folding energies at 0 °C of the 51 one-error mutants of X_0. At each position 1–15, the sequence of mutants is $\mathbf{N} \rightarrow \mathbf{A}$, $\mathbf{N} \rightarrow \mathbf{U}$, $\mathbf{N} \rightarrow \mathbf{G}$, and $\mathbf{N} \rightarrow \mathbf{C}$, where the trivial replacement leaving the sequence unchanged is omitted ($\mathbf{N} = \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$). The folding energy of the reference sequence is shown as *dotted line*, and the *color code* refers to the number of base pairs (see caption of Fig. 4). The lower histogram presents the numbers of structures in the mutant spectrum of X_0 at a given base pair distance d^{S}_{0k} from the reference structure $\mathsf{S}_0 = \Phi(\mathsf{X}_0)$

together nine times. As expected, there is appreciable scatter in folding energies between sequences, which form the same minimum free energy structures. Additional information on the ruggedness of the free energy landscape is provided by the correlation length of the folding energies, $\varrho(l)$ (Fontana et al. 1991, 1993): For a chain length $l = 17$ and the natural four-letter alphabet, a free energy correlation length of $\varrho(17) \approx 3.0$ is computed, which means that the energy values at Hamming distance d_{H} = 3 have a correlation coefficient of $1/e = 0.37$ with the energy of the reference sequence X_0, at d_{H} = 10, and this coefficient is only 0.036 implying that the neighborhood memory on X_0 has practically faded out and the statistics of the energy distribution is about the same as found at any randomly chosen point in sequence space. Eventually, we consider the relation between similarities between structures as expressed by the base pair distance d_{S} (Fig. 5, lower plot): The most frequently occurring distance is $d_{0k}^{\mathrm{S}} = 8(9\times)$, followed by $d_{0k}^{\mathrm{S}} = 1$ and $d_{0k}^{\mathrm{S}} = 9(7\times\ \text{each})$, and $d_{0k}^{\mathrm{S}} = 10$, $d_{0k}^{\mathrm{S}} = 5$, and $d_{0k}^{\mathrm{S}} = 6(6, 5\ \text{and}\ 2\times, \text{respectively})$.

The experimental approach to determine fitness landscapes of small RNA molecules has been profiting substantially from the availability of deep sequencing and high-throughput methods (Pitt and Ferré-D'Amaré 2010). We mention here only recent work that succeeded to explore almost the entire sequence space of a small RNA molecule of chain length $l = 24$ (Athavale et al. 2014; Jiménez et al. 2013) and refer to Chap. 3 (this volume) for details.

A comparison of the landscape obtained from mapping structures into folding energies (Fig. 5) with the Nk model is tantamount to estimating the value of k for $N = l$, the chain length of the RNA sequence. In other words, we need to answer the question: 'Mutations at how many positions along the sequence change the free energy of folding?' Considering the upper plot in Fig. 5, we see that mutations in the unpaired nucleotides of the hairpin loop leave ΔG_0^0 unchanged, and in addition, we find seven more mutants exhibiting values close to the reference value. In total, we have 16 out of 51 mutations leaving 35 cases of change. Normalizing to sites gives a vague estimate of $k = 11$ suggesting that on average, 11 positions out of 17 exert influence of the energy of folding. As expected, a realistic landscape built from sequence-dependent biopolymer properties is very rugged but not completely uncorrelated as would be an Nk model with $k = l - 1 = 16$. The correlation although weak comes from the regularities of mapping structures into folding energies, and more base pairs yield higher energies in absolute value, for example.

3 Mutations and Population Dynamics

Before the seminal paper by Watson and Crick (1953), the concept of mutation was nebulous and it required molecular insight in order to conceive appropriate models for the replication process. After the proposal of the structure of, b-DNA was on the table; however, one could immediately guess how nucleic acids replicate and mutate as the authors themselves stated: ‘It has not escaped our notice that the specific pairing we have postulated suggests a copying mechanism for the genetic material.’

Mutation models. Two concepts are currently prevailing, which originate from different mutation mechanisms: (i) the quasispecies model introduced in 1971 by [Eigen (1971), Eigen and Schuster (1977)] and (ii) the selection–mutation model attributed to Crow and Kimura (1970, p. 265, Eq. 6.4.1), which is also know as *paramuse* (parallel mutation and selection) model (Baake et al. 1997). Mutation in the quasispecies model is attributed to the reproduction process, and correct replication and mutations are visualized as different reaction channels of the same replication step (Chap. 1, Fig. 1), whereas mutation in the paramuse model is due to some external process independent of reproduction [Fig. 6; for reviews, see Baake and Wagner (2001), Burger (1998)]. Nevertheless, the kinetic equations resulting from both models are closely related. The difference in the mutation mechanism has biological consequences: The number of mutations is proportional to the number of reproduction events or generations in the quasispecies model, whereas proportionality with respect to time is predicted by the paramuse model provided the external driving forces causing mutation are constant. Observations on organisms of largely different genome size from viroids to higher eukaryotes reveal roughly constant spontaneous mutation rates for classes of organisms. The mutation rates per genome and replication event range from 1 found with viroids, RNA viruses, and also with sexual reproduction of higher eukaryotes to 1/300 for microbes with DNA-based chromosomes (Drake et al. 1998; Gago et al. 2009). Proportionality with respect to real time is the basis of the molecular clock model (Ho and Duchêne 2014; Lanfear et al. 2010), which apart from still to be explained vagaries seems to be correct for vertebrates. Accordingly, it is a matter of the problem under consideration whether quasispecies or paramuse is the model of choice.

At this point, we would like to mention that substantial insight into quasispecies and error thresholds were gained by showing that the value matrix W of the quasispecies equation is equivalent to the row transfer matrix of a 2D Ising model of magnetism (Leuthäeusser 1986, 1987). In particular, the analogy to spin systems

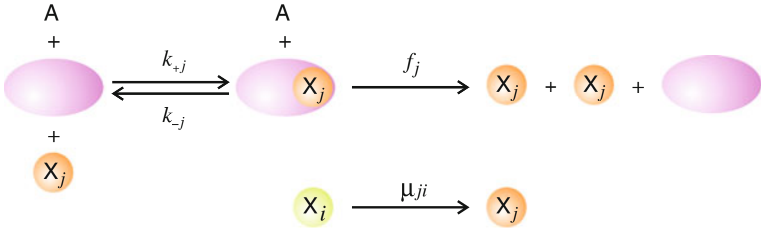


Fig. 6 The Crow–Kimura model of reproduction and mutation. The Crow–Kimura model combines error-free reproduction with replication-independent mutation. Although it leads to the same differential equation (7a) as the quasispecies replication–mutation model shown in Chap. 1 (Fig. 1), the interpretation of the parameters and the physical restrictions on them are different

allowed for the application of methods forms statistical physics and the generalization to spin glasses made it possible to show straightforwardly that the transitions on simple and more complex landscapes may fulfill the requirements of first-order phase transitions in the limit of infinite chain lengths l (Tarazona 1992). Similarly, it was shown that the paramuse model corresponds to the Hamiltonian of an Ising quantum chain (Baake et al. 1997; Baake and Wagner 2001) and methods from quantum statistical mechanics were successfully applied in the search for solutions of the replication–mutation problem [see Chap. 5 and, for example (Bratus et al. 2014; Galluccio 1997; Kang and Park 2008; Park et al. 2010; Saakian and Hu 2006; Saakian et al. 2004) as well as the review (Baake and Gabriel 1999)].

The kinetic differential equation of the quasispecies model [Chap. 1 (this volume), Eq. (3)], formulated in normalized variables $x_j = [\mathsf{X}_j]/\sum_{i=1}^{N}[\mathsf{X}_i]$ with $\sum_{i=1}^{N} x_i = 1$ can be easily written in matrix form,

$$\begin{aligned}\frac{\mathrm{d}x_j}{\mathrm{d}t} &= \sum_{i=1}^{N} Q_{ji} f_i x_i - x_j \bar{f}(t); \quad j = 1,\ldots,N \quad \text{or}\\ \frac{\mathrm{d}x}{\mathrm{d}t} &= (Q \cdot F - \bar{f}(t))x,\end{aligned} \tag{7a}$$

where $x = (x_1,\ldots,x_N)^t$ is the column vector of normalized concentrations $\sum_{i=1}^{N} x_i = 1, Q = \{Q_{ij}; i,j = 1,\ldots,N\}$, is the mutation matrix—with Q_{ij} being the frequency at which X_i is synthesized as a correct ($i = j$) or erroneous ($i \neq j$) copy of the template X_j—and the fitness values f_i are contained in the diagonal matrix $F = \{F_{ij} = f_i \delta_{i,j}; i,j = 1,\ldots,N\}$. The mean fitness of the population is denoted by $\bar{f}(t) = \sum_{i=1}^{N} f_i x_i(t)$. In case of the paramuse model, we obtain:

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = (f_j - \bar{f}(t))x_j + \sum_{j=1}^{N} \mu_{ji} x_j \quad \text{or} \quad \frac{\mathrm{d}x}{\mathrm{d}t} = (F + \boldsymbol{\mu} - \bar{f}(t))x. \tag{7b}$$

Herein, the mutation matrix is denoted by $\boldsymbol{\mu}$.

Both Eqs. (7a) and (7b) can be easily cast into identical form by introducing the value matrix W

$$\frac{\mathrm{d}x}{\mathrm{d}t} = (\mathrm{W} - \bar{f}(t))x \quad \text{with} \quad W = Q \cdot F \quad \text{or} \quad W = \boldsymbol{\mu} + F, \tag{7c}$$

respectively. The resulting Eq. (7a–7d) is mildly nonlinear and can be solved by means of an integrating factor transformation and the solution of the remaining eigenvalue problem (Eigen et al. 1988, 1989; Jones et al. 1976; Thompson and McBride 1974). The value matrix W has to be a primitive matrix in order to fulfill the conditions for the applicability of the Perron–Frobenius theorem (Seneta 1981, pp. 3–11 and 22–23),[6] which guarantees that (i) the largest eigenvalue is real, positive, and non-degenerate; and (ii) the largest eigenvector has only strictly positive components. Exact solutions of Eq. (7c) are obtained through diagonalization of the value matrix: $H.W.B = \Lambda$ where the diagonal matrix $\Lambda = \{\Lambda_{ii} = \lambda_i; i = 1,\ldots,N\}$ embodies the eigenvalues λ_i. The matrices $H = \{h_{ij}\}$ and $B = \{b_{ij}\} = H^{-1}$ fulfill the eigenvalue equations $H.W = \Lambda.H$ and $W.B = B.\Lambda$ and contain the left-hand and right-hand eigenvectors of the value matrix W, which in explicit form are the row vectors $\mathbf{h}_k = (h_{ki}, i = 1,\ldots,N)$ and the column vectors $\mathbf{b}_j = (b_{ij}; i = 1,\ldots,N)^t$. The solutions can now be expressed by

$$\begin{aligned} x_j(t) &= \frac{\sum_{k=1}^N b_{jk} \sum_{l=1}^N h_{kl} x_l(0) \exp(\lambda_k t)}{\sum_{i=1}^N \sum_{k=1}^N b_{ik} \sum_{l=1}^N h_{kl} x_l(0) \exp(\lambda_k t)} \\ &= \frac{\sum_{k=1}^N b_{jk} \beta_k(0) \exp(\lambda_k t)}{\sum_{i=1}^N \sum_{k=1}^N b_{ik} \beta_k(0) \exp(\lambda_k t)} \quad \text{with} \quad \beta_k(0) = \sum_{l=1}^N h_{kl} x_l(0), \end{aligned} \tag{7d}$$

wherein the eigenvalues λ_k are the rate parameters and the coefficients $\beta_k(0)$ encapsulate the initial conditions.

The difference between the two mutation models cannot be seen from these mathematical results and boils down to two issues: (i) The quasispecies model treats replication and mutation as parallel reaction channels of one reaction step, and accordingly, the value matrix is a product of the mutation and the fitness matrix, whereas reproduction and mutation are independent reaction steps in the paramuse case and the two matrices are added; and (ii) the mutation matrix Q of the quasispecies is a stochastic matrix, $\sum_{i=1}^N Q_{ij} = 1$, because a replication has to be either correct or error-prone, whereas the condition $\sum_{i=1}^N \mu_{ij} = 0$ is used in the paramuse model. In addition, mutation is commonly restricted to single point mutations in the paramuse model. As said before, apart from these technical details, the mutation mechanisms shown in Fig. 1, Chap. 1 (this volume), and Fig. 6 are dealing with entirely different situations. In the quasispecies model, mutation occurs during the reproduction process and this is the situation that is relevant for viruses (see Chaps. 7, 9, 12, and 14, this volume).

Deterministic and stochastic autocatalysis. In order to set the stage for a discussion of the dynamics of quasispecies formation, we consider first the simple autocatalytic chemical reaction $\mathsf{A} + \mathsf{X} \rightarrow 2\mathsf{X}$ in the well-defined and controllable environment of a flow reactor (Schmidt 2004, p. 87ff). In order to relate to the quasispecies concept, we interpret simple autocatalysis as a replication–mutation system in which all individual sequences are lumped together in one species: $\mathsf{X} = \mathsf{X}_1 \oplus \mathsf{X}_2 \oplus \ldots \oplus \mathsf{X}_N$, and hence, the stochastic and deterministic variables take on the form $\mathcal{C} = \sum_{i=1}^N \mathcal{X}_i$ and $c = \sum_{i=1}^N x_i$, respectively. A solution containing the

[6] A matrix W is primitive if (i) all the elements of matrix W are nonnegative and (ii) some finite power W^m is a positive matrix, which means that all entries of W^m are strictly positive.

compound A in concentration a_0 flows into the reactor with a rate parameter r [vol × time^{-1}], and the inflow is compensated by an outflow of the volume of reactor solution, thus yielding the reaction equations

$$* \xrightarrow{a_0.r} \mathsf{A}; \tag{8a}$$

$$\mathsf{A} + \mathsf{X} \xrightarrow{f} 2\mathsf{X}, \tag{8b}$$

$$\mathsf{A} \xrightarrow{r} \text{ø}; \quad \text{and} \tag{8c}$$

$$\mathsf{X} \xrightarrow{r} \text{ø}. \tag{8d}$$

The rate parameter f is the analogue to fitness in the biological models and has the dimension [vol × mole^{-1} × time^{-1}].[7] The kinetic differential equations are obtained straightforwardly

$$\begin{aligned} \frac{\mathrm{d}a}{\mathrm{d}t} &= (a_0 - a)r - fac \quad \text{and} \\ \frac{\mathrm{d}c}{\mathrm{d}t} &= fac - cr = (fa - r)c, \end{aligned} \tag{8e}$$

The equation is also valid for the simplified system with the lumped concentrations $c(t) = \sum_{k=1}^{N} x_k$ if we replace the parameter f by the function $\bar{f}(t) = \sum_{k=1}^{N} f_k x_k(t) / \sum_{k=1}^{N} x_k(t)$, which represents the mean fitness of the population. Strictly speaking, the mean fitness is a function of time, and for common initial conditions, it is a non-decreasing function of time, and then, evolution is tantamount to fitness optimization. For the purpose of illustration, however, we shall assume a constant mean fitness corresponding to the rate parameter: $\hat{f} = f$. Then, it is straightforward to analyze the stationary states: $\mathrm{d}a/\mathrm{d}t = 0$ and $\mathrm{d}c/\mathrm{d}t = 0$ give rise to two solutions, S_1 and S_2,

$$\begin{array}{lll} S_1 : \; \bar{a} = r \cdot \hat{f}^{-1} \bar{c}, \; = a_0 - r \cdot \hat{f}^{-1} & \text{asymp.stable for } a_0 > r \cdot \hat{f}^{-1}, \\ S_2 : \; \bar{a} = a_0, \; \bar{c} = 0 & \text{asymp.stable for } a_0 < r \cdot \hat{f}^{-1}. \end{array} \tag{8f}$$

State S_1 corresponds to virus infection with a non-vanishing stationary virus concentration in the host, whereas S_2 models a situation where the virus dies out and the host recovers from the disease. With respect to stability, the two states are mutually exclusive: S_1, the state of infection, requires a minimum amount of susceptible material—cells or other forms of *nutrients*—and is asymptotically stable in the range $a_0 > r/\hat{f}$, whereas the state of extinction S_2 is asymptotically stable if

[7] We remark that autocatalytic steps play the key role in models of theoretical epidemiology. Features of the mechanism (8a–8g) for autocatalysis in the flow reactor remind, for example, of dynamical properties of models for infectious diseases (see, e.g., Mollison 1995).

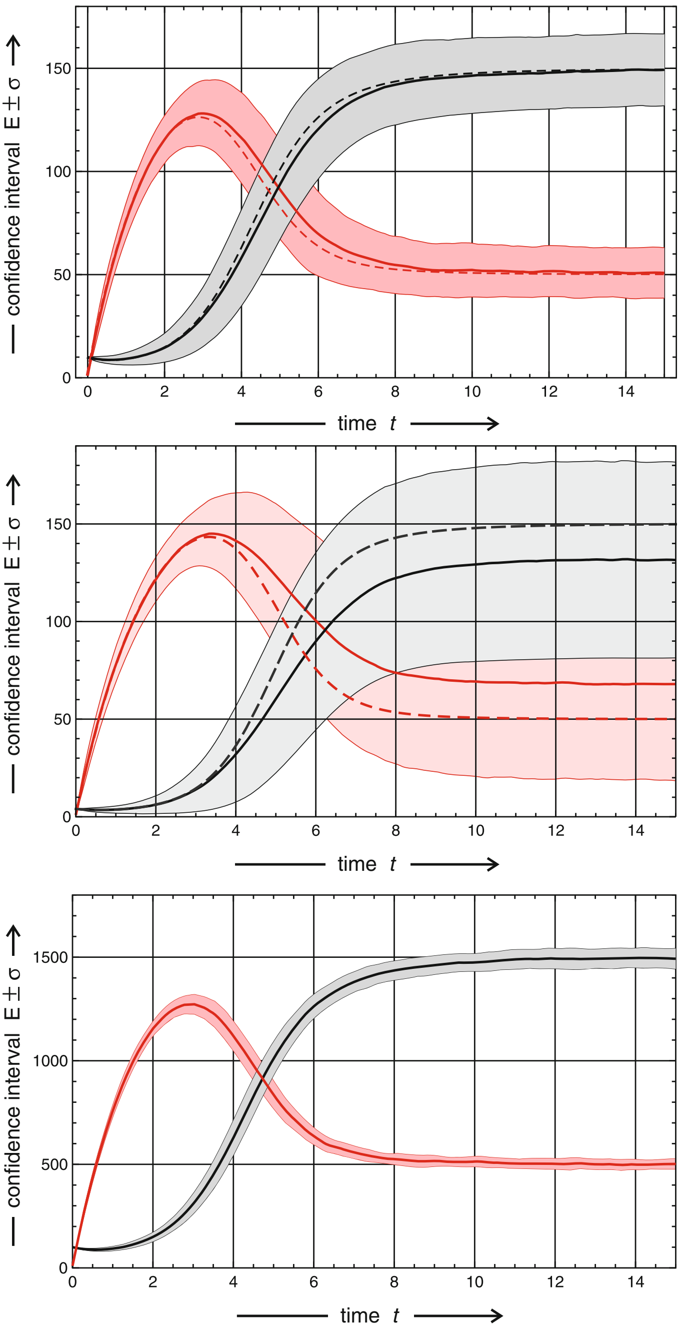
$a_0 < r/\hat{f}$. At the inflow concentration $a_0 = r/\hat{f}$, the system exhibits a transcritical bifurcation [see, e.g., (Strogatz 1994, pp. 50–52)].

