European Review, Vol. 17, No. 2, 281–319 © 2009 Academia Europæa
doi:10.1017/S1062798709000787 Printed in the United Kingdom

Genotypes and Phenotypes in the Evolution of Molecules*

PETER SCHUSTER

Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria. E-mail: pks@tbi.univie.ac.at Also, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

The landscape paradigm

Genotypes are DNA or RNA sequences that – together with epigenetic and environmental influences – determine the unfolding of the phenotype. Commonly, this process is extremely complicated and – at least for the time being – escapes rigorous mathematical analysis and serious computer modeling. Nevertheless, the relations between genotypes and phenotypes play a fundamental role in biology and in its applications to pharmaceutical research and medicine. In particular, many questions concerning evolution and its mechanisms cannot be answered without an understanding of the phenotypic consequences of changes in the genotypes. Neglecting epigenetics and environmental change for the moment, genotypes and phenotypes play clearly defined distinct parts in Darwinian evolution, which is understood as the interplay of variation and selection: all variations, mutations, recombination, and gene duplication, are changes in the polynucleotide sequences of the genotype whereas the phenotype is the target of selection.

Historically, the idea of encapsulating genotype–phenotype relations in the postulate of a landscape in the theory of evolution is due to Sewall Wright.[1] He used the landscape metaphor to illustrate optimization in the sense of Darwin's natural selection: populations climb a fitness landscape through optimization of mean fitness and in a stationary situation all species occupy local optima that correspond to the niches in an ecosystem. In the 1930s, one major problem of Wright's metaphor was that it remained unclear, in essence, what was to be plotted on the horizontal axes of the landscape given fitness is the vertical coordinate axis. A second, and even more substantial, criticism had been raised by Ronald Fisher: the metaphor is built upon

*Reprinted by courtesy from Schuster P. 2009. Genotypes and Phenotypes in the Evolution of Molecules. In Caetano-Anolles G. (in press). *Evolutionary Genomics and Systems Biology*. New York: John Wiley & Sons.

the assumptions of (i) constant fitness values and (ii) time invariant fitness landscapes, which had to be made in order to guarantee the applicability of the theorem of natural selection (see, for example, Fisher's 'Fundamental Theorem of Natural Selection'[2,3] and the recent analysis of it[4]).

Binary sequences

AUGC sequences

Molecular biology revealed the structures of nucleic acid and proteins and provided a basis for handling genotypes and phenotypes by means of sound theoretical concepts. The notion of sequence space has been introduced for nucleic acids[5] and for proteins.[6] The principle of sequence space construction is simple: A point '*i*' is assigned to every sequence X_i and the Hamming distance, $d_{ij} = d_H(X_i, X_j)$ serves as metric. (The Hamming distance counts the number of positions in which two aligned sequences differ.[7,8]. It is identical with the minimal numbers of (single) point mutations required to convert one sequence into the other.) The properties of sequence space are illustrated best by means of a build-up principle: the sequence space $\mathcal{Q}_n^{(\kappa)}\{X_\cdot; d_H\}$ is the set of all strings of length n over an alphabet of κ digits. $\mathcal{Q}_n^{(\kappa)}$ may be constructed recursively by joining κ spaces of strings of length $n-1$, $\mathcal{Q}_{n-1}^{(\kappa)}$ (Figure 1). The construction principle is the same for any alphabet – binary, three-letter, four-letter – but the objects obtained are difficult to describe except in the case of binary sequences where $\mathcal{Q}_n^{(2)}$ is a hypercube of dimension n. Sequence spaces, in general, are high dimensional objects – the dimension is $n \times (\kappa - 1)$ – and low-dimensional, in particular two-dimensional, illustrations are frequently misleading. Two often misconceived features are: (i) all sequences in sequence space are (topologically) equivalent and hence they have the same number of neighbors – there are no sequences in the interior of $\mathcal{Q}$; and (ii) distances in high-dimensional spaces are small compared with those in low-dimensional spaces with the same number of nodes. A trivial but nevertheless important feature of sequence spaces is their connectedness. From every genotype we can reach every arbitrarily chosen genotype through a series of successive point mutations whose number never exceeds n or, in other words, the Hamming distance between two arbitrarily chosen sequences fulfils: $d_H(X_i, X_j) \leq n$ forall $(X_i, X_j) \in \mathcal{Q}_n^{(\kappa)}$ (independently of κ).

Genotype–phenotype relations can be viewed as mappings from sequence space into a space of phenotypes $S_n\{S_\cdot; d_S\}$, which comprises all possible phenotypes and has some distance measure d_S as metric. Fitness and other properties of phenotypes are thought to be expressed quantitatively by some function, $f_k = \Phi(S_j)$ where S_j is the phenotype formed by some genotype $X_i : S_j = \Psi(X_i)$. Fitness values are represented by real numbers, $f_k \in \mathbb{R}^1$, with the common restriction to non-negative values.

The definition of genotype or sequence space is only a minor first step towards an understanding of genotype-phenotype relations. The complexity of

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

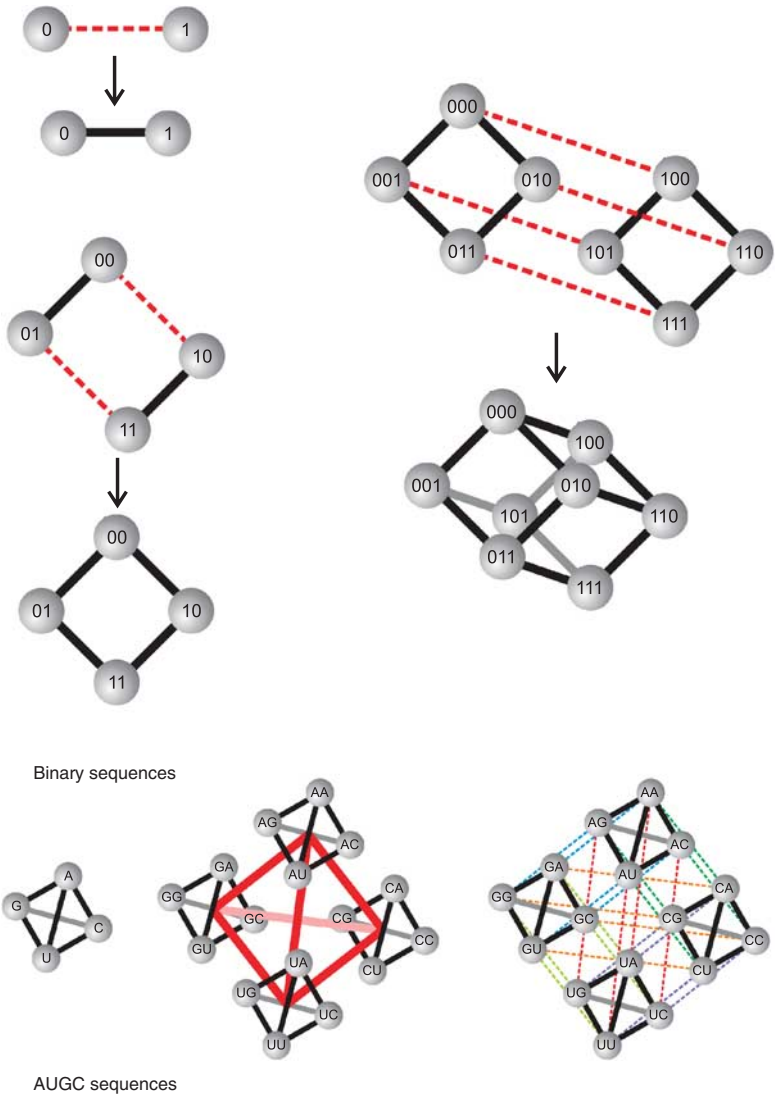


Figure 1. Sequence spaces. The properties of sequence spaces are illustrated by means of a recursive construction principle. The sequence space for strings of chain length $n + 1$, $\mathcal{Q}_{n+1}^{(\kappa)}$ is constructed from two sequence spaces for strings of chains length n, $\mathcal{Q}_{n}^{(\kappa)}$, which are obtained by adding one symbol, (0 or 1) or (A or U or G or C), respectively, on the LHS to the string. Joining all pairs of sequences with Hamming distance $d_H = 1$ by a straight line yields the sequence space $\mathcal{Q}_{n+1}^{(\kappa)}$. The upper part of the figure deals with binary sequences: $\mathcal{Q}_{n}^{(2)}$ is a hypercube of dimension n. The lower part of the figure indicates the same construction for natural four letter sequences. The single digit element, which is a straight line (and one-dimensional) for binary sequences, is a tetrahedron (and three-dimensional) in the four digit case. The sequence space $\mathcal{Q}_{2}^{(4)}$ for two letter AUGC-strings is a tetrahedron of tetrahedra (middle), a fairly complicated looking object in six-dimensional space

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

phenotypes and additional influences through epigenetics and environmental factors is currently prohibitive for useful constructions of phenotype spaces for whole cells or organisms. There are, however, examples of simpler mappings from sequence spaces into phenotypes or structures that are currently accessible by theory as well as by experiment (see also the special issue of the *Journal of Molecular Evolution*, on Experimental Evolution[9]). We mention two of them: (i) *in vitro* evolution of biomolecules with predefined functions, in particular nucleic acid molecules[10,11] and proteins[12,13] and (ii) virus evolution.[14] In both cases, the genotype is a polynucleotide that is short compared with the genoms of organisms. The numbers of possible genotypes in these relatively small sequence spaces are nevertheless huge compared with realistic population sizes: For chain lengths $n > 40$ the number of possible polynucleotide sequences exceeds Avogadro's number, in protein sequence space this happens already at chain length $n > 19$. The enormous size of sequence spaces and the principal accessibility of every genotype by mutation, in essence, set the stage for evolutionary optimization.

Molecular phenotypes

The notion of a molecular phenotype was used in the analysis and interpretation of the first evolution experiments of RNA molecules *in vitro*.[15,16] It is commonly understood as the structure of biomolecules and the properties derived from the structure. In the case of polynucleotide evolution, the situation is particularly simple because genotype and phenotype are different features of the same molecule, the nucleotide sequence and the molecular structures with its properties, respectively. In directed evolution of proteins,[13] the genotype is a DNA or RNA molecule and the phenotype is the protein molecule obtained by (transcription and) translation. In DNA display[17] the sequences are coupled to small molecules from a library that can be created, for example, by combinatorial chemistry and, in this case, the phenotype is the small molecule and its properties.

Most of the currently adopted attempts to predict function for DNA, RNA or protein sequences try to split the genotype–phenotype relation into two parts represented by mappings from sequence to structure and from structure to function:

$$\mathit{sequence} \overset{S_i=\Psi(X_i)}{\longrightarrow} \mathit{structure} \overset{f_i=\Phi(S_i)}{\longrightarrow} \mathit{function} \tag{1}$$

The rationale underlying the two-step approach is that both, prediction of structure from known sequence and prediction of function from known structure are less hard problems than the one-step prediction of function from sequence. Indeed, folding biopolymer sequences into molecular structures and inferring functions from structures follow principles from molecular physics, which are, in principle, known from structural chemistry and chemical kinetics, and thermodynamic stability.

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

Protein structures

Historically, the concept that protein folding follows a straightforward and reversible downhill process and ends at the thermodynamically stable (and therefore uniquely characterized) conformation of the molecule was derived from early work on the protein bovine pancreatic ribonuclease A.[18,19] The sequence of the small protein with a chain length of 124 amino acids was determined by Stanford Moore and William Stein[20,21] and only four years later the three-dimensional molecular structure of the protein had been determined.[22–25] The major breakthrough in understanding folding of ribonuclease A came from the work by Christian Anfinsen:[26–29] the protein was denatured through breaking four disulfide bonds by reduction and complete unfolding. On oxidation with air the molecule returned to its native conformation in extremely high yields. Anfinsen cast the findings on ribonuclease A folding and unfolding into three criteria called the *thermodynamic hypothesis* of protein structure: (i) *uniqueness* – the sequence has only one conformation of minimum free energy (m.f.e.) and no other energetically nearby lying state; (ii) *stability* – small changes in the surrounding environment cannot result in substantial changes in the m.f.e.-conformation; and (iii) *kinetic accessibility* – a smooth free energy path leads from the unfolded random coil to the folded state. In essence, a two-state model considering a folded and an unfolded state of the molecule is sufficient to describe the observations. In more recent years the kinetics of the catalytic mechanism of ribonuclease A has been studied in detail.[30] The most beautiful results obtained for ribonuclease suffer from the fact that biomolecules behaving like ribonuclease are rather a small minority in the universe of proteins. As a matter of fact, the majority of natural and artificial proteins behave differently and all three criteria of the thermodynamic hypothesis are rarely fulfilled. A stable protein structure requires a subtle balance between hydrophilic and hydrophobic interactions and, accordingly, the sequences from large sections of protein sequence space fail to from structures, because the polypeptides aggregate and the aggregates are insoluble in aqueous solution. Membrane proteins are an exception because they adopt their structures in a natural hydrophobic environment (for a review of the state of the art in membrane protein structure analysis see, for example, Ref. 31). Understanding protein structure and prediction of structures from known sequences turned out to be extremely hard, remained a major issue of biophysics for more than 30 years, and is still one of the hot topics.

