Phylogeny and evolution of structure

Tanja Gesell and Peter Schuster

Abstract

Darwin’s conviction that all living beings on Earth are related and the graph of relatedness is tree-shaped has been essentially confirmed by phylogenetic reconstruction first from morphology and later from data obtained by molecular sequencing. Limitations of the phylogenetic tree concept were recognized as more and more sequence information became available. The other path-breaking idea of Darwin, natural selection of fitter variants in populations, is cast into simple mathematical form and extended to mutation-selection dynamics. In this form the theory is directly applicable to RNA evolution in vitro and to virus evolution. Phylogeny and population dynamics of RNA provide complementary insights into evolution and the interplay between the two concepts will be pursued throughout this chapter. The two strategies for understanding evolution are ultimately related through the central paradigm of structural biology: sequence $\Rightarrow$ structure $\Rightarrow$ function. We elaborate on the state of the art in modeling both phylogeny and evolution of RNA driven by reproduction and mutation. Thereby the focus will be laid on models for phylogenetic sequence evolution as well as evolution and design of RNA structures with selected examples and notes on simulation methods. In the perspectives an attempt is made to combine molecular structure, population dynamics and phylogeny in modeling evolution.

Key words: Evolution of structure; multiple structures; phylogeny; quasispecies concept; sequence-structure mappings; sequence evolution; simulations.

Tanja Gesell
Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories (MFPL), University of Vienna, Medical University of Vienna and University of Veterinary Medicine, Dr. Bohr-Gasse 9, Vienna, 1030, Austria. School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA 02138, USA e-mail: tanja.gesell@univie.ac.at

Peter Schuster
Institute für Theoretische Chemie der Universität Wien, Währingerstrasse 17, A-1090 Vienna, and The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA e-mail: psk@tbi.univie.ac.at
1 Evolutionary thinking in mathematical language

James Watson, one of the researchers who discovered the structure of the DNA double helix, begins his textbook *Molecular Biology of the Gene* in 1965 with an euphoric introduction to evolution to underline its general importance. Today, almost fifty years after the discovery of the double helix and more than 200 years after Charles Darwin’s birthday in 1809, evolution is still at the center of biological thinking: the statement of geneticist Theodosius Dobzhansky [33], “Nothing in biology makes sense except in the light of evolution”, is frequently cited and yet, Paul Griffiths [75] has given voice to the widespread feeling that an evolutionary perspective is indeed necessary, but it must be a forward-looking perspective allowing for a general understanding of the evolutionary process, not only a backward-looking perspective dealing with the specific evolutionary history of the species.

Here, we shall review current mathematical models of evolution while focusing in particular on two mechanistic aspects of evolution, phylogeny and population dynamics with respect to structure evolution, trying to work out how they are related. The first section introduces the concept of the phylogenetic tree and continues with an attempt to translate Charles Darwin’s thoughts on natural selection into mathematical language with the knowledge of his contemporaries in mathematics. Section 2 elaborates on the state of the art in modeling phylogeny and continues with a review of evolutionary dynamics based on reproduction and mutation. Section 3 then deals with the translation of results from theory to applications in order to demonstrate the practical usefulness of evolutionary models. The notes (section 4) will be concentrating on computation and simulation methods of phylogenetic aspects of the evolutionary process. In the prospects section 5 we shall discuss the perspectives afforded by attempts to combine molecular structure and phylogeny in modeling evolution, and mention future developments in the experimental determination of fitness landscapes in the sense of Sewall Wright’s metaphor [177].

Fig. 1 An evolutionary tree by Charles Darwin. The ancestral species is at position ‘1’. Extant species are denoted by endpoint and letters, and the remaining pendant edges represent extinctions. On the margin of his sketch of a tree Darwin had written, ‘I think’, before expanding his idea in *The Origin of Species* [30]: “The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species...” (*First Notebook on Transmutation of Species*, 1837, courtesy of Cambridge University Library).
Fig. 2 Sequence, structure, and function in structural biology.

The relation between sequence, structure, and function is modeled by two successive mappings, \( \Phi \) from sequences into structures and \( \Psi \) from structures into real numbers being a quantitative measure of function.

\[
\begin{align*}
\Phi & : (Q,d_H) \rightarrow (Y,d_Y) \\
\Psi & : (Y,d_Y) \rightarrow \mathbb{R}^1 \\
S & \xrightarrow{\Phi} Y = \Phi(S) \xrightarrow{\Psi} f = \Psi(Y)
\end{align*}
\]

sequence \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \}
1.1 The phylogenetic tree

In present-day phylogenetic research – following Darwin’s ideas (Fig. 1) – we assume that \( n \) sequences \( S_n \) are related by an (unrooted or rooted) tree \( T \) whose leaves represent the aligned sequences. The tree \( T = (V, E) \) consists of a vertex set \( V \) and a branch set \( E \in V \times V \) [142], where the lengths of the branches of \( T \) are a measure of the extent of evolutionary change. Above all, in this book we are interested in the change nucleotide patterns of RNA molecules along a phylogenetic tree. The vertex set \( V \) contains the taxon set \( S \), which maps one-to-one onto the leaf set.

**Definition 1.** A phylogenetic tree \( T \) is a connected, undirected, acyclic graph whose leaves are labeled bijectively by the taxon set \( S \).

1. An unrooted phylogenetic tree \( T \) has no vertices of degree two.
2. A rooted phylogenetic tree \( T \) has an internal vertex, which may have degree two and which forms the root of the tree.
3. A star tree is a phylogenetic tree \( T \) with one internal vertex, which may have a cardinality degree of the taxon set \( S \) or in other words, all taxa have one common ancestor.

The tree-length \( \Lambda_T \) is the sum of the branch lengths.

\[
\Lambda_T = \sum_{e \in E} \lambda_e
\]  

where \( \lambda_e > 0 \) represents the length of a branch \( e \in E \). In commonly used phylogenetic methods the branch length is measured in numbers of substitutions per site. This genetic distance \( d \) estimated from nucleotide sequences are generally based on models of sequence evolution – as we shall discuss in section 2.1.1. By contrast, the

Fig. 3 Phylogenetic trees showing relatedness of sequences \((S_0)\). The numbers \( e_i \) symbolize the lengths of the branches measured in Hamming distance \( d_{H} \). The distance between any pair of sequences can be computed by adding up the lengths of the connecting branches. Lhs: Unrooted tree of five sequences \((S_1 \text{ to } S_5)\): this tree does not contain a node corresponding to the ancestor of all five sequences. Rhs: Rooted tree.
Hamming distance of two sequences, \( d_H(S_i, S_j) \), is the smallest number of substitutions, extensions, and deletions required to convert the two sequences into each other.

Charles Darwin’s great foresight concerning the existence of a tree of life has been confirmed by phylogenetic reconstruction of the evolution of multicellular eukaryotes by making extensive use of data from molecular sequencing [72]. The idea of a single root of the tree of life and a common universal ancestor, however, turned out to be much less clear. Although only a few doubts were raised regarding the single, non-recurring origin of life on Earth, the nature of the primordial prokaryote remained in the realm of speculation [37, 174], and for a long time it remained undecided whether such an ancestor had been a single species or a clan of genetically strongly interacting clones. As more and more sequence information became available on extensive horizontal gene transfer (see [16] and section 1.2) among early prokaryotic organisms, it became clear that the tree concept becomes more and more obscure the further one approaches the distant past [38, 39]. The current reconstruction of the early history of life on Earth by a plethora of genomic data is not supportive of a single tree of life for archaebacterial, eubacterial, and primitive eukaryotic species but seems to be converging towards a scenario with multiple species rapidly exchanging genetic information [6, 115, 146].

1.2 Limitations of the phylogenetic tree concept

Given the massive amount of sequence data observed today, a general goal in phylogeny is to reconstruct the evolutionary history of patterns of contemporary organisms, typically in the form of a phylogenetic tree. Accumulations of mutations by the copying process and environmental factors manifest the changes in the sequences called substitutions. Thus, given the vertical transmissions in time, the discrete character of mutations and a well-defined alphabet, the biopolymer sequences represent a unique memory of the phylogenetic past. The sequences may either come from DNA, protein, RNA or other character-based molecules with linear arrangements of monomers from several classes. Misinterpretations of data, however, are possible, e.g. through time-heterogenous processes. Moreover, present day sequence-based phylogenies of organisms are based on many different genes, which can lead to controversial interpretations of evolutionary relationships between the organisms. Evolution of genes is different from species evolution, and they should not be confused. Indeed, there are various reasons why phylogenetic inference is not always straightforward and potentially leads to misconceptions (see Figs. 4), examples of problem sources are random sorting of ancestral polymorphism [132], horizontal gene transfer [36, 100], and gene duplication or gene loss [123]. As mentioned before, horizontal gene transfer obscures phylogeny and the question arises of how to define species trees for a set of prokaryotic taxa and subsequently, of how such a species tree can be inferred? In current research, phylogenetic networks offer an alternative to phylogenetic trees (A good introduction to this problematic is given...
by [83]). In addition, the fields of phylogenetic analysis and population genetics will come closer together in the near future as complete genomic sequences for large numbers of individuals, strains and species will become available thanks to advanced sequencing technologies.

However, the reader should note that the application of phylogenetic methods goes beyond the reconstruction of phylogenetic trees for organisms. The phylogenetic analysis of molecular sequences using phylogenetic trees is an established field and several books (e.g., [55, 142]) offer detailed descriptions of the different approaches. In the present chapter, we shall present some applications of phyloge-

**Fig. 4 Limitations in phylogeny.** L.h.s: trees for individual characters (inner tree) can differ from the species tree (outer gray tree). On the left branches, random sorting of ancestral polymorphism at subsequent speciation events. On the right branches horizontal gene transfer, i.e. the lateral transfer of individual genes or sequences between species. R.h.s: a duplication event on the left branches and an instance of gene loss (x) on the right branch, both of which may lead to misconceptions of a species tree.
Phylogeny and evolution of structure  

netic substitution models in the context of RNA research and comparative genomics in section 3. Just as phylogenetic analysis, population genetics is a mature discipline that has been addressed in a number of books, including [77, 79]. Both phylogeny and population genetics remain active and continually evolving areas of research. In an attempt to interconnect the two disciplines for future RNA research, we shall focus on the question of how phylogeny and population genetics are related to structure evolution on the pages to come.

1.3 The mathematics of Darwin’s selection in populations

In 1838 the Belgium mathematician Pierre-François Verhulst published a kinetic differential equation, which presumably was not known to Darwin. The Verhulst or logistic equation describes population growth in ecosystems with finite resources:

\[
\frac{dN}{dt} = N \cdot r \left( 1 - \frac{N}{K} \right) \quad \text{and} \quad N(t) = N(0) \frac{K}{N(0) + (K - N(0)) e^{-rt}}.
\]

Fig. 5 Solution curves of the logistic equations (2.2').

Upper plot: The black curve illustrates growth in population size from a single individual to a population at the carrying capacity of the ecosystem. The red curve represents the results for unlimited exponential growth, \(N(t) = N(0) \exp(rt)\).

Parameters: \(r = 2, N(0) = 1, \) and \(K = 10000\).

Lower plot: Growth and internal selection is illustrated in a population with four variants. Color code: \(N\) black, \(N_1\) yellow, \(N_2\) green, \(N_3\) red, \(N_4\) blue.

Parameters: fitness values \(f_j = (1.75, 2.25, 2.35, 2.80), N_j(0) = (0.8888, 0.0888, 0.0020, 0.0004), K = 10000\).

The parameters were adjusted such that the curves for the total population size \(N(t)\) coincide (almost) in both plots.
The number of individuals $I$ in the population or the population size at time $t$ is denoted by $I_t = N(t)$, $N(0)$ is the population size at time $t = 0$, $r$ is the growth parameter or *Malthusian* parameter named after the English economist Robert Malthus, and $K$ is the *carrying capacity*, the maximal population size that can be sustained by the ecosystem. The interpretation is straightforward: Populations grow by reproduction and the population size increases proportionally to the number of individuals already present times the Malthusian parameter $-N(t) \cdot r$, growth requires resources and this is taken into account by the third factor $(1 - N(t)/K)$, which approaches zero when the population size reaches the carrying capacity and no further growth is possible. The solution of Verhulst equation is called the *logistic curve*. In the early phase of growth, $N(t) \ll K$ (Fig. 5; upper plot, red curve), the logistic curve represents unlimited exponential growth, the effect of finite resources is significant at populations sizes in the range of 20% saturation, $N(t) = 0.2K$, and larger.

The Verhulst model of constrained growth is dealing with *multiplication*, the first factor of Darwin’s principle. *Selection* follows straightforwardly from a partitioning into $n$ subpopulations \([139]\), $I_j$ ($j = 1, \ldots, n$) with $I_j = N_j$ and $N(t) = \sum_{j=1}^n N_j(t)$. Each variant has a specific growth parameter or *fitness value* denoted by $f_j; j = 1, \ldots, n$ and the result is the *selection equation* \((3)\) describing the evolution of the population:

\[
I_j(t) = \frac{N_j(t)}{\sum_{j=1}^n N_j(t)} \frac{N(t)}{K} \left(1 - \frac{N(t)}{K}\right)
\]

*Fig. 6 Solution curve of the selection equation* \((3)\). The system is studied at constant maximal population size, $N = K$, and the plots represent calculated changes of the variant distributions with time. The upper plot shows selection among three species $I_1$ (yellow), $I_2$ (green), and $I_3$ (red), and then the appearance of a fourth, fitter variant $I_4$ (blue) at time $t = 6$, which takes over and becomes selected thereafter. The lower plot presents an enlargement of the upper plot around the point of spontaneous creation of the fourth species ($I_4$). Parameters: fitness values $f_j = (1, 2, 3, 7)$; $x_j(0) = (0.9, 0.08, 0.02, 0)$ and $x_4(6) = 0.0001$. 

\[
\frac{N(t)}{K}
\]
\[ \frac{dN_j}{dt} = N_j \left( f_j - \frac{N}{K} \phi(t) \right); \quad j = 1, \ldots, n \quad \text{with} \quad \phi(t) = \frac{1}{N(t)} \sum_{i=1}^{n} f_i N_i(t). \quad (3) \]

Equation (3) is solved by means of normalized variables \( x_j(t) = \frac{N_j(t)}{N(t)} \):

\[ \frac{dx_j}{dt} = x_j \left( f_j - \phi(t) \right), \quad j = 1, \ldots, n; \quad x_j(t) = \frac{x_j(0) \cdot \exp(f_j t)}{\sum_{i=1}^{n} x_i(0) \cdot \exp(f_i t)}, \quad j = 1, \ldots, n. \quad (3a) \]

The size of the subpopulations is obtained through multiplication by the total population size

\[ N_j(t) = N(t) \cdot x_j(t) = N(t) \cdot \frac{N_i(0) \cdot \exp(f_j t)}{\sum_{i=1}^{n} N_i(0) \cdot \exp(f_i t)}; \quad j = 1, \ldots, n, \quad (3b) \]

and hence, the knowledge of the time dependence of population size is required. It is obtained by means of the integral \( \Phi(t) = \int_0^t \phi(\tau) d\tau \)

\[ N(t) = N(0) \frac{K}{N(0) + (K - N(0)) e^{-\Phi(t)}}. \quad (2a) \]

Herein \( \exp(-\Phi(t)) \) replaces \( \exp(-rt) \) in the solution of the Verhulst equation. The course of selection in the variables \( N_j \) and \( x_j \) is essentially the same and the restriction to constant population size, \( N = K \), was done only in order to simplify the analysis. Typical solution curves are shown in Fig.6. The interpretation of the solution curves (3b) is straightforward: For sufficiently long time the sum in the denominator is dominated by the term containing the exponential with the highest fitness value, \( f_m = \max \{ f_j; j = 1, \ldots, n \} \), the consequence is that all variables except \( x_m \) vanish and the fittest variant is selected: \( \lim_{t \to \infty} [x_m] = N \). Finally, the time derivative of the mean fitness, \( \frac{d\phi(t)}{dt} = \text{var}[f] \geq 0 \), encapsulates optimization in Darwinian evolution: \( \phi(t) \) is optimized during natural selection (see, e.g., [139]). The expressions for selection are rather simple, derivation and analysis are both straightforward, and everything needed was standard mathematics at Darwin’s time.

