RNA *In Silico*
The Computational Biology of RNA Secondary Structures

Christoph Flamm
Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
xtof@tbi.univie.ac.at

Ivo L. Hofacker
Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
ivo@tbi.univie.ac.at

Peter F. Stadler
*Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria; and
Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA
studla@tbi.univie.ac.at or stadler@santafe.edu
*Address for correspondence

(Version Time-stamp: "1999-04-28 22:05:33 ivo")

ABSTRACT. RNA secondary structures provide a unique computer model for investigating the most important aspects of structural and evolutionary biology. The existence of efficient algorithms for solving the folding problem, i.e., for predicting the secondary structure given only the sequence, allows the construction of realistic computer simulations. The notion of a "landscape" underlies both the structure formation (folding) and the (*in vitro*) evolution of RNA.

Evolutionary adaptation may be seen as hill climbing process on a fitness landscape which is determined by the phenotype of the RNA molecule (within the model this is its secondary structure) and the selection constraints acting on the molecules. We find that a substantial fraction of point mutations do not change an RNA secondary structure. On the other hand, a comparable fraction of mutations leads to very different structures. This interplay of smoothness and ruggedness (or robustness and sensitivity) is a generic feature of both RNA and protein sequence-structure maps. Its consequences, "shape space covering" and "neutral networks" are inherited by the fitness landscapes and determine the dynamics of RNA evolution. Punctuated equilibria at phenotype level and a diffusion like evolution of the underlying genotypes are a characteristic feature of such models. As a practical application of these theoretical findings we have designed an algorithm that finds conserved (and therefore potentially functional) substructures of RNA virus genomes from sparse data sets.

The folding dynamics of particular RNA molecule can also be studied successfully based on secondary structures. Given an RNA sequence, we consider the energy landscape formed by all possible conformations (secondary structures). A straight forward implementation of the Metropolis algorithm is sufficient to produce a quite realistic folding kinetics, allowing to identify meta-stable states and folding pathways. Just as in the protein case there are good and bad folders which can be distinguished by the properties of their energy landscapes.

KEYWORDS: RNA Secondary Structures, Fitness Landscapes, Energy Landscapes, Molecular Evolution, Punctuated Equilibria, Folding Kinetics, Folding Pathways.

1. Introduction

The relationships between the sequence and the (three-dimensional) structure of a biopolymer is a core issue in biochemistry and molecular biology. While most of the research on biopolymer folding is concerned with protein folding, the same questions can be posed for RNA molecules (Draper, 1996). An important advantage of RNA is that, on the level of secondary structure, the structure prediction problem can be solved with reasonable accuracy. Based on this observation, it is possible to construct detailed computer models of different aspects of the sequence-structure-function relationships ranging all the way from *in vitro* evolution to folding kinetics.

RNA secondary structures provide a discrete, coarse grained concept of structure similar in complexity to lattice models of proteins. In contrast to the latter, RNA secondary structures are a faithful coarse graining of the 3D structures. It should be noted, however, that there are examples of RNA molecules with significantly different secondary structure which exhibit similar 3D structures and the same function (Uhlenbeck, 1998). Secondary structures are routinely used to display, organize, and interpret experimental findings, they are oftentimes conserved over evolutionary times scales, and *in vitro* selection experiments with RNA more often than not yield families of selected sequences that share distinctive secondary structure features.

In this contribution we (briefly) review three aspects of the "computational biology of RNA secondary structures": (1) the solutions to the folding problem and its variants, (2) the generic properties of the sequence structure relations and their implications for the dynamics of RNA evolution, (3) the properties of the conformational energy function and its implications for the kinetics of RNA folding. The notion of a *landscape* plays a key role in our investigations.

2. RNA Secondary Structures and Their Prediction

We begin our discussion with the formal definition of a secondary structure: A *secondary structure* on a sequence is a list of base pairs $[i,j]$ with $i<j$ such that for any two base pairs $[i,j]$ and $[k,l]$ with $i \le k$ holds:

(i) $i=k$ if and only if $j=l$, and
(ii) $k<j$ implies $i<k<l<j$.