In the late 1980s a new concept for the interpretation of protein folding was developed, which made use of an energy landscape.[32,33] The energy of the protein is plotted upon conformation space of the protein sequence. Conformation space is commonly continuous in chemistry; the coordinates are all bond lengths, bond angles, and dihedral angles that determine the structure of the molecule.[34] The notion of an energy landscape describing energy as a function of

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

the coordinates of a molecule is a result of the Born–Oppenheimer approximation in quantum mechanics, which separates the motions of fast-moving electrons and slow nuclei. Accurate energy landscapes are accessible through computation for small molecules and small molecular aggregates. As more and more data become available, the empirical reconstruction of free energy landscapes of proteins at atomic resolution becomes within reach.[35] The numbers of degrees of freedom in conformation space are also hyperastronomical: considering only dihedral angles – in other words keeping bond lengths and bond angles constant – we estimate for a the chain length of ribonuclease A, $n = 124$, some 10^{90} angular degrees of freedom leading to about 10^{150} local minima of the energy landscape. Levinthal[36,37] formulated a paradox in view of these huge numbers of degrees of freedom: how can a protein manage to find the native conformation in a time interval as short as a millisecond when sequential sampling of conformation, one every picosecond, would take longer than the age of the universe? The answer is sketched in Figure 2: the folding landscape has the shape of a funnel, under folding conditions (almost) all random coil conformation have conformations of lower free energy in the neighborhood, a enormous large number of trajectories leads to the target conformation, and hence only a negligibly small fraction of conformation space is sampled along an individual trajectory. The *Anfinsen funnel* (Figure 2, left-hand side) describes the idealized case of a fast folding protein such as ribonuclease A, whereas most proteins are characterized by a rugged folding landscape with a great number of local (free) energy minima (Figure 2, right-hand side). Many proteins need assistance for folding that is provided *in vivo* by chaperonins being large protein assemblies with cavities, inside which the unfolded protein finds its way into the native conformation.[38]

In essence, the mechanism of protein folding is understood by now.[32,33,39,40] Conventionally, protein structure is described at four hierarchical levels: primary, secondary, tertiary, and quaternary structure. The primary structure is the amino acid sequence of the polypeptide chain, the secondary structure consists of regular structural elements formed through closure of hydrogen bonds of the polypeptide backbone, the tertiary structure is the 3D structure of a protein or a protein subunit and the quaternary structure, eventually, provides information on the numbers and spatial arrangement of protein subunits.[41] Two notions of structural units of proteins are used in addition to the presented classification of structure: (i) A *protein domain* consists of a part of protein sequence that can fold, exist, function, and evolve independently of the rest of the protein. Domains are highly variable with respect to chain lengths, which are typically lying between 25 to 500 residues. (ii) A *structural motif* is a structural 3D element or fold within the polypeptide chain, which is transferable from one protein to other proteins. In folding, the polypeptide chain passes a series of stages: (i) local interactions, in particular nuclei of α-helices and β or reverse turns, are introduced into the random coil at one of many – more or less equivalent – positions;

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

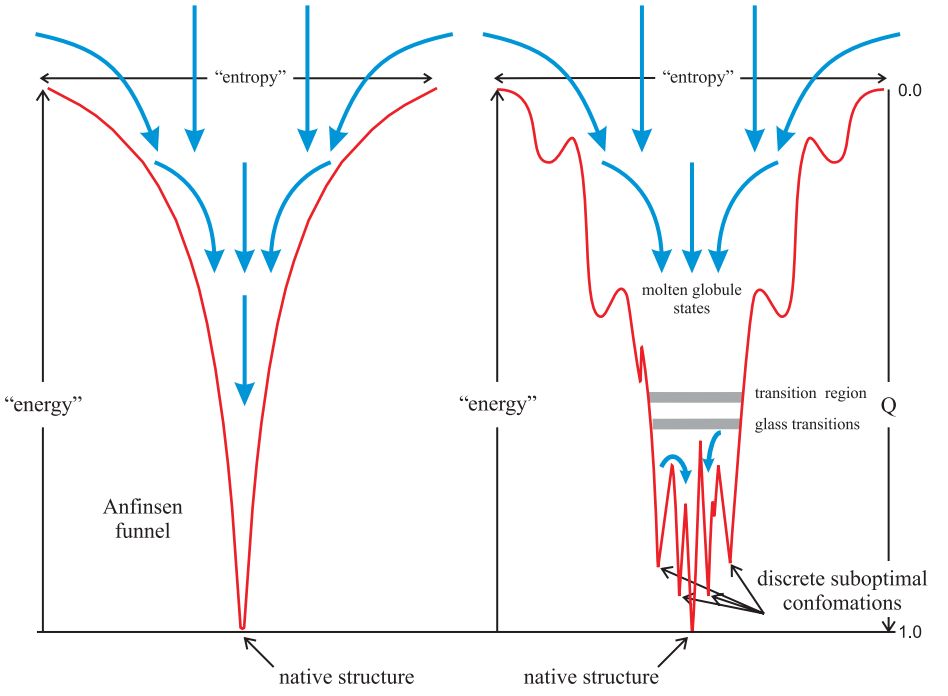


Figure 2. Energy landscapes of protein folding. The sketch of the landscape on the LHS corresponds to the *Anfinsen dogma* of protein folding: The unfolded random coil of the polypeptide sequence is converted smoothly into the unique and stable native structure as observed with ribonuclease A. The sketch of the folding funnel on the RHS represents the more common case as observed with most proteins [Onuchic et al., 1997]: The native structure is reached via various intermediates that are represented by molten globules, sometimes long lived glassy states and (discrete) suboptimal conformations, which act as *folding traps*. The abscissa axis in both sketches is an appropriate cross section of conformational space. The factor Q is the fraction of native like contacts. Typically $Q = 0.3$ for molten globules, $Q = 0.6$ in the transition region and $Q = 0.7$ in the range of glass transitions. 'Entropy' and 'energy' are put in quotation marks because they are just illustrations implying that a wide funnel sustains a larger ensemble of trajectories leading to the target state, and the depth of the funnel is a measure of the stability of the native state. The majority of entropic contributions is not encapsulated in the width of the funnel and commonly the quantity on the ordinate axis is not pure energy but Gibbs' free energy lacking entropy contributions from these degrees of freedom that are illustrated on the abscissa axis

(ii) secondary structures grow until about 30% of the contacts in the native state have been established and the molecules form a so-called molten globule – a partially ordered structure with still substantial flexibility; (iii) further loss of conformational freedom induces transitions to more rigid states – sometimes

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

of glassy nature; and (iv) confinement to one of the narrow deep values corresponding either to the native structure or to one of the suboptimal conformations, which are usually inactive and therefore addressed as misfolded states. Conversion from a suboptimal state to the native conformation may be fast or slow depending on the barrier separating the valleys. In is commonly assumed that an ordered and rigid structure is required for efficient catalysis, a recent protein engineering study, however, produced a molten globule with perfect catalytic performance that is practically the same as that of the natural counterpart.[42] Prediction of protein structure from a known sequence is still a very hard task. Progress is regularly monitored every two years by *Critical Assessment of Techniques for Protein Structure Prediction* (CASP) contests: the two latest prediction evaluation meetings of the committees were CASP 6 and CASP 7.[43,44] The progress within the last two years has been modest but two changes were significant: (i) the gap between human prediction groups and automatic servers has been closed, and (ii) an improvement has been observed with template-based models resulting from the usage of multiple templates, template free modeling in regions where no template is available, and refinement.

Nucleic acid structures

Folding of random coil polynucleotide chains into DNA and RNA structures has been studied less frequently by far than protein folding – there are more than 40,000 protein-only structures in the *Protein Data Bank* compared with 575 deposited RNA-only structures. Nevertheless, the current understanding of nucleic acid structures is not far behind our knowledge on proteins. This has mainly three reasons: (i) nucleic acids are polyelectrolytes and hence almost always soluble in water; (ii) the structures of nucleic acids fall into two distinct classes, double helical duplexes and single stranded structures; and (iii) the dominant contribution to the stability of structures is the interaction of base pairs in double helical stacks. Indeed, formation of stacked base pairs is the major driving force for folding single stranded nucleic acid molecules into structures as it is for the formation of duplexes. Although DNA in nature is almost always double stranded and RNA mostly single stranded, both nucleic acids can and do exist in both forms. Examples are deoxyribozymes that are single-stranded catalytically active DNA molecules,[45] double-strand RNA viruses, and double-stranded RNA in regulation of gene expression through RNA interference.[46] The most important issue of double stranded DNA is the sequence dependence of double helical (B-DNA) structures, which is the key to protein recognition. Empirical data based duplex structure prediction from known local DNA sequences has been successful.[47–49] Important issues of higher order structures in cyclic DNA concern supercoil, catenation, and other topological properties.[50] In this review, we shall not discuss duplex structures further but concentrate on conformations of single strand (RNA) molecules.

Polyelectrolytes require counter ions, which influence structure, and accordingly the structures of nucleic acids depend on ionic strength as well as the nature of the ions. This has been known from the early days of modeling DNA double helical structures from fiber diffraction data, and turned out to be particularly important for most full RNA structures, which are formed only when divalent $Mg^{2\oplus}$ is present in the solution. Metal ions are also known to occur as elements of protein structure – a well known example is $Zn^{2\oplus}$ in zinc fingers[51,52] – but more frequently they play an essential part in the catalytic function of proteins. Like in proteins, RNA structures can be partitioned into primary, secondary, tertiary, and quaternary structure elements. The primary structure is the nucleotide sequence, the secondary structure, in essence, is a listing of Watson-Crick and GU wobble base pairs and consists of a small number of motifs that can be combined with few restrictions only, the tertiary structure comprises additional interactions in RNA structure, which place the secondary structure elements in 3D space. These interactions are often characterized as tertiary structural motifs. Commonly, the introduction of tertiary interactions keeps secondary structures unchanged but in rare cases tertiary structure formation causes secondary structure rearrangements.[53] The quaternary structure is defined as in proteins but plays only a minor role except in RNA protein complexes, for example in virions or cellular complexes like the ribosome.

RNA structure analysis and prediction is facilitated by the existence of motifs at all structural levels.[54–56] Secondary structure motifs fall into four classes (Figure 3): (i) stacks, (ii) loops, (iii) joints, and (iv) free ends. In essence, the stacks provide the (only) stabilizing contributions to RNA structure, whereas the other elements are accompanied with positive free energy contributions. Loops are single-stranded elements attached to stacks, a hairpin loop to a single stack, a bulge or an internal loop to two stacks and a multi-loop to three or more stacks. Small hairpin loops commonly lead to large positive free energy contributions, because several degrees of freedom are frozen when the loop is closed. Exceptions, among others, are especially stable tetraloops, where a favorable geometry allows for additional base stacking.[57,58] Joints are single strands combining two otherwise independent motifs – in case the joint is cut the RNA is partitioned into two unconnected molecules. Free ends, eventually, are single stranded stretches at the 5′-end or the 3′-end of the RNA molecule. Joints and free ends are characterized by high conformational flexibility. As in proteins, composite motifs are also found in RNA. As an example we mention the kink-turn motif,[59] which is a combination of two stacks and a bulge or an internal loop between them. For certain constraints on loop size and RNA sequence the result is a sharp turn of the ribose-phosphate backbone and an acute angle formed by the axes of the double helices. The conventional definition of RNA secondary structure excludes pseudo-knots (see Figure 4 and the next section). RNA secondary structures are

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

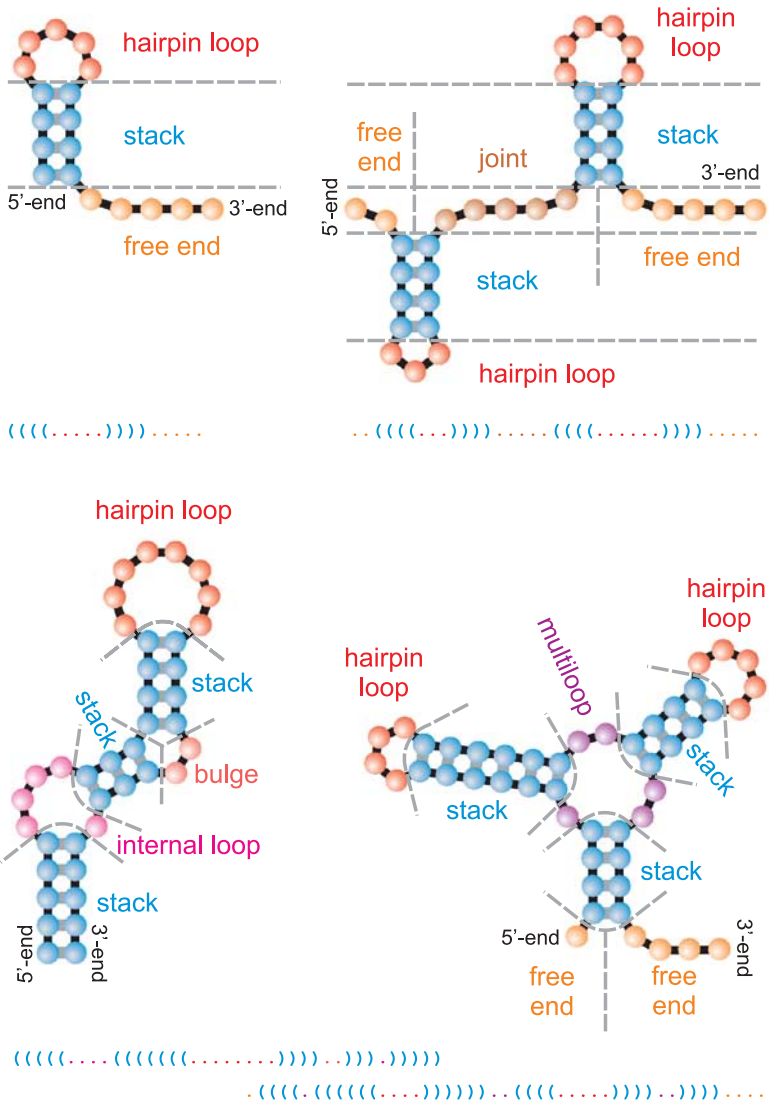


Figure 3. Modules of RNA secondary structures. Stacks (blue) consist of base pairs combined in Watson-Crick-type double helices. Hairpin loops (red) terminate stacks, bulges and internal loops (pink and magenta) are adjacent to two stacks, and multiloops (violet) combine three or more stacks. A joint (brown) is an element joining two otherwise independent parts of the structure and free ends (orange) are mobile single strands at the 5′- and/or the 3′-end of the RNA. Below the conventional representation of the secondary structures we show an equivalent representation of structures by parentheses and dots: parentheses symbolize base pairs – the opening parenthesis is nearer to the 5′-end, the closing parenthesis is nearer to the 3′-end – and the dots stand for unpaired nucleotides. As with sequences the 5′-end is on the LHS, the 3′-end on the RHS of the parentheses string. The assignment of parentheses to base pairs follow the mathematical notation