Mechanisms creating new variants are not part of the nineteenth century model of evolution. Recombination and mutation were unknown and new variants appear spontaneously in the population like the \textit{deus ex machina} in the ancient antique theater. Fig.6 illustrates selection at constant population size and the growth of a spontaneously created advantageous variant. Eventually the simple mathematical model described here encapsulates all three preconditions of Darwin’s natural selection and reflects the state of knowledge of the evolutionists in the second half of nineteenth century.
2 Mathematical models

Historically, the first synthesis of Darwin’s theory and Mendelian genetics was performed through mathematical modeling of evolution by the three scholars of population genetics, Ronald Fisher, J.B.S. Haldane, and Sewall Wright. Modeling of evolution and many other topics in biology was and still is mainly based on ordinary differential equations (ODEs) for essentially two reasons: (i) ODEs are at least moderately well suited for handling the problems and (ii) handling and analyzing ODEs is based on 350 years experience from physics and mathematics. Partial differential equations (PDEs) are used for the description of migration and spreading of populations in space, in particular in ecological and epidemiological models, and for modeling morphogenesis. Apart from ODE and PDE models difference equations are used to describe discrete generations, and stochastic modeling by means of Markov processes is applied to analyze problems in biology. Markov models are particularly important in phylogeny.

Classical phylogeny was based on morphological comparisons which, although successful, escaped quantitative analysis and handling. The advent of molecular biology and the fast growing facilities for determination of biopolymer sequences and structures changed the scene entirely. Sequencing was developed first for proteins and later for nucleic acids, it laid down the basis for sequence comparisons and initiated the field of molecular evolution. Since twenty years a true explosion of available DNA sequences is taking place and the dramatically increasing demands for electronic storage facilities and retrieval as well as the requirements for new tools for data analysis are challenging computer science and led to the development of the novel discipline of bioinformatics. The enormous increase in computing power and data storage capacities, eventually, revolutionized also biological modeling and simulation, computational biology became a field of research in its own right. Discrete mathematics, in particular combinatorics and graph theory are now indispensable for biological modeling not only in molecular phylogeny but also for nucleic acid structure predictions and several other fields.

Fig. 7 Sequence evolution. The evolution of sequences is described by substitutional changes of single positions during a certain time span $\Delta t$, measured in number of substitution per site $d$. For independent-sites models we assume that positions evolve independently (as indicated by the framed boxes) and according to the same process (see also the closely related uniform error rate model in section 2.3.4).
Fig. 8 Standard models. Independent evolution of sites, the Markov property, and continuous time are assumed, and the $|\mathcal{A}| \times |\mathcal{A}|$ instantaneous rate matrix is defined by $Q = \{(Q)_{ij}; i, j = 1, \ldots, |\mathcal{A}|\}$, and matrix, where $|\mathcal{A}| = \kappa$ is the number of character states. This chapter will mainly consider RNA evolution with $\mathcal{A} = \{A, C, G, U\}$, hence $|\mathcal{A}| = \kappa = 4$. For binary and nucleotide sequences, amino acid sequences in proteins and codon sequences in genes, the alphabet size is 2, 4, 20 and 64, respectively.

### 2.1 Phylogenetic models

Sequence structuring during a time span $\Delta t$ is described by the evolution of sequences along a tree through single nucleotide substitutions at independent positions (Fig. 7). Explicit formal sequence evolution models are most prominently analyzed by maximum likelihood approaches [e.g. 54, 91, 93, 157]. Clearly, the development of such models requires more or less well justified assumptions concerning the evolutionary process.

#### 2.1.1 Commonly used independent-site sequence evolution models

Molecular sequence data provide a recording of characters at a given instant, e.g. at time $t_0$, and contain no direct information on the rate at which they are evolving. During evolution, mutation and natural selection act upon molecules that are integrated in an organism, and molecules have no knowledge of their previous history. Such a lack of memory is one of the basic assumptions of phylogeny, and in the theory of stochastic processes this property is attributed to Markov processes. In other words, the evolutionary future depends only on the current state and not on any previous or ancestral state. Under the standard assumption of a stationary, time homogeneous and reversible Markovian substitution process one can, for every positive $\Delta t$, compute the probability of nucleobase $B_j$ at position $s_k$ under the condition that nucleobase $B_i$ has been at this position in the initial state ($t(0) = t_0$), which evolves into $B_j$ within the time span $\Delta t = t - t_0$. To define such a process

---

1. Each site of the sequence evolves independently.
2. The substitution process has no memory of past events (Markov property).
3. The process remains constant through time (homogeneity).
4. The process starts at equilibrium (stationarity).
5. Substitutions occur in continuous time.

**Table 1** Assumptions of the commonly used nucleotide substitutions models, of which some are collected in Tab. 2.
one only has to specify an instantaneous rate matrix $Q$, in which each entry $q_{ij} > 0$ stands for the rate of change from state $B_i$ to state $B_j$ during an infinitesimal period of time $dt$, illustrated in Fig. 8. The diagonal elements of matrix $Q$ are defined such that the individual rows sum up to zero by

$$q_{ii} = - \sum_{j=1, j \neq i}^{\kappa} q_{ij} \quad \text{and} \quad \sum_{j=1}^{\kappa} q_{ij} = 0. \quad (4)$$

This definition is readily interpreted: $q_{ij}$ is the rate determining the change at a given position $s_k$ from $B_i$ to $B_j$ and $\sum_{j=1, j \neq i}^{\kappa} q_{ij}$ is the rate for a change from $B_i$ to any other nucleobase at this position. The diagonal element $q_{ii} = - \sum_{j=1, j \neq i}^{\kappa} q_{ij}$, the so-called waiting time, thus determines how fast $B_i$ is replaced by another nucleobase.

The rate matrix $Q$ can be formulated in terms of time–dependent probabilities, which are elements of a probability matrix $P(t)$: $P_i(t)$ is the probability to find $B_i$ at time $t$ at a given position $s_k$ if $B_i$ was the nucleobase at time $t_0$ at this position. From the definition of the matrix $Q$ for an infinitesimal $dt$ and time homogeneity of the process follows

$$P(t) = e^{Qt} \cdot P(0) = e^{Qt}, \quad (5)$$

since $P(0)$ is the unit matrix for the definition of probabilities used above. Equ.(5) is rather a formal relation that cannot be evaluated analytically, because the expo-

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC69</td>
<td>*</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>*</td>
<td>β</td>
<td>α</td>
<td>β</td>
</tr>
<tr>
<td>K80</td>
<td>α</td>
<td>*</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>*</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>HKY</td>
<td>α</td>
<td>α</td>
<td>*</td>
<td>α</td>
<td>α</td>
<td>β</td>
<td>*</td>
<td>β</td>
</tr>
<tr>
<td>TN93</td>
<td>βπC</td>
<td>α1πGβπT</td>
<td>*</td>
<td>πCπGπT</td>
<td>*</td>
<td>απCβπG</td>
<td>βπCπT</td>
<td>*</td>
</tr>
<tr>
<td>F81</td>
<td>aπC</td>
<td>bπG</td>
<td>cπT</td>
<td>aπC</td>
<td>bπG</td>
<td>eπT</td>
<td>aπC</td>
<td>bπG</td>
</tr>
<tr>
<td>GTR</td>
<td>aπA</td>
<td>dπG</td>
<td>fπT</td>
<td>aπA</td>
<td>dπG</td>
<td>eπT</td>
<td>aπA</td>
<td>dπG</td>
</tr>
</tbody>
</table>

Table 2: Instantaneous rate matrices. The table contains six different rate matrices that are based on the assumptions summarized, for convenience, in Tab. 1. The first model was developed by Jukes and Cantor [91], and is specified by a single free parameter since it assumes equal mutation rates between all nucleobases. The other models ordered by increasing number of parameters are K80: Motoo Kimura’s two parameter model distinguishes between transitions and transversions [93], and the more general single substitution models, which consider different base compositions, in particular HKY: the Hasegawa-Kishino-Yano model [78], TN93: the Tamura-Nei model [156], F81: the Felsenstein 81 model [54], and GTR: the general time reversible model [127]. The diagonal elements (*) are given by Equ. 4.
The exponential function of a general matrix can only be expressed as an infinite series of matrix products or computed numerically in terms of the eigenvalues and eigenvectors of the matrix $Q$. In order to turn the probabilities into a dynamical model we define a normalized vector of nucleobase frequencies, $\pi(t) = (\pi_1(t), \ldots, \pi_κ(t))'$ with $\sum_i \pi_i(t) = 1$, which describes the nucleobase distribution at time $t$. The time development is therefore expressed in form of a differential equation

$$\frac{d\pi}{dt} = Q \cdot \pi \quad \text{and} \quad \pi(t) = e^{Qt} \cdot \pi(0).$$

The equation on the rhs is identical with Equ.(5) and again solutions are accessible only through numerical computation of the eigenvalue problem for $Q$ (see also section 2.3.4). Independently of the initial values $\pi(0) = \pi(0)$ the nucleobase distribution converges towards a stationary distribution $\bar{\pi} = (\bar{\pi}_1, \ldots, \bar{\pi}_κ)'$. The instantaneous rate matrix $Q$ can be multiplied by an arbitrary factor $f > 0$ that can be absorbed in time $\tau = t \cdot f^{-1}$, since time and rate are confounded and only their product can be inferred without extrinsic information [54]. We typically scale time such that the expected rate of substitutions per site is

$$-\sum_i \pi_i q_{ii} = 1. \quad (7)$$

Due the multiple substitutions we never observe $d$. We rather observe the number of differences $p$, that is computed as

$$p = 1 - \sum_i \pi_i p_{ii(t)}. \quad (8)$$

The most common models use independence along the sites as well as the other assumptions stated in Tab.1. It is worth noticing that not every stationary process is reversible. Reversibility, however, is a convenient and reasonable assumption made by the most commonly used models. A stationary Markov process is said to be timereversible if,

$$\pi_i q_{ij} = \pi_j q_{ji}. \quad (9)$$

Under the reversibility assumption, the twelve nondiagonal entries of rate matrix $Q$ can be described by six symmetric terms, the so-called exchangeability terms and four stationary frequencies. Thus, the general time reversible model (GTR) of Tab. 2 has 8 free parameters. The further assumption that the rate of substitution is the same for all nucleotides, can be relaxed by including rate heterogeneity, assuming that many of the complexities of molecular evolution are primarily manifested as a difference in the relative rate for sites changes, whilst maintaining all other aspects of the evolutionary process.

\[\text{For convenience we define } \pi \text{ as column vector but to save space we write it as a transposed row vector } \pi'.\]
Rate Heterogeneity

Each site has a defined probability of evolving at a given rate, independently of its neighbors. The distribution of relative rates $r$ is frequently assumed to follow a gamma distribution \[165, 178\]. Further developments, breaking the distribution into a pre-specified number of categories, make the model more computationally efficient \[179\]. Van de Peer et al. \[166\] used empirical pair-wise methods to infer site-specific rates of alignment positions. Based on this idea, Meyer et al. \[106\] introduced a maximum likelihood framework for estimating site-specific rates from pairs of sequences with an iterative extension to compute site-specific rates and the phylogenetic tree simultaneously. In the sense of reflecting different selective constraints at different sites, hidden Markov models (HMM) are used to assign rates of change to each site, according to a Markov process that depends on the rate of change at the neighboring site. These approaches model site dependencies through shared rate parameters while still assuming independent changes at the different sites \[e.g. 56\]. Since we are interested in the evolution of RNA structure, including different constraints such as energy values, it makes sense to focus on relaxing the assumption about independent evolving sites.

2.1.2 From independent to jointly modeling substitution events

For many years, various authors have attempted to overcome the assumption of independence across sites. In the simplest cases, a dependence structure with joint substitution events is modeled. A prominent example are codon models, which consider the dependence of neighboring sites within a codon \[71, 111\] rather than mononucleotide models in protein coding regions. The assumption of independence among sites is naturally relaxed in case the whole nucleobase triplet of a codon is taken as a unit of evolution. Evidently, this leads to larger rate matrices and makes computations more demanding, but for protein coding regions such models are comparatively more realistic.

In the same way, Schöninger and von Haeseler \[134\] have suggested modeling of joint substitution events in RNA helical regions. Clearly, the nucleotides in stem regions of RNA molecules cannot evolve independently of their base-pairing counterparts. Given the frequencies of admissible base-paired doublets, it is quite likely that substitution at a non-base-pairing doublet (fig.9b) will lead to a base-paired doublet within a relatively short time interval as we would expect in the case of so-called compensatory mutations (see fig.9). The sites are classified into two categories: (i) helical regions and (ii) loops, joints, and free ends – however, with the assumption of a fixed RNA secondary structure. While the units of loop regions are mononucleotides, as in the conventional independent models (Tab. 2), the units of helical regions are doublets (base-pairs). In the helical regions, the state space is extended to all 16 pair combinations, $B_i, B_j \in B = A \times A = \{AA, AC, AG, \ldots, GU, UU\}$. Correspondingly, the F81 model (Tab. 2) is extended to a $16 \times 16$ instantaneous rate matrix by taking into account the stationary fre-
Fig. 9 Compensatory mutation in an RNA double helix. A substitution within a double helical region destroys a base pair and creates an internal loop: (a)→(b). Further substitution at the internal loop may restore the double helical structure either by reverting the mutation, (b)→(c), or by creating other base pairs at this position (b)→(d) and (b)→(e). The gain in fitness accompanying base pair formation is mirrored by high probabilities for such events, which is tantamount to relatively short waiting times (see text).