In order to analyze the influence of stochasticity on the reaction scheme (8a–8g), we formulate a bivariate master equation in the two random variables $\mathcal{A}$ and $\mathcal{C} = \sum_{k=1}^{N} \mathcal{X}_k$, where $\mathcal{X}_k$ is the random variable associated with X_k, with the probabilities $P_A = P(\mathcal{A}(t) = A)$ and $P_C = P(\mathcal{C}(t) = C)$, respectively. This master equation is the probabilistic analogue to Eq. (8e),

$$\begin{aligned}\frac{\mathrm{d}P_A(t)}{\mathrm{d}t} &= a_0 r P_{A-1}(t) + (\hat{f}A \cdot C - a_0 r - rA)P_A(t)\\ &\quad - (\hat{f}(A+1)(C-1) - r(A+1))P_{A+1}(t)\\ \frac{\mathrm{d}P_C(t)}{\mathrm{d}t} &= \hat{f}(A+1)(C-1)P_{C-1}(t) - (\hat{f}A \cdot C + rC)P_C(t)\\ &\quad + r(C+1)P_{C+1}(t),\end{aligned} \tag{8g}$$

which describes the probabilistic development of populations starting from initial probability distributions $P_A(0)$ and $P_C(0)$. For practical purposes, sharp initial distributions, $P_A(0) = \delta_{A,A_0}$ and $P_C(0) = \delta_{C,C_0}$, are almost always applied, because they are technically simpler to handle and they allow for direct comparison of the solutions derived from the ODE (8e) and from the master Eq. (8g). It is straightforward to show that the initial condition $C(0) = C_0 = 0$ implies $C(t) = 0$ for all $t > 0$. Whenever the number of autocatalytic units has reached the value zero, it remains there or in other words, the state $S_2(C = 0)$ is an absorbing state or boundary. This fact represents also the major difference between the deterministic and the stochastic model of autocatalysis: Since S_2 is the only absorbing state of the system, all trajectories have to converge to this state in the limit of infinite time, $\lim_{t\to\infty} C(t) = 0$. Under conditions at which the state S_1 is asymptotically stable in the deterministic system (8e), $a_0 > r/\hat{f}$, the master equations support a *quasistationary* state (Nåsell 2011): The concentration of the autocatalyst approaches a constant value, and for a sufficiently large initial number of autocatalytic units C, $C(0) = C_0 > C_{\text{crit}}$, this value coincides with the value of c at $C(t) \approx \bar{c} = a_0 - r/\hat{f}$ and stays at this value for very long time, although the state S_2 will be reached with probability one at infinite time. For smaller values of C_0, the system converges to S_1 or goes extinct with a probability distribution depending on C_0. In Fig. 7, deterministic solution curves of (8e) are compared with the results of trajectory sampling for the master Eq. (8g).

There is one additional fundamental difference between the deterministic and the stochastic solution, which is also related to the fact that S_2 is absorbing. The deterministic equations are formulated in continuous variables, $a(t)$ and $x(t)$, which can become arbitrarily small without vanishing. This is not true for the stochastic variables, $\mathcal{A}$(t) and $\mathcal{X}(t)$, which are integers by definition and take on only the values $0, 1, \ldots$. For sufficiently small values of $\mathcal{X}(0)$, the system may die out in the early phase with a certain probability, which decreases with increasing $\mathcal{X}(0)$. The problem


confidence interval E ± σ
0
50
100
150
0
2
4
6
8
10
12
14
time t
confidence interval E ± σ
0
50
100
150
0
2
4
6
8
10
12
14
time t
confidence interval E ± σ
0
500
1000
1500
0
2
4
6
8
10
12
14
time t

◂ **Fig. 7** Autocatalysis in the flow reactor. The figure illustrates two different sources of stochasticity in autocatalytic systems: (i) Random fluctuations become important at small particle numbers for every chemical reaction and (ii) the stochastic autocatalytic reaction has an absorbing boundary for zero autocatalytic units as this may lead to significant differences between the stochastic and the deterministic system. The topmost plot shows the expectation values of concentrations within the one standard deviation confidence interval, $E \pm \sigma$, for the input material A and the autocatalyst X calculated from a sample of 1000 trajectories calculated by means of an algorithm attributed to Daniel Gillespie (1977). The expectation values are compared with the deterministic solution integrated from the ODE (Eq. (8e); *dotted lines*; $x(0) = 10$). The plot in the middle shows the same system with a different initial condition $(x(0) = 4)$. The change in the deterministic ODE integration concerns the initial phase, and both curves converge to identical stationary values, but the expectation value of the stochastic process leads to much smaller stationary amounts of autocatalyst when the initial value $x(0)$ was smaller. The plot at the bottom is dealing with the same reaction but with ten times larger particle numbers that give rise to smaller fluctuations relative to the expectation values. Shown are the expectation values within the one standard deviation confidence interval, $E \pm \sigma$ for the input material A and the autocatalyst X calculated from a sample of 100 trajectories. The deterministic solution curves coincide with the expectation value within the line width. Color code: $a(t)$ and $E(\mathcal{A}(t))$ *red*, and $x(t)$ and $E(\mathcal{X}(t))$ *black*. Choice of parameters for *upper* and *middle plot*: $a_0 = 200$, $r = 0.5$ and $f = 0.001$; initial conditions: $a(0) = 1$ and $x(0) = 4$ or $x(0) = 10$; choice of parameters for *lower plot*: $a_0 = 2000$, $r = 0.5$ and $f = 0.0001$; initial conditions: $a(0) = 10$ and $x(0) = 100$

is easily visualized by considering the probability densities $\mathrm{P}(\mathcal{A}(\mathrm{t}))$ and $\mathrm{P}(\mathcal{X}(\mathrm{t}))$, which are both bimodal for sufficiently long time and where the two modes correspond to the two states S_1 and S_2. Changing the initial condition, $\mathcal{X}(0)$ changes the relative weights of the two modes but not the (local) probability distributions around the modes themselves. In other words, for smaller initial numbers $\mathcal{X}(0)$, the probability to die out in the early phase is larger, more trajectories get absorbed in state S_2, and the expectation value $\mathrm{E}(\mathcal{X}(t))$ is diminished accordingly. This fact is nicely demonstrated by a comparison of the two plots at the top and in the middle of Fig. 7, which differ only in the initial condition $\mathcal{X}(0) = x(0)$ for which the values 10 or 4 were chosen. The deterministic solution curves converge to the same stationary values, whereas large differences in the stationary expectation values are observed. The phenomenon is important in virology and implies that initially small numbers of infectious units need not result in the development of disease.

Provided $\mathcal{X}(0)$ has been chosen large enough such that bifurcation in the early phase of the process plays no role, the stochastic expectation value follows the deterministic ODE solution except minor deviations, which disappear in the longtime limit when the (quasi-)stationary state is approached. Minor deviations between the stochastic expectation value and the deterministic solution are observed in full agreement with the analytic solution for the simple irreversible autocatalytic reaction $\mathsf{A} + \mathsf{X} \rightarrow 2\mathsf{X}$ (Arslan and Laurenzi 2008). Such small deviations have to be expected since the coincidence of the deterministic and the stochastic approach is true for linear systems only, in particular for first-order chemical reactions (van Kampen 2007, pp. 122–127). Autocatalysis in the flow reactor exhibits another feature: The fluctuations in the concentration of input material A meet the expectations for a conventional chemical systems and are near $\sqrt{N}$, whereas the fluctuations around the expectation values of the concentration of the autocatalyst X are larger in agreement with the mentioned analytical results.

Deterministic and stochastic quasispecies dynamics. Although total virus populations are commonly large, the importance of stochastic effects cannot be ruled out for reasons that are related to the existence of an absorbing boundary S_2. As we outlined for the simple autocatalytic process in the previous paragraph, initial phases have no influence on the deterministic longtime results but may bias the stochastic expectation values.

In order to be able to compare deterministic and stochastic results, we choose again the controllable experimental setup of a flow reactor. The model simplifies the set of materials required for replication by the assumption of a virtual compound A that flows into the reactor with a rate parameter r in the form of a solution with concentration a_0. As before, the inflow is compensated by an outflow of the volume of reactor solution resulting in reaction equations that have been analyzed by Schuster and Sigmund (1985):

$$* \xrightarrow{a_0 \cdot r} \mathsf{A}; \tag{9a}$$

$$\mathsf{A} + \mathsf{X}_k \xrightarrow{w_{jk}} \mathsf{X}_j + \mathsf{X}_k,\ j,k = 1,\ldots,N; \tag{9b}$$

$$\mathsf{A} \xrightarrow{r} \varnothing; \quad \text{and} \tag{9c}$$

$$\mathsf{X}_k \xrightarrow{r} \varnothing,\ k = 1,\ldots,N, \tag{9d}$$

where the parameters $w_{jk} = Q_{jk} f_k$ are the same as used in Eqs. (7a) and (7c). The kinetic differential equations are

$$\begin{aligned} \frac{\mathrm{d}a}{\mathrm{d}t} &= (a_0 - a)r - a\left(\sum\nolimits_{k=1}^{N} f_k x_k\right), \\ \frac{\mathrm{d}x_k}{\mathrm{d}t} &= a\left(\sum\nolimits_{j=1}^{N} Q_{kj} f_j x_j\right) - x_k r,\ k = 1,\ldots,N, \end{aligned} \tag{9e}$$

whereby we applied concentrations, $a = [\mathsf{A}]$ and $x_k = [\mathsf{X}_k](k = 1,\ldots,N)$, and made use of the condition $\sum_{i=1}^{N} Q_{ik} = 1$. Equation (9a–9e) can be modeled stochastically by means of a master equation that allows for numerical computation of trajectories by means of a simulation algorithm (see Fig. 8), which is attributed to Gillespie (1977, 2007).

The main issue of this section is a comparison of quasispecies formation according to (9a–9e) between the deterministic and the stochastic approach. Replication and mutation at constant total concentration, $c = \sum_{i=1}^{N} x_i(t) = \text{const}$, have been analyzed as a multi-type branching process (Demetrius et al. 1985), and the major result was that the longtime solutions of the ODE (7a) coincide with the stationary expectation values of the branching process. Since analytical results are available for the replication–mutation mechanism in exceptional cases only, the

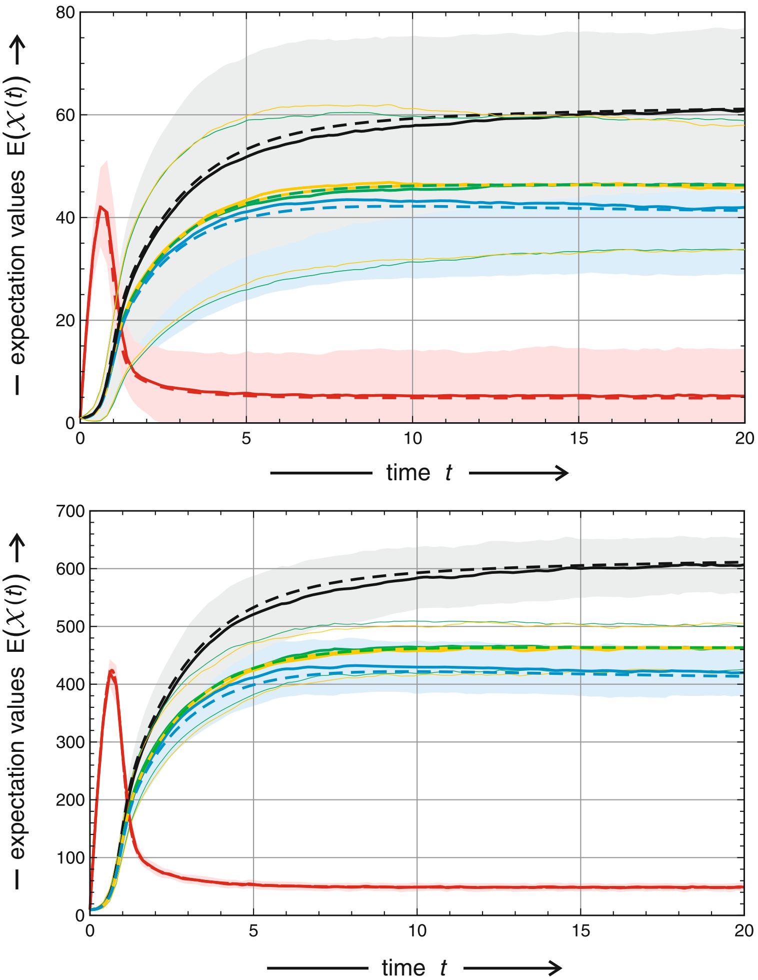


Fig. 8 Quasispecies formation in the flow reactor. The plots show the results of sampling 100 trajectories for the reaction mechanism (9a–9e) calculated by means of the Gillespie algorithm (Gillespie 1977). The smallest possible system was chosen: $l = 2$ with a master sequence X_1 (Γ_0) and three mutants X_2, X_3, and X_4, where X_2 and X_3 form the one-error class Γ_1 and yield identical deterministic solutions, and X_4 is the only sequence of the two-error class Γ_2. Shown are the expectation values within the one standard deviation confidence interval, $E \pm \sigma$, and the deterministic solutions obtained by integration of the ODE (9e) (*dashed lines*). Color code: $a(t)$ and $E(\mathcal{A}(t))$ *red* and *pink* confidence interval, $x_1(t)$ and $E(\mathcal{X}_1(t))$ *black* and *gray* confidence interval, $x_2(t)$ and $E(\mathcal{X}_2(t))$ *yellow* and confidence interval shown by *thin lines*, $x_3(t)$ and $E(\mathcal{X}_3(t))$ *green* and confidence interval shown by *thin lines*, and $x_4(t)$ and $E(\mathcal{X}_4(t))$ *blue* and *light blue* confidence interval. A single-peak landscape was used, and the uniform error rate model was applied. *Upper plot*, choice of parameters: $a_0 = 200$, $r = 0.5$, $f_m = 0.011$, $f = 0.010$ and $p = 0.1$ and initial conditions: $a(0) = x_1(0) = x_2(0) = x_3(0) = x_4(0) = 1$; *lower plot*, choice of parameters: $a_0 = 2000$, $r = 0.5$, $f_m = 0.0011$, $f = 0.0010$ and $p = 0.1$ and initial conditions: $a(0) = x_1(0) = x_2(0) = x_3(0) = x_4(0) = 10$

comparison of deterministic and stochastic quasispecies formation in the flow reactor is made here by means of numerical simulation. We choose two different initial conditions: (i) uniform distribution far away from the stationary distribution and (ii) the stationary distribution of the deterministic system. In the former case, the approach toward the stationary distribution is fast, and apart from some minor deviations, the expectation value obtained for the quasistationary distribution of the master equation is identical to the solution of the kinetic ODE (Fig. 8). The quasispecies in the flow reactor exhibits the same feature as autocatalysis: The fluctuations in the concentration of the input material A are near $\sqrt{N}$, whereas the fluctuations around the expectation values of concentrations for the members of the quasispecies, X_j, are larger, even larger than for autocatalysis. Again, minor deviation between the stochastic expectation value and the deterministic solution has to be expected and is observed indeed. Necessarily, the stochastic expectation and the deterministic result coincide at the stationary values.

A natural question to ask is whether or not the ranking of genotypes according to the frequency of occurrence in the population is changed through the action of fluctuations or, in other words, can the fittest genotype temporarily be outgrown by another sequence in the stochastic system. The answer is straightforward: The most important source of the fluctuations is self-enhancement of the replication process; the differences in the expectation values of the individual concentrations, $\mathrm{E}(\mathcal{X}_i(t))$, become smaller when the mutation rate increases; and thus, stochasticity may well interfere with the quasispecies structure in small populations or at high mutation rates. The two examples in Fig. 8 indicate two different scenarios: At the lower sample size $(a_0 = 200)$, the confidence intervals overlap and accordingly, we cannot expect that the most frequent sequence, which we isolate at some instant t_1, is the same as the most frequent sequence isolated at t_2, or in other words, the temporarily most frequent molecular species need not be fittest one. For the larger sample size $(a_0 = 2000)$, however, the confidence interval of the master sequence is well separated from the confidence intervals of the mutants and we can expect to find the master sequence almost always being present at the highest concentration irrespectively of fluctuations in the concentrations.

In summary, stochastic quasispecies formation meets all expectations from stochastic chemical kinetics. The most important difference to the deterministic approach concerns the fact that the quasispecies is quasistationary in the stochastic model, the only asymptotically stable state is the absorbing boundary, and every autocatalyst including mutant distributions such as quasispecies has to die out in the limit $t \to \infty$. In practice, this result is of academic interest only and has no consequences for real systems because the time to extinction is of hyper-astronomical length. A practical consequence, nevertheless, can arise from the bifurcation at short times: For small particle numbers, $C_0 = \sum_{k=1}^{N} \mathsf{X}_k(0) < 10$, the replication–mutation ensemble dies out with a non-negligible probability before it comes close to the quasistationary distribution. Apart from these specific effects, we obtained the general results that—not unexpectedly—concentration fluctuations are the more important and the higher the mutation rates, the smaller the population sizes are.

4 Quasispecies and Error Thresholds

Three kinds of studies on the dependence of quasispecies on mutation rates were performed: (i) analytical approximations using simple fitness landscapes and simplifications of the mutation matrix, for example, the *uniform error rate* model or the *zero mutational backflow* approximation; (ii) 'exact' numerical computations[8] on simple fitness landscapes with the full uniform error rate mutation matrix; and (iii) fully resolved fitness landscapes with the full uniform error rate mutation matrix. In general, we shall consider here stationary solutions of the replication–mutation ODE (7c) as functions of the error or mutation rate parameter per nucleotide site and replication denoted by p. In order to be able to handle the problem in a transparent way, we assume that the mutation rate is independent of the position on the sequence and characterize this simplifying assumption as the *uniform error rate model*. Then, the elements of the mutation matrix take on the simple form

$$Q_{ij}(p) = (1-p)^{l-\mathrm{d}^{H}_{ij}} p^{\mathrm{d}^{H}_{ij}} = (1-p)^{l} \varepsilon^{d^{\mathrm{H}}_{ij}} \quad \text{with} \quad \varepsilon = \frac{p}{1-p}, \tag{10}$$

wherein d^{H}_{ij} is the Hamming distance between the two sequences, X_i and X_j, and p is the mutation rate parameter. We shall assume further that we are dealing with binary sequences.[9] The concept of the error threshold is now introduced in three paragraphs reporting (i) 'exact' numerical results and two approximations, (ii) the zero mutational backflow assumption, and (iii) the phenomenological approach conceived by (Eigen 1971).

Solutions without approximation. 'Exact' solutions of the replication–mutation Eq. (7d) are obtained in terms of eigenvalues and eigenvectors of the value matrix W that, as said before, has to be a primitive matrix in order to fulfill Perron–Frobenius theorem (Seneta 1981, pp. 3–11 and 22–23). This theorem guarantees several important properties of the eigenvalues and eigenvectors of W. Two of them are of particular importance for the analysis of quasispecies and error thresholds: (i) The largest eigenvalue of W, λ_1 is non-degenerate, real and positive,

$$\lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_N|, \; \lambda_1 = |\lambda_1| > 0,$$

and (ii) all components of the right-hand eigenvector $\mathbf{b}_1$ associated with λ_1 are strictly positive. Both properties are required for physically meaningful results of the replication–mutation problem. Uniqueness of the solution means that the stationary mutant distribution is completely determined by the fitness landscape,

[8] By the notion 'exact,' we mean here 'without approximations.' In order to make clear that numerical computations can never be exact in the strict sense, we put *exact* between apostrophes.