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

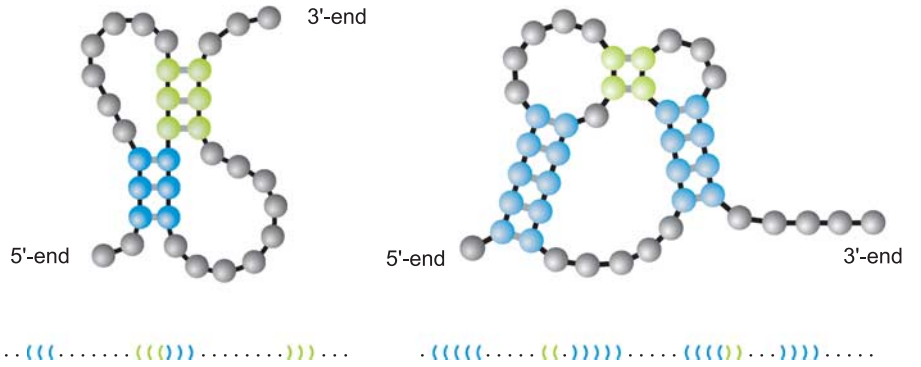


Figure 4. Pseudo-knots in RNA structures. Pseudo-knots are structures with Watson–Crick base pairs that cannot be cast into the parentheses representation without violating the mathematical notation. Parentheses cannot be assigned unambiguously to the base pairs without usage of colors. The figure sketches hairpins from two classes: (i) an hairpin-type (H-type pseudo-knot) (LHS) where a hairpin is involved in downstream base pairing, and (ii) the kissing loops motif (RHS) involving two hairpin loops forming a stack. Colored parentheses representations are shown below the figures

much more important than protein secondary structures, because every nucleotide is contained in a secondary structure motif and secondary structure formation commonly covers the major part of the free energy of folding.

Tertiary motifs are larger in number and richer in diversity than secondary structure motifs.[56] A systematic nomenclature of base pairs allows for a classification of non-Watson-Crick type nucleotide-nucleotide interactions[60] The search for tertiary RNA motifs has been very successful so far[61–63] and is still continued on a worldwide basis.[64] The most common and overall structure dominating motif is end-to-end base pair stacking of helices, also called continuous interhelical base stacking (COIN stacking). It combines stacks into elongated double helical stretches. A well known example is found in tRNAs – the 3D structure was first determined for phenylalanyl-tRNA ($tRNA^{phe}$)[65] – where the stacks terminated by the dihydro-U-loop and the anticodon loop form one extended helix and so do the stack of the TΨ-loop and the terminal stack carrying the CCA-end. The 'L'-shaped tRNA structure is stabilized by four $Mg^{2\oplus}$ ions binding to specific sites and a number of tertiary interactions involving a pseudo-knot, non-Watson-Crick base pairs, base intercalation, and binding to 2′-OH of the ribose moieties (Figure 5). Studies on randomized genes have shown that the reverse-Hogsteen base pair bridging the TΨ-loop (T54 = A58) is essential for the rigid and strong contact between the dihydro-U- and the TΨ-loop, and that the base pair is needed together with other interactions for the maintenance of the 'L'-shape.[66]

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

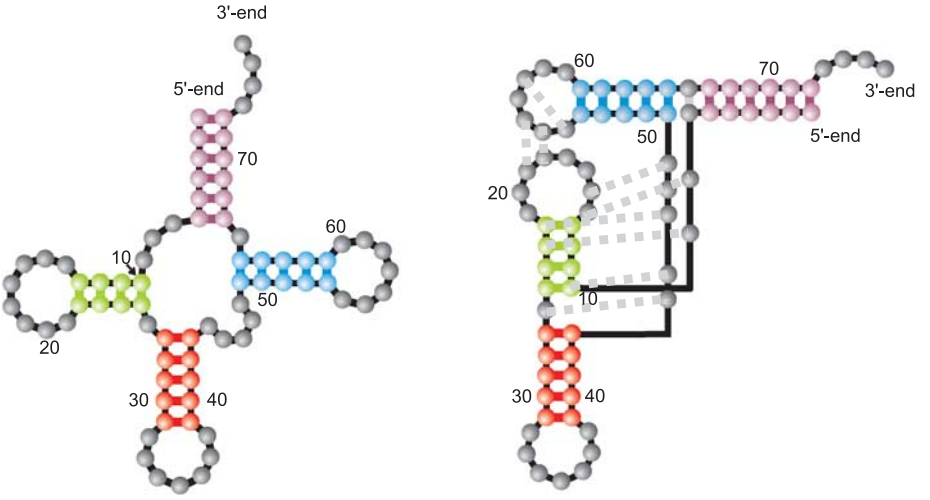


Figure 5. Tertiary interactions in tRNA structures. The figure on the LHS shows the conventional cloverleaf secondary structure of phenylalanyl transfer RNA ($tRNA^{phe}$). Continuous interhelical base (COIN) stacking shapes of the molecule into an 'L'. The stack closed by the dihydro-U-loop (green) associates end-on-end with the anticodon stack (red), the nucleotide between the two stacks, G26, forms a non-Watson-Crick base pair with A44. Similarly, the stack of the TΨ-loop (blue) is coaxial with the terminal stack (violet) with one regular AU base pair in between. Other tertiary interactions further stabilizing the 'L'-structure are shown as broken grey lines

Kinetic folding of RNA molecules follows similar principles as does kinetic protein folding. The process is initiated by local folding of structural nuclei of stacks at several positions of the RNA sequence, then the stacks grow until they form the still flexible secondary structure. The introduction of tertiary contacts and the addition of $Mg^{2\oplus}$ cations result in the full 3D structure of the molecule. Although the kinetic details of hairpin formation are quite involved,[67] the overall kinetics can be described well as a cooperative process[68–70] and modeled by straightforward algorithms[71] or computed by Arrhenius kinetics.[72] It is worth mentioning the highly promising single molecule techniques, which are steadily providing additional information on biopolymer structures and structure formation. Techniques successfully applied to RNA and protein folding are atomic force microscopy and fluorescence techniques.[73,74]

The RNA model

The landscape metaphor introduced in the first section requires either empirical data or a realistic model in order to test its applicability to RNA evolution and optimization of molecular properties. The RNA model is based on two different

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

inputs: (i) the kinetic theory of molecular evolution[5,75–77] provides the tool for the analysis of evolutionary dynamics at the molecular level, and (ii) the folding of RNA sequences into secondary structures yields simplified biomolecular structures that are suitable for the computation of parameters.[78] The relation between RNA sequences and secondary structures is used for modeling fitness landscapes of evolutionary optimization since secondary structures are physically well defined and meaningful and, at the same time, accessible to rigorous mathematical analysis.[79] In particular, RNA secondary structures allow for the introduction of most features of real structures in a straightforward and analyzable way.

RNA replication and mutation

The evolution of RNA molecules in the test tube represents the simplest system that fulfils the criteria for Darwinian evolution: (i) multiplication, (ii) variation, and (iii) selection. Evolutionary studies of RNA molecules in the test tube were initiated in the 1960s by Sol Spiegelman and his group[15,16] and has remained a highly active field ever since.[10,80] The kinetics of RNA replication by means of viral replicases has been studied in great detail.[81–83] Although RNA replication follows a complicated multistep reaction mechanism, the overall kinetics under suitable conditions consisting of excess replicase and nucleotide triphosphates can be described by simple exponential growth (Figure 6). In this phase, complementary replication

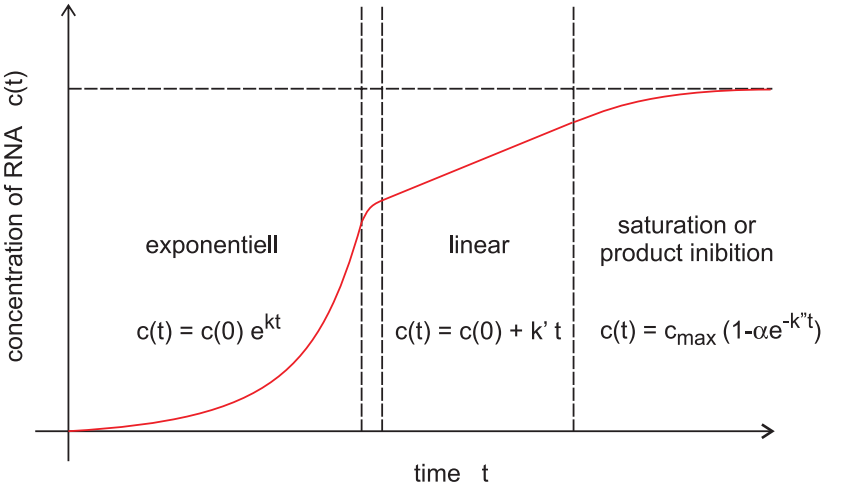


Figure 6. RNA replication by viral replicases. The growth curve of RNA concentration in a closed system with polymerase and excess nucleotide triphosphates is shown.[81] In the exponential phase the total concentration of RNA is smaller than the total concentration of replicase, in the linear phase RNA is present in excess and, eventually at high RNA concentration the growth curve levels off, since the enzyme is bound in inactive RNA-replicase complexes and RNA synthesis is blocked by product inhibition

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

sketched in Figure 7 can be represented by a simple two-step mechanism

$$\begin{aligned} X_+ &\xrightarrow{f_+} X_- + X_+ \text{ and} \\ X_- &\xrightarrow{f_-} X_+ + X_- \end{aligned} \tag{2}$$

The solution of the kinetic equations leads to two modes that describe fast internal equilibration and growth of the plus-minus ensemble with a rate parameter, $f = \sqrt{f_+ f_-}$, that is the geometric mean of the two rate constants:

$$\begin{aligned} \eta(t) &= \eta(0)e^{-ft}, \eta = \frac{\sqrt{f_+}x_+ - \sqrt{f_-}x_-}{f} \text{ and} \\ \zeta(t) &= \zeta(0)e^{+ft}, \zeta = \frac{\sqrt{f_+}x_+ + \sqrt{f_-}x_-}{f} \end{aligned} \tag{3}$$

with $x_+ = [X_+]$ and $x_- = [X_-]$ being the concentrations of plus and minus strands, respectively. Variation is introduced into ensembles of replicating RNA molecules by unprecise copying or mutation. Three classes of mutations are distinguished: (i) single nucleotide mismatch in the replication duplex leading to a point mutation (Figure 8), (ii) duplication of part of the RNA sequence leading to insertion of nucleotides, and (iii) deletion of nucleotides.

A molecular theory of evolution based on the kinetics of replication and mutation has been formulated by Manfred Eigen.[5] The concept is based on the reaction network,

$$X_j \xrightarrow{Q_{ij}f_j} X_i + X_j; i, j = 1, 2, \ldots, N \tag{4}$$

which is considered under the idealized conditions of excess nucleotide triphosphates and replicase. The rate parameter f_j refers to replications – correct and incorrect – of template X_j and the factor Q_{ij} represents the frequency of production of X_i as a copy of X_j. Since every copy has to be either correct or a mutant the conservation relation $\sum_{i=1}^{N} Q_{ij} = 1$ holds. The kinetic differential equations resulting from equation (4) with $x_i = [X_i]$ are linear

$$\frac{\mathrm{dx}}{\mathrm{d}t} = \mathrm{W} \cdot \mathrm{x} \text{ with } \mathrm{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \tag{5}$$

and can be solved in terms of eigenvalues and eigenvectors of the selection-mutation matrix W, which can be factorized into a product of the mutation matrix

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

Q and the diagonal matrix of replication rate parameters, F:

$$\mathrm{W} = \mathrm{Q} \cdot \mathrm{F} \text{ with } \mathrm{Q} = \begin{pmatrix} Q_{11} & Q_{12} & \dots & Q_{1N} \\ Q_{21} & Q_{22} & \dots & Q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{N1} & Q_{N2} & \dots & Q_{NN} \end{pmatrix} \text{ and}$$

$$\mathrm{F} = \begin{pmatrix} f_1 & 0 & \dots & 0 \\ 0 & f_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_N \end{pmatrix}.$$

Since all sequences X_i can be reached from everywhere in sequence space by a chain of successive point mutations, the matrix W^m has only strictly positive entries for sufficiently large m and the Perron-Frobenius theorem[84] holds: The eigenvector ζ_0 corresponding to the largest eigenvalue λ_0 has exclusively strictly positive components and all mutants X_i are present in the population after some time: $x_i(t) > 0$ for $t >> 0$. The eigenvalue λ_0 is positive and all components of the eigenvector are growing exponentially. The mutant distribution determined by the eigenvector ζ_0 is called *quasi-species* since it represents the genetic reservoir of an asexually replicating species. It is straightforward to introduce a constraint into equation (5) that limits population size $\sum_{i=1}^{N} x_i = c$ asymptotically to $\lim_{t\to\infty} c(t) = c_0$:

$$\frac{\mathrm{dx}}{\mathrm{d}t} = \mathrm{W} \cdot \mathrm{x} - \frac{\bar{f}}{c_0}\mathrm{x} = (\mathrm{W} - \frac{\bar{f}}{c_0}\mathbb{E}) \cdot \mathrm{x} \tag{5a}$$

Here, $\mathbb{E}$ is the unit matrix and $\bar{f} = \sum_{i=1}^{N} f_i x_i / c$ represents the mean replication rate parameter of the population. In the solutions of equation (5a) the population approaches indeed the stable stationary state $\lim_{t\to\infty} \sum_{i=1}^{N} x_i(t) = \sum_{i=1}^{N} \bar{x}_i = c_0$, which is determined by the components of the eigenvector ζ_0. Choosing $c_0 = 1$ yields relative or $\mathbb{L}^1$ normalized concentrations: $\sum_{i=1}^{N} x_1 = 1$ (which will be used in the rest of this section).