The six possible base pairs $\mathcal{B} = \{AU, UA, GC, CG, GU, UG\}$ are explicitly considered in a $6 \times 6$ instantaneous rate matrix $Q$ [160], or augmented by one state for all mismatch pairs in a $7 \times 7$ matrix [162]. Commonly, the simultaneous substitution rate of both nucleobases in a base pair are assumed to be zero. However, there are also models which allow doublet substitutions [133]. Since Motoo Kimura’s early works on the rates of nucleobase substitutions [93], the observation of compensatory mutations has been thought to explain the conservation of structure. Among further studies on the rate of compensatory mutations we mention [85, 150] here. Rate matrices for RNA evolution were also determined empirically from a large sample of related RNA sequences [148]. A more recent investigation [147] has shown that different secondary structure categories evolve at different rates.
2.1.3 Context dependent substitutions

Relaxing the assumption of independently evolving sequence fragments through the introduction of overlapping dependencies makes the models substantially more involved. Nevertheless, accounting for stacking interactions between pairs of adjacent residues in RNA helices and other energetic constraints in RNA structure (see chapters by Ivo Hofacker and Christian Zwieb) would definitely improve the relation between substitutions and RNA structure. Not unexpectedly, empirical studies have found indications that the assumption of independent evolution of sites including the dependency on flanking sequence patterns is too restrictive and does not match the experimental data [e.g. 24, 108]. Pioneering work of context dependent substitutions with respect to the CpG-deamination process has been done by Jens Jensen and Anne-Mette Pedersen [87] using a Markov model of nucleotide sequence evolution in which the instantaneous substitution rates at a site were allowed to depend on the states of a neighboring site at the instant of the substitution. So far they were able to consider pairs of sequences only. The model consists of a first component that depends on the type of change while the second component considers the CpG-deamination process. Later the model has been extended by accounting for arbitrary reversible codon substitutions with more flexibility concerning CpG-deamination [27]. In these approaches, inference is obtained using Markov chain Monte Carlo (MCMC) of expectation maximization (EM) based pseudo-likelihood estimation and phylogenetic reconstruction is still a problem [5]. Models reducing the context were developed that allow for context dependent substitutions but can be approximated by simple extensions of Joe Felsenstein’s original framework. These models impose certain limitations on the independence of sites but allow for exact inference without too much additional cost in computation efforts (an example is the approach described in [145]). Nevertheless, some of these models are analytically solvable as was recently shown [7] for a special subcase of a model suggested by Tamura [156] and extended by the inclusion of CpG doublets. These works take up
earlier ideas on the solvability of this class of models suggested in [40] and formally proven later [3].

Global context dependency models related to considering protein structure were first developed in Jeffrey Thorne’s group using a Bayesian MCMC [126]. They defined an instantaneous rate matrix that specifies rates of change from every possible sequence every other possible sequence with the common restriction of no more than one position being allowed to change at a particular instant. This approach has been modified in order to explore a possibility where the relative rate of sequence evolution is affected by the Gibbs free energy of RNA secondary structure [182]. The results are slightly ambiguous because medium size and large RNA molecules have natural or expected structures that differ substantially from the minimum free energy (mfe) structures and they may be constrained by structure and function rather than by thermodynamic stability (see section 2.3.1). Taking into account the possible dependence of structure on all sites, the rate matrix $Q$ is $4^l \times 4^l$ for sequence or chain length $l$, and any full analysis of $Q$ is not feasible even if $l$ is extremely small (for $l = 10$ the dimension of $Q$ is already $4^{10} \times 4^{10}$). To overcome this high dimensionality, they use a sequence path approach [87, 117].

It is a well known fact in optimization theory that too many additional parameters for modeling add noise and imply a risk of overfitting the data, and this equally true for site dependence models. Careful strategies for model building are required in order to balance additional parameters with the available empirical information. Such models call for extensive analysis of their mathematical behavior, for example stability and convergence to equilibrium, and for investigations of their ability to predict the statistical properties observed with biological sequences. As a consequence, computer simulation seems to be the most appropriate tool for identifying the necessary parameters for modeling site dependence.

A general simulation framework, which takes into account site-specific interactions mimicking sequence evolution with various complex overlapping dependencies among sites, has been introduced by Tanja Gesell and Arndt von Haeseler [66]. This framework is based on the idea of applying different substitution matrices at each site ($Q_k; k = 1, \ldots, l$), which are defined by the interactions with other sites in the sequence. To this end, a neighborhood system $\mathcal{N} = (N_k)_{k=1,2,\ldots,l}$ is introduced such that $N_k \subset \{1, \ldots, l\}$, $k \notin N_i$ for each $k$ and if $i \in N_k$ then $k \in N_i$ for each $i, k$. $N_k$ contains all sites that interact with site $s_k$. With $n_k$ they denote the cardinality of $N_k$, i.e. the number of sites that interact with $s_k$. Thus $Q = \{Q_k; k = 1, \ldots, l\}$ constitutes a collection of possibly different substitution models $Q_k$ acting on the sequence and an annotation of correlations among sites (Fig.11). The approach allows for modeling evolution of nucleotide sequences along a tree for user defined systems of neighborhoods and instantaneous rate matrices.
Fig. 11 A ribozyme domain with overlapping dependencies. Example of a sequence $S$ with overlapping dependencies on site 153. Such dependencies occur, for example, in ribozyme domains [26]. The substitution rate for the whole sequence $q(S)$ is obtained as the sum of the rates of each site $q(S) = \sum_{k=1}^{l} Q_k(B_j^{(k)}, B_j^{(k)})$. The nucleobase instantaneous substitution rate depends on the states of the neighborhood system of this site at the instant of the substitution, described in the instantaneous rate matrix $Q_k$. The dimension of $Q_k$ depends on the number of neighbors $n_k$ taking into account at this site $k$.

Fig. 12 Genetic and apparent distance. The genetic distance $d(g)$, with $g(t)$ being a monotonously increasing function of realtime, counts all substitutions whereas the observed distance $p(g)$ is reduced by double and multiple substitutions at the same site and hence $d \geq p$. The Jukes-Cantor model [91] is used.

2.2 Time in phylogeny and evolution

Evolutionary dynamics is commonly described in realtime or in other words, the time axis in Figs.5 and 6 is physical time. The morphological reconstruction of biological evolution makes extensive use of the fossil record and hence is anchored in stratigraphy and other methods of palaeontological time determination. Phylogenetic reconstructions, on the other hand, are counting substitutions and thus not dependent on absolute timing. Without losing generality we assume that substitutions are related to realtime by some monotonously increasing function $g(t)$. It is impor-
tant to distinguish a genetic distance \( d(g) \) being the number of substitutions and the observable distance \( p(g) \). Because of double and multiple substitutions, which remain undetectable by sequence comparisons, \( d \geq p \) is always fulfilled (Fig.12).

Indeed, \( p(g) \) converges to some saturation value \( \lim_{g \to \infty} p(g) = p_\infty \). At this point the two sequences carry no more information on their phylogenetic relatedness. Assuming independent substitution events, \( p \) depends on chain length \( l \) and in the limit of long chains \( p \) and \( d \) become equal: \( \lim_{l \to \infty} p(l) = d \).

The function \( d(g) \) relating substitutions to real time clearly depends on the evolutionary model, in particular on the rate matrix of substitutions (Table 2). In the simplest possible case, the Jukes-Cantor model we have \( p(g) = 3(1 - e^{-4g/3})/4 \) and \( d(g) = g \). At small values of \( d \) we have \( p \approx d \) as required for the model, and the saturation value is \( p_\infty = 3d/4 \). For more models we refer to the monograph [112, p.265ff.]. As an example of an analysis of very ancient RNA genes that are already close to the saturation limit at present we mention the attempt to construct a single phylogenetic tree for all tRNAs [50].

Historically, the first and also the most spectacular attempt to relate phylogeny and time was the hypothesis of a molecular clock of evolution [99, 107, 154], which is tantamount to the assumption of a linear relation \( g = \lambda t \): The numbers of mutational changes in informational macromolecules, proteins and nucleic acids, are constant through time and possibly independent of the particular lineage in the sense of a universal \( \lambda \) value. The molecular clock hypothesis seemed to be in excellent agreement with the neutral theory of evolution [94], although the lack of influence of the generation time has always been kind of a mystery. More and more accurate data, however, have shown that the pace of the molecular clock varies with the particular protein under consideration. In addition, the molecular clock was found to tick differently in different lineages and for comparisons of sequences that diverged recently or long time ago [4]. Although the usefulness of the molecular clock as an independent source of timing has never been seriously questioned by molecular evolutionary biologists [21], a deeper look at data and models more deeply revealed many deviations from the naïve clock hypothesis [20, 80, 128].

### 2.3 Evolution and design of RNA structures

The choice of RNA as model system has several reasons: (i) a large percentage of the free energy of folding comes from Watson-Crick base pairing, and base pairing follows a logic that is accessible to combinatorics, (ii) a great deal but definitely not nearly all principles of RNA function can be derived from easy to analyze and predict secondary structures, (iii) evolution of RNA molecules can be studied in vitro and this fact found direct application in the design of molecules with predefined properties, and (iv) viroids, RNA viruses, their life cycles and their evolution can be studied at molecular resolution and provide insight into a self-regulated evolutionary process. The knowledge on molecular details of RNA based entities with partly autonomous life cycles provides direct access to a world of increasing complexity
Fig. 13 RNA secondary structures viewed by thermodynamics and folding kinetics. An RNA sequence $S$ of chain length $l = 33$ nucleotides has been designed to form two structures: (i) the single hairpin mfe structure, $Y_0$ (red) and (ii) a double hairpin metastable structure, $Y_1$ (blue). The Gibbs free energy of folding ($\Delta G$) is plotted on the ordinate axis. The leftmost diagram shows the minimum free energy structure $Y_0$ being a single long hairpin with a free energy of $\Delta G = -26.3$ kcal/mole. The plot in the middle contains, in addition, the spectrum of the ten lowest suboptimal conformations classified and color coded with respect to single hairpin shapes (red) and double hairpin shapes (blue). The most stable – nevertheless metastable – double hairpin has a folding free energy of $\Delta G = -25.3$ kcal/mole. The rightmost diagram shows the barrier tree of all conformations up to a free energy of $\Delta G = -5.6$ kcal/mole where the energetic valleys for the two structures merge into one basin containing 84 structures, 48 of them belonging to the single hairpin subbasin and 36 to the double hairpin subbasin. A large number of suboptimal structures has free energies between the merging energy of the subbasins and the free reference energy of the open chain ($\Delta G = 0$).

from molecules to viroids, from viroids to viruses and so on – all being now an enormous source of molecular data that wait to be converted into comprehensive and analyzable models of, for example, host-pathogen interactions and their evolution. Nowhere else is the relation between phylogeny and functional evolution – sometimes even within one host – so immediately evident as with RNA viruses.

### 2.3.1 Notion of structures

Notion, analysis and prediction of RNA structures are discussed in other chapters of this edited volume, in particular, chapter $\text{\LaTeX}$ by Ivo Hofacker and chapter $\text{\LaTeX}$ by Christian Zwieb. Here, we want to concentrate on one aspect of RNA structures that became clear only within the last two decades: In the great majority of natural, evolutionary selected RNA-molecules we are dealing with sequences forming a single stable structure, whereas randomly chosen sequences generically form a great variety of metastable suboptimal structures in addition to the minimum free energy structure [138]. Important exceptions of the one sequence-one structure paradigm
are RNA switches fulfilling regulatory functions in nature [73, 103, 144, 173] and synthetic biology [17]. Such riboswitches are multiconformational RNA molecules, which are involved in posttranscriptional regulation of gene expression. The conformational change is commonly induced by ligand binding or ribozymic RNA cleavage. Multitasking by RNA molecules clearly imposes additional constraints on genomics sequences and manifests itself through a higher degree of conservation in phylogeny.

2.3.2 Sequences with multiple structures

The extension of the notion of structure to multiconformational molecules is sketched in Fig. 13. The RNA molecule forms a well defined minimum free energy (mfe) structure – being a perfect single hairpin (lhs of the figure). In addition to the mfe structure, the sequence $S$ like almost all RNA sequences$^3$ forms a great variety of other, less stable structures called suboptimal structures (shown in the middle of the figure). Structures, mfe and suboptimal structures, are related through transitions, directly or via intermediates, which in a simplified version can be represented by means of a barrier tree [58, 175] shown on the rhs of the figure. Kinetic folding introduces a second time scale into the scenario of molecular evolution.$^4$ Based on Arrhenius theory of chemical reaction rates,

$$k = A \cdot e^{-E_a/RT}, \tag{10}$$

the height of the barrier, $E_a$ determines the reaction rate parameter $k$ and thereby the half life of the conformation $t_{1/2} = \ln 2/k$. In equation (10), $A$ is the pre-exponential factor of the reaction, $R$ is the gas constant and $T$ the absolute temperature in °Kelvin. The two structures shown in Fig. 13 are connected by a lowest barrier of 20.7 kcal/mole that depending on the pre-exponential factors implies half lives of days or even weeks for the two conformations. In a conventional experiment with a time scale of hours the two conformations would appear as two separate entities. Barriers, nevertheless, can be engineered to be much lower and then an equilibrium mixture of rapidly interconverting conformations may be observed. Several constraints are required for the conservation of an RNA switch, and the restrictions of variability in sequence space are substantial.

The comparison of the two dominant structures $Y_0$ and $Y_1$ in Fig. 13 provides a straightforward example for the illustration of different notions of stability: (i) thermodynamic stability, which considers only the free energies of the mfe structures – $Y_0$ in Fig. 13 is more stable than $Y_1$ since it has a lower free energy, $\Delta G(Y_0) < \Delta G(Y_1)$, (ii) conformational stability, which can be expressed in terms of suboptimal structures or partition functions within a basin or a subbasin of an RNA sequence – a conformationally stable molecule has no low lying suboptimal

$^3$ Exceptions are only very special sequences, homopolynucleotides, for example.