The first condition simply means that each nucleotide can take part in at most one base pair. The second condition forbids knots and pseudo-knots. Secondary structures form a special type of graphs. In particular, a secondary structure graph is *outer-planar*, which means that it can be drawn in the plane in such a way that all vertices (which represent the nucleotides) are arranged on a circle, and all edges (which represent the bases pairs) lie inside the circle and do not intersect. While pseudo-knots are important in many natural RNAs (Westhof and Jaeger, 1992), they can be considered part of the tertiary structure for our purposes. The restriction to knot-free structures is necessary for efficient computation by dynamic programming algorithms. The recent algorithm by Rivas and Eddy (1999) is able to deal with a large class of pseudo-knotted structures, but is extremely costly. Moreover, the information about the energetics of pseudo-knots is still very limited (Gultyaev *et al.*, 1999).

Regarding secondary structures as special types of outer-planar graphs paves the way for a mathematical investigation of the structures. For instance, one can count the number of possible distinct secondary structures for a given chain

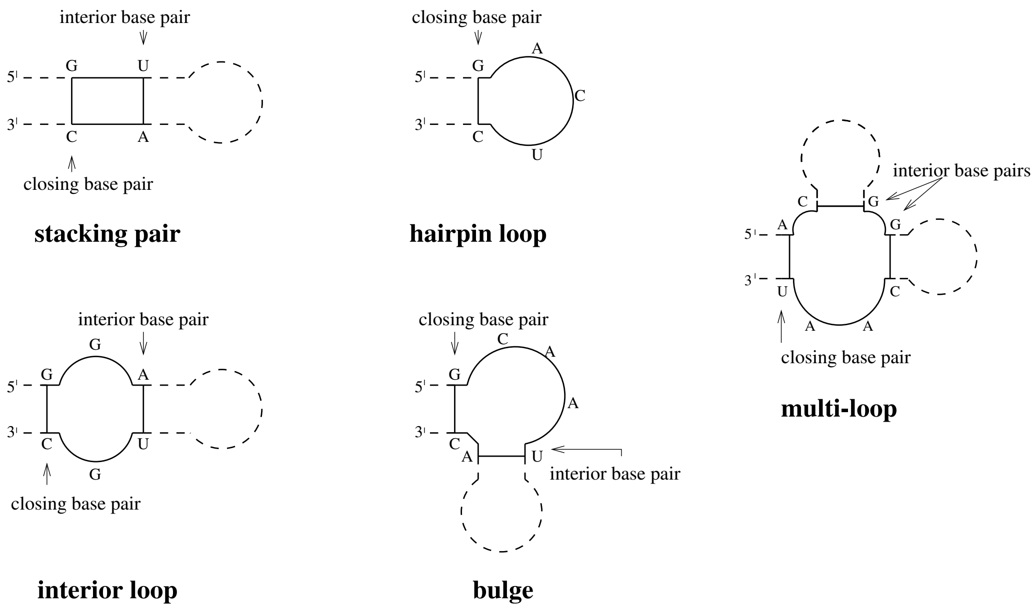

Figure 1. RNA secondary structure elements. Any secondary structure can be uniquely decomposed into these types of loops.

length n or the number of all secondary structures that can be formed by a particular sequence (Waterman, 1995; Hofacker *et al.*, 1998). For instance, the number of secondary structures with minimal stack length $s = 2$ grows like 1.86^n. If the most common types of pseudo-knots are included, there are about 2.35^n different structures (Haslinger and Stadler, 1999). Consequently, there are many more sequences, 4^n, than secondary structures. Before we explore the consequences of this observations, however, we consider the structure prediction problem.

Secondary structures can be uniquely decomposed into loops as shown in Figure 1 (note that a stacked base pair is considered a loop of size zero). The energy of an RNA secondary structure is assumed to be the sum of the energy contributions of all loops. Energy parameters for the contribution of individual loops have been determined experimentally, see, e.g., (Freier *et al.*, 1986; Jaeger *et al.*, 1989; Walter *et al.*, 1994) and depend on the loop type, loop size, and partly on its sequence. Usually, only Watson-Crick (**AU**, **UA**, **CG** and **GC**) and **GU** and **UG** pairs are allowed in computational approaches since non-standard base-pairs have in general context-dependent energy contributions that do not fit into the "nearest-neighbor model". It turns out that this standard energy model has a solid graph theoretical foundation (Leydold and Stadler, 1998): the loops form the unique minimal cycle basis of the secondary structure graph.