All entries of the mutation matrix Q can be derived from three parameters, the mutation rate p, the Hamming distance d_H (X_i, X_j), and the sequence length n, provided the uniform mutation rate model is adopted. This model is based on the assumption that mutation rates are independent of the position on the RNA sequence:

$$Q_{ij} = (1-p)^n \cdot \varepsilon^{d_H(X_i, X_j)} \text{ with } \varepsilon = \frac{p}{1-p} \tag{6}$$

The rate parameters f_i are derived from the mappings Ψ and Φ from sequence space into shape space and into real numbers as formulated in equation (1).

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

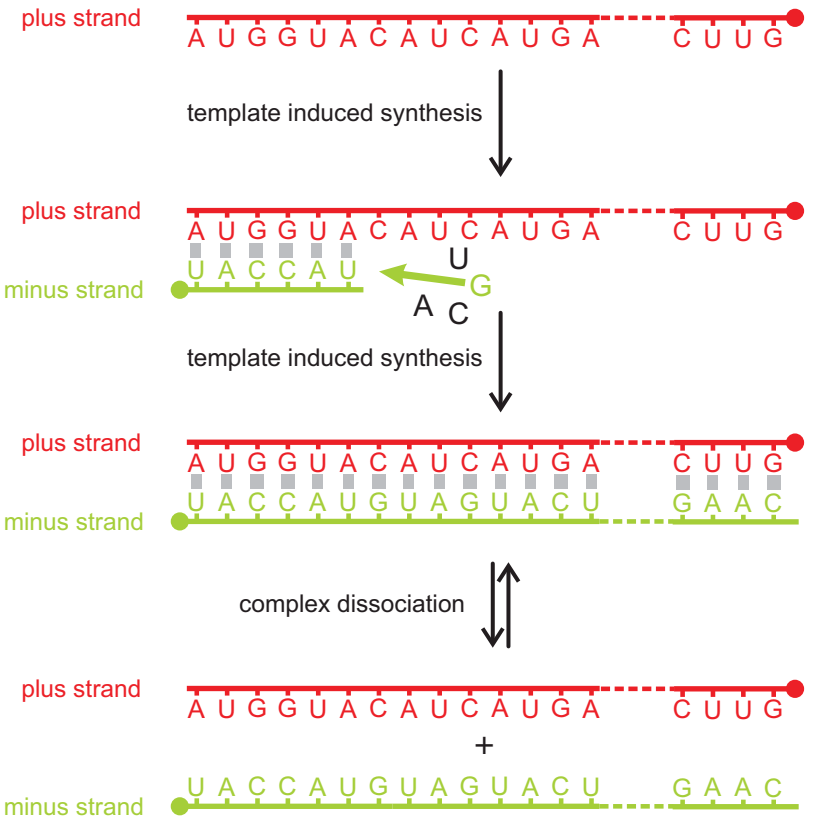


Figure 7. Complementary replication of RNA. Complementary replication consists of (i) duplex formation from single strands by template induced synthesis and (ii) dissociation of the duplex into a plus and a minus strand. The dissociation of the completed duplex is highly unfavorable because of the large negative free energy of duplex formation. Complex dissociation is facilitated by the enzyme, which separates the two strands on the fly in order to allow for independent structure formation and prevention of the formation of the complete duplex

In the absence of neutrality, the stationary distribution of sequences contains a master sequence, X_M, which is characterized in terms of the largest replication rate parameter: $f_M = \max\{f_i \forall i = 1, \ldots, N\}$. At sufficiently small mutation rates p, the stationary concentration of the master sequence is largest, $\bar{x}_M = \max\{\bar{x}_i \forall i = 1, \ldots, N\}$. A simple expression for stationary concentrations can be derived from the single peak model landscape. In this landscape, a higher replication parameter is assigned to the master and identical values to all others sequences: $f_M = \sigma_M \cdot f$ and $f_i = f$ for all $i \neq M$.[85–87] The (dimensionless) factor σ_m is called the superiority of the master. The assumption leading to the single peak landscape is in the spirit of mean field approximations, since all mutants

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

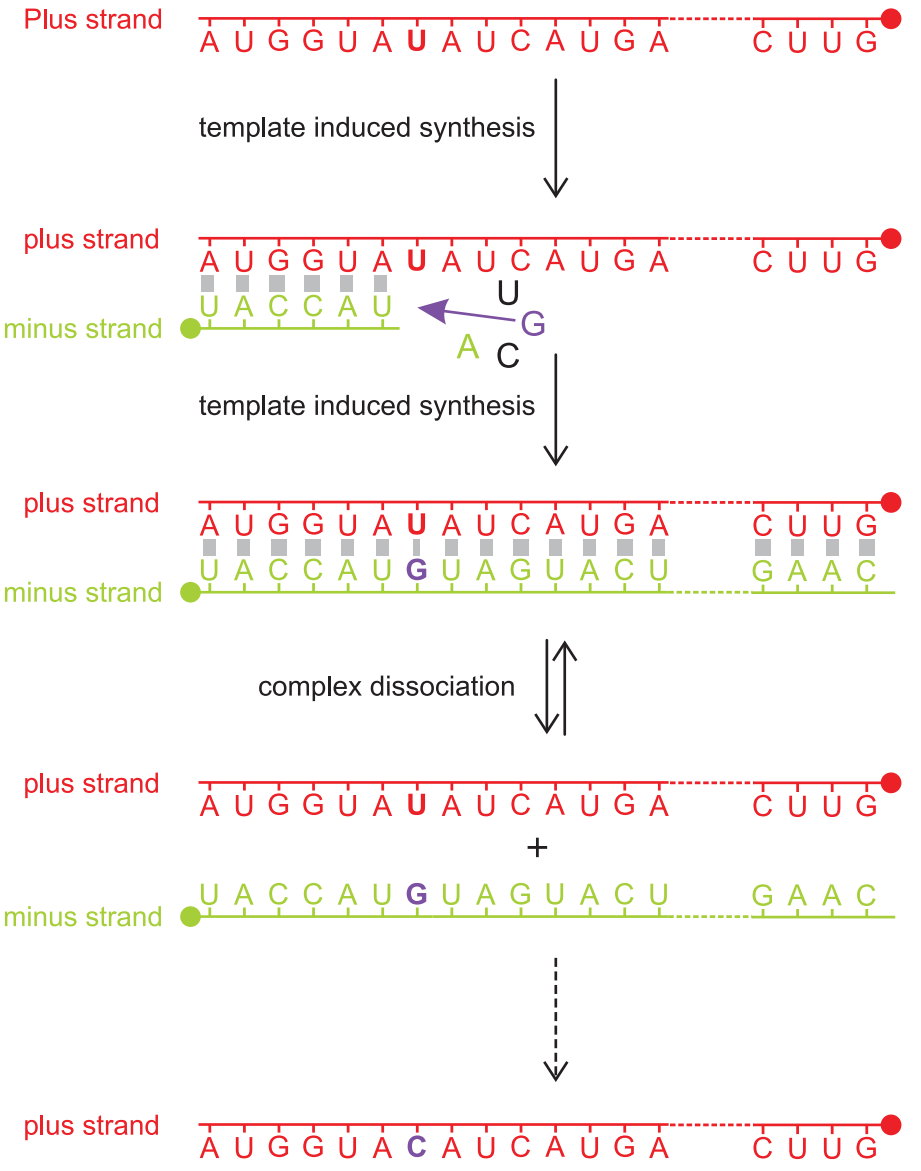


Figure 8. Point mutation in replication of RNA. Point mutation results from a mismatch in the replication duplex. The figure sketches the result of a U-G mismatch that leads to a point mutation of transition type: A→G and U→C

are lumped together into a single molecular species with average fitness. The concentration of the mutant cloud is simply $x_c = \sum_{j=1, j\neq \mathrm{M}}^{m} x_j = 1 - x_\mathrm{M}$ and the replication–mutation problem boils down to an exercise in a single variable, x_M, the frequency of the master. A mean-except-the-master replication rate parameter is defined $\bar{f} = \sum_{j\neq \mathrm{M}} f_j x_j/(1 - x_\mathrm{M})$ and then the superiority is of the

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

form: $\sigma_M = f_M/\bar{f}$. Neglecting mutational backflow we can readily compute the stationary frequency of the master sequence,

$$\bar{x}_M = \frac{f_M Q_{MM} - \bar{f}}{f_M - \bar{f}} = \frac{\sigma_M Q_{MM} - 1}{\sigma_M - 1} \tag{7}$$

Non-zero frequency of the master requires $Q_{MM} = \sigma_M^{-1} > Q_{min}$. Within the uniform error rate model an *error threshold*, defined by $\bar{x}_M = 0$ in the *no mutational backflow approximation*, occurs at a minimum single digit accuracy of

$$q_{min} = 1 - p_{max} = \sqrt[n]{Q_{min}} = \sigma_M^{-1/n} \text{ or } p_{max} = 1 - \sigma_M^{-1/n} \tag{8}$$

Figure 9 shows the stationary frequency of the master sequence, $\bar{x}_M$, as a function of the error rate. The exact solution of (5a) approaches the uniform distribution at mutation rates above error threshold. In other words, the concentrations of all molecular species in the population become identical. Such a state can never be achieved in real populations since population sizes N are always many orders of magnitude smaller than the numbers of sequences in sequence space – for a rather very large population size of $N = 10^{15}$ the chain length at which sequence space matches population size is about $n = 25$. Accordingly, the no mutational backflow approximation as well as the exact solution of the differential equation (5a) fail to describe replication–mutation dynamics at mutation rates above the error thresholds because of finite population size effects (see later). The error threshold phenomenon is used in virology for the design of new antiviral drugs.[14,88]

RNA secondary structures

An RNA secondary structure S of the sequence $X = (a_1, a_2, \ldots, a_n)$ where the nucleotides are chosen from an alphabet, e.g. $a_i \in \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$, is a planar graph with the nodes being the individual nucleotides a_i. The edges, $i \cdot j \in S$, are defined by the following criteria:

(i) for all nodes $i \leq (n-1)$ holds $i \cdot i+1 \in S$ (backbone),
(ii) for all nodes i exists maximal one $k \neq \{i+1, i-1\}$ such that $i \cdot k \in S$ (base pairs),
(iii) from $i \cdot j \in S$ and $k \cdot l \in S$ with $k < l$, and $i < k < j$ follows $i < k < l < j$ (no pseudo-knot rule), and
(iv) a criterion for structure formation, commonly minimization of free energy.

The backbone (i) represents the polynucleotide chain consisting of alternating phosphate and ribose moieties. The rule for base pairs (ii) defines all base pairs in structure S and excludes base triplets and other interaction involving more than two bases. The no pseudo-knot rule (iii) excludes structures shown in Figure 4.

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

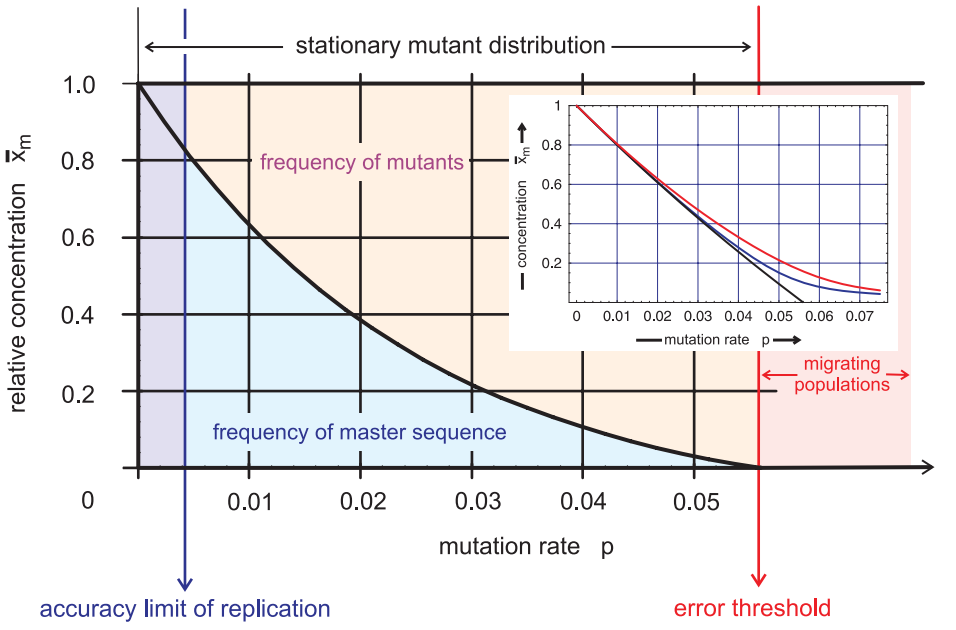


Figure 9. Error threshold in replication. The figure sketches the (relative) stationary concentration of the master sequence in the population as a function of the mutation rate $\bar{x}_M(p)$. It vanishes at the error threshold in the 'no mutational backflow' approximation. The insert shows curves obtained as the exact solution derived from the largest eigenvector of the matrix W (red), by an approximation based on equal concentrations of all mutants that corresponds to the population at mutation rates $p > p_{max}$ and becomes exact at $p = 0.5$ (blue), and by the no mutational backflow approximation (equation (7), black). The red curve and the blue curve approach each other above the error threshold and converge to the uniform distribution. The deterministic equation (5a) and its approximations fail to describe population dynamics at mutation rates above threshold. In addition, all replication processes in reality are bound by a minimum error rate, p_{min}, that represents the physical accuracy limit of replication

The m.f.e. condition, finally, defines the conditions under which S is as a possible structure of X. Thanks to these criteria the search for RNA secondary structures can be performed by means of dynamic programming.[89–93] Introducing the search for pseudo-knots into the search for optimal structures is possible in principle, but raises the computational demands enormously.[94] The situation with other tertiary motifs is similar. The currently used approach to predict tertiary structures starts from secondary structures and introduces tertiary contacts where sequence and structure make it possible.