$^4$ Timescale number one is the evolutionary process itself. In order to be relevant for evolutionary dynamics the second timescale has to be substantially faster than the first one.
Fig. 14 A sketch of the mapping of RNA sequences onto secondary structures. The points of sequence space (here 183 on a planar hexagonal lattice) are mapped onto points in shape space (here 25 on a square lattice) and, inevitably, the mapping is many to one. All sequences $S$ folding the same mfe structure form a neutral set, which in mathematical terms is the preimage of $Y_k$ in sequence space. Connecting nearest neighbors of this set – these are pairs of sequences with Hamming distance $d_H = 1$ – yields the neutral network of the structure, $G_k$. The network in sequence space consists of a giant component (red) and several other small components (pink). On the network the stability against point mutations varies from $\hat{\lambda} = 1/6$ (white points) to $\hat{\lambda} = 6/6 = 1$ (black point). We remark that the two-dimensional representations of sequence are used here only for the purpose of illustration. In reality, both spaces are high-dimensional – the sequence space of binary sequences $Q_2(l)$, for example, is a hypercube of dimension $n$ and that of natural four-letter sequences $Q_4(l)$ an object in $3n$ dimensional space.

conformations that can be interconverted with the mfe structure at the temperature of the experiment; for kinetic structures separated by high barriers the partition functions are properly restricted to individual subbasins [104], and (iii) mutational stability that is measured in terms of the probability with which a mutation changes the structure of a molecule (see Fig.14 and section 2.3.3). All three forms of stability are relevant for evolution and phylogeny, but mutational stability and the spectrum of mutational effects – adaptive, neutral or deleterious – are most important.

2.3.3 Mapping sequences onto structures

RNA secondary structures provide a simple and mathematically accessible example of a realistic mapping of biopolymer sequences onto structures [138, 140]. Here modeling will be restricted to the assignment of a single structure to each sequence (lhs diagram in Fig.13). Apart from a few exceptions, experimental information on conformational free energy surfaces and fitness landscapes is rather very limited but the amount of available data is rapidly growing. Realistic means here that the neighborhood relations are similar to the observations in reality: (i) realistic landscapes are rugged, since nearest neighbor, i.e. Hamming distance one ($d_H = 1$), sequences
may have entirely different or very similar properties, and (ii) realistic landscapes are characterized by \textit{neutrality} in the sense that different sequences may lead to similar or identical structures and may have indistinguishable properties for evolution. Both features are observed in mappings of RNA sequences onto secondary structures.

The mapping in the forward direction, i.e. from sequence space onto a space of structures called \textit{shape space}, is formalized by

\[ \Phi: \left( \mathcal{Q}_l^{(\kappa)}, d_{H} \right) \rightarrow (\mathcal{Y}_l, d_Y) \text{ or } Y = \Phi(S). \]  

As said in the introduction, both sequence and structure space are metric spaces having the Hamming distance \( d_H \) and an appropriate structure distance \( d_Y \) as metrics. The superscript ‘\( \kappa \)’ is the size of the nucleobase alphabet (\( \kappa = 4 \) for natural sequences). The subscript ‘\( l \)’ indicates that, for simplicity, we restrict the considerations here to sequences of the same lengths \( l \). A set \( \Gamma(Y_k) \) containing all sequences folding into the same structure \( Y_k \) is denoted as \textit{neutral set}, it represents the preimage of the structure in sequence space and it is defined by

\[ \Gamma_k = \Gamma(Y_k) = \Phi^{-1}(Y_k) = \{ S_j | Y_k = \Phi(S_j) \}. \]

The neutral network \( G_k \) is the graph obtained from the set \( \Gamma_k \) by connecting all pairs of nodes with Hamming distance \( d_H = 1 \) by an edge (see Fig.14). The probability of a change in structure as a consequence of a mutation of the sequence \( S_j \in \Gamma_k \) is measured by the local degree of neutrality

\[ \hat{\lambda}_k(S_j) = \frac{\# \text{ neutral mutations of } S_j}{\# \text{ all mutations of } S_j}. \]  

Accordingly, \( \hat{\lambda}_k(S_j) = 0 \) implies that every mutation of \( S_j \) leads to a new structure and \( \hat{\lambda}_k(S_j) = 1 \) expresses the fact that the structure \( Y_k \) is formed by all single point mutants of \( S_j \), which might be characterized as a \textit{mutation resistant} sequence. An average of the local degree of neutrality is taken over the entire neutral set \( \Gamma_k \) to yield \( \bar{\lambda}_k = \frac{1}{|\Gamma_k|} \sum S_j \in \Gamma_k \hat{\lambda}_k(S_j) \), the (mean or global) degree of neutrality, which is a measure for the mutational stability of the structure \( Y_k \) as discussed in section 2.3.2. Typical \( \bar{\lambda}_k \)-values for common RNA secondary structures – as found in nature or obtained by evolution experiments with RNA molecules – lie in the range \( 0.2 < \bar{\lambda}_k < 0.3 \), implying that roughly 25 % of all mutations are neutral with respect to structure. The degree of neutrality for complete three-dimensional structures is not known and additional definitions are required, because spatial structures are points in a continuous rather than discrete shape space. Coarse graining of structures is necessary in order to be able to distinguish alike from different.
Fig. 15 The quasispecies as a function of the point mutation rate $p$. The plot shows the stationary mutant distribution of sequences of chain length $l = 50$ as a function of the point mutation rate $p$. The upper part contains the approximation by perturbation theory according to equation (16) and is compared with the exact results presented in the lower part of the figure. Plotted are the relative concentration of entire mutant classes: $\bar{y}_0$ (black) is the master sequence, $\bar{y}_1$ (red) is the sum of the concentrations of all one error mutants of the master sequence, $\bar{y}_2$ (yellow) that of all two error mutants, $\bar{y}_3$ (green) that of all three error mutants, and so on. In the perturbation approach the entire population vanishes at a critical mutation rate $p_{cr}$ called the error threshold (which is indicated by a broken gray line at $p_{cr} = 0.04501$) whereas a sharp transition to the uniform distribution is observed with the exact solutions. Choice of parameters: $f_m = 10$, $f_i = 1 \forall i = 1, \ldots, n; i \neq m$.

### 2.3.4 Chemical kinetics of evolution

In order to introduce mutations into selection dynamics Manfred Eigen [42] conceived a kinetic model based on stoichiometric equations, which handle correct replication and mutation as parallel reactions

$$ (A) + S_i \xrightarrow{Q_{ji}} S_j + S_i; \ i, j = 1, \ldots, n .$$  

(14)

In normalized coordinates (14) corresponds to a differential equation of the form
The finite size constraint \( \phi(t) = \sum_{i=1}^{n} f_i \) is precisely the same as in the mutation-free case (3), and the same techniques can be used to solve the differential equation [86, 150]. Solutions of (15) are derived in terms of the eigenvectors of the \( n \times n \) matrix \( W = Q \cdot F = \{ W_{ji} = Q_{ji} f_j \} \). The mutation frequencies are subsumed in the matrix \( Q = \{ Q_{ji} \} \) with \( Q_{ji} \) being the probability that \( S_i \) is obtained as an error copy of \( S_j \). The fitness values are the elements of a diagonal matrix \( F = \{ F_{ji} = f_i \} \). Exact solution curves and stationary mutant distributions \((\bar{P})\) are obtained by numerical computation. The stationary populations have been called quasispecies since they represent the genetic reservoirs of asexually reproducing species.

A quasispecies exists if every sequence in sequence space can be reached from every other sequence along a finite-length path of single point mutations in the mutation network. In addition the quasispecies contains all mutants at nonzero concentrations \((\bar{x}_i > 0) \forall i = 1, \ldots, n\). In other words, after sufficiently long time a kind of mutation equilibrium is reached at which all mutants are present in the population. In absence of neutral variants the quasispecies consists of a master sequence, the fittest sequence \( S_m : \{ f_m = \max(f_i; i = 1, \ldots, n) \} \), and its mutants, \( S_j (j = 1, \ldots, n, i \neq m) \), which are present at concentrations that are, in essence, determined by their own fitness \( f_j \), the fitness of the master \( f_m \) and the off-diagonal element of the mutation matrix \( Q_{ji} \) that depends on the Hamming distance from the master sequence \( d_H(S_i, S_m) \) (see Equ. (18)).

The coefficient of the first term in Equ. (15) for any sequence \( S_j \) consists of two parts: (i) the selective value \( W_{jj} = Q_{jj} f_j \) and (ii) the mutational flow \( \omega_j = \sum_{i=1,i \neq j}^{n} Q_{ji} f_i \). For the master sequence \( S_j \) the two quantities, the selective value of the master, \( W_{mm} \) and the mutational backflow \( \omega_m \) are of particular importance as we shall see in the discussion of strong quasispecies (see below).

The error threshold.

Application of perturbation theory neglecting back mutations at zeroth order yields analytical approximations for the quasispecies [42, 44], which are valid for sufficiently small mutation rates [153], \( Q_{ji} \ll \{ Q_{ii}, Q_{jj} \} (i \neq j) \):

\[
\bar{x}_m \approx \frac{Q_{mm} - \sigma_m^{-1}}{1 - \sigma_m^{-1}} \quad \text{and} \quad \frac{\bar{x}_j}{\bar{x}_m} \approx \frac{W_{jm}}{W_{mm} - W_{jj}}, \quad j = 1, \ldots, n; j \neq m
\]

with \( \sigma_m = f_m / f_m^* \) and \( f_m^* = \sum_{k=1,k \neq m}^{n} f_j \bar{x}_j / (1 - \bar{x}_m) \).

---

5 By definition of fitness values, \( f_i \geq 0 \), and mutation frequencies, \( Q_{ji} \geq 0 \), \( W \) is a non-negative matrix and the reachability condition boils down to the condition: \( W^k \gg 0 \), i.e. there exists a \( k \) such that \( W^k \) has exclusively positive entries and Perron-Frobenius theorem applies [143].
The superiority $\sigma_m$ is a measure of the advantage in fitness the master has over the rest of the population, and $f_m$ is the mean fitness of this rest.\(^6\)

In order to gain basic insight into evolutionary dynamics we introduce the uniform error rate model: The point mutation rate $p(s_k)$ is expressed in terms of a (mean) mutation rate per site and reproduction event that is assumed to be independent of the site $s_k$ and the particular sequence $S_i$. Further simplification is introduced by the use of binary rather than four-letter sequences.\(^7\) Then, diagonal and off-diagonal elements of matrix $Q$ are of the simple form

\[
Q_{ii} = (1 - p)^l \quad \text{and} \quad Q_{ji} = (1 - p)^{l - d_{H}(S_j, S_i)} p^{d_{H}(S_j, S_i)} = (1 - p)^l \varepsilon^{d_{H}(S_j, S_i)}
\]

with $\varepsilon = \frac{p}{1 - p}$, \(^{(17)}\)

and insertion into Equation (16) yields the three parameter $(l, p, \sigma)$ expression

\[
\bar{x}_m \approx \frac{(1 - p)^l - \sigma^{-1}}{1 - \sigma^{-1}} \quad \text{and} \quad \bar{x}_j \approx \varepsilon^{d_{H}(S_j, S_m)} \frac{f_m}{f_m - f_j} \bar{x}_m .
\]

Equation (18) provides a quantitative estimate for the concentrations of mutants: For given $p$, $\bar{x}_j$ is the larger the smaller the Hamming distance from the master, $d_{H}(S_j, S_m)$, and the smaller the larger the difference in fitness, $f_m - f_j$ is. The stationary concentration of the master sequence, $\bar{x}_m$, vanishes at some critical mutation rate $p = p_{cr}$ called the error threshold \([10, 42, 45, 46, 153]\): \(^8\)

\[
p_{cr} = 1 - \sigma^{-1/l} \implies p_{max} \approx \frac{\ln \sigma}{l} \quad \text{and} \quad l_{max} \approx \frac{\ln \sigma}{p} .
\]

Fig.15 compares a quasispecies calculated by the perturbation approach with the exact solution and shows excellent agreement up to the critical mutation rate $p = p_{cr}$.

This agreement is important because quantitative applications of quasispecies theory to virology are essentially based on Equation (16) (see section 2.3.7 and [34]).

At constant chain length $l$ the error threshold defines a maximal error rate for evolution, $p \leq p_{max}$, and at constant reproduction accuracy $p$ the length of faithfully copied polynucleotides is confined to $l \leq l_{max}$ \([46, 48]\). The first limit of a maximal error rate $p_{max}$ has been used in pharmacology for the development of new antiviral strategies \([35]\), and the second limit entered hypothetical modeling of early

---

\(^6\) An exact calculation of $\bar{f}_{-m}$ is difficult because it requires knowledge of the stationary concentrations of all variants in the population: $\bar{x}_i; i = 1, \ldots, n$. For computational details see \([44, 45, 47, 139]\).

\(^7\) It should be noted that artificially synthesized two letter (DU; D = 2,6-diamino-purine) ribozymes have perfect catalytic properties \([124]\).

\(^8\) Zero or negative concentrations of sequences clearly contradict the exact results described above and are an artifact of the perturbation approach. Nevertheless, the agreement between the exact solutions and the perturbation results up to the error threshold as shown in Fig.15 is remarkable.
biological evolution where the accuracy limits of enzyme-free replication confine the lengths of polynucleotides that can be replicated faithfully [49].

Exact solutions and fitness landscapes.

Exact solutions of Equ.(15) computed numerically for a single-peak fitness landscape, \( f_m = f_0 \) and \( f_j = f_n \forall j = 1, \ldots, n; j \neq m \), show a remarkably sudden change in the population structure at the critical mutation rate, \( p = p_{cr} \), that manifests itself in three observations (Fig.15):

(i) the concentration of the master sequence \( \bar{x}_m \), becomes very small – zero in the perturbation approach (16),

(ii) an abrupt change in the population structure that sharpens with increasing chain length \( l \) in a phase transition like manner,

(iii) a transition to the uniform distribution, \( \bar{x}_i = \kappa^{-1} \forall i = 1, \ldots, \kappa^l \), which is the exact solution at the mutation rate \( p = \tilde{p} = \kappa^{-1} \).

The transition occurs already at small mutation rates far away from \( \tilde{p} \) (\( p_{cr} = 0.045 \) versus \( \tilde{p} = 0.5 \) in Fig.15). Undoubtedly, evolution is not possible at point mutation rates \( p > p_{cr} \), since the existence of a uniform distribution implies that reproduction is operating but inheritance is not: Too many errors are made in the copying process and no individual sequence can be stably maintained over many generations.