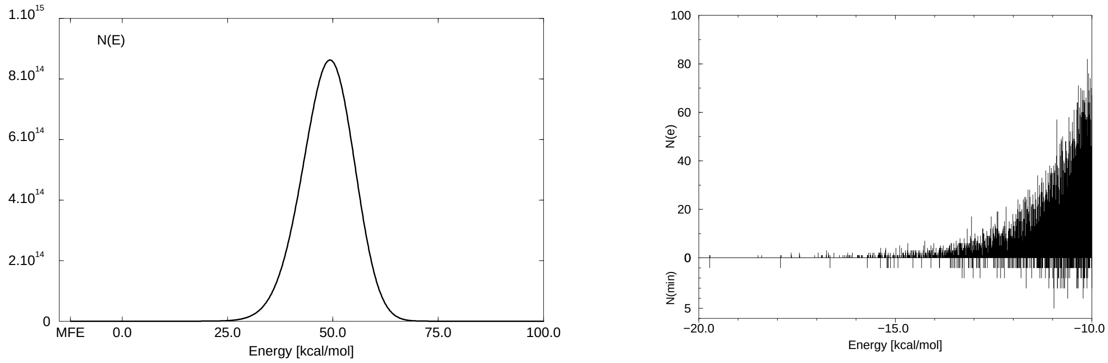

Figure 2. Density of states of the yeast tRNA$^{\rm phe}$. Top: Complete Density of States computed with an energy resolution of 0.1 kcal/mol, computed using the Density of state algorithm. The total number of structures is $14,995,224,405,213,184$. Less than 2 million structures have negative energy, the reference state being the open structure. The lower figure shows the density of states and the density of local minima in the region above the native state at higher resolution. For this plot all structures within 15kcal/mol the ground state were generated by suboptimal folding and tested for being local minima. The tRNA sequence with modified bases used here displays only a few suboptimal structures within a few kT above the native state.

The additive form of the energy model allows for an elegant solution of the minimum energy folding problem by means of a dynamic programming scheme that is similar to sequence alignment. This similarity was first realized and exploited by Waterman 1978, see also (Waterman and Smith, 1978), the first dynamic programming solution was proposed by Nussinov and Jacobson (1980), originally for the "maximum matching" problem of finding the structure with the maximum number of base pairs (Nussinov *et al.*, 1978). Zuker and coworkers (1981; 1984) formulated the algorithm for the minimum energy problem using the now standard energy model. Since then several variations have been developed: Michael Zuker (1989) devised a modified algorithm that can generate a subset of suboptimal structures within a prescribed increment of the minimum energy, see also (Schmitz and Steger, 1992). The algorithm will find any structure ψ that is optimal in the sense that there is no other structure ψ' with lower energy containing all base pairs that are present in ψ. John McCaskill (1990) noted that the partition function over all secondary structures $Q = \sum_\psi \exp(-\Delta G(\psi)/kT)$ can be calculated by dynamic programming as well. In addition his algorithm can calculate the frequency with which each base pair occurs in the Boltzmann weighted ensemble of all possible structures, which can be conveniently represented in a "dot-plot", see Figure 7.

The memory and CPU requirements of these algorithms scale with sequence length n as $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$, respectively, making structure prediction feasible even for large RNAs of about 10000 nucleotides, such as the genomes of RNA viruses (Hofacker *et al.*, 1996; Huynen *et al.*, 1996a).

McCaskill's work was extended in our group to yield an algorithm that computes the complete density of states of an RNA sequence at predefined energy resolution (Cupal *et al.*, 1996; Cupal, 1997). Another method for calculating the density of states, based on enumeration of structures, was proposed earlier by Higgs (1993). However, his algorithm is restricted to subset of structures containing no helices shorter than three and uses a simplified energy model. Still, our algorithm is rather demanding as it needs to store $\mathcal{O}(n^2m)$ entries and $\mathcal{O}(n^3m^2)$ operations to compute them, where m is the number of energy bins used. Thus it is applicable only to sequences up to some 100 nucleotides.

Most recently, a program has been designed by the Vienna group that can generate *all* secondary structures within some interval of the minimum energy based on dynamic programming and multiple backtracking (Wuchty *et al.*, 1999; Wuchty, 1998). The performance of the algorithm depends mainly on the number of structures found. Since the number of possible structures grows exponentially with chain length, the energy range that can be considered shrinks with increasing chain length. In practice, suboptimal folding can handle about a few million structures, corresponding, e.g., to an energy range of, say, 12 kcal/mol at a chain length of 100 bases. An example application is shown in Figure 2.

Most of these algorithms are part of the `Vienna RNA Package` (Hofacker *et al.*, 1994), which is freely available from `http://www.tbi.univie.ac.at/`.