[9] The use of binary sequences ($\kappa = 2$) facilitates several operations and implies no loss of generality. Natural four-letter sequences ($\kappa = 4$) can be encoded by binary sequences of twice the chain length.

$\mathcal{L} = \{f_k; k = 1,\ldots,N\}$, and the matrix of mutation frequencies, Q. Exclusively positive components of the eigenvector $\mathbf{b}_1$ implies that all mutants are present in the mutant distribution and no mutant can vanish as a consequence of consecutive replication and mutation events. The second issue has been discussed already in Chap. 1 (this volume). It is necessary to distinguish between the deterministic approach, which allows small concentrations down to any fraction of single molecules, and the stochastic approach where random variables are restricted to positive integers, although probabilities and moments of distributions can be arbitrarily small. In reality, a mutant distribution will consist always of a core of mutants, which are permanently present, and a fluctuating periphery.

After sufficient long time, the solutions of the replication–mutation Eq. (7d) are dominated by the largest eigenvalue λ_1:

$$x_j(t) \approx \frac{b_{j1}\beta_1(0)\exp(\lambda_1 t)}{\sum_{i=1}^{N} b_{i1}\beta_1(0)\exp(\lambda_1 t)} = \frac{b_{j1}}{\sum_{i=1}^{N} b_{i1}} = \bar{x}_j \quad \text{for large } t.$$

The longtime solution is independent of time t and initial conditions $\beta_k(0)$. It is fully determined by the fitness landscape and the mutation matrix, and it represents the genetic reservoir of an asexually reproducing species and has been characterized as *quasispecies* (Eigen and Schuster 1977, pp. 541, 549 ff.): 'A quasispecies is defined as a given distribution of macromolecular species with closely interrelated sequences dominated by one or several (degenerate) master copies.' Here, we can make it more precise by saying that the concentration ratios of the individual components are given by the largest eigenvector $\mathbf{b}_1$ of the value matrix W. A quasispecies contains one fittest genotype X_m—or in case of neutrality, several fittest genotypes—surrounded by a cloud of closely related mutants. The dominant genotype X_m is characterized as *master sequence*. The relative stationary concentrations of individual mutants, $\bar{x}_j$, are determined by their own fitness f_j and by their Hamming distance from the master sequence, $d^H_{\mathsf{X}_j\mathsf{X}_m} = d^H_{jm}$.

The computation of 'exact' numerical solutions is facilitated enormously by using single-peak fitness landscapes (3d) and adopting the uniform error rate approximation. Then, all mutants in a given class are described by the same ODE and their concentrations can be lumped together into a *class concentration*: $y_k(t) = \sum_{i=1}^{N_k} x_i$ with $\mathsf{X}_i \in \Gamma_k$ and $N_k = \binom{l}{k}$ (Nowak and Schuster 1989; Swetina and Schuster 1982). Figure 9 shows the mutant distribution of a quasispecies expressed in class concentrations as a function of the mutation rate parameter p. Starting at $p = 0$, where the master sequence represents the selected genotype, the relative concentration of the master sequence in the quasispecies decreases gradually and mutants gain in relative amount. At some critical mutation rate, $p = p_{\mathrm{tr}}$, the quasispecies distribution changes abruptly, and within a short interval Δp, the sequence distribution approaches the uniform distribution, $\mathcal{U} = \bar{x}_i = 1/l^2 \forall i = 1,\ldots,N$ (Swetina and Schuster 1982), which is the exact solution at $p = \frac{1}{2}$. The (approximate) uniform distribution then remains the stationary solution of the ODE (7a) in

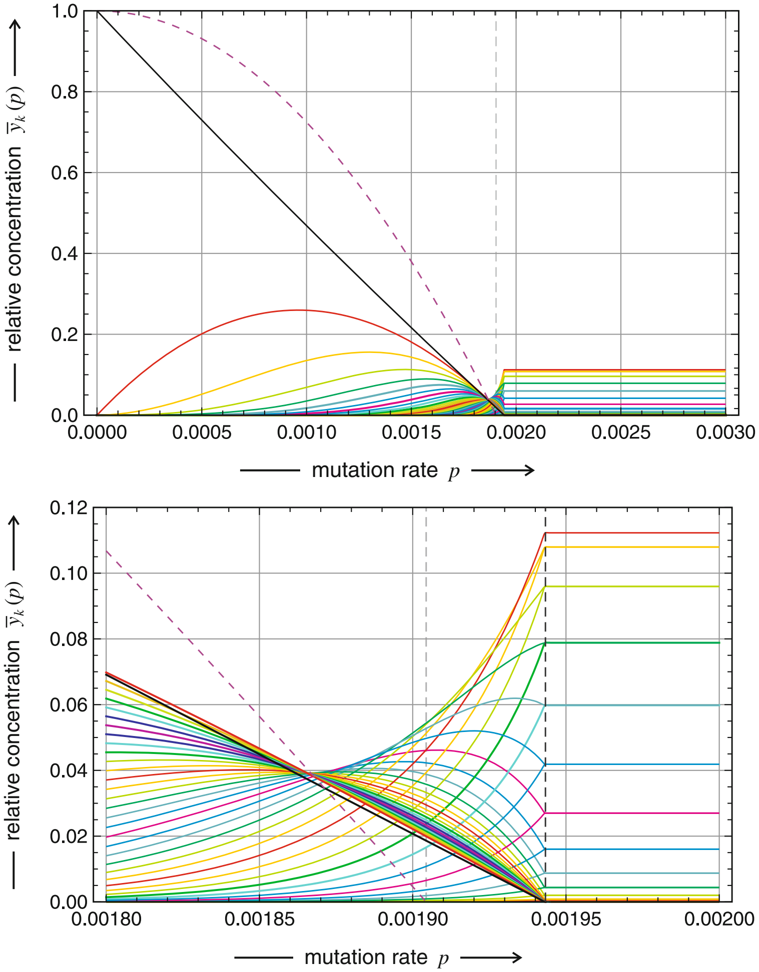


Fig. 9 Quasispecies as a function of the error rate. The *upper plot* shows numerically computed 'exact' curves. The *dashed gray line* indicates the error threshold at $p_{\mathrm{cr}} = 0.001904$ obtained by the phenomenological approach, and the *dashed violet curve* is the phenomenological total concentration $\hat{c}^{(0)}(p)$ obtained from (13). The *lower plot* is an enlargement and shows the error threshold derived from the mergence of the concentration curves for complementary classes, Γ_k and Γ_{l-k} (*dashed black line* at $p_{\mathrm{mg}}^{(\theta)} = 0.001943$). According to our knowledge, the work of (Swetina and Schuster 1982) was the first publication showing this shape of an error-induced transition in quasispecies. The positions of the error threshold calculated from level crossing are as follows: $p_{\mathrm{tr}}^{\vartheta} = 0.00192229, 0.00194101, 0.00194288, 0.00194307,$ and 0.00194308 for $\vartheta = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$, and 10^{-6}. Parameters: $l = 50, f_m = 1.1, \bar{f}_{-m} = f = 1.0$

the entire range $p_{\mathrm{tr}} \le p \le \frac{1}{2}$, and at $p = \frac{1}{2}\mathcal{U}$, it becomes the exact solution. The transition at $p = p_{\mathrm{tr}}$ increases in sharpness with increasing chain lengths l and reminds of cooperative transitions known in the theory of polymers (Lifson 1961;

Schwarz 1968; Zimm 1960). Exploiting the analogy of the quasispecies approach and equilibrium statistical mechanics of a two-dimensional spin lattice Tarazona (1992) was able to show that the error threshold on the single-peak fitness landscape corresponds to a first-order phase transition in the limit of infinite chain lengths l.

Two quantitative measures for the location of the error threshold are obtained by straightforward and simple numerical procedures:

(i) *Level crossing* determines the p-value at which the curve for the stationary concentration of the master sequence, $\bar{x}_m(p)$, crosses a predefined concentration level, $\bar{x}_m(p_{\mathrm{tr}}^{(\vartheta)}) = \vartheta$ with ϑ being a threshold value that has to be chosen small enough for a given chain length l.[10] In the example shown in Fig. 9, the convergence of $p_{\mathrm{tr}}^{(\vartheta)}$ with decreasing values of ϑ is very fast. Appropriately, the converged limit is taken as the position of the transition. We remark that we are dealing here with *semiconvergence* because the curve bends off to the at still lower ϑ-values in order to reach the point $\bar{x}_m(\frac{1}{2}) = 1/2^l$.

(ii) *Complementary class mergence* makes use of the fact that the uniform distribution implies coalescence of the concentrations, $\bar{y}_k = \sum_{i=1}^{N_k} \bar{x}_i,\ \mathsf{X}_i \in \Gamma_k$, for complementary classes (Γ_k, Γ_{l-k}), since $\binom{l}{k} = \binom{l}{l-k}$ and $\bar{y}_k = \bar{y}_{l-k} = \binom{l}{k}/2^l$. Accordingly, the p-values $p = (p_{\mathrm{mg}}^{(\theta)})_k$ at which the difference $\Delta_k = |\bar{y}_k - \bar{y}_{l-k}|$ becomes as small as some predefined value $(\Delta_k)_{\mathrm{cr}} = \theta$ can be taken as the kth coalescence error rate. Then, a measure for the sharpness of the transition is given by the width of the band spanned by the different locations of $(p_{\mathrm{mg}}^{(\theta)})_k; k = 0, \ldots, \lfloor \frac{l}{2} \rfloor$, i.e., $\Delta p_{\mathrm{mg}}^{(\theta)} = \max((p_{\mathrm{mg}}^{(\theta)})_k) - \min((p_{\mathrm{mg}}^{(\theta)}))_k$ with $k = 0, \ldots, \lfloor \frac{l}{2} \rfloor$. In the cases studied here, the values $(p\frac{(\theta)}{\mathrm{mg}})_k$ did not change monotonously with k but increased from $k = 0$ up to some maximum value but further on decreased until $k = \lfloor \frac{l}{2} \rfloor$ has been reached (examples for class mergence are presented in the paragraph *error thresholds on simple landscapes*). It is important to stress that both measures for the location of the transition, the converged value from level crossing as well as the value from complementary class mergence, yield very similar results for realistic chain lengths ($l \gg 50$) as shown in Fig. 9 for a single-peak landscape. Despite the fact that we have no analytical expression for p_{tr} and p_{mg}, the numerically calculated values are nicely confirming the existence of the error threshold as a transition phenomenon of the cooperative transition type.

[10] For sufficiently long sequences, the particular choice ϑ = 0.01, 0.001 or 0.0001 is unimportant because the results for small values are very similar and converge to a limit (see Fig. 9), but for short chains, the concentration values of the uniform distribution $\mathcal{U}$ set a lower limit for $\bar{x}_m(p)$. For example, in case of $l = 10$, the value $\bar{x}_m(\frac{1}{2}) = 1/2^l = 1/1024$ is compatible only with the choice $\vartheta = 1/100$ because $\vartheta = 1/1000$ is too close to $\bar{x}_m(\frac{1}{2})$.

The zero mutational backflow approximation. The notion *mutational backflow* concerns mutations from the mutant cloud back to the master sequence:

$$\Phi_{\mathsf{X}_m \leftarrow \mathsf{X}_{(j)}} = \sum_{j=1,j\neq m}^{N} Q_{mj} f_j \bar{x}_j = \sum_{j=1,j\neq m}^{N} w_{mj} \bar{x}_j. \tag{11}$$

Zero mutational backflow and the consistent neglect of the mutational flow between mutants imply that all off-diagonal elements in the mutation matrix Q are zero except those describing the mutations from the master sequence to the mutant cloud. In other words, Q contains only the elements in the column of the master sequence, $Q_{jm}; j = 1, \ldots, N; j \neq m$, and the diagonal terms, Q_{ii}. The replication–mutation Eq. (7a) is modified and becomes much simpler:

$$\frac{\mathrm{d}x_m^{(0)}}{\mathrm{d}t} = (Q_{mm} f_m - \phi) x_m^{(0)}, \tag{12a}$$

$$\frac{\mathrm{d}x_j^{(0)}}{\mathrm{d}t} = (Q_{jj} f_j - \phi) x_j^{(0)} + Q_{jm} f_m x_m^{(0)}. \tag{12b}$$

The superscript ‘(0)’ indicates the approximation. The flow by definition is adjusted to compensate for the net growth and accordingly takes on the form

$$\begin{aligned} \phi(\mathrm{x}^{(0)}(t)) &= \frac{1}{c^{(0)}} \left(\sum_{i=1}^{N} Q_{ii} f_i x_i^{(0)} + \sum_{j=1,j\neq m}^{N} Q_{jm} f_m x_m^{(0)} \right) \\ &= \frac{1}{c^{(0)}} \left(f_m x_m^{(0)} + Qf(c^{(0)} - x_m^{(0)}) \right), \end{aligned} \tag{12c}$$

where we applied a single-peak landscape with $\bar{f}_{-m} = f$ and the uniform error approximation $Q = (1-p)^l$. Stationary solutions of the ODE are readily calculated since Eqs. (12a) and (12c) contain only the variable $x_m^{(0)}$:

$$\bar{x}_m^{(0)} = \bar{c}^{(0)} \frac{Q(1-\sigma_m^{-1})}{1-Q\sigma_m^{-1}} \quad \text{and} \quad \bar{x}_j^{(0)} = \bar{c}^{(0)} \frac{Q\varepsilon^{d^{\mathrm{H}}_{jm}}}{1-Q\sigma_m^{-1}}. \tag{12d}$$

The input coming from the fitness landscape, $\sigma_m = f_m / \bar{f}_{-m}$, has been called the superiority of the master sequence, and it weights the fitness of the master sequence, f_m, against the mean fitness of all sequences except the master sequence: $\bar{f}_{-m} = \sum_{i=1,i\neq m}^{N} f_i x_i / (c - x_m)$. The assumption of zero mutational backflow is a fairly good approximation for small mutation rates (Swetina and Schuster 1982) and can be used as a reasonably accurate estimate of the stationary concentration of the master sequence, $\bar{x}_m(p)$, and the one-error class, $\bar{y}_1(p)$ (see Fig. 11), but fails to model quasispecies at larger mutation rates, in particular, near the error threshold.

The phenomenological approach. In the seminal paper on self-organization of macromolecules, Eigen (1971) introduced a variant of the zero mutational approximation that also allows for the derivation of analytical solutions for the stationary concentration. Eigen addressed his approach as *phenomenological theory of selection,* and therefore, we shall characterize it here as *phenomenological* as well. The approximation is only introduced into the growth term, and the change is not compensated in the flow ϕ. Accordingly, the condition of *constant organization* or constant population size of Eq. (7a) is violated, and hence, the total concentration, $\hat{c}^{(0)}$, will be a function of the mutation rate parameter p. The modified equations are identical with (12a) and (12b), but the flow term is different:

$$\phi(\mathrm{x}^{(0)}(t)) = \frac{1}{c^{(0)}} \sum_{i=1}^{N} f_i x_i^{(0)} = \frac{1}{c^{(0)}} \left(f_m x_m^{(0)} + \bar{f}_{-m}(c^{(0)} - x_m^{(0)})\right). \qquad (12c□)$$

Again, the problem is reduced to an ODE in a single variable and the stationary solution can be obtained straightforwardly[11]

$$\begin{aligned} \hat{x}_m^{(0)} &= \frac{Q - \sigma_m^{-1}}{1 - \sigma_m^{-1}}, \ \hat{x}_j^{(0)} = \varepsilon d_{jm}^{\mathrm{H}} (Q - \sigma_m^{-1})(1 - \sigma_m^{-1})^2, \quad \text{and} \\ \hat{c}^{(0)} &= \frac{(1 - Q\sigma_m^{-1})(Q - \sigma_m^{-1})}{Q(1 - \sigma_m^{-1})^2}. \end{aligned} \qquad (13)$$

The normalized concentrations of the phenomenological approach,

$$\frac{\hat{x}_m^{(0)}}{\hat{c}^{(0)}} = \frac{Q - \sigma_m^{-1}}{1 - \sigma_m^{-1}} \cdot \frac{Q(1 - \sigma_m^{-1})^2}{(1 - Q\sigma_m^{-1})(Q - \sigma_m^{-1})} = \frac{Q(1 - \sigma_m^{-1})}{1 - Q\sigma_m^{-1}} = \bar{x}_m^{(0)}$$

and $\hat{x}_j^{(0)}/\hat{c}^{(0)} = \bar{x}_j^{(0)}$ are identical to the solutions of the zero mutational backflow approximation. This is to be expected from a previously derived very general result, which states that normalized or relative concentrations are independent of the flow term $\phi(t)$ as long as the growth functions—here $\sum_{i=1}^{N} Q_{ji} f_i x_i$—are linear and the population size $c(t)$ does not vanish (Eigen and Schuster 1978, p. 13). The factor $Q - \sigma_m^{-1}$, which is common to all concentrations in the phenomenological approach, decreases with increasing mutation rate p and eventually becomes zero at the critical mutation rate $p = p_{\mathrm{cr}} = 1 - \sigma^{-1/l}$. At this point, the total concentration becomes zero and hence, the whole quasispecies vanishes. The key relation for survival of the quasispecies is the relation

$$Q \cdot \sigma_m \geq 1, \qquad (14)$$

[11] The stationary concentrations of the phenomenological approach are denoted by the 'hat' symbol: $\hat{x}_m^{(0)}$, $\hat{x}_j^{(0)}$, $\hat{y}_k^{(0)}$, $\hat{c}^{(0)}$, etc.

Again the problem is reduced to an ODE in a single variable and the stationary solution can be obtained straightforwardly[11]

$$\begin{aligned} \hat{x}_m^{(0)} &= \frac{Q-\sigma_m^{-1}}{1-\sigma_m^{-1}}\,, \ \hat{x}_j^{(0)} \ = \ \frac{\varepsilon^{\mathrm{d}^{\mathrm{H}}_{jm}}(Q-\sigma_m^{-1})}{(1-\sigma_m^{-1})^{\,2}}\,, \quad \text{and} \\ \hat{c}^{(0)} &= \frac{(1-Q\sigma_m^{-1})(Q-\sigma_m^{-1})}{Q(1-\sigma_m^{-1})^2}\ . \end{aligned} \tag{13}$$

The normalized concentrations of the phenomenological approach,

$$\frac{\hat{x}_m^{(0)}}{\hat{c}^{(0)}} \ = \ \frac{Q-\sigma_m^{-1}}{1-\sigma_m^{-1}} \cdot \frac{Q(1-\sigma_m^{-1})^2}{(1-Q\,\sigma_m^{-1})(Q-\sigma_m^{-1})} \ = \ \frac{Q(1-\sigma_m^{-1})}{1-Q\,\sigma_m^{-1}} \ = \ \overline{x}_m^{(0)}$$

and $\hat{x}_j^{(0)} \big/ \hat{c}^{(0)} = \overline{x}_j^{(0)}$ are identical to the solutions of the zero mutational backflow approximation.