Early works [172] have shown that the occurrence of error thresholds depends on the distribution of fitness values in sequence space. On simple landscapes, which are distinguished from realistic, complex landscapes, all genotypes of equal distance to the master have identical fitness values. Some smooth and simple fitness landscapes, in particular the additive, \( f_m = f_0 \) and \( f_j = f_0 - \alpha \cdot d_H(S_j, S_m) \forall j \neq m \), and the multiplicative landscape, \( f_m = f_0 \) and \( f_j = f_0 \cdot \beta d_H(S_j, S_m) \forall j \neq m \), with \( \alpha > 0 \) and \( \beta < 1 \) – which are both highly popular in population genetics – exhibit a gradual change from the homogeneous population, \( \bar{x}_m = 1 \), at \( p = 0 \) to the uniform distribution at \( p = \tilde{p} \) in contrast to the single-peak landscape that shows the error threshold phenomenon (Fig.15).

As said above (section 2.3.3), two features are characteristic for realistic landscapes: (i) ruggedness and (ii) neutrality [62, 84, 140]. Ruggedness can be modeled by assigning fitness differences at random within a band of fitness values with adjustable width \( d \). The highest fitness value is assigned to the master sequence, \( f_m = f_0 \) and all other fitness values are obtained by means of the equation

\[
f(S_j) = f_n + 2d(f_0 - f_n) \left( \eta_j^{(s)} - 0.5 \right), \quad j = 1, \ldots, \kappa^l; j \neq m,
\]

where \( \eta_j^{(s)} \) is the \( j \)-th output random number from a pseudorandom number generator with a uniform distribution of numbers in the range \( 0 \leq \eta_j^{(s)} \leq 1 \) that has been
Fig. 16 Quasispecies on model and realistic landscapes. The plots show exact solution curves for individual variants in the stationary distribution as a function of the mutation rate, $\bar{x}_i(p)$. The chain length is $l = 10$ corresponding to 1024 binary sequences on a fitness landscape defined by equation (20). The upper part of the figure was calculated with $d = 0$ corresponding to a single-peak landscape, whereas the lower part represents the results for a band width $d = 0.9375$ close to full randomness of fitness values. Choice of parameters: $f_m = 2$, $f_i = 1 \forall i = 1, \ldots, n; i \neq m, s = 491$, $d = 0$ (upper part) and $d = 0.9375$ (lower part); color code: master sequence (black), 10 individual single point mutants (red), and 45 individual double point mutants (yellow).

Typical results for the stationary mutant distribution are shown in Fig. 16: At $d = 0$ the landscape becomes the single-peak landscape and all mutants in a given mutant class have the same stationary concentration, $\bar{x}_i(p)$, the error threshold manifests itself by the sharp decrease in the stationary concentration of the master, $\bar{x}_m(p)$. For practical purposes it is useful to define a minimum value of the stationary concentration in order to compute the threshold value of the mut-

---

9 The seed $s$ indeed defines all details of the landscape that in turn is completely defined by $s$ and the particular type of the pseudorandom number generator.
tation rate: \( \bar{x}_{\min}(p_{\text{cr}}) = x_{\min} \) for the calculations reported here \( x_{\min} = 0.01 \) was used. The introduction of ruggedness has two effects: (i) The stationary solutions of the sequences belonging to one class split and form a band, which becomes broader as the width of the random scatter, \( d \), increases and (ii) the error threshold becomes sharper and the critical mutation rate \( p_{\text{cr}} \) moves towards smaller values of \( p \).

**Strong quasispecies**

Considering the entire range from \( p = 0 \) to the error threshold at \( p = p_{\text{cr}} \) it may happen that the one quasispecies, \( \bar{\Pi}_0 \), is replaced by another one, \( \bar{\Pi}_1 \) at some critical mutation rate \( p_{\text{tr}} \) or, in other words, \( \bar{\Pi}_0 \) is the stable stationary mutant distribution in the range \( 0 \leq p \leq p_{\text{tr}} \) and \( \bar{\Pi}_1 \) is stable in the range \( p_{\text{tr}} \leq p \leq p_{\text{cr}} \) – provided \( \bar{\Pi}_1 \) is not becoming unstable in this range and then replaced by another quasispecies. The transition between the two quasispecies described in this paper [141, figure 11, p. 657] is extremely sharp and reminds of a phase transition. In the limit \( l \to \infty \) the transition between two quasispecies is indeed a phase transition in sequence space [45]. Eleven years later the same phenomenon was observed in computer simulations based on digital organisms and called survival of the flattest.

The cause of the transition between two quasispecies is readily visualized: At \( p = 0 \) all mutation rates \( Q_{ji} \) are zero and selection in the sense of survival of the fittest, the master sequence \( S_m \) with \( W_{mm} = f_m = \max \{ f_i; i = 1, \ldots, n \} \) according to Equ. (3a) is observed. For \( p > 0 \) the effective fitness contains in addition to the selective value also the contribution from the mutational backflow. For a sequence \( S_k \) with somewhat smaller fitness, \( f_k < f_m \), or selective value, \( W_{kk} < W_{mm} \), a more efficient mutational backflow can overcompensate the fitness difference. An straightforward approximation [141] considers only the single point mutation neighborhoods of the two sequences \( S_m \) and \( S_k \), which are assumed to be sufficiently distant in sequence space. In order to be able to derive an analytical expression all one-error mutants are assumed to have the same fitness \( f_{m+1} \) and \( f_{k+1} \).\(^\text{10}\) respectively, with \( f_{k+1} > f_{m+1} \) – the less fit sequence \( S_k \) is situated in a neighborhood with higher fitness and the fitter sequence \( S_m \) has a less fit neighborhood. Next we consider the two \( 2 \times 2 \) matrices for the two independent blocks:

\[
M = (1-p)^l \begin{pmatrix} f_m & f_{m+1} l \varepsilon \\ f_m e & f_{m+1} (1 + (l+1)\varepsilon^2) \end{pmatrix} \quad \text{and} \\
K = (1-p)^l \begin{pmatrix} f_k & f_{k+1} l \varepsilon \\ f_k e & f_{k+1} (1 + (l+1)\varepsilon^2) \end{pmatrix}.
\]

The largest eigenvalues of the two matrices, \( \lambda_m(p) \) and \( \lambda_k(p) \) are readily calculated as functions of the mutation rate \( p \). The condition \( \lambda_m(p_{\text{tr}}) = \lambda_k(p_{\text{tr}}) \) defines

\(^{10}\) Alternatively one could think of these fitness values as mean values for the entire one error classes.
Fig. 17 Quasispecies on a realistic landscape supporting a quasispecies with multiple phase transitions. The plots show exact solution curves for individual variants in the stationary distribution as a function of the mutation rate, $\bar{\chi}(p)$. The chain length is $l = 10$ corresponding to 1024 binary sequences on a fitness landscape defined by equation (20). The upper part of the figure was calculated with $d = 0.5$ and shows typical quasispecies behavior with an error threshold at $p_{ct} = 0.0105$, the lower part with fully developed scatter $d = 1.0$ shows three distinct phase transitions marked as $\text{tr}_1$ ($p_{tr1} = 0.000655$), $\text{tr}_2$ ($p_{tr2} = 0.001792$) and $\text{tr}_3$ ($p_{tr3} = 0.002790$), and eventually an error threshold at $p_{ct} = 0.0083$ (outside the plot). Choice of parameters: $f_m = f_0 = 1.1$, $f_n = 1.0$, $s = 637$; color code: master sequence (black), 10 individual single point mutants (red), and 45 individual double point mutants (yellow), six error mutants (magenta), seven error mutants (chartreuse), eight error mutants (yellow), nine error mutants (red) and eventually the ten error mutant (black).
Fig. 18  **Quasispecies on a realistic landscape supporting a strong quasispecies.** The plots show exact solution curves for individual variants in the stationary distribution as a function of the mutation rate, $\bar{x}(p)$. The chain length is $l = 10$ corresponding to 1024 binary sequences on a fitness landscape defined by equation (20). The upper part of the figure was calculated with $d = 0.5$ and shows simple quasispecies behavior with an error threshold at $p_{cr} = 0.0110$, the lower part with fully developed scatter at $d = 1.0$ shows an error threshold at $p_{cr} = 0.0078$ (outside the plot). Choice of parameters: $f_m = f_0 = 1.1$, $f_n = 1.0$, $s = 919$; color code: master sequence (black), 10 individual single point mutants (red), and 45 individual double point mutants (yellow).
the position of the transition. The equation for $p_{tr}$ is rather complicated but it is straightforward to proceed in several steps:

$$\alpha = 1 - \frac{f_m - f_k}{f_{k+1} - f_{m+1}}, \quad \beta = 1 - \frac{f_m f_k (f_{k+1} - f_{m+1})}{f_{m+1} f_{k+1} (f_m - f_k)}$$

$$\gamma = \frac{(f_m f_{m+1} - f_k f_{k+1})^2}{(l-1) f_{m+1} f_{k+1} (f_m - f_k) (f_{k+1} - f_{m+1})},$$

$$\zeta_{tr} = \frac{1}{2} (\alpha + \beta - \gamma + \sqrt{(\alpha + \beta - \gamma)^2 - 4 \alpha \beta}),$$

$$\varepsilon_{tr} = \sqrt{1 - \frac{\zeta_{tr}}{l-1}} \quad \text{and} \quad p_{tr} = \frac{\varepsilon_{tr}}{1 + \varepsilon_{tr}}. \quad (21)$$

This approximation – like the perturbational approach to the position of the error threshold – yields astonishingly good results. For $l = 50$, $f_m = f_0 = f_{(0)} = 10^{11}$, $f_{m+1} = f_{(1)} = 1, f_k = f_{(50)} = 99$, and $f_{k+1} = f_{(49)} = 2$ perturbation theory yields $p_{tr} = 0.0366$ and $p_{cr} = 0.0450$ whereas the values obtained by numerical solution of full eigenvalue problem are $p_{tr} = 0.0362$ and $p_{cr} = 0.0454$, respectively [141].

The nature of a quasispecies and the appearance of phase transitions clearly depend on the particular fitness landscape. We illustrate by means of two extreme cases, examples of which are characterized by the random seeds $s = 637$ and $s = 919$: For low and moderate scatter of fitness values ($d = 0.5$) a landscape specific spreading of the curves of individual sequences within one error class is observed and otherwise the pictures are very similar as a comparison of the two upper plots in Figs. 17 and 18 shows. Increasing the scatter to the maximal value of $d = 1$ shows dramatic differences between the two landscapes. In the first case ($s = 637$) a phase transition appears at $d = 0.950$, two consecutive phase transitions are observed at $d = 0.995$ and eventually at full scatter ($d = 1$) three phase transitions are found and the corresponding master sequences for individual ranges are: 0 for $0 \leq p < 0.000655$, 1003 for $0.000655 \leq p < 0.001792$, 923 for $0.001792 \leq p < 0.002790$, and 247 for $0.002790 \leq p < 0.00832 = p_{cr}$.

The second case ($s = 919$) exhibits no phase transitions at all in the entire range $0 \leq p < 0.00839 = p_{cr}$ and has been characterized as a strong quasispecies. The special stability of the master sequence, which is not replaced by any other sequence of high fitness can be explained by means of the distribution of fitness values: The master sequence 0 ($f_0 = 1.1$), the fittest one error mutant 4 ($f_4 = 1.0966$) and the fittest two error mutant 516 ($f_{516} = 1.0970$) form a cluster that is coupled by intensive mutational flow and backflow (Fig. 19), and this highly efficient clan of sequences is so strong that it cannot be outperformed by another group of sequences surrounding a master sequence. A diagnostic tool for predicting strong quasispecies is shown in Fig. 19: The fittest mutant of the two error class in the (Hamming distance one) neighborhood of the fittest one error mutant has to have its fittest neighbor not in the

---

11 The subscript in parentheses stands for the error class, e.g., for $l = 50$ $f_{(1)}$ is the fitness of all 50 one error sequences or $f_{(50)}$ is the fitness of the single 50 error sequence.
Fig. 19 Mutation flow in strong quasispecies. The sketch focuses on the sequence space around the master sequence (0; chain length \( l = 10 \)) and the sequences with the highest fitness values are indicated by boldface numbers. The numbers are decimal values of binary sequences, e.g., 516 = 1000000100. Normal mutation flow is shown in gray thin lines, strong mutation flow in both directions, 0 -- 4 -- 516, is indicated by black thick lines and the related but weak (and therefore unimportant) flow, 0 -- 512 -- 516, by black dotted lines. Detailed values: \( s = 919 \); fitness values: \( f_0 = 1.1000 \), \( f_4 = 1.0966 \), \( f_{512} = 0.9296 \), \( f_{516} = 1.0970 \); color code: master sequence (black), one error class with 10 single point mutants (red), two error class with 45 double point mutants (yellow), and three error class with 120 triple mutants (green).

Is there an evolutionary advantage for strong quasispecies? Apparently, under conditions of shifting mutation rates phase transitions can destabilize populations when they drift into regimes where other quasispecies are stable. Whether we are dealing with a strong quasispecies or not depends on the fitness landscape and a rough estimate how likely a randomly chosen region in sequence space would be suitable for a strong quasispecies can be found yields about 20%. This is not a small fraction and populations would not need to drift very far to reach an area in sequence space that supports stable mutant distributions. A prediction to be checked experimentally would be therefore that evolution had found such regions and natural populations should represent strong quasispecies.
Neutrality.

Neutrality can be introduced into random landscapes in a straightforward way by means of a predefined degree of neutrality, $\lambda$. Then the fitness landscape is of the form

$$
\begin{align*}
    f(S_j) &= \begin{cases} 
        f_0 & \text{if } \eta_j^{(s)} \geq 1 - \lambda, \\
        f_n + \frac{2d}{1-\lambda} (f_0 - f_n) (\eta_j^{(s)} - 0.5) & \text{if } \eta_j^{(s)} < 1 - \lambda,
    \end{cases}
\end{align*}
$$

with the two limiting cases: (i) $\lim \lambda \to 0$ yielding the non-neutral random landscape (20) and (ii) $\lim \lambda \to 1$ leading to the fully neutral case as modeled and analyzed by Motoo Kimura [94]. Evolution on neutral landscapes is described by neutral networks (section 2.3.3) formed from sequences of identical fitness. Depending on the Hamming distance between neutral master sequences they form either a group of sequences coupled by selection dynamics or random selection takes place and only one sequence survives in the sense of Kimura’s theory. The case of vanishing mutation rates, $\lim p \to 0$, has been analyzed for two neutral sequences $S_j$ and $S_k$ and different Hamming distance $d_H(S_j, S_k)$:

1. $d_H = 1$: $\lim_{p \to 0} \bar{x}_j = 1$ or $\lim_{p \to 0} \bar{x}_k = 0.5$,
2. $d_H = 2$: $\lim_{p \to 0} \bar{x}_j = \alpha$ or $\lim_{p \to 0} \bar{x}_j = \alpha/(1 + \alpha)$, $\lim_{p \to 0} \bar{x}_k = 1/(1 + \alpha)$, with some value $0 \leq \alpha \leq 1$, and
3. $d_H \geq 3$: $\lim_{p \to 0} \bar{x}_1 = 1$, $\lim_{p \to 0} \bar{x}_2 = 0$ or $\lim_{p \to 0} \bar{x}_1 = 0$, $\lim_{p \to 0} \bar{x}_2 = 1$.