Secondary structures can be represented as strings consisting of dots, and left and right parenthesis related by mathematical convention (Figure 3) without losing information. This fact provides an upper bound for the number of possible

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

secondary structures: $N_s(n) << 3^n$ since acceptable mathematical parentheses notation is a severe restriction. Application of combinatorics yields a remarkably good approximation for sufficiently long sequences:[78,95]

$$N_S(n) \approx 1.4848 \times n^{-3/2}(1.84892)^n$$

Accordingly, the number of sequences, $N = 4^n$, is always larger – commonly much larger – than the number of secondary structures and we are dealing therefore with neutrality.

Folding RNA sequences into conventional secondary structures with minimal free energies provides a suitable model system for studying realistic sequence-structure maps of biopolymers for several reasons: (i) almost all RNA sequences form some base pairs and structures are found everywhere in sequence space, (ii) RNA folding follows a simple base pairing logic and hence it is accessible by mathematics and computation, and (iii) RNA secondary structures are physically meaningful and provide a basis for discussing RNA function. These three properties that are not fulfilled in the case of proteins, and the capability of multiplication in simple replication assays make RNA a suitable model for studies of evolution *in vitro* and *in silico*.

Neutrality and its consequences

The mappings defined in equation (1) provide the theoretical basis for both, rational and evolutionary design of biomolecules. Since we are dealing with orders of magnitude, more sequences than structures and a multitude of structures serving the same task, both mappings Ψ and $\mathbf{\Phi}$ are non-invertible in the sense that many sequences form the same m.f.e. structure and many different structures may have the same function. The mapping Ψ is sketched in Figure 10. The inversion of the mapping $S = \Psi(X)$ generally results in a set of sequences $G(S)$ defining the pre-image of structure S in sequence space:

$$G(S) = \mathbf{\Psi}^{-1}(S) \doteq \{X | \mathbf{\Psi}(X) = S\} \tag{9}$$

It is a subset of the compatible set of structure S:[96] $G(S) \subseteq C(S)$. Since every sequence X_k maps onto some structure S_k, the union of all neutral sets covers the entire sequence space: $\bigcup_k G(S_k) = \mathcal{Q}$.

Global properties of neutral networks can be derived from random graph theory.[97] The characteristic quantity for a neutral network is the degree of neutrality $\bar{\lambda}$, which is obtained by averaging the fraction of Hamming distance-one neighbors that fold into the same m.f.e. structure, $\lambda_X = n_{\text{ntr}}^{(1)}/(n \cdot (\kappa - 1))$ – with $n_{\text{ntr}}^{(1)}$ being the number of neutral single nucleotide exchange neighbors – over the whole network, $G(S)$:

$$\bar{\lambda}(S) = \frac{1}{|G(S)|} \sum_{X \in G(S)} \lambda_X \tag{10}$$

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

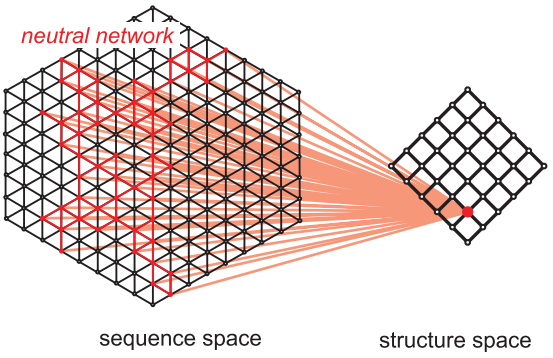


Figure 10. Neutral networks and compatible sequences. The set of sequences folding into the same m.f.e. structure S is denoted by $G(S)$. It defines the nodes of the neutral network of structure S in sequence space. Connecting all pairs of sequences with Hamming distance $d_H = 1$ yields the neutral network $\mathbf{\Gamma}(S)$ (the graph drawn in red). A neutral network is embedded in the set of compatible sequences $C(S)$, $G(S) \subseteq C(S)$. A compatible sequence of structure S, $X_C(S)$, forms S either as its m.f.e. structure or as one of its suboptimal conformations

where $|G(S)|$ is the number of sequences forming the neutral network. Connectedness of neutral networks, among other properties, is determined by the degree of neutrality:[98]

$$\text{With probability one a network is} \begin{cases} \text{connected} & \text{if } \bar{\lambda} \geq \lambda_{\text{cr}} \\ \text{not connected} & \text{if } \bar{\lambda} < \lambda_{\text{cr}} \end{cases} \tag{11}$$

where $\lambda_{\text{cr}} = 1 - \kappa^{-\frac{1}{\kappa-1}}$. Interestingly, this threshold value depends exclusively on the number of digits in the nucleotide alphabet. Calculation yields the critical values $\lambda_{\text{cr}} = 0.5$, 0.423, and 0.370 for two, three, and four letter alphabets, respectively. Random graph theory predicts a single largest component for non-connected networks, i.e. networks below threshold, which is commonly called the *giant component*.

Real neutral networks derived from RNA secondary structures sometimes deviate significantly from the prediction of random graph theory. In particular, they can have two or four equally sized largest components. This deviation is readily explained by non-uniform distribution of the sequences belonging to $G(S_k)$ over sequence space, which is caused by specific properties of the structure S_k.[99,100] For example, structures that allow for closure of additional base pairs at the ends of stacks are more likely to be formed by sequences that have an excess of one of the two bases forming a base pair than by those with an ideally balanced distribution ($n_{\text{G}} = n_{\text{C}}$ and $n_{\text{A}} = n_{\text{U}}$). For **GC** sequences the neutral network of such a structure is less dense in the middle part of sequence space

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

($n_G = n_C$) than above ($n_G > n_C$) or below ($n_G < n_C$), and we find two equally sized largest components, one at excess **G** and one at excess **C**.

Neutrality in sequence space has consequences for the selection process. The scenario of neutral evolution has been investigated in great detail by Motoo Kimura.[101,102] In the absence of differences in fitness values the distribution of neutral genotypes or sequences drifts randomly in sequence space until one particular genotype becomes fixed. Kimura's theory yields two highly relevant results: (i) the average time of replacement of one genotype by another is the reciprocal rate of mutation, $\tau_{\text{subst}} = 1/p$, and hence independent of population size, and (ii) the time of fixation of a mutant is proportional to the population size, $\tau_{\text{fix}} = 4N_e$ (with N_e being the effective population size). Neutrality can be introduced into model fitness landscapes, the corresponding selection-mutation equation (5a) is solved straightforwardly, and yields at the limit of small mutation rates for two sequences depending on the Hamming distance:[103]

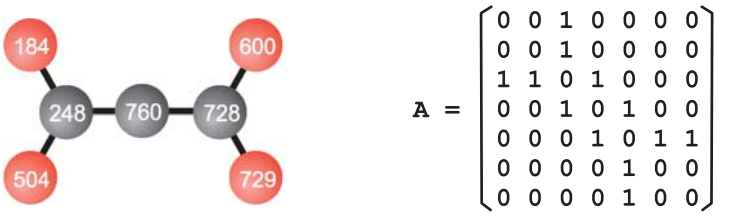
$$d_H(X_i, X_j)\begin{cases} = 1: & \lim_{p\to 0} x_i = 0.5 \text{ and } \lim_{p\to 0} x_j = 0.5 \\ = 2: & \lim_{p\to 0} x_i = \alpha \text{ and } \lim_{p\to 0} x_j = 1-\alpha \\ \geq 3: & \begin{cases} \lim_{p\to 0} x_i = 0 \text{ and } \lim_{p\to 0} x_j = 1 \text{ or} \\ \lim_{p\to 0} x_i = 1 \text{ and } \lim_{p\to 0} x_j = 0 \end{cases} \end{cases} \tag{12}$$

A pair of fittest neutral nearest neighbor sequences appears in the stationary mutant distribution strongly coupled at equal concentrations; two sequences, X_i and X_j, with Hamming distance $d_H(X_i, X_j)=2$ form a strongly coupled pair with a concentration ratio $\alpha/(1-\alpha)$; and for Hamming distance $d_H(X_i, X_j) \geq 3$ the Kimura scenario holds: either of the two sequences is selected depending on initial conditions (and/or random fluctuations). The group of two or more neutral sequences that is selected is called the *core* of the quasi-species and replaces the master sequence of the non-neutral case. For more than two neutral nearest neighbor sequences the core of the quasi-species is derived straightforwardly: we consider the selection-mutation matrix W and neglect all terms $O(\varepsilon^2)$. Without changing the eigenvectors of W we set $f = f(1-p)^n$ and $\varepsilon = 11$, and obtain the adjacency matrix A. The core is then computed as the largest eigenvector of A. An example is shown in Figure 11. Increasing mutation rates $p > 0$ lead to small or moderate changes in the relative concentrations of sequences in the core, in fortunate cases ratios of concentrations hold almost up to the error threshold.

Stochastic effects in RNA evolution

Stochasticity becomes important when particle numbers are small and this is certainly the case for rare mutations in evolution. For RNA molecules the number of possible single-point mutations is $3n$, and this increases like binomial coefficients with the Hamming distance. A related source of stochastic effects

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press



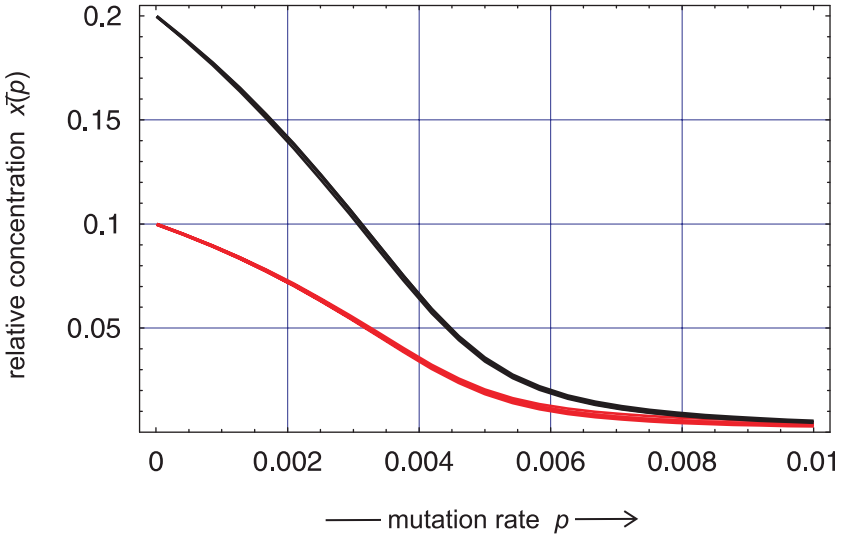


Figure 11. Neutral networks and quasi-species. An example of a quasi-species core for a degree of neutrality $\lambda = 0.1$. Fitness values f_i were assigned randomly to all 1024 binary (GC) sequences of chain length $n = 10$ with the constraint of 10% having the highest fitness value. The numbers on the sequences represent the decimal equivalent of the binary sequence, e.g. the two sequences $X_{184} \equiv$ CCGCGGGCCC and $X_{248} \equiv$ CCGGGGGCCC with Hamming distance $d_H(X_{184}, X_{248}) = 1$. The selected neutral network (upper part, LHS) comprises seven sequences. The relative concentrations in the limit of vanishing mutation rates, $\lim p \to 0$, are given by the largest eigenvector of the adjacency matrix A (upper part, RHS): $e_0 = (\bar{x}_{184}, \bar{x}_{248}, \bar{x}_{504}, \bar{x}_{600}, \bar{x}_{728}, \bar{x}_{729}, \bar{x}_{760}) = (0.1, 0.2, 0.1, 0.1, 0.2, 0.1, 0.2)$. As the computed curves $\bar{x}_i(p)$ show, the ratio of the individual stationary in the limit is also a good approximation for finite mutation rates almost up to the error threshold

concerns the smallness of all real populations compared to sequence space: in molecular evolution experiments, the numbers of RNA molecules in an experiment can hardly exceed 10^{15}, which is practically nothing compared to 1.6×10^{60}, the number of sequences in $\mathcal{Q}_{100}^{(4)}$ and therefore quasi-species are always truncated at a certain distance from the population center.