Another approach to RNA structure prediction is to take into account the dynamics of the folding process. Such kinetic folding algorithms are the topic of section 5. In the case of functional RNAs, and provided a sufficient number of related sequences is available, the structure can be inferred from co-variations. This phylogenetic approach is beyond the scope of this review, but see e.g. (Gutell, 1993).

3. The Sequence-Structure Map

We have already mentioned above that there are many more sequences than structures. Hence, many sequences must fold into the same secondary structure. Moreover, extensive computer simulations have shown that there are only few common secondary structures and many rare ones, see Figure 3.

A structure ψ is *common* if it is formed by more sequences than the average structure. Data from both large samples of long sequences ($n \gg 30$) (Schuster *et al.*, 1994; Schuster, 1995) and from exhaustive folding of all short sequences (Grüner *et al.*, 1996a; 1996b) support two important observations: (i) the common structures represent only a small fraction of all structures and this fraction decreases with increasing chain length; (ii) the fraction of sequences

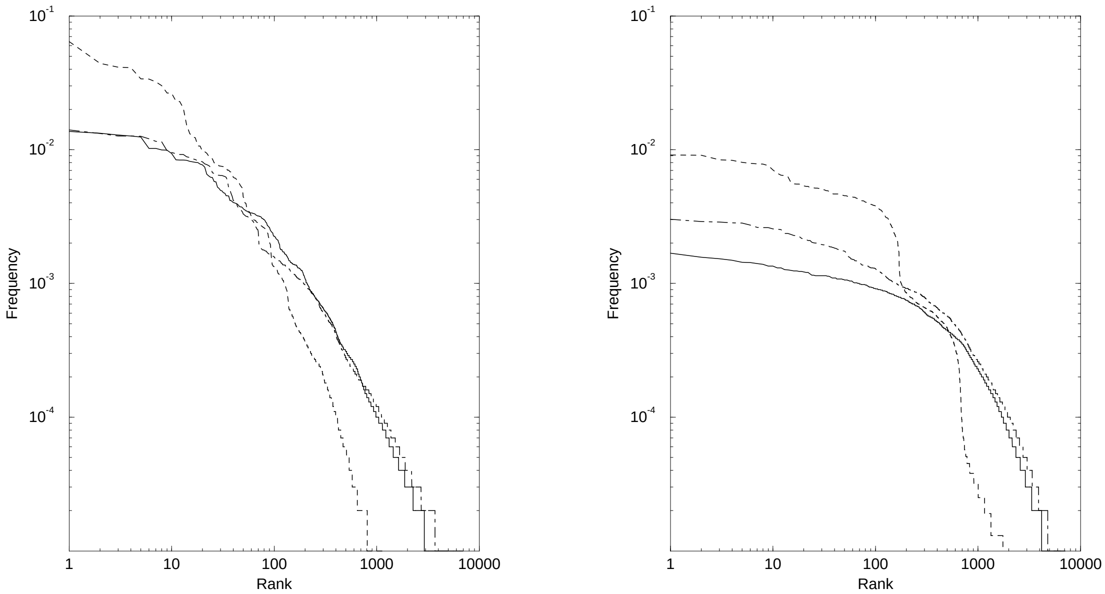

Figure 3. Zipf's law for coarse grained RNA secondary structures. The structures of 100,000 random sequences are ranked according to their frequencies. The ranking yields a distribution which follows a generalized Zipf's law $f(r) = a(r+b)^{-c}$, where r and $f(r)$ are the rank and the frequency of the corresponding structure, respectively. The constant a is a normalization factor, b can be interpreted as the number of "very frequent" structures. The constant c determines the slope of the tail of the distribution. We found distributions following this form of generalized Zipf's law for all algorithms, parameter sets, and alphabets. Full line: Minimum energy structures computed with the an up-to-date parameter set. Dash-dotted line: Minimum energy structures computed with an older parameter set. Dashed line: Deterministic kinetic folding algorithm. L.h.s.: **AUGC** alphabet, r.h.s.: **GC** alphabet. For the details of the coarse graining procedure and the parameter sets see (Tacker *et al.*, 1996).

folding into common structures increases with chain length and approaches 100% in the limit of long chains. Thus, for sufficiently long chains almost all RNA sequences fold into a small fraction of the secondary structures. The effective ratio of sequences to structures is therefore even larger than the combinatorial estimate. Furthermore, only common structures are likely to play a role in natural evolution and in evolutionary biotechnology (Schuster, 1995; Bacher and Ellington, 1998). RNA and proteins, despite their different chemistry, apparently share fundamental properties of their sequence-structure maps: the repertoire of stable native folds seems to be highly restricted or even vanishingly small (Chothia, 1992).