In full agreement with the exact result we find that two fittest sequences of Hamming distance $d_H = 1$ are selected as a strongly coupled pair with equal frequency of both members.

Numerical results show that strong coupling does not occur only for small mutation rates but extends over the whole range of $p$-values from $p = 0$ to the error threshold $p = p_{cr}$ (Fig.20). Examples for case 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 manifest itself by the lack of a unique consensus sequence of the population has an important consequence for phylogeny reconstruction (see Fig.21). As shown in the sketch in Fig.14 neutral networks may comprise several sequences and then, all neutral nearest neighbor sequences form a strongly coupled cluster in reproduction where the individual concentrations are determined by the largest eigenvector of the adjacency matrix of the network.
Fig. 20 Quasispecies on a realistic landscape with neutral sequences. Two neighboring sequences of highest fitness form a master pair of Hamming distance $d_H = 1$, which is surrounded by 18 single point mutations. The plot shows the dependence of the joint quasispecies on the point mutation rate $p$. The concentrations of the two master sequences are practically identical for all mutation rates. Choice of parameters: $f_m = 1.1, f_n = 1.0, d = 0.5, \lambda = 0.1$; color code: master sequences (black and red), 18 individual single point mutants of both master sequences (orange).

Fig. 21 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 ($d_H = 1$). The consensus sequence is not unique 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 $\alpha$ that is determined by the fitness values of other neighboring sequences, and the nucleobases corresponding to the two master sequences appear with the same ratio $\alpha$. 

---

**Mathematical Formulations**

- Master sequence 1
  - Master sequence 2
  - Consensus sequence

- Intermediate I
  - Intermediate II
  - Master sequence 2
  - Consensus sequence

---
Advantages and deficits of the quasispecies concept

The conventional synthetic theory of evolution considers reproduction of organisms rather than molecules [105]. In contrast, the kinetic theory of evolution [42, 46, 47] is dealing with the evolutionary processes at the molecular level. Correct reproduction and mutation are implemented as parallel reactions and various detailed mechanisms of reproduction can be readily incorporated into the kinetic differential equations. An extensively investigated example is RNA replication by means of an enzyme from the bacteriophage Qβ [11, 12, 13, 171]. The detailed kinetic analysis reveals the condition for the applicability of the mutation selection model (15): The RNA concentration has to be smaller than the concentration of the replication enzyme. Provided the mechanism of reproduction is known it is straightforward to implement replication kinetics in a system of differential equations [120, pp.29-75]. This is also true for epigenetic effects, which may require the use of delay differential equations. Furthermore, the kinetic theory rather than the conventional genetics approach is the appropriate basis for molecular evolution and, in particular, molecular phylogeny since these concepts deal with genes and genomes within organisms and not with the organisms themselves.

Explicit consideration of mutations as parallel reactions to error-free reproduction is manifested in the structure of stationary populations, no ad hoc assumptions are required for the appearance of mutants. In addition, high mutation rates as found, for example, in test tube evolution experiments [9, 14, 149] and in virus reproduction [34] do not represent a problem and are handled equally well as low mutation rates. The kinetic model relates evolutionary dynamics directly to fitness landscapes, which have a straightforward physical interpretation and can be measured. Unclear and ambiguous results as obtained with simple models of fitness landscapes demonstrate that an understanding of evolution is impossible without sufficient knowledge on the molecular basis of fitness. In simple systems fitness is a property that can be determined by the methods of physics and chemistry, and thus independently of evolutionary dynamics. The current explosion of harvested data provides a new source of molecular information that can be used for the computation of fitness values under sufficiently simple conditions.

As a starting point the phylogenetic approach focusses on the evolution of individual sites $s_i^{(j)}$ in a sequence $S_i = (s_1^{(i)} s_2^{(i)} \cdots s_l^{(i)})$ whereas the kinetic model considers full sequences initially as expressed by the mutation frequencies $Q_{ji}$ and fitness values $f_i$. The independent site model of theoretical phylogeny is augmented by dependencies on other sites with increasing complexity eventually ending up at accounting all sites in the sequence. The kinetic approach progresses in opposite direction from sequences to smaller entities but eventually encounters the same parameter problem as the phylogenetic method. The uniform error model, for example, assumes independent mutation of and equal frequencies of mutation at all sites. A major difference between the two approaches concerns handling of mutations. In phylogeny the matrix $Q$ is a rate matrix whereas the kinetic approach separates mutation expressed by the matrix $Q$ and structural or functional effects represented by the fitness values in matrix $F$. This has the advantage that all fitness related energetic
effects, for example base pairing in double helical structure fragments are entering F rather than Q. Of course, there are sequence dependent effects on pure mutation like the occurrence of hot spots with higher mutation frequencies than at the other sites of the sequence, which commonly depend on RNA structure. Again taking into account site and context dependent point mutation rates increases the number of parameters enormously and considering to much detail encounters the same problem as in phylogeny. In the era of genomics the whole sequence approach appears more appropriate and the wealth of currently available data are more easily introduced into the kinetic approach.

The quasispecies is the stationary solution of a deterministic, differential equation based model that, in principle, is bound to constant population size. Analysis of the basic ODEs, however, has shown that the results in relative concentrations \( x_i; \sum_{i=1}^{n} x_i = 1 \) are generally valid as long as the population does neither die out \( \left( \sum_{i=1}^{n} N_i = 0 \right) \) nor explode \( \left( \sum_{i=1}^{n} N_i = \infty \right) \) [47]. Then, the particle number are not normalizable and the structure of the population cannot be predicted from solutions of Equ.(15). Nevertheless, quasispecies may still exist for vanishing populations but any rigorous treatment has to start out from the original kinetic equations with population size being treated as a variable. Variants with zero fitness are compatible with the error threshold phenomenon [155, 158] but the prerequisites for the conventional calculation of the quasispecies are no longer fulfilled – not every sequence can be reached from every sequence by a finite chain of point mutations (section 3.2.2).

Another question is hard to answer at present: Do the populations in nature ever reach a stationary state? In vitro evolution experiments can be carried out in such a way that stationarity or mutation equilibrium is achieved, but is this true also in nature. In virus infections specific mutants appear also within the infected host but on the other hand the effect of infection and the course of disease is rather similar with comparable hosts indicating that viruses are in a comparable state at the beginning of an infection.

Two other problems are quite general in biological modeling in particular on the molecular level: (i) Most models assume spatial homogeneity whereas cells are highly structured objects with limited diffusion, active transport, and spatial localization of molecular players, and (ii) many results are derived from differential equations, which are based on the use of continuous variables, and thereby it is implicitly assumed that populations are very large. In principle, the definition of continuous space and time requires infinite population size what is a reasonable and well justified assumption in chemistry but not in biology where sample sizes may be very small. Examples are the often extremely small concentrations of regulatory molecules. Systematic studies on the very small bacterium Mycoplasma pneu-moniae in the spirit of systems biology [76, 98, 183] have shown that in extreme cases only one molecule per hundreds or even thousands of bacterial cells is present at a given instant in the population. Also the assumption of a homogeneous space is questionable: There is very little free diffusion in real cells and even bacterial cells have a very rich spatial structure. Stochasticity plays an important role and discrete stochastic rather than deterministic continuous variables should be applied.
Finally we mention a general problem for evolutionary models. Using conventional modeling with ODEs all populations would extend over whole sequence space. A drastic example is the uniform distribution of sequences is predicted at mutation rates above error threshold. Coverage of sequence space can never occur in a finite world: Even for small RNA molecules of tRNA size we would need a population size of $N = 10^{46}$ individuals in order to have one molecule for every possible sequence whereas the largest populations in \textit{in vitro} experiments with RNA hardly exceed $N = 10^{15}$ molecules. What we have instead are clones of sequences migrating through sequence space (see, for example, [62, 63, 84, 119]. Truncation of fitness landscapes has been suggested recently as a possible solution to the problem [130]. Alternatively, one could leave the full landscape and truncate populations through setting all concentrations less than one molecule per reaction volume equal to zero and eliminate the corresponding variables. New variables come into play when their concentration exceeds this truncation threshold similarly as occurring in stochastic processes.

\subsection*{2.3.6 Evolutionary optimization of structure}

In order to simulate the interplay between mutation acting on the RNA sequence and selection operating on the phenotypes, here the RNA structures, the sequence-structure map has to be an integral part of the model [60, 61, 62, 63]. 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, see section 4.2) and uses an algorithm developed by Daniel Gillespie [68, 70]. In target search problems the replication rate of a sequence $S_k$, representing its fitness $f_k$, is chosen to be a function of the structure distance between the mfe structure formed by the sequence, $Y_k = \Phi(S_k)$ and the target structure $Y_T$.

$$f_k(Y_k, Y_T) = \frac{1}{\alpha + d_S(Y_k, Y_T)/l}$$

which increases when $Y_k$ approaches the target ($\alpha$ is an adjustable parameter that was commonly chosen to be 0.1). A trajectory is completed when the population reaches a sequence that folds into the target structure. The simulated stochastic process has two absorbing barriers, the target and the state of extinction. For sufficiently large populations ($N > 30$ molecules) the probability of extinction is very small, for population sizes reported here, $N \geq 1000$, extinction has been never observed.

A typical trajectory is shown in Fig.22. In this simulation a homogenous population consisting on $N = 1000$ molecules with the same random sequence and corresponding mfe structure is chosen as initial condition (Fig.23). The target structure is the well-known clover leaf of phenyl-alanyl-transfer RNA (tRNA\textsubscript{phe}). The mean distance to target of the population decreases in steps until the target is reached.

\footnote{Several measures for the distance between structures can be applied. Here we have chosen the Hamming distance between the parentheses notation of structures, $d_S$.}
Phylogeny and evolution of structure

Fig. 22 A trajectory of evolutionary structure optimization. The topmost plot presents the mean distance to the target structure of a population of 1000 molecules (The initial random structure and the target structure are shown in Fig. 23. The plot in the middle shows the width of the population in Hamming distance $d_H$ 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 [84]). A synchronization is observed at the end of each quasi-stationary plateau where a new adaptive phase in the approach towards the target is initiated. The synchronization is caused 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 Eqn.(23).

[60, 63, 137]. Short adaptive phases are interrupted by long quasi-stationary epochs, the latter falling into two different scenarios: (i) The structure is constant and we observe neutral evolution in the sense of Kimura [94]. The numbers of neutral mutations are proportional to the numbers of replications and the evolution of the population can be understood as a diffusion process on the corresponding neutral network [84]. (ii) The process during the stationary epoch involves several structures with identical replication rates and the evolutionary process is a kind of random walk in the space of these neutral structures.
Fig. 23 Initial and target structure of the computer optimization experiment in Fig. 22. Structure $Y_0$ is the mfe structure of a random sequence with chain length $l = 76$. The secondary structure of phenyl-alanyl-transfer RNA (tRNA$^{\text{Phe}}$) was chosen to be the target structure $Y_{44}$. Stacking regions are shown in color. The two sequences, $S_0$ and $S_{44}$, are shown below; they differ in 46 positions. Structures were computed with the Vienna RNA Package, Version 1.8.5.

The diffusion of the population on the neutral network is illustrated by the plot in the middle of Fig. 22 showing the width of the population as a function of time [137, 138]. The population width increases during the quasi-stationary epoch and sharpens almost instantaneously after a sequence that allows for the start of a new adaptive phase in the optimization process had been produced by mutation. 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 the interpretation: The drift is almost always very slow unless the population center ‘jumps’ from one point in sequence space to another point from which the new adaptive phase is initiated. A closer look at the figure reveals the coincidence of the three events: (i) beginning of a new adaptive phase, (ii) collapse-like narrowing of the population, and (iii) jump-like migration of the population center.

2.3.7 From sequences and structures to genotypes and phenotypes

Genotypes or genomes are RNA or DNA sequences. The phenotype comprises structures and functions, both of which are dependent on the specific experimental or environmental setup. The simplest system capable of selection and mutation consists of RNA molecules in a medium that sustains replication. Then, genotype and phenotype are simply the RNA sequence and structure, respectively. In case of
the Qβ evolution experiments the functional requirements are a result of the replication mechanism [11, 12, 13]. In order to be replicated the Qβ-virus RNA molecules must carry an accessible recognition site for binding to the enzyme Qβ-replicase [8, 171] and the replication involves a complementary or plus-minus strand copying mechanism with \( f_j^{(+)} \) and \( f_j^{(-)} \) being the replication rate parameters for plus-strand and minus strand synthesis. After internal equilibration the plus-minus ensemble grows exponentially with an overall fitness constant, which is the geometric mean of the fitness values of both strands: \( f_j = \sqrt{f_j^{(+)} f_j^{(-)}} \). Thus, the phenotype in the case of Qβ-replication in the test tube is the ensemble consisting of both strands.

Outside plant cells viroids are naked, cyclic, and especially stable RNA molecules whose sequences are the viroid genotypes. Viroid RNAs are multiplied through transcription by the host cell machinery and carry specific recognition sites at which transcription is initiated. Replication of Potato spindle tuber viroid, for example, starts predominantly at two specific sites with the closely related sequences GGAGCGA at position A111 and GGGCGGA at position A325 of the viroid RNA with a chain length of \( l = 359 \) nucleotides [53] – the two positions are almost on opposite sides of the cyclic RNA, 214 or 145 nucleotides apart. Viroid RNAs have many loops and bulges that serve two purposes: (i) They allow for melting of the viroid RNA since a fully double stranded molecule would be too stable to be opened, and (ii) they carry the recognition sites and motifs for replication and system trafficking [32, 184], which have also been studied on the 3D structural level [185]. Although viroid reproduction in nature requires a highly specific host cell and there was a common agreement that viroids in general are highly species specific, recent attempts to replicate Avocado sunblotch viroid in yeast cells have been successful [31]. The viroid phenotype is already quite involved: It has a structural component that guarantees high RNA stability outside the host cell, but at the same time the structure is sufficiently flexible in order to be opened and processed inside the cell. Like the Qβ RNA, viroid RNA carries specific recognition sites for initiation and control of the reproduction cycle.