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

Therefore, stochastic effects are particularly important in molecular evolution under several conditions:

(i) In the regime of sufficiently accurate replication the master sequence or the core of a quasi-species is surrounded by a cloud of mutants. Near the truncation distance from the population center mutations become very rare and the mutants cannot reach stationarity but remain fluctuating elements.
(ii) At mutation rates above threshold mutations to distant sequences gain sufficiently high probability to destroy inheritance and all mutants become equally frequent in the deterministic approach. Since the population cannot cover the whole sequence space, it spreads and starts to migrate through the sequence space.
(iii) Populations on neutral networks drift in the sense of Kimura's neutral evolution. In particular, the population spreads and breaks up into different clones, which migrate through sequence space.

Scenarios (ii) and (iii) are similar but arise from two completely different origins: Scenario (ii) results from low accuracy that manifests itself in the elements of the Q-matrix and gives rise to migration of the population because of frequent mutations. The error threshold has been interpreted also as a localization threshold of the quasi-species in sequence space.[104] Scenario (iii) is tantamount to random drift in sequence space because of a degeneracy of the largest entries of matrix F.[105]

In order to simulate selection-mutation dynamics of RNA at the stochastic level, a realistic model based on chemical reactions in a flow reactor was conceived.[106–108] The sequence-structure map is an integral part of this model in the sense that sequences are converted into m.f.e. secondary structures by means of an RNA folding mechanism. Structures are evaluated to yield replication rate parameters or fitness values f_i. The simulation tool starts from a population of RNA molecules and simulates chemical reactions corresponding to replication and mutation in a continuously stirred flow reactor (CSTR) by using Gillespie's algorithm.[109–111] In target search the replication rate parameter of a sequence X_i, f_i, is chosen to be a function of the distance between the m.f.e. structure formed by the sequence, $S_i = f(X_i)$ and the target structure S_T,[112]

$$f_i(S_i, S_T) = \frac{1}{\alpha + d_H(S_i, S_T)/n} \tag{13}$$

which increases when S_i approaches the target structure S_T (α is an adjustable parameter that was chosen to be 0.1). A trajectory is completed when the population reaches a sequence that folds into the target structure. Accordingly, the simulated stochastic process has two absorbing barriers, the target and the

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

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 Figure 12. In this simulation, a homogeneous population consisting of N molecules with the same randomly chosen sequence is applied as an initial condition. The target structure is the well-known secondary structure of phenylalanyl-transfer RNA ($tRNA^{phe}$, see Figure 5). The distance to target averaged over the entire population decreases stepwise until the target is reached.[78,107,108] The process occurs on two timescales: Short adaptive phases are interrupted by long quasi-stationary epochs. Transitions between two structures S_i and S_j can be classified according to the nearness of their neutral networks $G(S_i)$ and $G(S_j)$.[113,114] Inspection of the sequence record during a quasi-stationary epoch on a given plateaus provides hints for the distinction of two scenarios:

(i) The structure is constant because of neutrality in the map Ψ and we observe neutral evolution. In particular, the number of neutral mutations accumulated is proportional to the number of replications on the population level. Evolution is a random walk of the population on a neutral network.
(ii) The process during the stationary epoch involves several structures with the same replication rate parameters. Because of neutrality in the map from structure to function, $\mathbf{\Phi}$, the population performs a kind of random walk in the space of neutral structures.

The random walk or the diffusion of the population on neutral networks is illustrated by the plot in the middle of Figure 12 showing the width of the population as a function of time.[78] The population width increases during the quasi-stationary epoch and sharpens almost instantaneously after mutation has created a sequence that allows for the start of a new adaptive phase. The scenario at the end of the plateau corresponds to a 'bottleneck' of evolution. The bottom part of the figure shows a plot of the migration rate or drift of the population center in sequence space and confirms this interpretation: the drift is almost always negligibly slow unless the population center jumps from one point in sequence space to another point in sequence space where the molecule initiating the new adaptive phase is located. A closer look at the figure reveals the coincidence of three events: (i) collapse-like narrowing of the population width, (ii) jump-like migration of the population center, and (iii) beginning of a new adaptive phase.

In Table 1 numerical data obtained from sampling evolutionary trajectories under identical conditions[115] are presented. The individual trajectories show enormous scatter in the time or the number of replications required to reach the target. Mean values and the standard deviations were obtained from statistics of

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

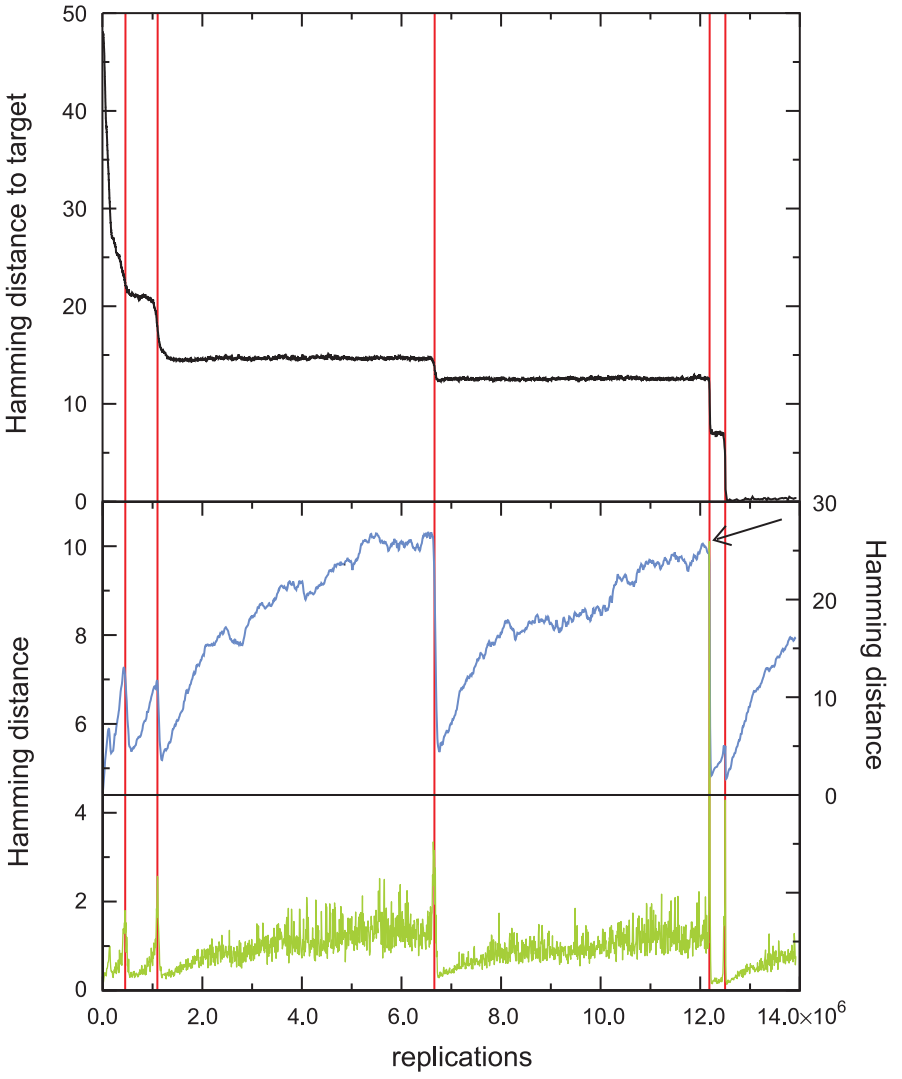


Figure 12. A trajectory of evolutionary optimization. The topmost plot presents the mean distance to the target structure of a population of 1000 molecules. The plot in the middle shows the width of the population in Hamming distance between sequences and the plot at the bottom is a measure of the velocity with which the center of the population migrates through sequence space. Diffusion on neutral networks causes spreading on the population in the sense of neutral evolution.[105] A remarkable synchronization is observed: at the end of each quasi-stationary plateau a new adaptive phase in the approach towards the target is initiated, which is accompanied by a drastic reduction in the population width and a jump in the population center (the top of the peak at the end of the second long plateau is marked by a black arrow). A mutation rate of $p = 0.001$ was chosen, the replication rate parameter is defined in equation (13), and initial as well as target structure are shown Table 1

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

Table 1. Statistics of the optimization trajectories. The table shows the results of sampled evolutionary trajectories leading from a random initial structure S_I to the structure of $tRNA^{phe}$, S_T as target.* Simulations were performed with an algorithm introduced by Gillespie.[109,110,125] The time unit is here undefined. A mutation rate of $p = 0.001$ per site and replication was used. The mean and standard deviation were calculated under the assumption if a log-normal distribution that fits well the data of the simulations

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

*The following structures S_I and S_T were used in the optimization:

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

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

trajectories under the assumption of a log-normal distribution. Despite the scatter three features are unambiguously detectable:

(i) The search in GC sequence space takes about five time as long as the corresponding process in AUGC sequence space in agreement with the difference in neutral network structure.
(ii) The time from initial conditions to target decreases with increasing population size.
(iii) The number of replications required to reach the target from initial conditions increases with population size.

Combining items (ii) and (iii) allows for a clear conclusion concerning requirements in time and resources of the optimization process: fast optimization requires large populations whereas economic use of material and/or energy suggests working with small population sizes just large enough to avoid extinction.

Systematic studies on the parameter dependence of RNA evolution were reported in a recent simulation.[116] Increase in mutation rate leads to an error threshold phenomenon that is closely to one observed with quasi-species on a single-peak landscape as described above.[77] Evolutionary optimization becomes more efficient[117] with increasing error rate until the error threshold is reached. A further increase in the error rate leads to an abrupt breakdown of the success in optimization. As expected, the distribution of replication rates or fitness values f_i in sequence space is also highly relevant: a steep decrease of fitness with the distance from the fittest master sequence (forming the target structure) leads to the sharp error threshold behavior as observed with single-peak landscapes, whereas flat landscapes show a broad maximum of optimization efficiency without an indication of threshold-like behavior.

Beyond the one sequence–one structure paradigm

So far it has been assumed implicitly that every RNA sequence gives rise to one unique structure. This is almost always true when the notion of structure is restricted to a well defined thermodynamic or process determined folding criterion, m.f.e. or *in situ* folding during RNA synthesis. In general, the number of structures S_k that are compatible with a given sequence X are commonly quite large and form the set of compatible structures $C(X)$, which consists of the m.f.e. structure together with all suboptimal structures. Efficient algorithms for the computation of suboptimal structures are available.[118,119] Because the numbers of suboptimal structures are almost always too large to be computed, stored and retrieved, the computational procedures use restrictions: in Ref. 118, certain common but less important classes of structures are neglected, and in Ref. 119 all structures are computed that lie within an predefined energy band above the m.f.e. (Figure 13). Alternatively, using the partition function of the states S_k,

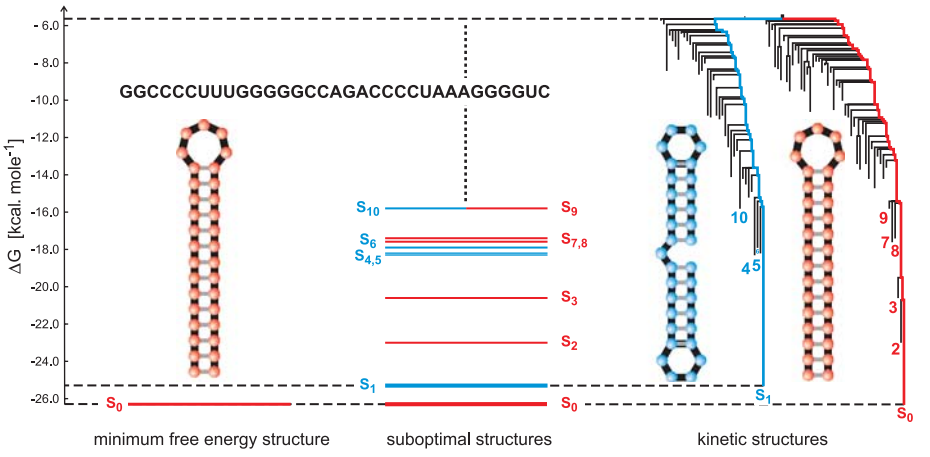


Figure 13. RNA structures. The m.f.e. structure of an RNA sequence is accompanied by a large number of suboptimal structures. The sequence GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC folds into a single hairpin structure S_0 with m.f.e. of -26.3 kcal/mole. The first suboptimal structure of this molecule, S_1, is a double hairpin with a free energy of -25.3 kcal/mole. The figure shows the m.f.e. structure (LHS; red), the spectrum of suboptimal structures (middle; suboptimal conformations related to S_0 are shown in red, those related to S_1 in blue), and the barrier tree of the sequence (RHS) with two major basins for S_1 (blue) and S_0 (red)

the superposition of all Boltzmann weighted structures can be calculated with little more computational effort than needed for the computation of the m.f.e. structure.[91,120] Yes-or-no pairing between two nucleotides is then replaced by a base pairing probability.