[11] The stationary concentrations of the phenomenological approach are denoted by the 'hat' symbol: $\hat{x}_m^{(0)}$, $\hat{x}_j^{(0)}$, $\hat{y}_k^{(0)}$, $\hat{c}^{(0)}$, etc.

which can be interpreted easily in qualitative terms: The loss of master copies due to error-prone replication, $Q = (1-p)^l$, has to be overcompensated by the higher fitness of the master as expressed by the superiority σ_m.

Beyond this error rate, $p > p_{\mathrm{cr}}$, genetic information cannot be transferred to future generations and therefore, the phenomenon has been characterized as *error threshold*. The equation for p_{cr} can be elegantly translated into a maximum error rate or a maximum chain length condition for successful transfer of genetic information to future generations. Simplified equations,

$$p_{\max} = p_{\mathrm{cr}} = 1 - \sigma^{-1/l} \approx \frac{\ln \sigma_m}{l} \quad \text{and} \quad l_{\max} \approx \frac{\ln \sigma_m}{p}, \tag{15}$$

were discussed in Chap. 1 (this volume) and are frequently applied to problems in virology, cancer research, and prebiotic evolution. In virology and cancer research, the key issue concerns the possibility to extinguish infections or stop proliferation by driving populations of viruses or cells into extinction by increasing the mutation rate. Two processes are fundamental for achieving this goal, either replication is pushed above the error threshold where the genetic information is lost or a large percentage of lethal variants is produced and the population becomes extinct (see Tejero et al. (2010) and Chap. 7, this volume). In prebiotic evolution, the phenomenological equation for the error threshold sets a limit to the chain l length of polynucleotides and thereby also to the information content, which can be faithfully transferred to future generation on the population level (see, e.g., Eigen and Schuster 1982).

The most remarkable property of the phenomenological approach is the quality of the results: As shown in Fig. 9 for a rather short chain length $l = 50$, the position of the transition to the uniform distribution, p_{tr}, is very close to the critical error rate p_{cr} where the quasispecies vanishes in the phenomenological approach, and the approximation becomes even better for increasing chain lengths. Here, we shall analyze the mathematical background of the approximations by considering the entire range of mutation rate parameters, $0 \le p \le \frac{1}{2}$. As shown in Fig. 11, the continuation of $\hat{x}_m^{(0)}(p)$ beyond the error threshold converges in the range of negative concentrations for $p \to \frac{1}{2}$ to the value

$$\hat{x}_m^{(0)}\left(\frac{1}{2}\right) = \frac{2^{-l} - \sigma_m^{-1}}{1 - \sigma_m^{-1}} \approx -\frac{1}{\sigma_m - 1}$$

(for binary sequences and $2^l \gg \sigma_m$). The curve $\hat{x}_m^{(0)}(p)$ is close to the exact curve $\bar{x}_m(p)$ up to the error threshold but then extends to negative values. The comparison with the consistent zero mutational backflow approximation $\bar{x}_m^{(0)}(p)$ shows three interesting features:

(i) Because the zero mutational backflow approximation fulfills the condition of constant population size, it reaches a positive value at $p = \frac{1}{2}$, which lies below the exact value of $\bar{x}_m = 2^{-l}$,

$$\bar{x}_m^{(0)}\left(\frac{1}{2}\right) = \frac{1-\sigma_m^{-1}}{2^l - \sigma_m^{-1}} \approx \frac{1-\sigma_m^{-1}}{2^l}.$$

Apart from very small mutation rates, the curve of the phenomenological approach $\hat{x}_m^{(0)}(p)$ lies closer to the exact curve $\bar{x}_m$ than the zero mutational backflow curve $\bar{x}_m^{(0)}(p)$ in the whole range $0 \le p \le \frac{1}{2}$.

(ii) Qualitatively, the downshift of the curve $\hat{x}_m^{(0)}(p)$ relative to the zero mutational backflow curve $\bar{x}_m^{(0)}(p)$ is easily explained: The flux ϕ is larger in the phenomenological approach than in the zero backflow approximation, and, other things being equal, this shifts the curve to lower values.

(iii) No only the relative stationary concentration $\hat{x}_m^{(0)}(p)$ is an excellent approximation for the master sequence but also the curve for the one-error mutants fits the exact curve very well, $\hat{y}_1^{(0)}(p) \approx \bar{y}_1(p)$, whereas the approach is very poor for all other sequences with two or more mutations (Fig. 11). This result is readily explained in terms of the mutation flow (Fig. 10): The sequences of the one-error class have the master sequence and $(l-1)$ sequences from the two-error class in their immediate neighborhood. The master sequence is described fairly correctly, and for small mutation rates, the sequences in the two-error class are present in very small concentrations only, and accordingly, also the absolute error in the stationary concentration is small. Sequences in the two-error class and in the higher error classes, however, receive mutational input in the zero backflow approximation only from the master sequence, and the larger inputs from the next lower error class are neglected.

All results calculated by the phenomenological approach are readily understood in qualitative terms, the high accuracy obtained in the approximation of the position of the error threshold, which makes the approach to an extremely useful and easy-to-handle tool, still waits for an explanation.

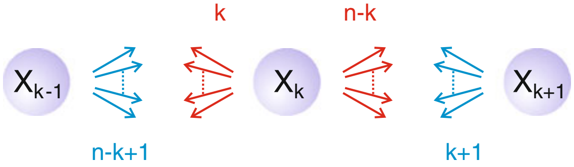


Fig. 10 Mutational flow in binary sequence space. The figure sketches the mutational flow on a hypercube. Every sequence has l Hamming distance one neighbors, k neighbors are situated in the class Γ_{k-1}, and $l-k$ neighbors in the class Γ_{k+1}. This implies that a sequence in Γ_k produces one-error mutants for k sequences in class Γ_{k-1} and for $l-k$ mutants in class Γ_{k+1}

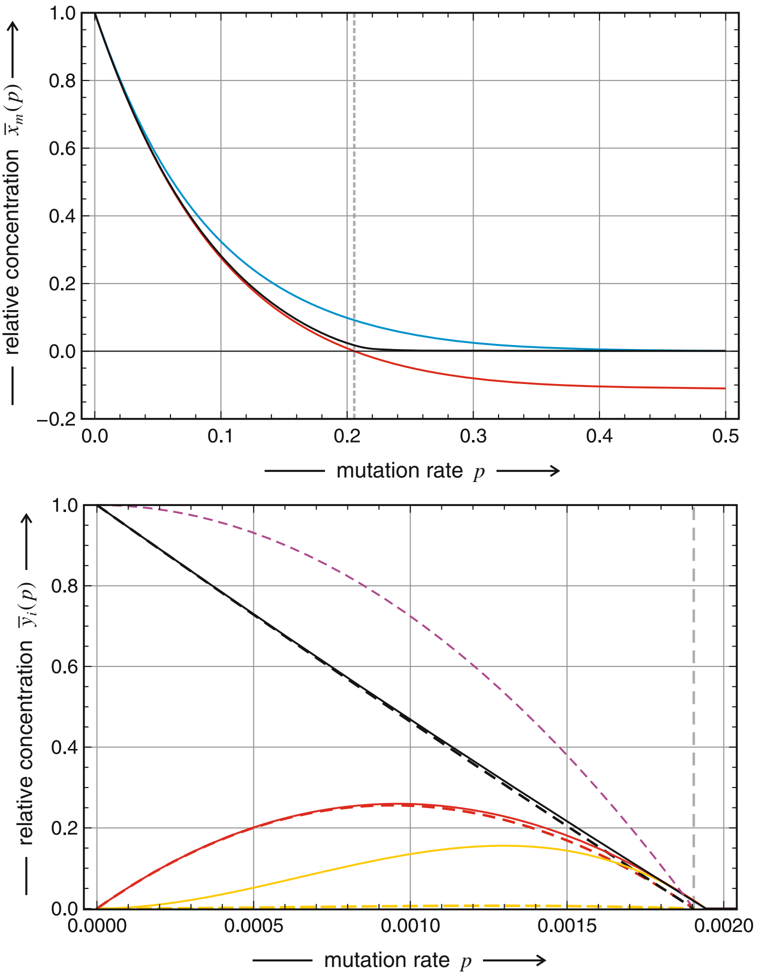


Fig. 11 The zero mutational backflow approximation and the phenomenological approach. In the *upper part* of the figure, the exact stationary concentration $\bar{x}_m(p)$ (*black*) is compared with the zero mutational backflow approximation $\bar{x}_m^{(0)}(p)$ (*blue*) and the phenomenological approach $\hat{x}_m^{(0)}(p)$ (*red*). Choice of parameters: $l = 10$, $f_m = 10.0$, $\bar{f}_{-m} = f = 1.0$, yielding an error threshold $p_{\mathrm{cr}} = 0.2057$. The *lower plot* demonstrates the excellent agreement of the phenomenological approach (*dashed lines*) with the exact stationary concentrations of the master $\bar{x}_m(p) = \bar{y}_0(p)$ (*black*) and the one-error class $\bar{y}_1(p)$ (*red*). As outlined in the text, the agreement is poor as expected for the two-error class (*yellow*) and all other higher mutational classes (Note that the *dashed yellow curve* is hardly distinguishable from the abscissa axis). The error threshold is indicated by a gray dashed line and the total concentration of the phenomenological approach is shown in violet. Choice of parameters: $l = 50$, $f_m = 1.1$, $\bar{f}_{-m} = f = 1.0$, and $p_{\mathrm{cr}} = 0.001904$

Error threshold on simple landscapes. Quasispecies on different simple fitness landscapes have been compared previously in several publications (see, e.g., Wiehe 1997; Schuster 2011). Here, we summarize only the most relevant findings. Some smooth landscapes, for example, the linear (3a) and the multiplicative landscape (3b), do not exhibit a cooperative transition like abrupt change of the quasispecies distribution in the $(\bar{y}_k, p)$-plot $(k = 0, \dots, l)$. In other words, the quasispecies changes smoothly from the selection of the master sequence, $\varUpsilon(0) = (\bar{x}_m = 1, \bar{x}_j = 0; j = 1, \dots, N, j \neq m)$, to the uniform distribution, $\varUpsilon(\frac{1}{2}) = \mathcal{U}$. The error threshold on the single-peak landscape (3d) has been discussed in great detail in the preceding paragraphs: It supports an error threshold near the position $p_{\mathrm{cr}} = (Q - \sigma_m^{-1})/(1 - \sigma_m^{-1})$. The hyperbolic landscape (3c) shows a cooperative transition, but it looks different from the error threshold on the single-peak landscape since it does not directly lead to the uniform distribution $\mathcal{U}$. The single-peak linear landscape (3e) eventually shows an error threshold provided the position of the step, h, is located at a sufficiently low class number k. Interestingly, the error threshold occurs at a higher mutation rate p separated from the decline of the stationary concentration of the master sequence.

All simple landscapes can be readily classified by resolving the error threshold phenomenon into three features: (i) a decrease of the stationary concentration of the master sequence to very small values—still above the uniform concentration $(\bar{x}_m = 2^{-l})$, (ii) a sharp transition of the quasispecies from the characteristic fitness and Hamming distance determined distribution of mutants to a different distribution that is characteristic for high mutation rates, and (iii) the nature of the high mutation rate distribution that often but not always is the uniform distribution $\mathcal{U}$. Quantitative measures for the first two criteria have been given in the paragraph on approximation-free solutions. For feature (i), this is the p-value at which the curve for the stationary concentration of the master sequence crosses a predefined concentration level, $\bar{x}_m(p_{\mathrm{tr}}^{(\vartheta)}) = \vartheta$, and for features (ii) and (iii), we recall the mergence of the stationary concentrations of complementary classes, $|\bar{y}_k(p_{\mathrm{mg}}^{(\theta)}) - \bar{y}_{l-k}(p_{\mathrm{mg}}^{(\theta)})| = \theta$, where the spectrum of $\left(p_{\mathrm{mg}}^{(\theta)}\right) k$-values defines both the position and the width of the transition. It is worth remembering that for the examples presented in Fig. 9 and Table 1 $(h = 0)$, both quantitative measures give the same result, $p_{\mathrm{tr}}^{(\vartheta)} \approx p_{\mathrm{mg}}^{(\theta)}$ for $\vartheta = \theta$.[12]

The single-peak linear landscape (3e) with different h-values provides an excellent study case for the quantitative evaluation of error thresholds (Fig. 12). The width of the error threshold transition for sequences with l = 10 is compared for the single-peak landscape and the single-peak linear landscapes with h = 2, 3, and 4.[13]

[12] This agreement is not accidental as a simple consideration shows: The lowest mutation rate for merging two classes is $(p_{\mathrm{mg}}^{(\theta)})_0$, the p-value where $\varDelta_0 = |\bar{y}_0 - \bar{y}_l| = |\bar{x}_m - \bar{x}_{-m}| = \theta$. Since the concentration of the complementary sequence of the master sequence with $\mathrm{d}^{\mathrm{H}}_{\mathsf{X}_m \mathsf{X}_{-m}} = l$ is commonly very small, $\bar{x}_{-m} \ll \bar{x}_m$, we find for $\vartheta = \theta$: $\varDelta_0 \approx \bar{x}_m$ and $p_{\mathrm{tr}}^{(\vartheta)} \approx \min(p_{\mathrm{mg}}^{(\theta)})_k = (p_{\mathrm{mg}}^{(\theta)})_0$.

[13] The single-peak linear landscape with $h = 1$ is identical with the single peak fitness landscape. The error threshold for h = 5 extends almost to $p = \frac{1}{2}$, and landscapes with $h > 5$ do not support error thresholds at all.

Table 1 Concentration level crossing and complementary class mergence near the error threshold

h	Level crossing $p_{\mathrm{tr}}^{(\vartheta)}$			Class mergence $p_{\mathrm{mg}}^{(\theta)}$	
	$\vartheta = 1/100$	$\vartheta = 1/1000$	$\vartheta = 1/10000$	$\theta = 1/1000$	$\Delta p_{\mathrm{mg}}^{(0.01)}$
0	0.1067	0.1103	0.1110	0.1103–0.1111	0.0008
2	0.1097	0.1227	0.1252	0.1227–0.1282	0.0055
3	0.0999	0.1342	0.1428	0.1342–0.1758	0.0416
4	0.0811	0.1365	0.1626	0.1365–0.3360	0.1995
5	0.0638	0.1244	0.1777	0.1244–0.4453	0.3209
6	0.0513	0.1053	0.1787	–	–
7	0.0426	0.0876	0.1650	–	–
8	0.0364	0.0737	0.1449	–	–

The decline of the master class, $\bar{y}_0 = \bar{x}_0$, at p-values near the error threshold p_{cr} is illustrated by means of the points $p_{\mathrm{tr}}^{(\vartheta)}$ where the curves cross the level $\bar{x}_0(p) = \vartheta$. Complementary class mergence is characterized quantitatively by the band between the lowest and the highest $(p_{\mathrm{tr}}^{(\vartheta)})_k$-value. The lowest value is always observed with $k = 0$ (see Fig. 12). Parameters: $l = 20$, $f_0 = 10.0$, and $f_n = 1.0$ yielding an error threshold at $p_{\mathrm{cr}} = 0.1088$

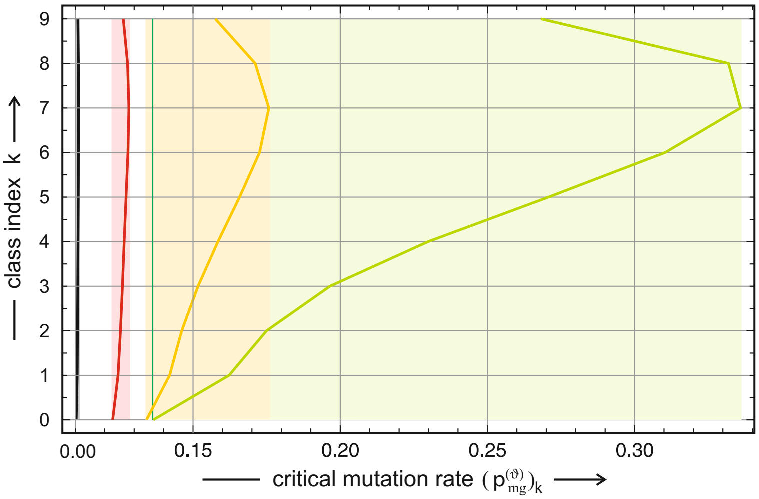


Fig. 12 The error threshold on single-peak linear landscapes. Shown are the critical mutation rates at which the curves for the stationary class concentrations approach each other up to a predefined difference, $(p_{mg}^{(\theta)})_k = |\bar{y}_k - \bar{y}_{l-k}| = \theta$ with $k = 0, 1, \ldots, \lfloor \frac{l}{2} \rfloor$. The areas in light colors represent the widths of the transitions. Parameter choice: $l = 20, f_0 = 10.0, f_n = 1.0$, $h = 0$ (*black*), $h = 2$ (*red*), $h = 3$ (*yellow*), and $h = 4$ (*chartreuse*)

Computed values for level crossing and complementary class mergence are shown in Table 1. The excellent agreement between the lower limit of the complementary class mergence values, $\left(p_{\mathrm{mg}}^{(\theta)}\right)_0$, and the level crossing value for the same value, $p_{\mathrm{tr}}^{(\vartheta)}$

Table 2 Complementary class mergence on single-peak and additive landscapes

$(p_{\mathrm{mg}}^{(0.01)})_k$	Additive landscape f_k (3a)		Single-peak landscape f_k (3d)	
k	$v = 10$	$v = 20$	$v = 10$	$v = 20$
0	0.01630	0.002552	0.01164	0.004969
1	0.06791	0.004363	0.01210	0.004977
2	0.17233	0.007967	0.01261	0.004983
3	0.24174	0.012590	0.01282	0.004990
4	0.22508	0.027993	0.01230	0.004997
5	–	0.064427	–	0.005005
6	–	0.113894	–	0.005011
7	–	0.153431	–	0.005013
8	–	0.162072	–	0.005009
9	–	0.120962	–	0.004990
$\Delta p_{\mathrm{mg}}^{(0.01)}$	0.22544	0.15952	0.00118	0.000045
$p_{\mathrm{tr}}^{(0.01)}$	0.01634	0.002552	0.01175	0.004969

Complementary class mergence characterized quantitatively by the bandwidth between the lowest and the highest $\left(p_{\mathrm{mg}}^{(\vartheta)}\right)_k$-value for $\vartheta = 0.01$ is compared for single-peak and linear landscapes with chain lengths $v = 10$ and $v = 20$. In all cases, the lowest value is always observed with $k = 0$ (see Fig. 12). In addition, the values for level crossing of the master class at $p_{\mathrm{tr}}^{(\theta)}$-values with $\theta = 0.01$ are given. Parameters: $v = 10$ and 20, $f_0 = 1.1$, and $f_n = 1.0$ for the single peak and $f_n = 0.9$ for the linear landscape yielding error thresholds on the single-peak landscape at $p_{\mathrm{cr}} = 0.00949$ and $p_{\mathrm{cr}} = 0.00475$, respectively

with $\theta = \vartheta$, is remarkable in all four cases ($h = 0,2,3,4$).[13] For $h = 5$, the band of complementary class mergence becomes so broad—$0.1244 \leq p_{\mathrm{mg}}^{(1/1000)} \leq 0.4453$ in the example shown in Table 1—that it extends almost to the limit $p = \frac{1}{2}$. For $h \geq 6$, no threshold is observed.