Viruses, in essence, are like viroids but the complexity of the life cycle is increased by three important factors: (i) virus DNAs or RNAs carry genes that are translated in the host cell, yield virus specific factors, and accordingly viruses have genetic control on the evolution of these coding regions,(ii) the virus capsid may contain functional protein molecules, for example replicases, in addition to the virus specific genetic material, and (iii) viruses are coated by virus specific proteins or proteins and membranes (For a recent treatise of virus evolution see [34]).

Bacterial phenotypes are currently to complex for a comprehensive analysis at the molecular level. An exception are the particularly small and cell-wall free bacteria of the genus Mycoplasma. In particular, Mycoplasma genitalium is a parasitic bacterium and was considered to be the smallest organism for quite some time. Its genome consists of one circular chromosome with 582,970 base pairs and 521 genes of which 482 encode for proteins. Extensive studies aiming at full systems biology

---

13 The smallness record is currently hold by Nanoarchaeum equitans with a genome size of 490,885 base pairs.
of a cellular organism were performed on the somewhat larger species *Mycoplasma pneumoniae* with a genome size of \( l \approx 816,000 \) base pairs [76, 98, 183]. For larger organisms – ordinary bacteria of about tenfold size and small eukaryotic cells – flux balance analysis [114] rather than full molecular systems biology has been performed (An early example of flux balance analysis of *Escherichia coli* is found in [41]). Another approach to complex phenotypes is based on completely annotated genomes and information on gene interactions mainly coming from proteomics. A gene interaction network has been worked out recently for a small eukaryotic cell with roughly 6000 genes [29] and it is overwhelmingly complex.

### 3 From theory to applications

Here, we shall illustrate how the theory of evolution is applicable to problems of practical interest. We present a few selected examples of applications from phylogeny and evolutionary dynamics, although the field has become so rich that one could easily fill books with examples.

#### 3.1 Applications of phylogenetic sequence evolution models

An in-depth knowledge of how sequences evolve may evidently help us to improve the reconstruction of phylogenetic trees based on sequences. This classical application is described in section 3.1.1. Regardless of this ongoing debate, sequence evolution models in recent years have also addressed the fields of RNA research from a broader viewpoint. Here, we only mention two further applications: the value of phylogenetic simulation for distinguishing ancestral and functional correlations, in section 3.1.2 and the application of phylogenetic complex models to genomic ncRNA screens in section 3.1.3.

##### 3.1.1 Phylogenetic tree inference

There are currently four main methods of phylogenetic inference: methods based on the parsimonious principle, i.e. maximum parsimony [57], statistical methods such as maximum likelihood [54] or Bayesian inference [122], and distance-based methods like neighbor-joining [131]. For further examples of tree reconstruction methods and detailed descriptions we refer to [55].

Even though, work has been done on the development of RNA base-pair substitution models with non-overlapping tuples, it has not yet been widely adopted by the scientific community. One reason could be that a priori knowledge about the molecular structure is necessary. The other reason could be that the improvements in phylogenetic inference afforded by these models are not significant enough in
comparison to independent models with rate heterogeneity. The bias introduced in sequence analysis by ignoring heterogeneous rates among sites has been studied in population genetics [cf. 2] and phylogenetic reconstruction [180], where it has been shown that the inclusion of $\Gamma$-distribution usually improves the estimation of other evolutionary parameters, including tree topology. Analyzing the efficiency of three reconstruction methods when sequence sites are not independent, Schöniger and von Haeseler [135] demonstrated that the inferred tree is not greatly affected by the presence of these kinds of correlations. In a comparative study Savill et al. [133] have shown that models permitting a non-zero rate of double substitutions performed better than those restricting - as is usually done - the number of allowed substitutions to one per unit time. A software package, called PHASE [82], for phylogenetic inference of RNA sequences with a range of base-pairs substitution models is available. A simulation study has been recently published that recommends to include RNA secondary structure during phylogenetic inference [92].

Direct combination, such as in the case of thermodynamic nearest neighbor models that follow a phylogenetic approach, is considerably harder to implement [182]. Standard methods can no longer be used for likelihood computation and parameter estimation if Markov random fields arise. Therefore, a lot of work is currently being

Fig. 24 Phylogenetic simulations and estimations under constraints. Whenever we analyze a set of homologous sequences, we have to take their evolutionary history of the observed sequences into account. Simulating sequence evolution, while taking site-specific interactions into account, is helpful for investigating the performance of both tree-building methods and structure prediction methods. Furthermore, they are of value in themselves because they contribute to our understanding of the interrelation between structure and the substitution process.
done with the purpose of covering the required technical skills, even though these methods are still not practical on a wider scale.

3.1.2 Phylogenetic simulations: Ancestral and functional correlations

The models used for simulation need to be more accurate and complex descriptions of nature than those used for inference. Taking molecular structure into consideration while simulating sequence evolution is helpful for investigating the performance of both tree-building methods and structure prediction methods. Furthermore, these simulations are of value in themselves because they contribute to our understanding of the intertwined relationship between structure and the substitution process. For example, the use of supervised sequence evolution allows us to control and study the extent of structural and sequence conservation as RNA structure stability or the influence of phylogenetic diversity. The necessity of a structure definition from a phylogenetic viewpoint can be illustrated by the ambiguous definition of RNA families \( \text{LTeX} \). Although methods using phylogeny and structure already exist, an explicit definition from a phylogenetic viewpoint was missing. A phylogenetic structure has since been introduced [65]. This definition is based on a simulation framework that includes a model which constitutes a range of different substitution models acting on a sequence and an annotation of correlations among sites, a so-called neighborhood system as described in section 2.1.3. A phylogenetic structure (PS) is then defined by a neighborhood system, a substitution model and a phylogenetic tree (Fig. 25). The substitution model specifies the evolutionary process of nucleotide evolution. However, the model is influenced by the neighborhood system that defines the interactions among sites in a sequence. The phylogenetic tree introduces an additional dependency pattern in the observed sequences. A PS

![Diagram](image.png)

Fig. 25 An example of a phylogenetic structure. Left: an example of a thermodynamically improbable neighborhood system was chosen for didactic reasons. Middle: model \( Q \) constitutes a collection of possibly different substitution models Right: example of a phylogenetic tree with three extant taxa.
appears in a set of sequences at different instants. These can be transformed, for example, into the minimum free energy secondary structure. A realization of a PS is then a relational object at instant $t$. Fig. 26 is an illustration of what has just been described, taking the example of a phylogenetic structure of Fig. 25 in relation to a minimum free energy (mfe) structure. For didactic reasons, a thermodynamically improbable neighborhood system with a helix of such length and a long part of independent sites is used. However, this thermodynamically improbable neighborhood system influences the collection of different substitution models acting on the sequence. As part of this concept, it is possible to mimic sequence evolution under the structural constraint of the PS, e.g. given the condition of compensatory mutations with an extended Felsenstein model. Due to the stochastic nature of the substitution process, however, sequences will probably observed in the course of evolution that at least temporarily - exhibit predicted structures, deviating to some extent from the neighborhood system. Fig. 26 shows the predicted mfe-structure of a generated sequence as one possible realization of the phylogenetic structure (Fig. 25) at time $d = 0.42$ on branch $a$. In comparison to the neighborhood system of the PS, the long helix of the neighborhood system maps well onto the helix of realizations. In contrast the upper independent part is folded due to the thermodynamic impossibility of such a long loop. In a nutshell: although the neighborhood system is thermodynamically improbable, it is transformed in possible thermodynamic realizations given an evolutionary history. Fig. 27 illustrate the diversity of realizations of a PS as defined in Fig. 25: the stem region of the realization should similarly be defined by the constraint through the neighborhood system and the substitution model, while

**Fig. 26 PS transformation:** A Phylogenetic Structure (PS) of Fig. 25 is summarized on the left and is transformed at time $d = 0.42$ in a mfe realization and suboptimal mfe realizations on the right. By comparing the realization with the neighborhood system most base pairs are the same. However, due to the thermodynamical improbability of such a long loop, the upper independent part is folded.
the upper loop region has no site-specific interactions and should lead to different realizations through the mfe folding. In the upper part, however, we can observe similar mfe folds between D and E at time $d = 0.31$, based on the previous ancestral state at C. With time, the differences between D and E become progressively more apparent. In the uppermost frame, the species B and D have more structural similarities than either has with E, although D and E are more closely related. Thus, the phylogenetic history suggests that D and E are a family. However, according to the observed secondary structure realization B and D form a class at time $d = 0.48$.

Following a PS, we have to distinguish between different constraints for interactions observed among the sites. A structural constraint defines the evolutionary strength of structuring sequences at different instants $t$ and can be differentiated as; ancestral constraint and neighborhood constraint. Neighborhood constraints are site-specific interactions acting on the sequence along the evolutionary process. In this sense, the interactions observable between the sites through this neighborhood constraint are called neighborhood or functional correlations. Following a PS, the evolution of nucleotides must furthermore be taken into account. The states at the

---

**Fig. 27** Observations of secondary-structure on realization levels while evolutionary history is neglected can mislead assignments of families, illustrated by red and gray rectangles. Left: shortly after a speciation event, we have similarities based on the realization at the speciation event C and not on the neighborhood system of Fig. 25 in the upper part. Right: after more time has elapsed, B and D have more structural similarities than either of them has with E, although D and E are more closely related in the phylogenetic tree. Thus, the phylogenetic history suggests that D and E are a family, while according to the observed secondary structure realizations, both B and D (gray rectangle) and D and E (red rectangle) can form a class.
internal nodes of the phylogeny are important because of the likelihood of the state remaining unchanged after only a short period of time. This depends on the model; it is an ancestral constraint that defines the influence of ancestral nucleotide distribution at an alignment site and can be associated with observable ancestral correlations in sequences. In general, term associations, correlations or dependencies are used to represent measurements from sequences via different estimation methods. If we estimate correlations from homologous sequence data, e.g. from an alignment, they are related through their evolutionary history and common ancestral states. Thus, ancestral as well as functional correlations can occur. So far, structure prediction methods have mostly been interested in predicting dependencies that result from neighborhood constraints.

3.1.3 Phylogenetic background models for genomic screens

The consensus structure prediction program Pfold [96] uses a $16 \times 16$ RNA base-pair substitution model combined with a context-free grammar similar to the genfinder programs EvoFold [118] and QFold [125]. For assessing the significance of predicted structures, e.g. estimating the false discovery rate in a genomic screen for ncRNAs, the genomic predictions should be compared to the results obtained from randomized data with the same dinucleotide content. In the case of single sequences, well known and widely used algorithms are available for generating dinucleotide controlled random sequences either by shuffling or by first Markov chain simulations [1, 28]. One approach is to simulate the alignments of a given dinucleotide content [67]. SISSIz is based on a complex substitution model that captures the neighbor dependencies and other important alignment features except for the signal in question. In addition, it directly combines the phylogenetic null model with the RNAalifold [81] consensus folding algorithm, thus yielding to a new variant of a thermodynamic structure based RNA gene finding program that is not biased by the dinucleotide content. For further details on ncRNA gene finding programs we refer to chapter LTEX.

3.2 Controlled evolution and evolutionary design

Evolution under natural conditions is much too complex for modeling and the lack of detailed understanding at the molecular level is prohibitive for applications of evolution to solve practical design. Controlled environmental conditions and reduction of the complexity of the evolving unit, however, allow for the utilization of evolutionary methods in biotechnology. We choose here three prominent examples: (i) applied molecular evolution, (ii) lethal mutagenesis in viral pathology, and (iii) controlled bacterial evolution.
3.2.1 Evolutionary design of molecules

Experimental [149] and theoretical studies [42] showed that evolution through mutation and selection is not bound to the existence of cellular life but occurs readily with RNA molecules in test tube experiments (For details we refer to a recent review on RNA evolution in laboratory experiments [90]). The known mechanistic details of *in vitro* evolution encouraged the development of techniques exploiting the evolutionary principle for the production of designed molecules with given functions. All these methods are based on sequence variation and selection, whereby variation has become the easy part since the degree of sequence changes can be varied widely from mutation with low mutation rates to random synthesis of polynucleotides. Selection for predefined function is the tricky part of directed evolution and commonly represents a challenge for the intuition of the experimenter.

Selection by exponential enrichment (SELEX) has become a popular and routinely used method [52, 163]. Targets for binding are attached to the stationary phase of a chromatographic column and a solution containing a variety of RNA molecules is poured through the column. Molecules with high affinity to bind to the target, called *aptamers* are first retained at the column, then washed out by another solvent, and the whole procedure is repeated several times with mutated samples derived from the best binders at this moment. Binding constants almost as large as those observed with the strongest natural aggregates can be obtained after some twenty selection cycles [95].

Evolutionary design has been applied to a great variety of other problems too. A view examples of recent review articles are given here: (i) directed evolution of nucleic acid enzymes [89], (ii) design of proteins by evolutionary methods [18, 86], and (iii) applications to design of low molecular weight compounds [176].

3.2.2 Virus evolution - error thresholds and lethal mutagenesis

Virus evolution is a broad field with many aspects. The idea to relate the properties of viruses to the structures viral RNA is almost 40 years old: In a pioneering paper Charles Weissmann [171] related the life cycle of the RNA bacteriophage Qβ to the secondary structure of its RNA. The structural difference between newly synthesized and mature RNA is exploited as an important regulatory element. Mutations on the RNA genotype may have direct consequences for both RNA structures and viral life cycles [34]. Here only the practical aspect of developing antiviral medication is mentioned. Many antiviral drugs are powerful because they increase the mutation rate and drive virus populations to extinction but for a satisfactory molecular explanation of the mechanism the required information on the fitness landscape is still missing. Nevertheless, *lethal mutagenesis* is an important phenomenon and simplified models providing phenomenological explanations have been developed.\[^{14}\]

\[^{14}\] An early paper [169] claimed that zero fitness values are incompatible with the existence of quasispecies and error threshold. The result, however, turned out to be an artifact of a rather naive
Manfred Eigen and Esteban Domingo originally explained lethal mutagenesis caused by pharmaceutical compounds increasing the mutation rate through driving populations beyond the error threshold [35, 43]: At mutation rates above the error threshold replication becomes random\(^{15}\) and the result is a complete loss of the genetic information and eventually the viral life cycle breaks down. Later on James Bull and Claus Wilke studied the dynamics of lethal mutagenesis in more detail and claimed that population extinction is a phenomenon independently of catastrophic error accumulation [22, 23, 152]. Recently, Francisco Montero and coworkers [158] modeled lethal mutagenesis by means of three-species replication-mutation kinetics with neglect of back mutation.