Rules defining nearest neighbors in shape space and a measure of distance between structures are required for the construction of a free energy surface that identifies the (meta)stable conformations as local minima and the transitions states for conformational changes as saddle points. Such rules form the move set of allowed elementary transitions between structures and represent individual steps in models for folding kinetics. An acceptable move set guarantees that every structure can be reached from every structure in shape space by a sequence of moves.[121] Opening and closing of single base pairs forms a move set fulfilling the condition. Empirical evidence suggests also including a shift move that can be understood as a specific combination of base pair opening and base pair closing into one move:

$$\cdot\cdot(((\cdot\cdot\cdot\cdot\cdot\cdot)))\cdot \rightarrow \cdot\cdot((((\cdot\cdot\cdot\cdot))))\cdot \quad \text{base pair closure,}$$

$$\cdot\cdot((((\cdot\cdot\cdot\cdot))))\cdot \rightarrow \cdot\cdot(((\cdot\cdot\cdot\cdot\cdot\cdot)))\cdot \quad \text{base pair opening, and}$$

$$((\cdot\cdot(((\cdot\cdot\cdot)))))\cdot \rightarrow (((\cdot\cdot((\cdot\cdot\cdot)))))\cdot \quad \text{base pair shift.}$$

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

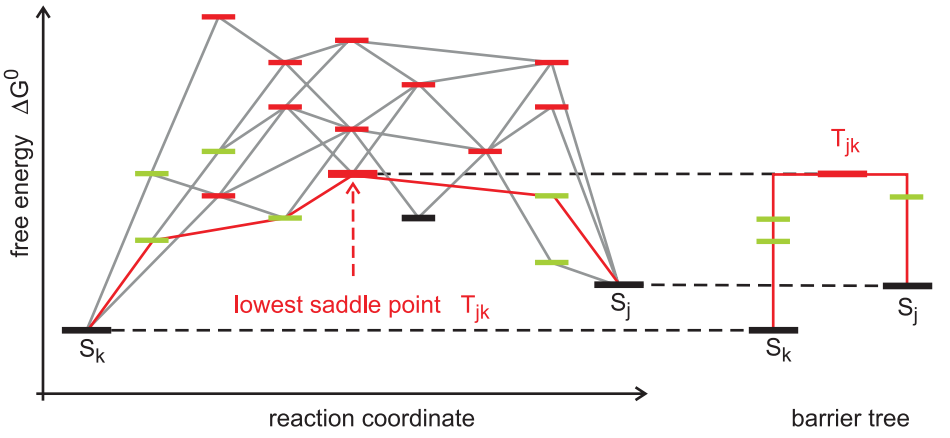


Figure 14. Conformation space and barrier tree. RNA secondary structures formed by one sequence fall into three classes: (i) local minima of the energy surface (black) are surrounded exclusively by suboptimal structures with higher free energies; (ii) saddle points (red) have two (or more) nearest neighbors in shape space that belong to two distinct basins; and (iii) (fully) unstable structures that are neither local minima or saddle points (green). The reaction coordinate is a path in shape space, which leads from one local minimum (conformation S_k) to another local minimum (conformation S_j). The barrier tree[71,126] is constructed by discarding all structures except local minima of the free energy surface and the lowest saddle points connecting them (an example is shown in Figure 13)

The move set defines the nearest neighbors of a given structure and allows for classification (Figure 14): a structure that is surrounded by structures of higher free energy represents a local minimum of the free energy surface and corresponds to a (meta)stable conformation. The conformation S_k corresponding to a local minimum of the free energy surface has a uniquely defined basin of attraction that is defined by the set of all structures from which downhill walks end uniquely in S_k. In addition to local minima, the saddle points of free energy surfaces are required for folding kinetics. A saddle point is defined by a locally lowest point in shape space, which has (two or more) nearest neighbors in shape space that belong to two distinct basins of attraction. All structures except those corresponding to local minima and saddle points are (fully) unstable structures.[122] It is straightforward to show that the inclusion of the shift move may change the nature of structures: some local minima are turned into unstable states.

The barrier tree is a coarse-grained simplification of the free energy surface of an RNA molecule. It discards all (fully) unstable structures and retains only (meta)stable conformations and saddle points. The barrier tree, nevertheless, allows for an identification of the basins of attraction (see the example shown in Figure 13). Small basins of attraction can be united to form larger ones until we

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

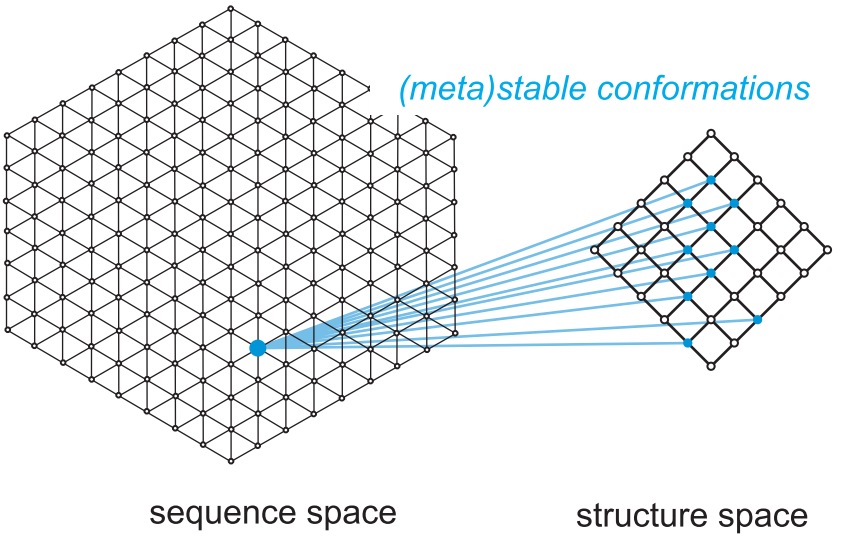


Figure 15. Suboptimal and compatible structures. Metastable conformations $S_k(X)$ of sequence X are defined by two conditions: (i) $\Delta G < 0$ for folding and (ii) conformation $S_k(X)$ is a local minimum of the free energy surface. These conformations form the set $G(X)$ in shape space. This set is embedded in the set of all structures that are compatible with sequence X, $G(X) \subseteq C(X)$. This compatible set $C(X)$ contains all structures of shape space that are compatible with sequence X. For the consideration of kinetic folding it is useful to include the set of saddle point structures $\tilde{G}(X)$ in the set of metastable structures forming thereby the set of structures of sequence X that is needed for the construction of barrier trees: $G(X) = G(X) \oplus \tilde{G}(X) \subseteq C(X)$

end up with a few major conformations each defining a large basin, and this procedure can be continued until only very few basins are retained or a single conformation remains. RNA molecules with several dominant basins of attraction corresponding to two or more (meta)stable conformations are called riboswitches, can be designed *in silico*[123] and occur also *in vivo*.[124] Conformational changes in natural riboswitches are commonly triggered by binding of small molecules and have regulatory function in metabolism. The barrier tree has also been used to compute Arrhenius-type folding kinetics of RNA molecules. The results are in good agreement with the exact computations of the folding kinetics on the computed conformational energy landscape unless there are many transition states whose energies lie close by.[72]

Finally, RNA suboptimal structures can also be considered in the context of sequence-structure mappings.[79] The set of structures that are compatible with a given sequence, $C(X)$ considered in Figure 15, is in a way *inverse* to the set of compatible sequences ($C(S)$ shown in Figure 10) since it deals with a non-invertible mapping in the opposite direction, from shape space into sequence

https://doi.org/10.1017/S1062798709000787 Published online by Cambridge University Press

space. A subset of the compatible structures, $G(X) \subseteq C(X)$, which contains all local minima of the free energy surface and the saddle points connecting the basins corresponding to (meta)stable conformations, provides the basis for the construction of barrier trees. All structures that are neither local minima nor saddle points are neglected. Local minima with positive free energies relative to the open chain, $\Delta G > 0$, and saddle points leading into their basins are also excluded. RNA evolution on neutral networks considered as a process with structure conservation and likewise kinetic RNA folding in conformation space is a process with conservation of sequence.[79]

Conclusions and outlook

The current state of the art in computation and empirical determination of fitness landscapes for evolution does not allow for predictions, because the accessible data are still rudimentary. The most promising areas of application are evolutionary design of molecules *in vitro* and virus evolution, where genotype spaces are large but accessible through extensive data collection. The greatest challenge for the future, presumably, is the same as in computational systems biology: despite an enormous wealth of data, only a small fraction is comparable because most of the currently accessible information is widely scattered in the literature and has been measured under incomparable condition. Further progress in reliability and predictive power of models depends, among other things, on validation and standardization of data.

Mathematical and computational tools are nevertheless available and can be implemented and used as soon as reliable information on the structure of landscapes becomes available. Evolution can be formally described and properly modeled as a process in sequence space as kinetic folding is visualized in shape space. The RNA model serves as a kind of tool kit that provides fundamental insights into basic structures and dynamics, which will later also be encountered in the real world.

References and Notes

1. S. Wright (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: D. F. Jones (ed), *Proceedings of the Sixth International Congress on Genetics*, Vol. 1 (Ithaca, NY: Brooklyn Botanic Garden), pp. 356–366.
2. R. A. Fisher (1930) *The Genetical Theory of Natural Selection* (Oxford, UK: Oxford University Press).
3. R. A. Fisher (1941) Average excess and average effect of a gene substitution. *Ann. Eugenics*, **11**, 53–63.
4. S. Okasha (2008) Fisher's fundamental theorem of natural selection – A philosophical analysis. *Brit. J. Phil. Sci.*, **59**, 319–351.

5. M. Eigen (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, **58**(10), 465–523.
6. J. Maynard Smith (1970) Natural selection and the concept of protein space. *Nature*, **225**, 563–564.
7. R. W. Hamming (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.*, **29**, 147–160.
8. R. W. Hamming (1989) *Coding and Information Theory*, 2nd edn (Englewood Cliffs, NJ: Prentice Hall).
9. N. Lehmann (2005) Special issue on Experimental Evolution. *J. Mol. Evol.*, **61**(2).
10. G. F. Joyce (2004) Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem.*, **73**, 791–836.
11. S. Klussmann (ed) (2006) *The Aptamer Handbook. Functional Oligonucleotides and Their Applications* (Weinheim, Germany: Wiley-VCh Verlag).
12. S. Brakmann and K. Johnsson (2002) *Directed Molecular Evolution of Proteins or How to Improve Enzymes for Biocatalysis* (Weinheim, Germany: Wiley-VCH).
13. C. Jäckel, P. Kast and D. Hilvert (2008) Protein design by directed evolution. *Annu. Rev. Biophys.*, **37**, 153–173.
14. E. Domingo, C. Parrish and J. Holland (eds) (2008) *Origin and Evolution of Viruses*, 2nd edn (San Diego: Academic Press).
15. D. R. Mills, R. L. Peterson and S. Spiegelman (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. USA*, **58**, 217–224.
16. S. Spiegelman (1971) An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, **4**, 213–253.
17. S. J. Wrenn and P. B. Harbury (2007) Chemical evolution as a tool for molecular discovery. *Annu. Rev. Biochem.*, **76**, 331–349.
18. R. T. Raines (1998) Ribonuclease A. *Chem. Rev.*, **98**, 1045–1065.
19. G. R. Marshall, J. A. Feng and D. J. Kuster (2008) Back to the future: Ribonuclease A. *Biopolymers (Peptide Science)*, **90**, 259–277.
20. D. G. Smyth, W. H. Stein and S. Moore (1963) Sequence of amino acid residues in bovine pnacreatic ribonuclease – revisions and confirmations. *J. Biol. Chem.*, **238**, 227–234.
21. S. Moore and W. H. Stein (1973) Chemical structures of pancreatic ribonuclease and deoxyribonuclease. *Science*, **180**, 458–464.
22. G. Kartha, J. Bello and D. Harker (1967) Tertiary structure of ribonuclease. *Nature*, **213**, 862–865.
23. H. W. Wyckoff, K. D. Hardman, N. M. Allewell, T. Inagami, L. N. Johnson and F. M. Richards (1967) The structure of ribonuclease-S at 3.5 Å resolution. *J. Biol. Chem.*, **242**, 3984–3988.
24. H. W. Wyckoff, K. D. Hardman, N. M. Allewell, T. Inagami, D. Tsernoglou, L. N. Johnson and F. M. Richards (1967) The structure of ribonuclease-S at 6 Å resolution. *J. Biol. Chem.*, **242**, 3749–3753.
25. E. E. Kim, R. Varadarajan, H. W. Wykoff and F. M. Richards (1992) Refinement of the crystal structure of ribonuclease S. Comparison with

and between the various ribonuclease A structures. *Biochemistry*, **31**, 12304–12314.
26. F. H. White Jr. (1961) Regeneration of native secondary and tertiary structure by air oxidation of reduced ribonuclease. *J. Biol. Chem.*, **236**, 1353–1360.
27. C. B. Anfinsen and E. Haber (1961) Studies on the reduction and re-formation of protein disulfide bonds. *J. Biol. Chem.*, **236**, 1361–1363.
28. C. B. Anfinsen, E. Haber, M. Sela and F. H. White (1961) Studies on the reduction and re-formation of protein disulfide bonds. *Proc. Natl. Acad. Sci. USA*, **47**, 1309–1314.
29. C. B. Anfinsen (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
30. C. Park and R. T. Raines (2003) Ribonuclease A is limited by the rate of substrate association. *Biochemistry*, **42**, 3509–3518.
31. A. Elofson and G. von Heijne (2007) Membrane protein structure: prediction versus reality. *Annu. Rev. Biophys.*, **76**, 125–140.
32. J. N. Onuchic, Z. Luthey-Schulten and P. G. Wolynes (1997) Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.*, **48**, 545–600.
33. C. M. Dobson, A. Šali and M. Karplus (1998) Protein folding: a perspective from theory and experiment. *Angew. Chem. Int. Ed.*, **37**, 868–893.
34. In the case of model proteins on lattices but also for nucleic acid secondary structures discrete conformation spaces are appropriate.
35. M. Vendruscolo and C. M. Dobson (2005) Towards complete description of free-energy landscapes of proteins. *Phil Trans. Roy. Soc. A*, **363**, 433–452.
36. C. Levinthal (1968) Are there protein folding pathways? *J. Chim. Phys.*, **65**, 44–45.
37. C. Levinthal (1969) How to fold graciously. In: P. Debrunner, J. C. M. Tsibris and E. Münck, (eds) *Mössbauer Spectroscopy in Biological Systems* (Urbana, IL: University of Illinois Press), pp. 22–24.
38. A. L. Horwich, W. A. Fenton, E. Chapman and G. W. Farr (2007) Two families of chaperonin: physiology and mechanism. *Annu. Rev. Cell Dev. Biol.*, **23**, 115–145.
39. K. A. Dill, S. B. Ozkan, M. S. Shell and T. R. Weikl (2008) The protein folding problem. *Annu. Rev. Biophys.*, **37**, 289–316.
40. R. F. Service (2008) Problem solved* (*sort of). *Science*, **231**, 784–786.
41. Subunits of proteins are defined as independent polypeptide chains, in other words, each subunit of a protein is characterized by a separate chain.
42. K. Vamvaca, B. Vögeli, P. Kast, K. Pervushin and D. Hilvert (2004) An enzymatic molten globule: efficient coupling of folding and catalysis. *Proc. Natl. Acad. Sci. USA*, **101**, 12860–12864.
43. E. E. Lattman (2005) Sixth meeting on the critical assessment of techniques for protein structure prediction. *Proteins*, **61**(S7), 1–2.
44. T. L. Trapane and E. E. Lattman (2007) Seventh meeting on the critical assessment of techniques for protein structure prediction. *Proteins*, **69**(S8), 1–2.