Equipped with the quantitative diagnostic tools for the detection of error thresholds, we return to the comparison of additive (3a) and single-peak landscapes (3d) in Table 2. The quantitative indicators reflect perfectly the visual inspection of the $\bar{y}_k(p)$-curves: For the chain length $l = 10$, the width of complementary class merging, $\Delta p_{\mathrm{mg}}^{(0.01)}$, for the additive landscape is 200 times as broad as for the single-peak landscape, and for $l = 20$, this factor is even 3500. Indeed, the error threshold has become very narrow already for short sequences of length $l = 20$ and shrinks further with increasing l. In addition, the relation between the two measures, $p_{\mathrm{tr}}^{(0.01)} \approx (p_{\mathrm{mg}}^{(0.01)})_0$, is already fulfilled up to 10^{-6} for sequences with $l = 20$.

In order to provide a hint on the prerequisites for the existence of an error threshold, we consider the derivative of the simple fitness landscapes with respect to the class index k (Fig. 13). All landscapes, which have a slope or derivative $|\partial f(k)/\partial k| > \alpha_{\mathrm{cr}}$, support error thresholds, whereas all less steep landscapes shown

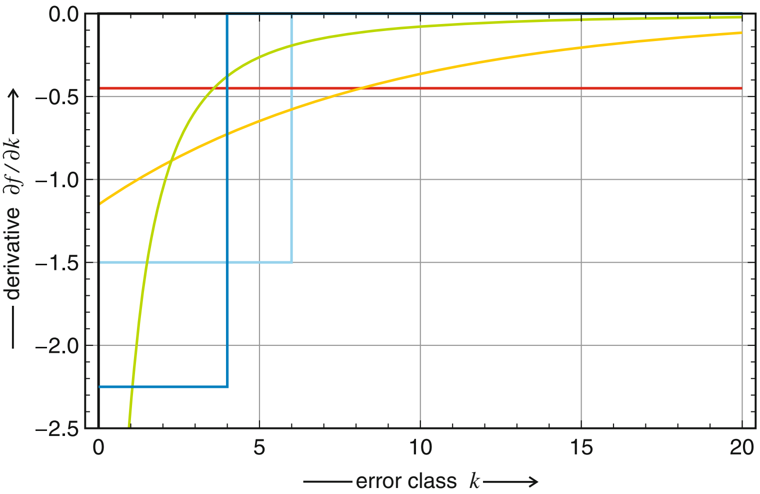


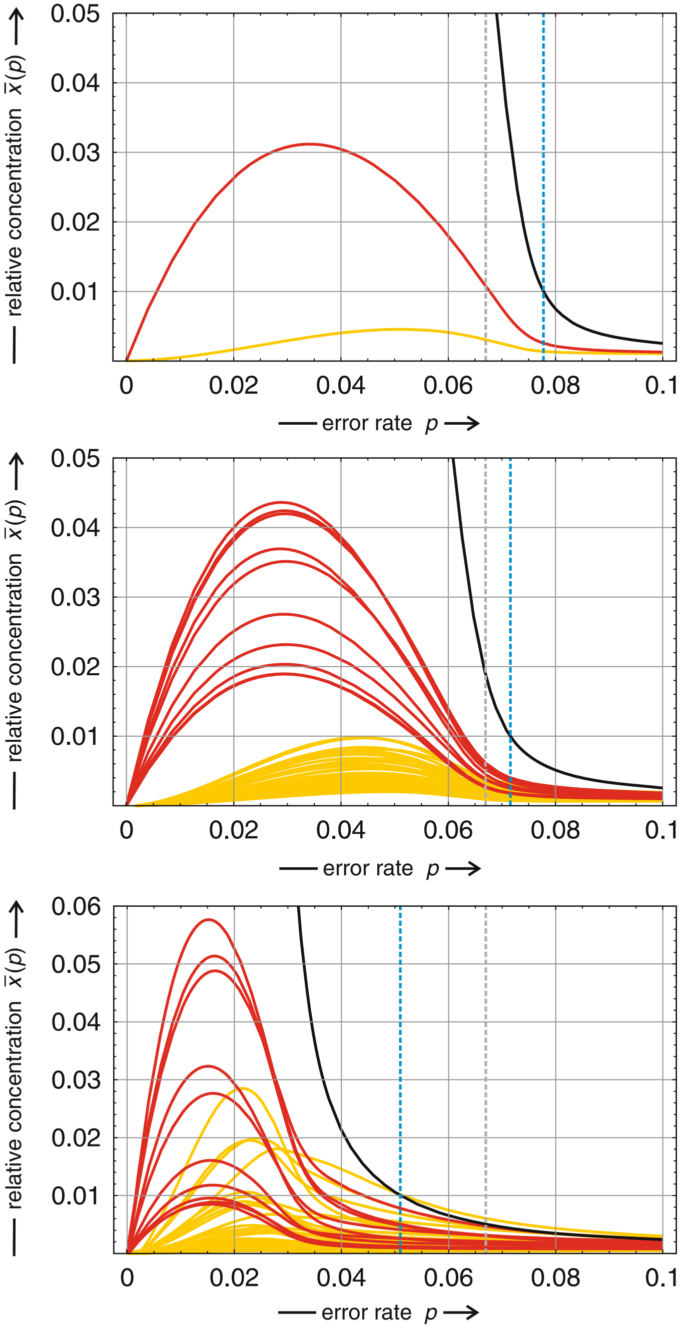
Fig. 13 The derivative of simple fitness landscapes. Shown are the derivatives, $\partial f/\partial k$, of the simple fitness landscapes (3a)–(3e). Choice of parameters: $l = 20$, $f_0 = 10.0$, and $f_n = 1.0$. *Color code* additive fitness (3a) in *red*, multiplicative fitness (3b) in *yellow*, hyperbolic fitness (3c) in *chartreuse*, single-peak step liner fitness (3e) $h = 6$ in *light blue* and $h = 4$ in *blue*, and the single-peak fitness (3d) in *black*. Error thresholds are found on the single peak, the single-peak linear with $h = 4$ and the hyperbolic fitness landscapes. On all other landscapes, smooth transitions from $p = 0$ to $p = \frac{1}{2}$ are observed

smooth transitions. For the examples given in the figure, this threshold values lie somewhere in the range $1.5 < \alpha_{\mathrm{cr}} < 2.25$.

Out of all the simple landscapes analyzed here, only the single-peak landscape supports an error threshold that fulfills simultaneously the three conditions: (i) fast decrease of the concentration $\bar{x}_m$ slightly below p_{tr}, (ii) a sharp transition at $p_{\mathrm{tr}}^{(\vartheta)}$ diagnosed by all $\left(p_{\mathrm{mg}}^{(\theta=\vartheta)}\right)_k$-values lying in a narrow interval, and (iii) the uniform distribution being the high mutation rate distribution.

Error threshold on realistic landscapes. There are families of smooth landscapes in which no error thresholds occur and this raises the question what we can expect to happen on realistic landscapes. For this goal, quasispecies as functions of the mutation rate p were calculated on about twenty different random realistic landscapes (RRL, 5a) for sequences of chain length $l = 10$.[14] Two results are relevant for our purpose here: (i) Quasispecies on realistic random landscapes show error thresholds and (ii) the position of level crossing as a measure for the error threshold

[14] Numerical computations of eigenvalues and eigenvectors become highly demanding with respect to CPU time and memory above $l = 10$. For $l = 20$, the diagonalization of the W-matrix with about the size $10^6 \times 10^6$ requires certain tricks (Niederbrucker and Gansterer 2011), and for $l = 50$, the dimension of W is more than $10^{15} \times 10^{15}$ and diagonalization is far beyond current technical capacities.

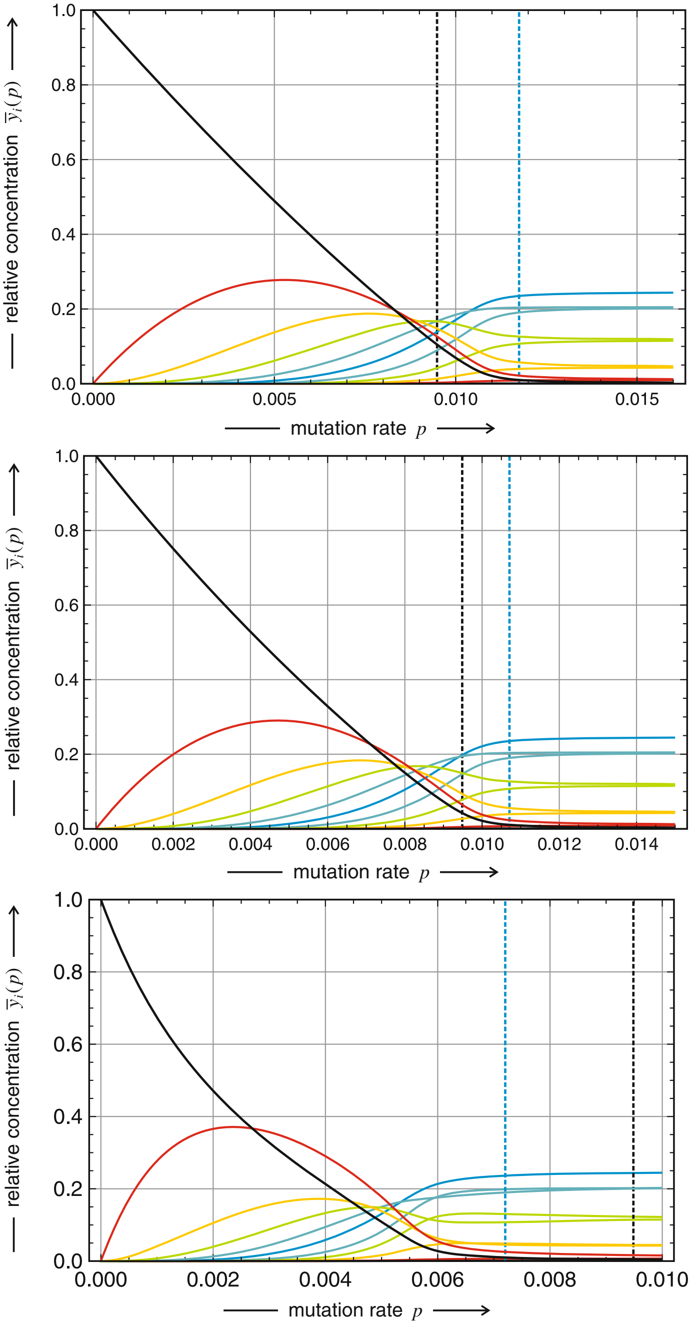

relative concentration $\bar{x}(p)$
0.05
0.04
0.03
0.02
0.01
0
0.02
0.04
0.06
0.08
0.1
error rate p
relative concentration $\bar{x}(p)$
0.05
0.04
0.03
0.02
0.01
0
0.02
0.04
0.06
0.08
0.1
error rate p
relative concentration $\bar{x}(p)$
0.06
0.05
0.04
0.03
0.02
0.01
0
0.02
0.04
0.06
0.08
0.1
error rate p

◀ **Fig. 14** Quasispecies on a realistic model landscape (RRL) with different random scatter d. Shown are the stationary concentrations $\bar{x}_j(p)$ on landscapes $\mathcal{L}(10, 2, 2.0, 1.0; 0.0, d, 491)$ for $d = 0$ (*upper plot*), $d = 0.5$ (*middle plot*), and $d = 0.9375$ (*lower plot*) for the classes Γ_0 (*black*), Γ_1 (*red*), and Γ_2 (*yellow*). In the topmost plot, the curves for all single point mutations $\mathsf{X}_j \in \Gamma_1$ and double point mutations $\mathsf{X}_j \in \Gamma_2$ coincide because of zero scatter, $d = 0$. The error threshold calculated by the phenomenological approach lies at $p_{\mathrm{cr}} = 0.06697$ and is indicated by a *dashed gray line*, and the positions of the $p_{\mathrm{tr}}^{(0.01)}$ values are 0.0778, 0.0716, and 0.0510 for $d = 0.0$, 0.5, and 0.9375, respectively (*dashed blue lines*)

moves to smaller mutation rates $p_{\mathrm{tr}}^{(\vartheta)}(d)$ when the ruggedness of the landscape given by the parameter d is increased. An illustrative example is shown in Fig. 14 where we find $p_{\mathrm{tr}}^{(0.01)}(0) = 0.0778$, $p_{\mathrm{tr}}^{(0.01)}(0.5) = 0.0716$, and $p_{\mathrm{tr}}^{(0.01)}(0.9375) = 0.0510$ for the cases shown in the plots. Figure 15 reports the movement of the position of $p_{tr}^{(m)}(d)$ toward lower mutation rate parameters with increasing scatter and illustrates the validity of the uniform distribution criterion independently of the extent of ruggedness as expressed by the parameter d: Despite the small chain length $l = 10$, convergence toward the uniform distribution at transition points far away from $p = \frac{1}{2}$ can be observed for all d-values and the $p_{\mathrm{tr}}^{(0.01)}$-value computed from level crossing is a good indicator for the positions at which merging of the stationary concentrations for the complementary classes, $|\bar{y}_k(p_{\mathrm{mg}}^{(\theta)}) - \bar{y}_{l-k}(p_{\mathrm{mg}}^{(\theta)})| = \theta$, occurs. Based on the level crossing criterion for the location of the error threshold, we find that the transition migrates to smaller p-values for higher scatter of the fitness values. This observation is readily interpreted: An increase in scatter implies that the difference in fitness values between the master sequence and the sequence with the next highest fitness value becomes smaller, and a smaller difference in fitness other factors being unchanged causes the transition to occur at a smaller p-value.

Random scatter of fitness values introduces fitness differences among the sequences within a given mutant class Γ_k, and instead of a single curve as found for $d = 0$, we obtain a bundle of curves for the individual sequences $\mathsf{X}_{(k)}$ belonging to this class. For small d-values, corresponding to small scatter of fitness values and for p-values sufficiently lower than the error threshold, $p \ll p_{\mathrm{cr}}$, the curves for the sequences belonging to given mutant classes form well-separated bands, which overlap at higher p- or higher d-values (Fig. 14). As seen in the class concentration plots (Fig. 15), the transition to the uniform distribution becomes somewhat irregular at high d-values. For example, in the bottom plot, the curves for ($k = 4$, $l - k = 6$) and for ($k = 3$, $l - k = 7$) cross first before they merge, which is due to the different spectra of f_j-values within the error classes. Nevertheless, also in these cases, the uniform distribution $\mathcal{U}$ is approached although at slightly higher p-values.

In proceeding toward the maximum scatter at the value $d = 1$, other transitions apart from the error thresholds are observed for the majority of RRLs (for exceptions, see next paragraph). These transitions, positioned at transition mutant rates, $p = (p_{\mathrm{tr}}^{q})$, mark dramatic changes in the stationary mutant distributions, and the primary quasispecies Υ_m, which is the stable distribution from the selection state


1.0
0.8
0.6
0.4
0.2
0.0
relative concentration $\bar{y}_i(p)$
0.000
0.005
0.010
0.015
mutation rate p
1.0
0.8
0.6
0.4
0.2
0.0
relative concentration $\bar{y}_i(p)$
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
mutation rate p
1.0
0.8
0.6
0.4
0.2
0.0
relative concentration $\bar{y}_i(p)$
0.000
0.002
0.004
0.006
0.008
0.010
mutation rate p

◀ **Fig. 15** Error thresholds on a *realistic* model landscape with different random scatter d. Shown are the stationary concentrations of classes $\bar{y}_j(p)$ on the landscapes $\mathcal{L}(10, 2, 1.1, 1.0; 0.0, d, 023)$ with $d = 0$ (*upper plot*), $d = 0.5$ (*middle plot*), and $d = 0.95$ (*lower plot*). The error threshold calculated by the phenomenological approach lies at $p_{\mathrm{cr}} = 0.009486$ (*black dotted line*), and the positions for level crossing decrease with the width of random scatter d and are situated at $p_{\mathrm{tr}}^{(0.01)} = 0.01175$, 0.01079, and 0.00720, respectively (*blue dotted lines*). For the analogous plots for fully developed randomness ($d = 1.0$), see Fig. 17

($p = 0$) onwards, is replaced at $p = (p_{\mathrm{tr}}^{q})_{m,k}$ by another quasispecies Υ_k. The mechanism by which quasispecies replace each other has been worked out analytically (Schuster and Swetina 1988) and is easily interpreted:[15] The stationary mutational backflow stabilizes master sequences through adding a positive term to the production function

$$W(\mathsf{X}_m) = w_m = Q_{mm} f_m \bar{x}_m + \sum_{j=1, j \neq m}^{N} Q_{mj} f_j \bar{x}_j = Q_{mm} f_m \bar{x}_m + \Phi_{m \leftarrow (j)},$$

and likewise, we have for a potential master sequence X_k,

$$W(\mathsf{X}_k) = w_k = Q_{kk} f_k \bar{x}_k + \Phi_{k \leftarrow (j)}.$$

In general, the first term decreases and the second term increases with increasing p. The necessary—but not sufficient—condition for the existence of a transition is $\Delta\Phi = \Phi_{m \leftarrow (j)} - \Phi_{k \leftarrow (j)} < 0$. In other words, the mutational backflow to the original master sequence of Υ_0 has to be weaker than the backflow to the sequence X_k in Υ_k. Since the fitness value f_m is the largest by definition, we have $f_m > f_j \; \forall j = 1, \ldots, n$, and at sufficiently small mutation rates p, the differences in the selective values, $\Delta\Psi = Q_{mm} f_m - Q_{kk} f_k > 0$, will always outweigh the difference in the backflow, $\Delta\Phi > |\Delta\Psi|$, and the quasispecies Υ_m is stable. With increasing values of p, however, the replication accuracy and $\Delta\Phi$ will decrease because of the term $Q_{mm} = Q_{kk} \approx (1 - p)^l$ in the uniform error approximation. At the same time, $\Delta\Psi$ will increase in absolute value and provided $\Delta\Psi < 0$ there might exist a mutation rate $p = p_{\mathrm{tr}}^{(q)}$ smaller than the error threshold value $p_{\mathrm{tr}}^{(q)} < p_{\mathrm{cr}}$ at which the condition $\Delta\Phi + \Delta\Psi = 0$ is fulfilled and consequently, the quasispecies Υ_k is the stable stationary solution of equation at $p_{\mathrm{tr}}^{(q)} < p < p_{\mathrm{cr}}$ provided it is not replaced by another quasispecies in another transition.

The influence of a distribution of fitness values instead of the single value f of the single-peak landscapes can be predicted straightforwardly: Since f_m is independent of the fitness scatter d and f_k is increasing with increasing scatter, the difference

[15] Thirteen years after this publication, the phenomenon has been observed in quasispecies of digital organisms (Wilke et al. 2001) and was called *survival of the flattest*.

$f_m - f_k$ will decrease with increasing scatter d. Accordingly, the condition for a transition between quasispecies can be fulfilled at lower p-values and we expect to find one or more transitions below the error threshold p_{cr}. No transition can occur on the single-peak landscape, but as d increases the difference $\Delta\Phi$ becomes smaller, and it becomes more and more likely that the difference in backflow $\Delta\Psi$ becomes sufficiently strong for a replacement of Υ_m by Υ_k below p_{cr}.