\(^{15}\) Random replication expresses the fact that error accumulation destroys the relation between template and copy and inheritance is no longer possible.
\[
\frac{dc_m}{dt} = (f_m(1-p)^l - \vartheta) c_m, \\
\frac{dc_k}{dt} = f_m(1-p)^d (1 - (1-p)^{(l-d)}) c_m + (f_k(1-p)^d - \vartheta) c_k, \text{ and} \\
\frac{dc_j}{dt} = f_m(1-(1-p)^d) c_m + f_k(1-(1-p)^d) c_k - \vartheta c_j; \quad c = \sum_{i=1}^{n} c_i,
\]

where \(S_m\) is the master sequence with the concentration \([S_m] = c_m\) and the replication rate parameter \(f_m\), \(c_k\) and \(f_k\) refer to the class of non-lethal mutants, and \(c_j\) to the class of lethal mutants, \(\vartheta\) is the uniform degradation rate parameter for all sequences, \(l\) is the chain length, \(d\) the number of positions at which mutation yields a lethal variant and, eventually \(p\) the single point mutation rate. This model [158] is characterized by two features: (i) The concentration of the material consumed in the reproduction process is assumed to be constant, \([A] = a_0\), and \(a_0\) is absorbed in the fitness parameter \(f_i\) \((i = 1, \ldots, n)\), and (ii) a degradation rate \(\vartheta\) is introduced for all species. In contrast to the selection-mutation equation (15) and the flow reactor discussed in section 4.2, the model system Eq. (24) does not approach a stationary state but the total concentration \(c\) either grows infinitely or goes extinct. Fig. 28 illustrates the result concerning lethal mutagenesis. There are two different scenarios of quasispecies development with increasing mutation rate \(p\), which depend on the degree of lethality that is expressed in the number of lethal sites \(d\): (i) At low lethality the quasispecies reaches first the error threshold at \(p = p_{\text{cr}}\), passes a range of \(p\)-values and then becomes extinct at \(p = p_{\text{ext}}\), and (ii) at sufficiently high degree of lethality the error threshold merges with the extinction threshold and the quasispecies dies out directly at \(p = p_{\text{ext}}\). It is worth noticing that the stability of the quasispecies against mutation increases with increasing degree of lethality corresponding to a shift of the error threshold towards higher mutation rate. Lethal mutagenesis is understood at the phenomenological level but when it comes to molecular details, more experimental data and a comprehensive molecular theory is required. Studies based on more realistic landscapes including lethal variants into model landscapes in the sense of (20) and (22) are still missing.

### 3.2.3 What we learn from bacterial evolution

Controlled bacterial evolution has been and still is studied in a long time experiment by Richard Lenski. He and his groups started twelve parallel experiments derived from a single clonal \(ara^{-16}\) mutation \(ara^+ \leftrightarrow ara^-\) is neutral [15]. strain of \textit{Escherichia coli} in February 1988. The original clone underwent an immediate revertant mutation to \(ara^+\), and six \(ara^+\) and six \(ara^-\) population were chosen for the series experiment. Every day the cultures are propagated by transfer of samples into new growth medium that has been intentionally chosen as poor in glucose, \footnote{The variants \(ara^+\) and \(ara^-\) differ in a single point mutation and in the capacity to utilize arabinose as nutrient. In growth media free of arabinose the}
probes are taken, isolated and deep frozen at regular intervals of 500 generations [101, 102]. Until now the twelve populations have passed about 53,000 generations. Three findings are of direct relevance for this contribution: (i) the early adaptation to the changed environment [51], (ii) the phylogeny of controlled bacterial evolution [116], and (iii) contingency and repeatability in evolution [15].

Early adaptation to new environmental conditions occurs in steps rather than continuously [51] reminding of the course of structural optimization shown in Fig.22. Although optimization of the phenotype exhibits punctuated appearance, mutations at the level of DNA sequences occur with a fairly constant rate per generation. The population evolves by forming clones that become separate in sequence space building thereby phylogenetic trees [116]. After 31,500 generations one of the twelve populations produced a mutant, which had the capability for citrate uptake from the growth medium [15]. This clone had an instantaneous advantage and started to grow much faster and showed a dramatic increase in population size, because it had conquered a hitherto unexploited niche with a new nutrient. The main question to ask about contingency in evolution concerns the probability of the adaptive event: Was it (i) a highly improbable singular event or was it (ii) an ordinary event of common probability, which needed a preparation in the sense that the clone had to migrate into some region of sequence space before? In the first case there would be no chance to repeat the event, whereas in the second case the invention of a citrate channel should be repeatable if one started from some earlier isolate and ran the tape a second time. Richard Lenski and coworkers could find an answer: Scenario (ii) is what happens in the Escherichia coli experiments and indeed samples isolated as early as generation 20,000, i.e. 11,500 generations before the cit\(^+\) mutation had happened in the original population, were able to develop advantageous citrate variants, whereas none was found with earlier isolates. The experiment is a beautiful demonstration of contingency in controlled evolution: Migration of the population in sequence space (in the sense of Fig.22) sets the stage for mutation events.

4 Notes

In our notes we focus on simulation programs. The concept of inverse folding as used by the program RNAinverse has already been described in Chapter \texttt{TEX} which searches for sequences folding into a predefined structure. However, these programs do not take any phylogenetic relationship into account. The next section 4.1 deals with sequence evolution models under constraints along phylogenetic trees, including notes for both user and developer. In section 4.2 we introduce a physical setup, the flow reactor, which is equally well suited for computer simulation and experimental implementation.

\footnote{Most \textit{Escherichia coli} strains are unable to live on citrate buffer because they have no mechanism for uptake of citrate or citric acid into the cell. The growth medium used by Lenski \textit{et al.} in the long time evolutions experiment contained citrate buffer for pH control.}
4.1 Phylogenetic simulation methods under constraint

Generating synthetic data is a significant task. Simulated data have to be generated with the same underlying parameters and statistics as the real data to which the tool will eventually be applied. Parameter estimation is a topic in its own right and we refer to other literature as well as chapter \LaTeX{} in this book. Given a sequence evolution model, there are at least two ways of simulating the substitution of nucleotides along phylogenetic trees. First, employing the matrix of substitution probabilities for any time interval or second, using a rate matrix for an infinitesimally short time interval. The first approach requires the transition probability matrix, which is, for example, calculated by numerical computation of the eigenvalues and eigenvectors of the rate matrix $Q$ (see section 2.1.1 and 2.3.4). If the number of substitution is large, the first probability matrix approach is faster since its computing time is independent of the number of substitutions. The second rate matrix approach, however, provides a way of simulating sequences under more complex models.

So far, different programs have been designed to simulate nucleotide sequences and protein sequences along a tree [136, 121, 74, 181, 151, 113, 164, 97]. One of the most commonly used programs Seq-Gen [121] has implemented a wide range of independent nucleotide substitution models. The PHASE package [82] has implemented base-paired substitution models, but is specifically designed for RNA sequences with secondary structure without taking any energy parameters into account. From a state-of-the-art perspective, the most important aspect is to be flexible in terms of simplicity and complexity. We have to allow both general well-known structural constraints such as the mfe secondary structure and other constraints, e.g. from specific families, motifs or other RNA or Protein interactions. A method that focuses on a unifying framework for simulating sequence evolution with arbitrary complexity was therefore developed, implemented in the program SISSI (Simulating Site-Specific Interactions) [66]. Beside the model and the tree, the input file is a general neighborhood file for a user-defined neighborhood system or a ct-file. The neighborhood system can be transformed into another known structure file such as a grammar or motifs file. The framework also allows us to define a different substitution matrix for each site. Several other sequence simulators including indels exit such as DAWG [25] and INDELible [59]. However, none of these programs takes site-specific interactions into account. A first algorithm for a simulation program including an indel process and site-specific interactions is described in [65].

By way of a last general note, the random generator, as with all simulations, should be chosen carefully. Furthermore, if a large number of simulations is run in fast succession, it is highly recommended to improve the resolution of the random number generator’s automatic seeding by adding some milliseconds to it. Alternatively, most programs offer the option of specifying a seed for the random number generator. This is important for allowing the replication of results, e.g., while testing and debugging, or for repeating the simulation.
Fig. 29 The flow reactor as a device for simulating deterministic and stochastic kinetics. A stock solution containing all materials for RNA replication ([A] = a₀) including an RNA polymerase flows continuously at a flow rate r into a well stirred tank reactor (CSTR) and an equal volume containing a fraction of the reaction mixture ([\star] = \{a, b, c\}) leaves the reactor (For different experimental setups see Watts [170]). The population of RNA molecules in the reactor (S₁, S₂, ..., Sₙ present in the numbers N₁, N₂, ..., Nₙ with $N = \sum_{i=1}^{n} N_i$) fluctuates around a mean value, $N \pm \sqrt{N}$. RNA molecules replicate and mutate in the reactor, and the fastest replicators are selected. The RNA flow reactor has been used also as an appropriate model for computer simulations [60, 61, 84, 120]. There, other criteria for selection than fast replication can be applied. For example, fitness functions are defined that measure the distance to a predefined target structure and mean fitness increases during the approach towards the target [63].

4.2 The flow reactor and its applications

Models based on differential equations yield solutions also in case of marginal stability. A famous example are the well known oscillations of the Lotka-Volterra system [109, pp.79-118]. Stochastic models commonly require a precise physical setup and a well defined environment in order to yield stable and meaningful solutions. The flow reactor provides a defined and experimentally controllable environment for deterministic kinetics but it is also a suitable simulation tool for stochastic approaches to chemical reactions based on the chemical master equation [64, 69, 70]. Straightforward simulations are limited by the maximal population sizes ($< 10^9$) that can be handled in actual computations. These populations are too small for modeling chemical reactions and special techniques were developed that allow for separation of deterministic and stochastic components [167, 168]. For many biological applications, however, the tractable population sizes are sufficient.
The sketch presented in Fig.29 shows an experimental setup that is suitable for analysis and computer simulation for various experimental implementations aiming at studies of evolution in the laboratory see, e.g., [170]. Two assumptions are commonly made: (i) The flow reactor is at thermal equilibrium with a controllable heat bath and (ii) the contents of the reactor is well-stirred in order to guarantee spatial homogeneity. Non-equilibrium conditions are created by a flux of rate $r$ that regulates influx and provides the source for the material consumed in the reactor and an outflux of reactor content to compensate for the change in volume. Thermodynamic equilibrium can be studied in the limit ($r \to 0$, $t \to \infty$). The reactor in the sketch (Fig.29) illustrates the optimization of RNA molecules through mutation and selection as described in section 2.3.6 as well as Eqs.(14). Further details can be found in the literature [61, 60, 84] and [120, pp.9-17].

5 Prospects of evolution

The present increase of molecular knowledge in the life sciences ranging from biopolymers to whole organisms, populations, and ecosystems is phenomenal. The advances in technology allow nowadays for harvesting data in large quantities that were unaccessible twenty years ago. Whole genome sequences as well as protein interaction maps for whole cells are now readily available. Still there is a long way to go from our present day data to a full understanding of cellular life. In particular, characterization of biomolecular structures and analysis of biochemical functions are indispensable for the bottom-up approach in the sense of systems biology. The impressive work on the mini-bacterium *Mycoplasma pneumoniae* [76, 98, 183] has demonstrated the need for biochemical analysis very clearly. The wealth of new data requires also a novel kind of theoretical biology [19], which sets the stage for modeling and computer simulation. As far as structure and phylogeny is concerned the new developments suggest to extend the initially presented paradigm of structural biology (Fig.2) by introduction of the evolutionary aspect of structure and function being the target of selection, and the role of phylogeny that can be visualized as a coarse-grained mapping of evolution onto sequence space.

Phylogeny of RNA is particularly well suited for modeling evolution, because it comprises fairly simple systems like evolution *in vitro* of RNA molecules and allows for straightforward stepwise progression in complexity: molecules → viroids → RNA viruses and retroviruses. Sufficient data for comprehensive models of viroid or RNA virus life cycles are not yet available but there are no real technical obstacles for harvesting them and we can expect a lot of progress in the near future. Most RNA viruses mutate with high rates and retrieving phylogenies within and between hosts is a hot topic in clinical studies and in epidemiology. Apart from cancer viral infections are a field that is predestined for personal medicine since the spectrum of individual immunological responses to viral infections is rather broad. At present the role of RNA in the cell seems to be a never ending story adding more and more
important features and regulatory tasks in genetics and epigenetics to this for long
time underestimated class of molecules.

Full understanding of the evolution of bacteria and eukaryotes on the molecular
level is still a program for the future but we have learned from other disciplines that
new techniques may change the situation completely in very short time – genome
sequencing serves as the most spectacular example. The key towards such an un-
derstanding is the genotype-phenotype mapping as encapsulated in the fitness land-
scape. Fitness landscapes for in vitro evolution of RNA are available or at least
accessible. The future challenge is to progress towards the more complex cases of
viroid and RNA virus evolution and even further to free living organisms.

Biologists familiar with the quite sophisticated tools in bioinformatics might ask,
whether simple models like the ones presented and discussed here can play a future
role in the era of extensive computer simulations. We think, the answer is definitely
yes. Our small number of examples have been sufficient to demonstrate this and
we expect further progress through combining the fields of phylogenetic and evolu-
tionary dynamics described in this chapter. Only sufficiently simple concepts and
theories can provide insights into complex systems and they define appropriate ref-
ence states. The complexity of the real world is then introduced similarly as in
physics by means of intellectually comprehensible perturbations of the idealized
cases. Admittedly, simple reference models are very often not yet known, complex
networks may serve as a familiar example, and their development is an important
future task for theorists in biology. Biology and chemistry are currently merging
and there is legitimate hope that the common strategies of physicists and chemists
become more popular in biology.

Acknowledgements  The authors wish to express their gratitude towards Carolin Kosiol for help-
ful discussions. T.G is funded by a fellowship of the Austrian genome research program GEN-AU
and the GEN-AU project "Bioinformatics Integration Network III". Financial support to the CIBIV
institute from the the Wiener Wissenschafts-, Forschungs- and Technologiefonds (WWTF, to Arndt
von Haeseler) is greatly appreciated.

References

alignments: a method for random sequence permutation that preserves dinu-
mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol.
[3] P. F. Arndt, C. B. Burge, and T. Hwa. DNA sequence evolution with neighbor-
56 Tanja Gesell and Peter Schuster