45. R. R. Breaker (1999) Catalytic DNA: in training and seeking employment. *Nature Biotechnology*, **17**, 422–423.
46. M. T. McManus and P. A. Sharp (2002) Gene silencing in mammals by small interfering RNAs. *Nature Rev. Genetics*, **3**, 737–747.
47. M. J. Packer, M. P. Dauncey and C. A. Hunter (2000) Sequence-dependent DNA structure: Dinucleotide comformational maps. *J. Mol. Biol.*, **295**, 71–83.
48. M. J. Packer, M. P. Dauncey and C. A. Hunter (2000b) Sequence-dependent DNA structure: Tetranucleotide comformational maps. *J. Mol. Biol.*, **295**, 85–103.
49. E. J. Gardiner, C. A. Hunter, M. J. Packer, D. S. Palmer and P. Willett (2003) Sequence-dependent DNA structure: A database of octamer structural parameters. *J. Mol. Biol.*, **332**, 1025–1035.
50. C. J. Benham and S. P. Mielke (2005) DNA mechanics. *Annu. Rev. Biomed. Engineering*, **7**, 21–53.
51. A. Klug and D. Rhodes (1987) 'Zinc fingers': a novel protein motif for nucleic acid recognition. *Trends Biochem. Sci.*, **12**, 464–469.
52. A. Klug (1999) Zinc fingers peptides for the regulation of gene expression. *Trends Biochem. Sci.*, **293**, 215–218.
53. M. Wu and I. Tinoco Jr. (1998) RNA folding causes secondary structure rearrangement. *Proc. Natl. Acad. Sci. USA*, **95**, 11555–11560.
54. P. B. Moore (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **67**, 287–300.
55. D. K. Hendrix, S. E. Brenner and S. R. Holbrook (2005) RNA structural motifs: building blocks of a modular biomolecule. *Quart. Rev. Biophys.*, **38**, 221–243.
56. S. R. Holbrook (2008) Structural principles from large RNAs. *Annu. Rev. Biophys.*, **37**, 445–464.
57. V. P. Antao, S. Y. Lai and I. Tinoco Jr. (1991) A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.*, **19**, 5901–5905.
58. V. P. Antao and I. Tinoco Jr. (1992) Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic Acids Res.*, **20**, 819–824.
59. D. J. Klein, T. M. Schmeing, P. B. Moore and T. A. Steitz (2001) The kink-turn: A new RNA secondary structure motif. *EMBO J.*, **20**, 4214–4221.
60. N. Leontis and E. Westhof (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
61. N. B. Leontis and E. Westhof (2003) Analysis of RNA motifs. *Curr. Op. Struct. Biol.*, **13**, 300–308.
62. A. Lescoute, N. B. Leontis, C. Massire and E. Westhof (2005) Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Res.*, **33**, 2395–2409.
63. A. Lescoute and E. Westhof (2006) The interaction network of structured RNAs. *Nucleic Acids Res.*, **34**, 6587–6604.
64. N. B. Leontis, R. B. Altmann, H. M. B. Berman, S. E. Brenner, J. W. Brown, D. R. Engelke, S. C. Harvey, S. R. H. Fabrice Jossinet,

S. E. Lewis, F. Major, D. H. Mathews, J. S. Richardson, J. R. Williamson and E. Westhof (2006) The RNA ontology consortium: An open invitation to the RNA community. *RNA*, **12**, 533–541.
65. A. Rich and U. L. RajBhandary (1976) Transfer RNA: Molecular structure, sequence, and properties. *Annu. Rev. Biochem.*, **45**, 805–860.
66. E. I. Zagryadskaya, N. Kotlova and S. V. Steinberg (2004) Key elements in maintenance of the t-RNA L-shape. *J. Mol. Biol.*, **340**, 435–444.
67. S. Chen (2008) RNA-Folding: Conformational statistics, folding kinetics, and ion electrostatics. *Annu. Rev. Biophys.*, **37**, 197–234.
68. D. Pörschke and M. Eigen (1971) Co-operative non-enzymic base recognition. III Kinetics of the helix-coli transition of the oligoribouridylic-oligoriboadenylic acid systems and of oligoriboadenylic acid alone at acidic pH. *J. Mol. Biol.*, **62**, 361–381.
69. D. Pörschke (1974) Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix. *Biophys. Chem.*, **1**, 381–386.
70. D. Pörschke (1977) Elementary steps of base recognition and helix-coli transitions in nucleic acids. In: I. Pecht and R. Rigler (eds) *Chemical Relexation in Molecular Biology* (Springer Verlag), pp. 191–218.
71. C. Flamm, W. Fontana, I. L. Hofacker and P. Schuster (1999) Elementary step dynamics of RNA folding. *RNA*, **6**, 325–338.
72. M. T. Wolfinger, W. A. Svrcek-Seiler, C. Flamm, I. L. Hofacker and P. F. Stadler (2004) Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.*, **37**, 4731–4741.
73. A. Borgia, P. M. Williams and J. Clarke (2008) Single-molecular studies of protein folding. *Annu. Rev. Biophys.*, **77**, 101–125.
74. P. T. X. Li, J. Vieregg and I. Tinoco Jr. (2008) How RNA unfolds and refolds. *Annu. Rev. Biochem.*, **77**, 77–100.
75. M. Eigen and P. Schuster (1977) The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*, **64**, 541–565.
76. M. Eigen and P. Schuster (1978) The hypercycle. A principle of natural self-organization. Part B: The abstract hypercycle. *Naturwissenschaften*, **65**, 7–41.
77. M. Eigen, J. McCaskill and P. Schuster (1989) The molecular quasispecies. *Adv. Chem. Phys.*, **75**, 149–263.
78. P. Schuster (2003) Molecular insight into the evolution of phenotypes. In: J. P. Crutchfield and P. Schuster (eds) *Evolutionary Dynamics – Exploring the Interplay of Accident, Selection, Neutrality, and Function* (New York: Oxford University Press), pp. 163–215.
79. P. Schuster (2006) Prediction of RNA secondary structures: from theory to models and real molecules. *Reports on Progress in Physics*, **69**, 1419–1477.
80. A. Watts and G. Schwarz, (eds) (1997) *Evolutionary Biotechnology – From Theory to Experiment*, Vol. 66/2-3 *of Biophysical Chemistry* (Amsterdam: Elsevier), pp. 67–284.
81. C. K. Biebricher, M. Eigen and J. W. C. Gardiner (1983) Kinetics of RNA replication. *Biochemistry*, **22**, 2544–2559.

82. C. K. Biebricher, M. Eigen and J. W. C. Gardiner (1984) Kinetics of RNA replication: plus-minus asymmetry and double-strand formation. *Biochemistry*, **23**(14), 3186–3194.
83. C. K. Biebricher, M. Eigen and J. W. C. Gardiner (1985) Kinetics of RNA replication: competition and selection among self-replicating RNA species. *Biochemistry*, **24**(23), 6550–6560.
84. E. Seneta (1981) *Non-negative Matrices and Markov Chains*, 2nd edn (New York: Springer-Verlag).
85. J. Swetina and P. Schuster (1982) Self-replication with errors – a model for polynucleotide replication. *Biophys. Chem.*, **16**, 329–345.
86. P. Tarazona (1992) Error-thresholds for molecular quasi-species as phase transitions: From simple landscapes to spinglass models. *Phys. Rev. A [15]*, **45**, 6038–6050.
87. D. Alves and J. F. Fontanari (1996) Population genetics approach to the quasispecies model. *Phys. Rev. E*, **54**, 4048–4053.
88. E. Domingo (ed.) (2005) Virus entry into error catastrophe as a new antiviral strategy. *Virus Research*, **107**(2), 115–228.
89. M. Zuker and P. Stiegler (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
90. M. Zuker and D. Sankoff (1984) RNA secondary structures and their prediction. *Bull. of Math. Biol.*, **46**(4), 591–621.
91. I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker and P. Schuster (1994) Fast folding and comparison of RNA secondary structures. *Mh. Chemie*, **125**, 167–188.
92. I. L. Hofacker (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
93. R. Svobodová Vařeková, I. Bradáč, M. Plchút, M. Škrdla, M. Wacenovsky, M. H. Mahr, G. Mayer, H. Tanner, H. Brugger, J. Withalm, P. Lederer, H. Huber, G. Gierlinger, R. Graf, H. Tafer, I. Hofacker, P. Schuster and M. Polčík (2008) www.rnaworkbench.com: a new program for analyzing RNA interference. *Computer Methods and Programs in Biomedicine*, **90**, 89–94.
94. E. Rivas and S. R. Eddy (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
95. I. L. Hofacker, P. Schuster and P. F. Stadler (1998) Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, **89**, 177–207.
96. Compatibility of sequences and structures is defined in the following way: a sequence X is compatible with structure S if and only if for every base pair in S, the sequence X contains pairable nucleotides in the two positions forming the pair. Similarly a structure S is comparable with a sequence X when the same relation – and, obviously, not its inversion – holds.
97. B. Bollobás (1985) *Random Graphs* (London: Academic Press).
98. C. Reidys, P. F. Stadler and P. Schuster (1997) Generic properties of combinatory maps. Neutral networks of RNA secondary structure. *Bull. Math. Biol.*, **59**, 339–397.

99. W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker and P. Schuster (1996) Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering. *Mh. Chemie*, **127**, 375–389.
100. W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker and P. Schuster (1996) Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Mh. Chemie*, **127**, 355–374.
101. M. Kimura (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.
102. M. Kimura (1983) *The Neutral Theory of Molecular Evolution* (Cambridge, UK: Cambridge University Press).
103. P. Schuster and J. Swetina (1988) Stationary mutant distribution and evolutionary optimization. *Bull. Math. Biol.*, **50**, 635–660.
104. J. S. McCaskill (1984) A localization threshold for macromolecular quasispecies from continuously distributed replication rates. *J. Chem. Phys.*, **80**, 5194–5202.
105. M. A. Huynen, P. F. Stadler and W. Fontana (1996) Smoothness within ruggedness. The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, **93**, 397–401.
106. W. Fontana and P. Schuster (1987) A computer model of evolutionary optimization. *Biophys. Chem.*, **26**, 123–147.
107. W. Fontana, W. Schnabl and P. Schuster (1989) Physical aspects of evolutionary optimization and adaptation. *Phys. Rev. A*, **40**, 3301–3321.
108. W. Fontana and P. Schuster (1998) Continuity in evolution. On the nature of transitions. *Science*, **280**, 1451–1455.
109. D. T. Gillespie (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.*, **22**, 403–434.
110. D. T. Gillespie (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.
111. D. T. Gillespie (2007) Stochastic simulation of chemical kinetics. *Ann. Rev. Phys. Chem.*, **58**, 35–55.
112. The measure for the distance between two structures S_i and S_j applied here is the Hamming distance between the two parentheses representations: $d_H(S_i, S_j)$.
113. W. Fontana and P. Schuster (1998) Shaping space. The possible and the attainable in RNA genotype-phenotype mapping,. *J. Theor. Biol.*, **194**, 491–515.
114. B. R. M. Stadler, P. F. Stadler, G. P. Wagner and W. Fontana (2001) The topology of the possible: formal spaces underlying patterns of evolutionary change. *J. Theor. Biol.*, **213**, 241–274.
115. Identical means here that everything was kept constant except the seeds for the random number generators.
116. A. Kupczok and P. Dittrich (2006) Determinants of simulated RNA evolution. *J. Theor. Biol.*, **238**, 726–735.

117. Efficiency of evolutionary optimization is measured by average and best fitness values obtained in populations after a predefined number of generations.
118. M. Zuker (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
119. S. Wuchty, W. Fontana, I. L. Hofacker and P. Schuster (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
120. J. S. McCaskill (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
121. A move set in sequence space that fulfils this condition is point mutation.
122. A saddle point is also unstable at least in one direction but it is locally stable in at least one other direction.
123. C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler and M. Zehl (2001) Design of multi-stable RNA molecules. *RNA*, **7**, 254–265.
124. R. K. Montange and R. T. Batey (2008) Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.*, **37**, 117–133.
125. D. T. Gillespie (1977) Concerning the validity of the stochastic approach to chemical kinetics. *J. Stat. Phys.*, **16**, 311–318.
126. C. Flamm, I. Hofacker, P. Stadler and M. Wolfinger (2002) Barrier trees of degenerate landscapes. *Z. Phys. Chem.*, **216**, 155–173.