As an example, we describe the development of quasispecies transitions on a typical RRL, $\mathcal{L}(10, 2, 1.1, 1.0; 0.0, d, 637)$, between the single-peak scenario ($d = 0$) and the full-band random landscape ($d = 1$). Starting form the unspecific error threshold scenario (Fig. 9), the error classes unfold into first separated and later overlapping bands until a scatter of $d = 0.85$ where the indication of a first $p_{\mathrm{tr}}^{(q)}$-transition appears near the error threshold. At $d = 0.925$, a transition $\Upsilon_0(\mathsf{X}_0) \rightarrow \Upsilon_1(\mathsf{X}_{247})$ can be identified, and this transition is the only transition at the d-values 0.95 and 0.975. Then, at $d = 0.995$, a second transition appears (see Fig. 16), and finally, we are dealing with three transitions at full randomness, $d = 1$. In addition to the quite common scenario of multiple transitions described here, we found also landscapes with a single transition at fully developed randomness (see Fig. 17) and landscapes sustaining no transition at all (see *strong quasispecies* in the next paragraph).

An important question is whether or not transitions between quasispecies have an influence on the convergence toward the uniform distribution $\mathcal{U}$ above threshold. Intuitively, we might suggest that this is not the case, but it is safer to consider a specific example. The RRL $\mathcal{L}(10, 2, 1.1, 1.0; 0.0, 1.0, 023)$ is chosen, because it exhibits a single transition, $\Upsilon_0(\mathsf{X}_0) \rightarrow \Upsilon_1(\mathsf{X}_{910})$, below the error threshold (Fig. 17). The middle plot shows the quasispecies Υ_1 centered around the master sequence X_{910} that is surrounded by three high-fitness sequences in the one-error class: $\mathsf{X}_{906}, \mathsf{X}_{926}$, and X_{942}, and one additional high-fitness sequence in the two-error class, X_{927}, which is directly attached to X_{926}. This example illustrates the role of mutational backflow Φ in stabilizing quasispecies above the transition. The lower plot shows the change of the class concentrations $\bar{y}_k(p)$ at the transition $\Upsilon_0(\mathsf{X}_0) \rightarrow \Upsilon_1(\mathsf{X}_{910})$ and at the error threshold, which leads to the uniform distribution $\mathcal{U}$. As expected, the fully developed random scatter smoothens the error threshold, shifts the lower boundary, $\min((p_{\mathrm{mg}}^{(\theta)})_k)$, of complementary class concentration merging to slightly smaller p-values but does not change the basic property of merging the concentrations of complementary classes (for a concrete quantitative example, see Table 3).

Finally, we remark that transitions between quasispecies provide a handicap for evolution because a small change in the mutation rate or in the fitness value may destabilize stationary mutant distributions, and we conjecture that natural systems will be driven toward landscapes with stable quasispecies in the sense that no transitions between quasispecies occur.

Strong quasispecies. A certain fraction of landscapes gives rise to scenarios for quasispecies as a function of the mutation rate p that are substantially different from the one discussed above: No transitions between quasispecies are observed not even

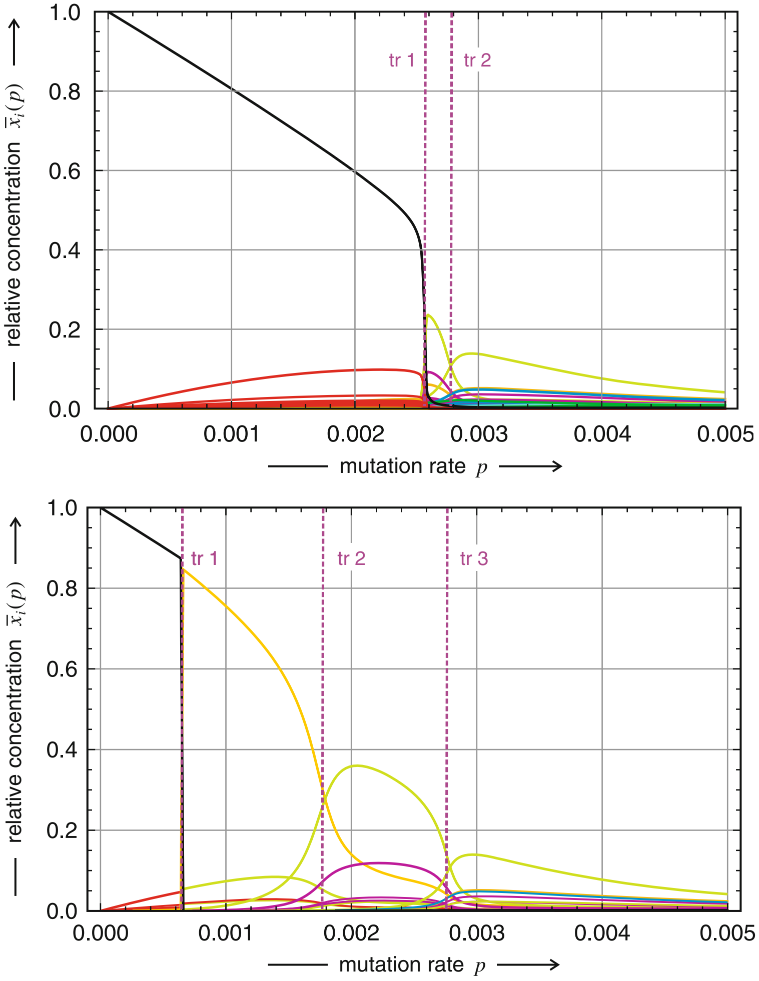
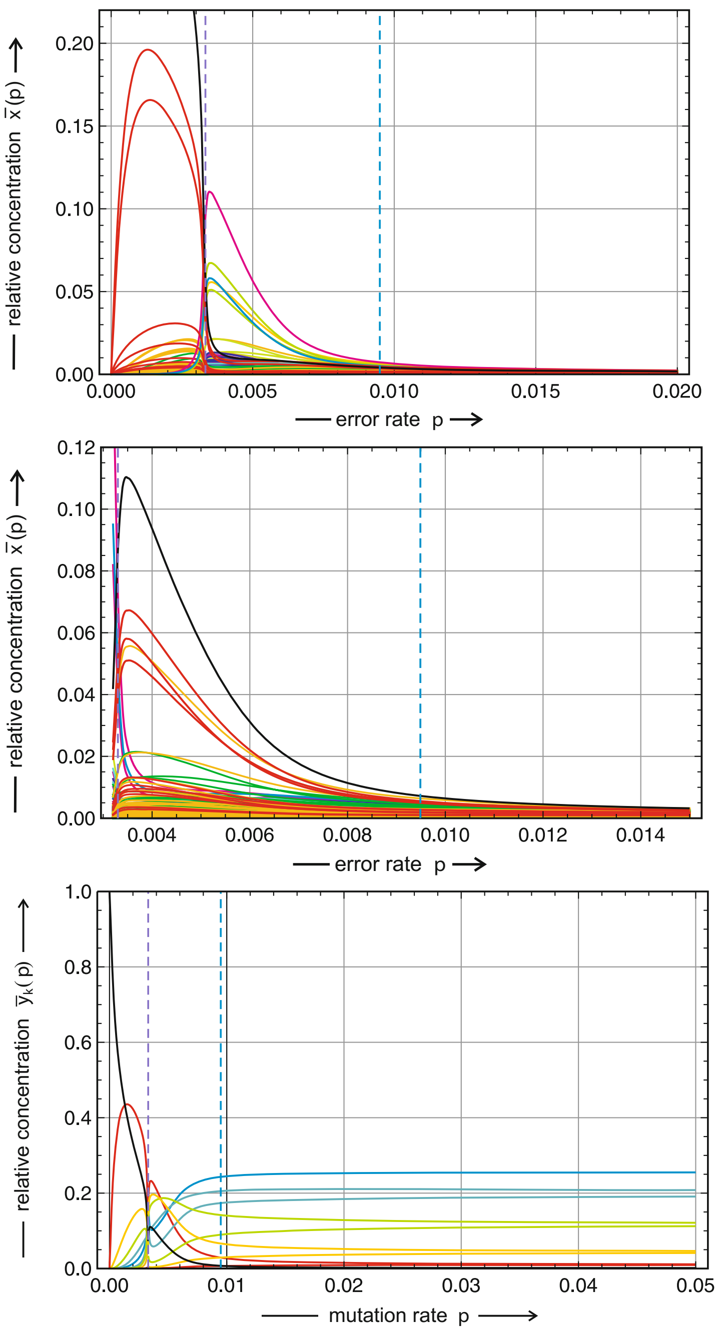


Fig. 16 A model landscape with multiple transitions between quasispecies. The plots present the stationary concentrations $\bar{x}_j(p)$ on landscapes $\mathcal{L}(10,2,1.1,1.0;0.0,d,637)$ with $d = 0.995$ (*upper plot*) and with fully developed scatter $d = 1.0$ (*lower plot*). The following master sequences are involved in the transitions between quasispecies at $d = 1.0$: tr$_1$: $\Upsilon_0(\mathsf{X}_0) \rightarrow \Upsilon_1(\mathsf{X}_{1003})$; tr$_2$: $\Upsilon_1(\mathsf{X}_{1003}) \rightarrow \Upsilon_2(\mathsf{X}_{923})$; tr$_3$: $\Upsilon_2(\mathsf{X}_{923}) \rightarrow \Upsilon_3(\mathsf{X}_{247})$. The Hamming distances at the transitions are $d^{\mathrm{H}}_{(0)},(1003) = 7$, $d^{\mathrm{H}}_{(1003),(923)} = 3$, and $d^{\mathrm{H}}_{(923),(247)} = 6$, respectively. For $d = 0.995$, the first transition does not exist; instead, we find $\Upsilon_0(\mathsf{X}_0) \rightarrow \Upsilon_1(\mathsf{X}_{923})$ and tr$_3$ becomes tr$_2$

at fully developed scatter $d = 1.0$ (Fig. 18). The most relevant feature of the quasispecies on these special landscapes concerns the classes to which the most frequent sequences belong. On the landscape defined by $s = 919$, these sequences are the master sequence (X_0; black curve), the one-error mutant (X_4; red curve), and


relative concentration $\bar{x}(p)$
0.20
0.15
0.10
0.05
0.00
0.000
0.005
0.010
0.015
0.020
error rate p
relative concentration $\bar{x}(p)$
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.004
0.006
0.008
0.010
0.012
0.014
error rate p
relative concentration $\bar{y}_k(p)$
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.01
0.02
0.03
0.04
0.05
mutation rate p

◀ **Fig. 17** Error threshold and transition between quasispecies. The landscape $\mathcal{L}(10,2,1.1,1.0;0.0,d,023)$ supports a transition $\Upsilon_0(\mathsf{X}_0)\rightarrow\Upsilon_1(\mathsf{X}_{910})$ at $(p_{\mathrm{tr}}^{q})_{0,910}\approx 0.00330$ (*violet dashed line*) and the error threshold computed from level crossing of X_{910} with $\vartheta=0.01$ at $p_{\mathrm{tr}}^{(0.01)}\approx 0.00837$. Above the error threshold, which lies at $p_{\mathrm{cr}}=0.00949$ (*blue dashed line*) in the corresponding single-peak landscape ($d=0.0$), the quasispecies converges to the uniform distribution $\mathcal{U}$ as immediately seen from the mergence of complementary class concentration curves. The quasispecies above the transition at p_{tr}^{q} is centered around the sequence X_{910} corresponding to Υ_1. It is worth noticing that the one-error class Γ_1 (*red*) has a class concentration that exceeds the master sequence by a factor two. This is mainly due to three sequences of high fitness, X_{906}, X_{926}, and X_{942}

Table 3 Concentration level crossing and complementary class mergence on landscapes with random scatter of fitness values

$(p_{\mathrm{mg}}^{(\theta)})_k$	Random scatter d						
k	0.0	0.5	0.7	0.9	0.95	0.975	1.0
0	0.01164	0.01097	0.01016	0.00884	0.00836	0.00809	0.00776
1	0.01210	0.01173	0.01123	0.01056	0.01039	0.01030	0.01022
2	0.01261	0.01292	0.01256	0.01161	0.01124	0.01103	0.01080
3	0.01282	0.01601	0.01768	0.01933	0.01972	0.01991	0.02009
4	0.01213	0.01283	0.01199	0.00962	0.00821	0.00757	0.00680
$p_{\mathrm{tr}}^{(\vartheta)}$	0.01175	0.01108	0.01028	0.00895	0.00848	0.00820	0.00788
$\Delta p_{\mathrm{mg}}^{(\theta)}$	0.00118	0.00504	0.00725	0.01049	0.01151*	0.01234*	0.01329*

Quantitative diagnostic tools are applied to the landscape $\mathcal{L}(10,2,1.1,1.0;0.0,d,919)$. The decline of the master class, $\bar{y}_0=\bar{x}_0$, at p-values near the error threshold $p_{\mathrm{cr}}=0.00948$ is illustrated by means of the points $p_{\mathrm{tr}}^{(\vartheta)}$ where the curves cross the level $\bar{x}_0(p)=\vartheta=0.01$. Complementary class mergence is characterized quantitatively by the band between the lowest and the highest $(p_{\mathrm{mg}}^{(\theta)})_k$-value ($\theta=0.01$). The lowest value is observed at $k=0$ for $d=0.0$, 0.5, 0.7, and 0.9. For the three highest random scatters, the lowest value is recorded for $k=4$ (indicated by an asterisk '*')

the two-error mutant (X_{516}; yellow curve).[16] Coming from neighboring classes, the three special sequences are situated close-by in sequence space—Hamming distances $d_{(0),(4)}^{\mathrm{H}}=d_{(4),(516)}^{\mathrm{H}}=1$ and $d_{(0),(516)}^{\mathrm{H}}=2$—and form a cluster, which is dynamically coupled by means of strong mutational flow (Fig. 19). As it turns out, such a quasispecies is not likely to be replaced in a transition by another one that is centered around a single master sequence and accordingly, we called such clusters *strong quasispecies*. The problem that ought to be solved now is the prediction of the occurrence of strong quasispecies from know fitness values.

A heuristic is mentioned that serves as an (almost perfect) diagnostic tool for detecting whether or not a given fitness landscape gives rise to a strong quasispecies: (i) For every mutant class, we identify the sequence with the highest fitness

[16] Naïvely, we would expect a band of one-error sequences at higher concentration than the two-error sequence.

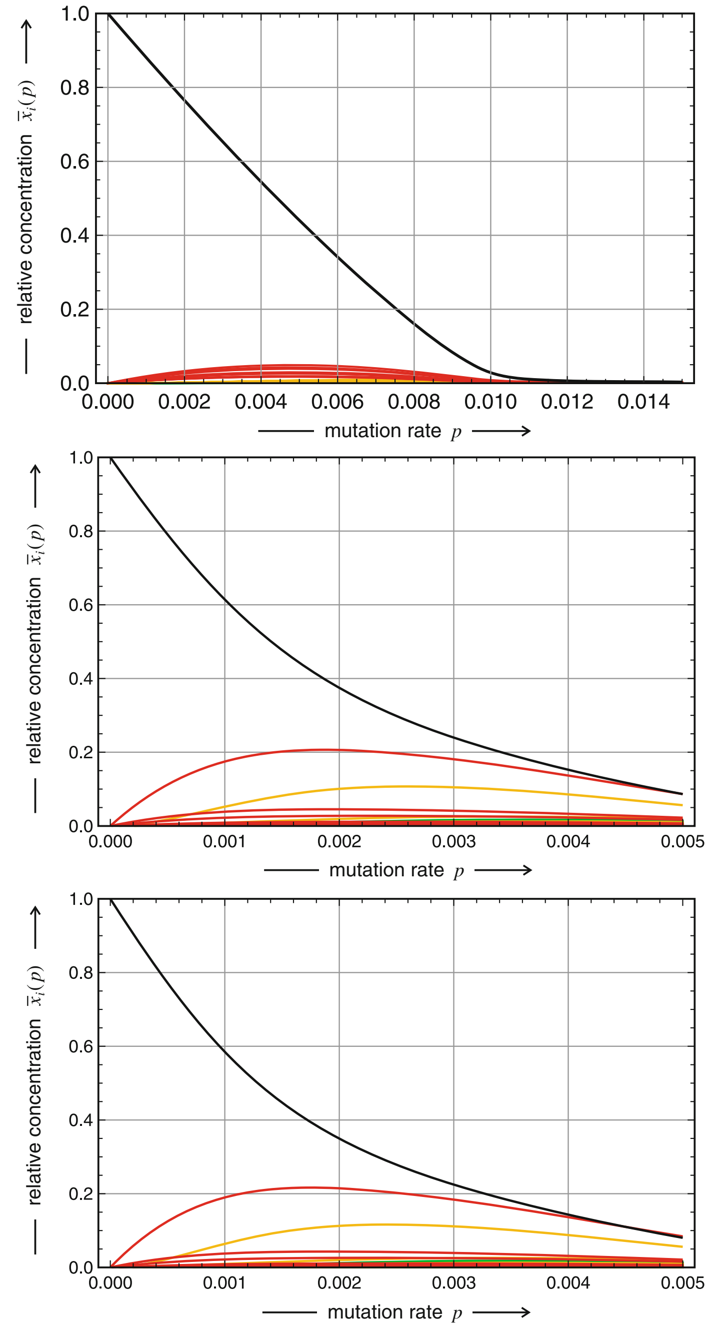


Fig. 18 Error thresholds on a model landscape with random scatter d and no transitions between quasispecies. The landscape $\mathcal{L}(10, 2, 1.1, 1.0; 0.0, d, 919)$ is computed and analyzed. Shown are the stationary concentrations $\bar{x}_j(p)$ for $d = 0.5$ (*upper plot*), for $d = 0.995$ (*middle plot*) and for fully developed random scatter $d = 1.0$ (*bottom plot*)