Naturally, we ask how sequences folding into the same (common) secondary structure are distributed in sequence space. In the following we review the results for minimum energy folding, i.e., we assume that the *folding map* assigns to each sequence the most stable secondary structure. The results reported below, however, have been shown to depend very little on the choice of algorithm (including various approaches to kinetic folding) and parameter sets (Tacker *et al.*, 1996).

Some notation is in order here: We call the set $S(\psi)$ of all sequences (genotypes) folding into phenotype ψ the *neutral set* of ψ. (For a mathematician S is the pre-image of ψ w.r.t. the folding map f).

Inverse folding can be used to determine $S(\psi)$. Naturally, a sequence x can fold into a given secondary structure ψ only if each pair of sequence positions that is paired in ψ is realized by one of the six possible base pairs. The set of all such sequences forms $C(\psi)$, the set of *compatible* sequences. Clearly, we have $S(\psi) \subseteq C(\psi)$. Note that many sequences in $C(\psi)$ will not have ψ as their most stable or kinetically most accessible structure. Thus the neutral set of ψ (for a particular folding map) will in general be only a small subset of the compatible set.

For RNA secondary structures an efficient inverse folding algorithm is available (Hofacker *et al.*, 1994). It was used to show that sequences folding into the same structure are (almost) randomly distributed within the set $C(\psi)$ of compatible sequences. A similar result was obtained for "protein space" (Babajide *et al.*, 1997) using so-called potentials of mean force (Sippl, 1990; 1993a;1993b).

The shape or topology of neutral sets has important implications for the evolution of both nucleic acids and proteins and for *de novo* design: For example, it has frequently been observed that seemingly unrelated protein sequences have essentially the same fold (Holm and Sander, 1997; Murzin, 1994; Murzin, 1996). Similarly, the genomic sequences of closely related RNA viruses show a large degree of sequence variation while sharing many conserved features in their secondary structures (Hofacker *et al.*, 1996; Rauscher *et al.*, 1997; Mandl *et al.*, 1998; Hofacker *et al.*, 1998).

Another well known example is the clover leaf secondary structure of tRNAs: The sequences of different tRNAs have little sequence homology (Eigen *et al.*, 1988) but nevertheless fold into the same secondary structure motif. Whether similar structures with distant sequences may have originated from a common ancestor, or whether they must be the result of convergent evolution, depends on the geometry of the neutral sets $S(\psi)$ in sequence space.

The local properties of the sequence-structure map can be investigated by considering pairs of RNA sequences that differ only by a single point mutation. A variety of methods is available to compare secondary structures and to quantify their differences by a (metric) distance, from counting the number of differing base pairs, to sophisticated alignment-like procedure such as tree editing.

It was noticed already in early work on RNA secondary structures (Fontana and Schuster, 1987) that a substantial fraction of point mutations are neutral, i.e., that many sequences differing only in a single position fold into the same sec-