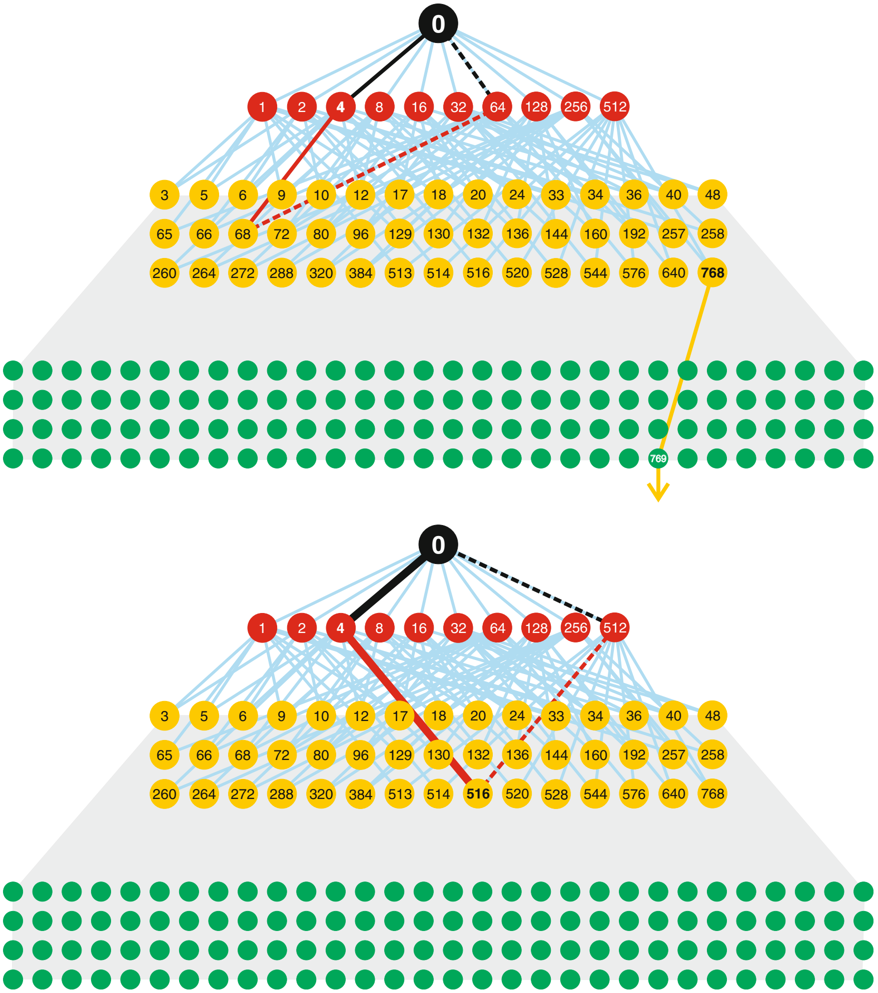


Fig. 19 Mutation flow in quasispecies. The sketch shows two typical situations in the distribution of fitness values in sequence space. In the *upper diagram* ($s = 637$), the fittest two-error mutant, X_{768}, has its fittest nearest neighbor, X_{769}, in the three-error class Γ_3. The fittest sequence in the one-error neighborhood of the fittest one-error mutant, X_4, is X_{68} and not $_{768}$, and hence, the mutational flow is not sufficiently strong for coupling the three sequences X_0, X_4, and X_{68} to a strong cluster, and transitions between different quasispecies are observed (Fig. 16). The lower diagram ($s = 919$) shows the typical fitness distribution for a strong quasispecies: The fittest two-error mutant, X_{516}, has its fittest nearest neighbor, X_4, in the one-error class Γ_1, and it coincides with the fittest one-error mutant. Here, the three sequences (X_0, X_4, and X_{516}) are strongly coupled by mutational flow, and a strong quasispecies is observed (Fig. 18)

value, f_m, $(f_{(1)})_{\max} = f(\mathsf{X}_{m(1)})$, $(f_{(2)})_{\max} = f(\mathsf{X}_{m(2)}), \ldots,$, and call them *class-fittest* sequences. Next, we determine the fittest sequences in the one-error neighborhood of the class-fittest sequences. Clearly, for the class k-fittest sequence $\mathsf{X}_{m(k)}$, this sequence lies either in class $k - 1$ or in class $k + 1$.[17] Simple combinatorics is favoring classes closer to the middle of sequence space because they have more members, $\binom{l}{k}$, in number. Any sequence in the two-error class, for example, has two nearest neighbors in the one-error class but $l - 2$ nearest neighbors in the three-error class (see Fig. 10). To be a candidate for a strong quasispecies requires that—against probabilities—the fittest sequence in the one-error neighborhood of $\mathsf{X}_{m(2)}$ lies in the one-error class: $(f_{(\mathsf{X}_{m(2)})_{m(1)}})_{\max}$ with $(\mathsf{X}_{m(2)})_{m(1)} \in \varGamma_1$ and preferentially, this is the fittest one-error sequence, $(\mathsf{X}_{m(2)})_{m(1)} \equiv \mathsf{X}_{m(1)}$. Since all mutation rates between nearest neighbor sequences in neighboring classes are the same —$(1-p)^{n-1}p$ within the uniform error model—the strength of mutational flow is dependent only on the fitness values, and the way in which the three sequences were determined guarantees optimality of the flow: If such a three-membered cluster was found, it is the one with the highest internal mutational flow for a given landscape. Figure 19 (lower picture, s = 919) shows an example where such three sequences form a strongly coupled cluster. There is always a fourth sequence—here X_{512}—belonging to the cluster, but it may play no major role because of low fitness. The heuristic presented here was applied to all 21 fitness landscapes with different random scatter, and three strong quasispecies (s = 401, 577, and 919) were observed. How many would be expected by combinatorial arguments? The probability for a sequence in $\varGamma_2$ to have a neighbor in $\varGamma_1$ is 2/10 = 0.2, and, since the sequence $\mathsf{X}_{m(1)}$ is fittest in $\varGamma_1$ and hence also in the one-error neighborhood of $\mathsf{X}_{m(2)}$, this is also the probability for finding a strong quasispecies. The sample that has been investigated in this study comprised 21 landscapes, and hence, we expect to encounter 21/5 = 4.2 cases, which is—with respect to the small sample size—in full agreement with the three cases that we found. The suggestion put forward in the heuristic mentioned above—a cluster of sequences coupled by mutational flow that is stronger within the group than to the rest of sequence space because of frequent mutations and high-fitness values—has been analyzed and tested through the application of perturbation theory (Schuster 2012). We dispense here from details since we shall not make further use of the corresponding expressions.

In order to study the influence of random scatter on the numerical computation of the location of the error threshold, we apply the two criteria, level crossing and complementary class mergence to the strong quasispecies on the landscape $\mathcal{L}(10, 2, 1.1, 1.0; 0.0, d, 919)$. The results are shown in Table 3. As already shown for other RRLs, the position of the crossing of $\bar{x}_0$ migrates to smaller mutation rates $p_{\mathrm{tr}}^{(\vartheta)}$ with increasing d. At the same time, the width of the transition increases by

[17] For class $k = 1$, we omit the master sequence X_m, which trivially is the fittest sequence in the one-error neighborhood, and search only in class $k = 2$.

about one order of magnitude from $\varDelta p_{\mathrm{tr}}^{(\theta)} = 0.0012{-}0.013$. Nevertheless, the quantitative diagnostic tools for the detection of the error threshold on complex landscapes works perfectly, and in contrast to doubts raised in the literature (Baake and Wagner 2001; Charlesworth 1990), even the landscapes with fully developed randomness ($d = 1.0$) sustain perfect error thresholds.
Selective neutrality. The second property of realistic fitness landscapes mentioned in Sect. 2 is *neutrality,* and in Eq. (5b), we made a proposal how neutrality can be implemented together with ruggedness. The resulting rugged and neutral fitness landscape (RNL) is characterized by two landscape specific parameters: (i) The random scatter is denoted by d as in the RRL landscape and (ii) a degree of neutrality λ. The value $\lambda = 0$ means absence of neutrality and $\lambda = 1$ describes the completely flat landscape in the sense of Kimura's *neutral evolution* (Kimura 1983). The result of the theory of neutral evolution that is most relevant here concerns *random selection*: Although fitness differences are absent, one randomly chosen sequence is selected by means of the autocatalytic replication mechanism, $\mathsf{X} \rightarrow 2\mathsf{X}$ and $\mathsf{X} \rightarrow \varnothing$. For most of the time, the randomly replicating population consists of a dominant genotype and a number of neutral variants at low concentration. An important issue of the landscape approach is the random positioning of neutral master sequences in sequence space, which is achieved by means of the same random number generator that is used to compute the random scatter of the other fitness values.

The RNL is the complete analogue to the rugged fitness landscape (RRL) under the condition that several master sequences exist, which have the same highest fitness values f_0. The fraction of neutral mutants is determined by the fraction of random numbers, which fall into the range $1 - \lambda < \eta \leq 1$, and apart from statistical fluctuations, this fraction is λ. At small values of the degree of neutrality λ, isolated peaks of highest fitness f_0 will appear in sequence space. Increasing λ will result in the formation of clusters of sequences of highest fitness. Connecting all fittest sequences of Hamming distance $d_{\mathrm{H}} = 1$ by an edge results in a graph that has been characterized as *neutral network* (Reidys et al. 1997; Reidys and Stadler 2002). Neutral networks were originally conceived as a tool to model, analyze, and understand the mapping of RNA sequences into secondary structures (Grüner et al. 1996a, b; Schuster et al. 1994). The neutral network in RNA sequence structure mappings is the preimage of a given structure in sequence space, and these networks can be approximated in zeroth order by random graphs (Erdős and Rényi 1959, 1960). Whereas neutral networks in RNA sequence structure mappings are characterized by a relatively high degree of neutrality around $\lambda \approx 0.3$ and sequence space percolation is an important phenomenon, we shall be dealing here with lower λ-values.

The two smallest clusters of mutationally coupled fittest sequences have Hamming distances $d_{\mathrm{H}} = 1$ and $d_{\mathrm{H}} = 2$ (Fig. 20). In the former case, we are dealing with the minimal neutral network of two neighboring sequences; in the latter case, the Hamming distance of two sequences are coupled through two intermediate sequences similarly as in the core of strong quasispecies. An exact

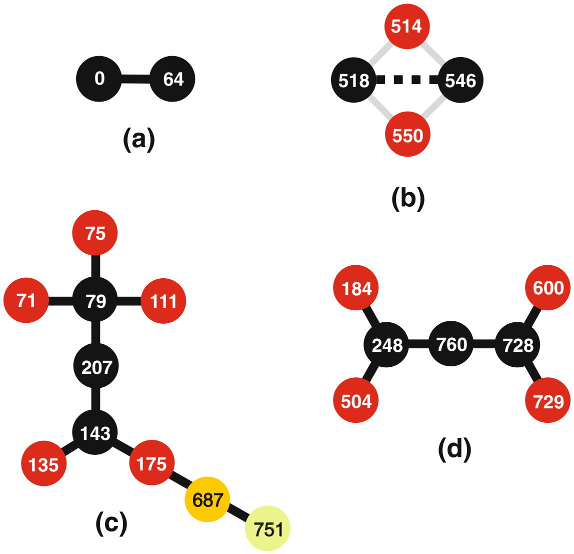


Fig. 20 Neutral networks in quasispecies. The sketch presents four special cases that were observed on rugged neutral landscapes defined in Eq. (5b). Part **a** shows the smallest possible network consisting of two sequences of Hamming distance $d_H = 1$ observed with $s = 367$ and $\lambda = 0.01$. Part **b** contains two sequences of Hamming distance $d_H = 2$, which are coupled through two $d_H = 1$ sequences; it was found with $s = 877$ and $\lambda = 0.01$. The neutral network in part **c** has a core of three sequences, surrounded by five one-error mutants, one of them having a chain of two further mutants attached to it; the parameters of the landscape are $s = 367$ and $\lambda = 0.1$. Part **d** eventually shows a symmetric network with three core sequences and four one-error mutants attached to it, observed with $s = 229$ and $\lambda = 0.1$. Choice of further parameters: $l = 10, f_0 = 1.1$, $f = 1.0$, and $d = 0.5$. Color code: core sequences in *black*, one-error mutants in *red*, two-error mutants in *yellow*, and three-error mutants in *green*

mathematical analysis is possible for both cases in the limit of vanishing mutation rates, $\lim p \rightarrow 0$ (Schuster and Swetina 1988). It yielded two results that are different from Kimura's neutral theory:

$$\lim_{p\to 0} \bar{x}_{\mathrm{I}} = \frac{1}{2}, \lim_{p\to 0} \bar{x}_{\mathrm{II}} = \frac{1}{2} \quad \text{for} \quad d^{\mathrm{H}}_{\mathsf{X}_{\mathrm{I}}\mathsf{X}_{\mathrm{II}}} = 1, \tag{16a}$$

$$\lim_{p\to 0} \bar{x}_{\mathrm{I}} = \frac{\alpha}{1+\alpha}, \lim_{p\to 0} \bar{x}_{\mathrm{II}} = \frac{1}{1+\alpha} \quad \text{for } d^{\mathrm{H}}_{\mathsf{X}_{\mathrm{I}}\mathsf{X}_{\mathrm{II}}} = 2, \tag{16b}$$

$$\begin{aligned} &\lim_{p\to 0} \bar{x}_{\mathrm{I}} = 1, \lim_{p\to 0} \bar{x}_{\mathrm{II}} = 0 \qquad \text{or} \qquad \lim_{p\to 0} \bar{x}_{\mathrm{I}} = 0, \lim_{p\to 0} \bar{x}_{\mathrm{II}} = 1, \\ &\qquad \text{for} \quad d^{\mathrm{H}}_{\mathsf{X}_{\mathrm{I}}\mathsf{X}_{\mathrm{II}}} \geq 3. \end{aligned} \tag{16c}$$

If the two neutral fittest sequences, X_{I} and X_{II}, are nearest neighbors in sequence space, $d^{\mathrm{H}}_{\mathsf{X}_{\mathrm{I}}\mathsf{X}_{\mathrm{II}}} = 1$, they are present at equal concentrations in the

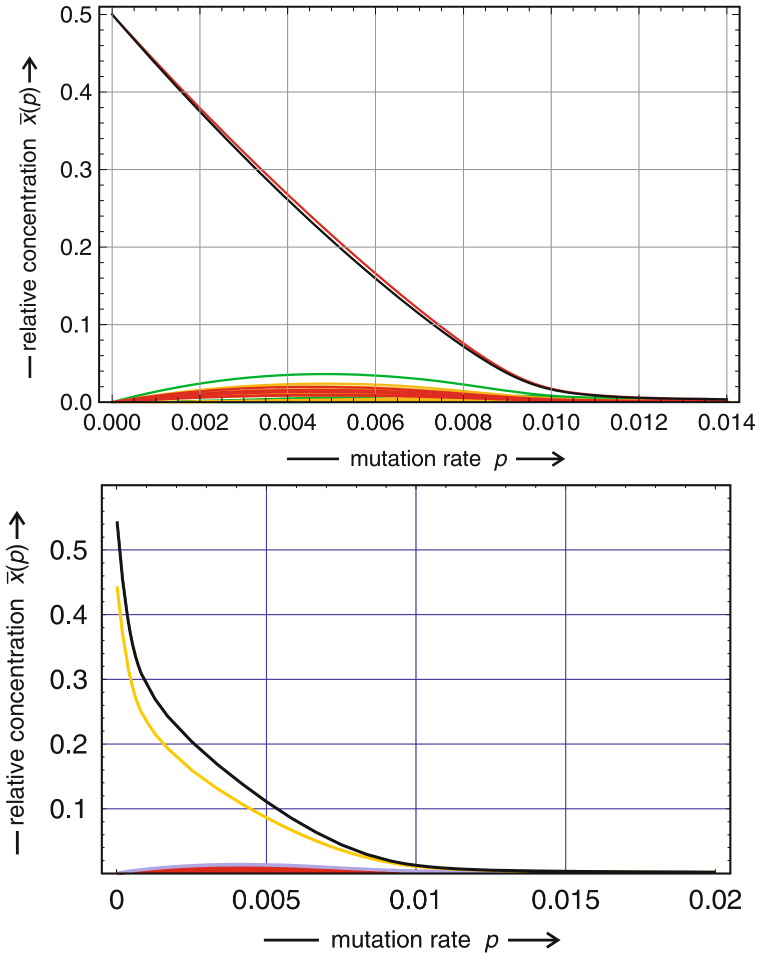


Fig. 21 Cluster on a weakly neutral rugged model landscapes. The plot at the *top* shows the quasispecies on the RNL $\mathcal{L}(10,2,1.1,1.0;0.1,0.5,637)$. The cluster in the core of the quasispecies is shown in Fig. 20a and consists of two Hamming distance $d_{\mathrm{H}} = 1$ master sequences, X_0 and X_{64}, which are present in equal concentrations from $p = 0$ to the error threshold. Further, we show their one-error neighborhoods, and the third fittest neutral sequence X_{324} at Hamming distance $d^{\mathrm{H}}_{(0),(324)} = 3$ (*green*). The *bottom plot* presents the quasispecies on the RNL $\mathcal{L}(10,2,1.1,1.0;0.1,0.5,877)$. The master pair X_{518} and X_{546} has Hamming distance $d_{\mathrm{H}} = 2$ and appears at roughly constant concentration ratio in the quasispecies over the entire range, $0 \le p < p_{\mathrm{cr}}$

low mutation rate limit, and in case they are next nearest neighbors in sequence space, $d^{\mathrm{H}}_{\mathsf{X}_{\mathrm{I}}\mathsf{X}_{\mathrm{II}}} = 2$, they are observed at some ratio α, and in both cases, none of the two sequences vanishes. Only for Hamming distances $d^{\mathrm{H}}_{\mathsf{X}_{\mathrm{I}}\mathsf{X}_{\mathrm{II}}} \ge 3$, Kimura's scenario of random selection occurs. It is important to stress a difference between the two scenarios, the deterministic ODE approach leading to clusters of neutral sequences and the random selection phenomenon of Kimura: In the quasispecies, we have strong mutational flow within the cluster of neutral sequences—which is not substantially different from the flow within the non-neutral clusters discussed in the previous paragraph—and this flow outweighs fluctuations. For Hamming distances d_H of three and more, the mutational flow is too weak to counteract random drift. In the random replication scenario, mutations do not occur and the only drive for change in particle numbers is random fluctuations.

In order to check the role of the predictions for the limit $p \to 0$ in the case of nonzero mutation rates, we search for appropriate test cases by inspection of RNL landscapes according to (5b). For small degrees of neutrality, we found indeed suitable neutral clusters on the landscapes ($s = 637, \lambda = 0.01$ and $s = 877, \lambda = 0.01$ both shown in Fig. 21). In full agreement with the exact result, we find that two fittest sequences of Hamming distance $d_\mathrm{H} = 1$ are selected as a strongly coupled pair with equal frequency of both members and numerical results show that

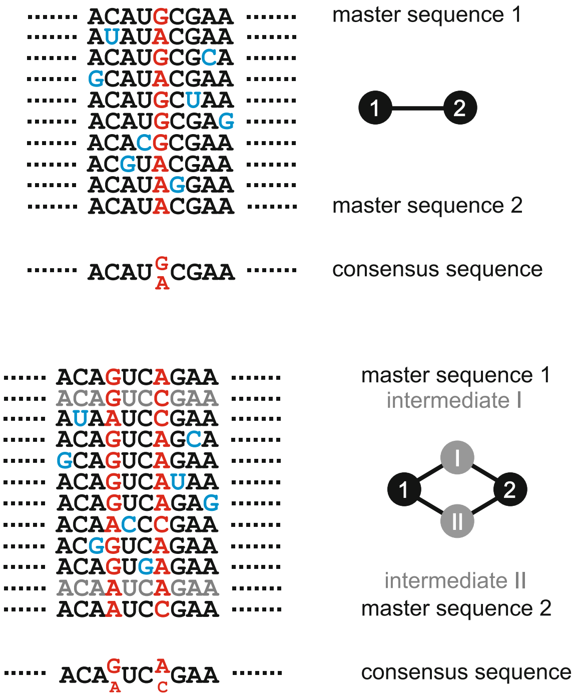

.... ACAUGCGAA .... master sequence 1
.... AUAUACGAA ....
.... ACAUGCGCA ....
.... GCAUACGAA ....
.... ACAUGCUAA ....
.... ACAUGCGAG ....
.... ACACGCGAA ....
.... ACGUACGAA ....
.... ACAUAGGAA ....
.... ACAUACGAA .... master sequence 2
1
2
.... ACAU^G_ACGAA .... consensus sequence
...... ACAGUCAGAA .... master sequence 1
...... ACAGUCCGAA .... intermediate I
...... AUAAUCCGAA ....
...... ACAGUCAGCA ....
...... GCAGUCAGAA ....
...... ACAGUCAUAA ....
...... ACAGUCAGAG ....
...... ACAACCCGAA ....
...... ACGGUCAGAA ....
...... ACAGUGAGAA ....
...... ACAAUCAGAA .... intermediate II
...... ACAAUCCGAA .... master sequence 2
I
II
...... ACA^G_AUC^A_CGAA .... consensus sequence

Fig. 22 Quasispecies and consensus sequences in case of neutrality. The *upper part* of the figure shows a sketch of sequences in the quasispecies of two fittest nearest neighbor sequences ($d_\mathrm{H} = 1$). The consensus sequence is not unique and differs in a single position where both nucleotides appear with equal frequency. In the *lower part,* the two master sequences have Hamming distance $d_H = 2$ and differ in two positions. The two sequences are present at some ratio α that is determined by the fitness values of other neighboring sequences, and the nucleobases corresponding to the differences in the two master sequences appear with the same ratio α

strong coupling does not occur only for small mutation rates but extends over the whole range of p-values from $p = 0$ up to the error threshold $p = p_{\mathrm{cr}}$. Examples for the case $d_{\mathrm{H}} = 2$ are also found on random neutral landscapes, and again, the exact result for vanishing mutation rate holds up to the error threshold. The existence of neutral nearest and next nearest neighbors manifests itself in the lack of unique consensus sequences of populations and has important consequences for the reconstruction of phylogenies (Fig. 22).

Neutral networks may comprise several sequences, and then, all neutral nearest neighbor sequences form a strongly coupled master cluster in reproduction. The distribution of individual members of the cluster in the limit $p \longrightarrow 0$ is readily obtained by diagonalization of the adjacency matrix.[18] The components of the largest eigenvector are proportional to the concentrations of elements of the replication network. Increasing the degree of neutrality λ gives rise to the formation of larger neutral networks. Commonly, there is a *giant cluster* and many small clusters as predicted by random graph theory.

5 Conclusions and Perspectives

The landscape concept was shown to be applicable to asexually reproducing virus populations, although fully empirically determined examples are not achievable at the current state of the art. Realistic landscapes are characterized by two global features: (i) ruggedness and (ii) neutrality. At sufficiently low mutation rates, all these landscapes support stationary mutant distribution called quasispecies no matter how large the random scatter of the individual fitness values is. The frequencies of individual genotypes in quasispecies are determined by their fitness values and the distances to the master sequence. Contradicting previous conjectures, such realistic landscapes exhibit error thresholds in the sense that the mutant distributions change abruptly at a certain critical mutations rate, which can be fully characterized by quantitative criteria. Above threshold, the conventional deterministic description by means of kinetic differential equations yields the uniform distribution of sequences as stationary solution and hence cannot provide a correct picture of replication–mutation dynamics. Under these conditions, populations drift randomly through sequence space in the sense of Kimura's neutral theory of evolution.

Although the kinetic equations allow for the derivation of general solutions in terms of eigenvalue problems, they are limited in reality because numerical computations are facing unsurmountable difficulties even for relatively small sequence lengths ($l \approx 50$). A phenomenological approach originally proposed by Eigen

[18] The adjacency matrix of a graph, A, is a symmetric square matrix that has an entry $a_{jk} = a_{kj} = 1$ whenever the graph has an edge between the nodes for X_j and X_k and zero entries everywhere else.

introduces simplifications, which allow for straightforward handling of long genotypes. This approach cannot be deduced from the original equations in a consistent way but represents an enormously successful heuristic for the calculation of error thresholds in real-world situations, and fortunately, the results become more accurate for longer polynucleotide sequences.

A problem for future research concerns the classification of landscapes in view of the mutation–selection dynamics upon them. We have sketched here two examples where the quasispecies dynamics can be predicted from the distribution of fitness values: (i) landscapes supporting strong quasispecies and (ii) landscapes with a tunable degree of neutrality. These studies make several predictions, and the next natural step is to test them experimentally and to initiate thereby a dialogue between theorists and experimentalists. Precisely, this dialogue made physics so successful, but unfortunately, it is still underdeveloped in biology.

6 Color Code for Sequences and Classes

The individual curves in plots of quasispecies as functions of the mutation rate p are color coded in order to make them better distinguishable. Most plots refer to a chain length $l = 10$ and adopted the following color code. For concentrations of classes rather than individual sequences, we use a different color code in order to make merging of complementary classes better visible.

Class no.	Color	
	Sequences	Classes
0	Black	Black
1	Red	Red
2	Yellow	Yellow
3	Green	Green
4	Sea green	Sea green
5	Blue	Blue
6	Magenta	Sea green
7	Chartreuse	Green
8	Yellow	Yellow
8	Red	Red
10	Black	Black

Acknowledgements The calculations reported here were done with the package *Mathematica 9.0*. For the simulations based on Gillespie's algorithm, we made use of the open-source SSA-program within X-Cellerator package. We are grateful to Bruce Shapiro for making this software public.

The author wishes to acknowledge support by the University of Vienna, Austria, and the Santa Fe Institute, Santa Fe, USA. A number of colleagues have helped in discussions. I am particulary grateful to Reinhard Bürger, Esteban Domingo, Christoph Flamm, Leticia Gonzalez-Herrero, Ivo Hofacker, Markus Oppel, David Saakian, and Peter Stadler.

References

Altenberg L (1997) NK fitness landscapes. In: Bäck T, Fogel DB, Michalevicz Z (eds) Handbook of evlutionary computation, chapter B 2.7.2. Oxford University Press, Oxford, UK, pp 2.7.5–2.7.10

Arslan E, Laurenzi IJ (2008) Kinetics of autocatalysis in small systems. J Chem Phys 128:e015101

Athavale SS, Spicer B, Chen IA (2014) Experimental fitness landscapes to understand the molecular evolution of RNA-based life. Curr Opin Chem Biol 22:35–39

Baake E, Baake M, Wagner H (1997) Ising quantum chain is equivalent to a model of biological evolution. Phys Rev Lett 78:559–562

Baake E, Gabriel W (1999) Biological evolution through mutation, selection, and drift: an introductory review. In: Stauffer D (ed) Annual review of computational physics VII. World Scientific, Singapore, pp 203–264

Baake E, Wagner H (2001) Mutation-selection models solved exactly with methods of statistical mechanics. Genet Res Camb 78:93–117

Beerenwinkel N, Pachter L, Sturmfels B, Elena SF, Lenski RE (2007) Analysis of epistatic interactions and fitness landscapes using a new geometric approach. BMC Evol Biol 7:e60

Betancourt AJ, Bollback JP (2006) Fitness effects of beneficial mutations: the mutational landscape model in experimental evolution. Curr Opin Genet Dev 16:618–623

Biebricher CK (1983) Darwinian selection of self-replicating RNA molecules. In: Hecht MK, Wallace B, Prance GT (eds) Evolutionary biology, vol 16. Plenum Publishing Corporation, New York, pp 1–52

Biebricher CK, Eigen M, William C, Gardiner J (1983) Kinetics of RNA replication. Biochemistry 22:2544–2559

Biebricher CK, Eigen M, William C, Gardiner J (1984) Kinetics of RNA replication: plus-minus asymmetry and double-strand formation. Biochemistry 23:3186–3194

Biebricher CK, Eigen M, William C, Gardiner J (1985) Kinetics of RNA replication: competition and selection among self-replicating RNA species. Biochemistry 24:6550–6560

Bratus AS, Novozhilov AS, Semenov YS (2014) Linear algebra of the permutation invariant Crow-Kimura model of prebiotic evolution. Math Biosci 256:42–57

Bürger R (1998) Mathematical properties of mutation-selection models. Genetica 102(103):279–298

Charlesworth B (1990) Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet Res Camb 55:199–221

Chou H-H, Delaney NF, Draghi JA, Marx CJ (2014) Mapping the fitness landscape of gene expression uncovers the cause of antagonism and sign epistasis between adaptive mutations. PLoS Genet 10:e1004149

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York (Reprinted at The Blackburn Press, Caldwell, NJ, 2009)

Demetrius L, Schuster P, Sigmund K (1985) Polynucleotide evolution and branching processes. Bull Math Biol 47:239–262

Drake JW, Charlesworth B, Charlesowrth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148:1667–1686

Edwards SF, Anderson PW (1975) Theory of spin glasses. J Phys F 5:965–974

Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523

Eigen M, McCaskill J, Schuster P (1988) Molecular quasispecies. J Phys Chem 92:6881–6891

Eigen M, McCaskill J, Schuster P (1989) The molecular quasispecies. Adv Chem Phys 75:149–263

Eigen M, Schuster P (1977) The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. Naturwissenschaften 64:541–565

Eigen M, Schuster P (1978) The hypercycle. A principle of natural self-organization. Part B: the abstract hypercycle. Naturwissenschaften 65:7–41

Eigen M, Schuster P (1982) Stages of emerging life—five principles of early organization. J Mol Evol 19:47–61

Elena SF, Sanjuán R (2007) Virus evolution: insights from an experimental approach. Annu Rev Ecol Evol Syst 58:27–52

Erdös P, Rényi A (1959) On random graphs. I. Publ Math 6:290–295

Erdös P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61

Fisher RA (1941) Average excess and average effect of a gene substitution. Ann Eugenics 11:53–63

Fontana W, Griesmacher T, Schnabl W, Stadler PF, Schuster P (1991) Statistics of landscapes based od free energies, replication and degradation rate constants of RNA secondary structures. Mon Chem 122:795–819

Fontana W, Konings DAM, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. Biopolymers 33:1389–1404

Fontana W, Schnabl W, Schuster P (1989) Physical aspects of evolutionary optimization and adaptation. Phys Rev A 40:3301–3321

Fontana W, Schuster P (1987) A computer model of evolutionary optimization. Biophys Chem 26:123–147

Fontana W, Schuster P (1998a) Continuity in evolution. On the nature of transitions. Science 280:1451–1455

Fontana W, Schuster P (1998b) Shaping space. The possible and the attainable in RNA genotype-phenotype mapping. J Theor Biol 194:491–515

Gago S, Elena SF, Flores R, Sanjuan R (2009) Extremely high mutation rate of a hammerhead viroid. Science 323:1308

Galluccio S (1997) Exact solution of the quasispecies model in a sharply peaked fitness landscape. Phys Rev E 56:4526–4539

Gavrilets S (1997) Evolution and speciation on holey adaptive landscapes. Trends Ecol Evol 12:307–312

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340–2361

Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58:35–55

Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Schuster P (1996a) Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. Mon Chem 127:355–374

Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Schuster P (1996b) Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering. Mon Chem 127:375–389

Hamming RW (1950) Error detecting and error correcting codes. Bell Syst Tech J 29:147–160

Hamming RW (1986) Coding and information theory, 2nd edn. Prentice-Hall, Englewood Cliffs, NJ

Ho SYW, Duchêne S (2014) Molecular-clock methods for estimating evolutionary rates and timescales. Mol Ecol 23:5947–5965

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. Mon Chem 125:167–188

Jain K, Krug J (2007) Adaptation in simple and complex fitness landscapes. In: Bastolla U, Porto M, Eduardo Roman H, Vendruscolo M (eds) Structural approaches to sequence evolution. Molecules, networks, populations, chapter 14. Springer, Berlin, pp 299–339

Janet A (1895) Considérations méchaniques sur l'évolution et le problème des espèces. In: Comptes Rendue des 3me Congrès International de Zoologie. 3me Congres International de Zoologie, Leyden, pp 136–145

Jiménez JI, Xulvi-Brunet R, Campbell GW, amd Irene A, Chen RT (2013) Comprehensive experimental fitness landscape and evolutionary network for small RNA. Proc Natl Acad Sci USA 110:14984–14989

Jones BL, Enns RH, Rangnekar SS (1976) On the theory of selection of coupled macromolecular systems. Bull Math Biol 38:15–28

Kang Y-G, Park J-M (2008) Survival probability of quasi-species model under environmental changes. J Korean Phys Soc 53:868–872

Kauffman S, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. J Theor Biol 128:11–45

Kauffman SA (1993) The origins of order. Self-organization and selection in evolution. Oxford University Press, New York

Kauffman SA, Weinberger ED (1989) The N-k model of rugged fitness landscapes and its application to the maturation of the immune response. J Theor Biol 141:211–245

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kingman JFC (1978) A simple model for the balance between selection and mutation. J Appl Probab 15:1–12

Kingman JFC (1980) Mathematics of genetic diversity. Society for Industrial and Applied Mathematics, Philadelphia

Kouyos RD, Leventhal GE, Hinkley T, Haddad M, Whitcomb JM, Petropoulos CJ, Bonhoeffer S (2012) Exploring the complexity of the HIV-1 fitness landscape. PLoS Genet 8:e1002551

Kouyos RD, von Wyl V, Hinkley T, Petropoulos CJ, Haddad M, Whitcomb JM, Böni J, Yerly S, Cellerai C, Klimkait T, Günthard HF, Bonhoeffer S, The Swiss HIV Cohort Study (2011). Assessing predicted HIV-1 replicative capacity in a clinical setting. PLoS Pathog 7: e1002321

Lanfear R, Welch JJ, Bromham L (2010) Watching the clock: studying variation in rates of molecular evolution between species. TREE 25:495–503

Leuthäusser I (1986) An exact correspondence between Eigen's evolution model and a two-dimensional ising system. J Chem Phys 84:1884–1885

Leuthäusser I (1987) Statistical mechanics of Eigen's evolution model. J Stat Phys 48:343–360

Lifson S (1961) On the theory of helix-coil transitions in polypeptides. J Chem Phys 34:1963–1974

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA package 2.0. Algorithms Mol Biol 6:e26

McCoy JW (1979) The origin of the "adaptive landscape" concept. Am Nat 113:610–613

McGhee GR Jr (2007) The geomerty of evolution: adaptive landscapes and theoretical morphospaces. Cambridge University Press, Cambridge

Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci USA 58:217–224

Mollison D (ed) (1995) Epidemis models: their structure and relation to data. Cambridge University Press, Cambridge

Nåsell I (2011) Extiction and quasi-stationarity in the stochastic logistic SIS model, vol 2022. Lecture Notes in Mathematics. Springer, Berlin

Niederbrucker G, Gansterer WN (2011) Efficient solution of evolution models for virus populations. Procedia Comput Sci 4:126–135

Nowak M, Schuster P (1989) Error thresholds of replication in finite populations. Mutation frequencies and the onset of Muller's ratchet. J Theor Biol 137:375–395

Park J-M, noz EM, Deem MW (2010) Quasispecies theory for finite populations. Phys Rev E 81: e011902

Pitt JN, Ferré-D'Amaré AR (2010) Rapid construction of empirical RNA fitness landscapes. Science 330:376–379

Provine WB (1986) Sewall wright and evolutionary biology. University of Chicago Press, Chicago

Reidys C, Stadler PF, Schuster P (1997) Generic properties of combinatory maps. Neutral networks of RNA secondary structure. Bull Math Biol 59:339–397

Reidys CM, Stadler PF (2001) Neutrality in fitness landscapes. Appl Math Comput 117:321–350

Reidys CM, Stadler PF (2002) Combinatorial landscapes. SIAM Rev 44:3–54

Ruse M (1996) Are pictures really necessary? The case of Sewall Wright's 'adaptive lanscapes'. In: Baigrie BS (ed) Picturing knowledge: historical and philosophical problems concerning the use of art in science. University of Toronto Press, Toronto, pp 303–337

Saakian DB, Hu C-K (2006) Exact solution of the Eigen model with general fitness functions and degradation rates. Proc Natl Acad Sci USA 113:4935–4939

Saakian DB, Hu C-K, Khachatryan H (2004) Solvable biological evolution models with general fitness functions and multiple mutations in parallel mutation-selection scheme. Phys Rev E 70: e041908

Sanjuán R, Moya A, Elena SF (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci USA 101:8396–8401

Schmidt LD (2004) The engineering of chemical reactions, 2nd edn. Oxford University Press, New York

Schuster P (2003) Molecular insight into the evolution of phenotypes. In: Crutchfield JP, Schuster P (eds) Evolutionary dynamics—exploring the interplay of accident, selection, neutrality, and function. Oxford University Press, New York, pp 163–215

Schuster P (2006) Prediction of RNA secondary structures: from theory to models and real molecules. Rep Prog Phys 69:1419–1477

Schuster P (2011) Mathematical modeling of evolution. Solved and open problems. Theory Biosci 130:71–89

Schuster P (2012) Evolution on "realistic" fitness landscapes. Phase transitions, strong quasispecies, and neutrality. Working Paper 12-06-006, Santa Fe Institute, Santa Fe, NM

Schuster P (2013) Present day biology seen in the looking glass of physics of complexity. In: Rubio RG, Ryazantsev YS, Starov VM, Huang G, Chetverikov AP, Arena P, Nepomnyashchy AA, Ferrus A, Morozov EG (eds) Without bounds: a scientific canvas of nonlinearity and complex dynamics. Springer, Berlin, pp 589–622

Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc Roy Soc Lond B 255:279–284

Schuster P, Sigmund K (1985) Dynamics of evolutionary optimization. Ber Bunsenges Phys Chem 89:668–682

Schuster P, Swetina J (1988) Stationary mutant distribution and evolutionary optimization. Bull Math Biol 50:635–660

Schwarz G (1968) General theoretical approach to the thermodynamic and kinetic properties of cooperative intramolecular transformations of linear biopolymers. Biopolymers 6:873–897

Seneta E (1981) Non-negative matrices and Markov chains, 2nd edn. Springer, New York

Sherrington D, Kirkpatrick S (1975) Solvable model of spin glasses. Phys Rev Lett 35:1792–1796

Skipper RA Jr (2004) The heuristic role of Sewall Wright's 1932 adaptive landscape diagram. Philos Sci 71:1176–1188

Spiegelman S (1971) An approach to the experimental analysis of precellular evolution. Quart Rev Biophys 4:213–253

Strogatz SH (1994) Nonlinear dynamics and chaos. With applications to physics, biology, chemistry, and engineering. Westview Press at Perseus Books, Cambridge

Swetina J, Schuster P (1982) Self-replication with errors—a model for polynucleotide replication. Biophys Chem 16:329–345

Tarazona P (1992) Error thresholds for molecular quasispecies as phase transitions: from simple landscapes to spin glasses. Phys Rev A 45:6038–6050

Tejero H, Marín A, Moran F (2010) Effect of lethality on the extinction and on the error threshold of quasispecies. J Theor Biol 262:733–741

Thompson CJ, McBride JL (1974) On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules. Math Biosci 21:127–142

Toulouse G (1977) Theory of frustration effect in spin-glasses. I. Commun Phys 2:115–119

Toulouse G (1980) The frustration model. In: Pękalski A, Przystawa JA (eds) Modern trends in the theory of condensed matter, vol 115., Lecture Notes in PhysicsBerlin, Springer, pp 195–203

van Kampen NG (2007) Stochastic processes in physics and chemistry, 3rd edn. Elsevier, Amsterdam

Walsh B, Blows MV (2009) Abundant variation + strong selection = multivariate genetic constraints: a geometric view of adaptation. Annu Rev Ecol Evol Syst 40:41–59

Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171:737–738

Weinreich DM (2011) High-throughput identification of genetic interactions in HIV-1. Nat Genet 41:398–400

Wiehe T (1997) Model dependency of error thresholds: the role of fitness functions and contrasts between the finite and infinite sites models. Genet Res Camb 69:127–136

Wilke CO, Wang JL, Ofria C (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature 412:331–333

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Jones DF (ed) International proceedings of the sixth international congress on genetics, vol 1. Brooklyn Botanic Garden, Ithaca, NY, pp 356–366

Wright S (1988) Surfaces of selective value revisited. American Naturalist 131:115–123

Zimm BH (1960) Theory of "melting" of the helical form in double chains of the DNA type. J Chem Phys 33:1349–1356

Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244:48–52