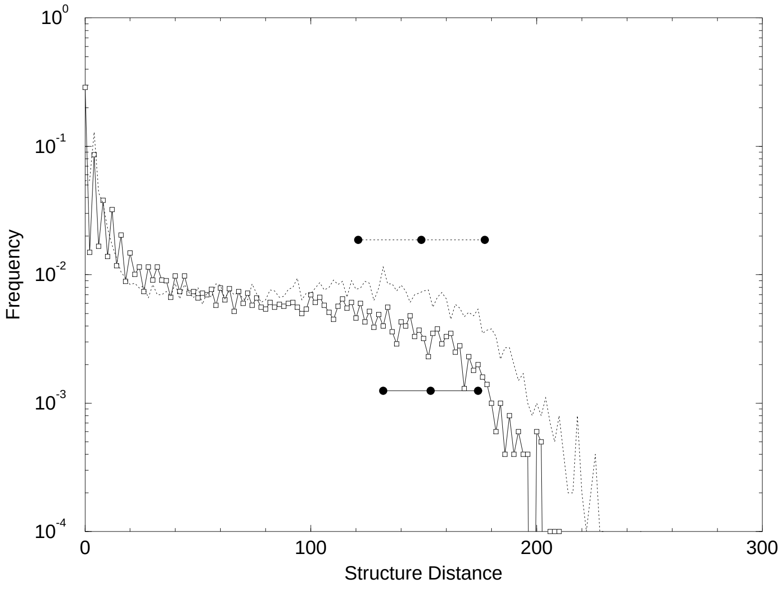

Figure 4. Distribution of structure distances between RNA sequences differing by a single point mutation, $n = 200$. Full line: natural **GCAU** alphabet, dotted line: **GC** alphabet. About 30% of the sequence pairs fold into the same structure. This high degree of neutrality implies the existence of connected neutral networks. On the other hand, a substantial fraction of point mutations leads to structure distances comparable to the structure distances between random sequences (mean and one standard deviation are indicated by circles). The structure distance is defined as edit distance on the tree representations of secondary structure graphs, see (Fontana *et al.*, 1993a) and (Hofacker *et al.*, 1994) for details.

ondary structure, see Figure 4. On the other hand, a comparable fraction of point mutants folds into secondary structures that have at best a vague resemblance with their parents' structures.

A simple mathematical model (based on random graph theory) of the sequence-structure map can be built on the following three observations: (i) inverse folded sequences are randomly distributed in $C(\psi)$, (ii) there is a large fraction λ of neutral mutations, (iii) non-neutral mutations often yield very different structures (Reidys *et al.*, 1997; Reidys, 1997). This model makes two rather surprising predictions:

(1) The connectivity of neutral sets changes drastically when λ passes the threshold value:

$$\lambda_{cr}(\alpha) \;=\; 1 - \sqrt[\alpha-1]{\frac{1}{\alpha}}\,, \tag{3.1}$$

where α is the size of the alphabet. Neutral sets consist of a single component that span the sequence space if $\lambda > \lambda_{cr}$ and below threshold, $\lambda < \lambda_{cr}$, the network is partitioned into a large number of components, in general, a giant component and many small ones. In the first case we refer to $S(\psi)$ as the *neutral network* of ψ. For RNA it is necessary to split the random graph into two factors corresponding to unpaired bases and base pairs and to use a different value of λ for each factor. For $\alpha = 2$ we find $\lambda_{cr} = 0.5$. For natural RNA sequences we have $\alpha = 4$ for the unpaired regions and $\alpha = 6$ for the paired regions. The critical values are $\lambda_{cr}(4) \approx 0.37$ and $\lambda_{cr}(6) \approx 0.301$, respectively. The fraction of neutral neighbors is much larger than these critical values for common RNA secondary structures, hence the neutral sets $S(\psi)$ form connected neutral networks within the sets $C(\psi)$ of compatible sequences (Reidys *et al.*, 1997). The situation appears to be similar for proteins (Babajide *et al.*, 1997).

(2) There is *shape space covering*, that is, in a moderate size ball centered at any position in sequence space there is a sequence x that folds into any prescribed secondary structure ψ. The radius of such a sphere, called the *covering radius* r_{cov}, can be estimated from simple probability arguments (Schuster, 1995)

$$r_{cov} \approx \min\{h \mid B(h) \geq S_n\}, \tag{3.2}$$

with $B(h)$ being the number of sequences contained in a ball of radius h. The covering radius is approximately 10-15% of the diameter of the sequence space. The covering sphere represents only a small connected subset of all sequences but contains, nevertheless, all common structures and forms an evolutionary representative part of shape space.

Extensive sample statistics (Schuster *et al.*, 1994) and exhaustive folding of all **GC**-sequences with given chain length $n \leq 30$ (Grüner *et al.*, 1996b) have so far been in excellent agreement with the random graph theory.

The existence of extensive neutral networks meets a claim raised by Maynard-Smith (Maynard-Smith, 1970) for protein spaces that are suitable for efficient evolution. The evolutionary implications of neutral networks are explored in detail in (Huynen *et al.*, 1996b; Huynen, 1996) and will be reviewed in the following section. Empirical evidence for a large degree of *functional* neutrality in protein space was presented recently by Wain-Hobson and co-workers (Martinez *et al.*, 1996).

4. Fitness Landscapes and Evolutionary Dynamics

Since Sewall Wright's seminal paper (Wright, 1932) the notion of a *fitness landscape* underlying the dynamics of evolutionary optimization has proved to be one of the most powerful concepts in evolutionary theory. Implicit in this idea is a collection of genotypes arranged in an abstract metric space, with each genotype next to those other genotypes which can be reached by a single mutation, as well as a fitness value assigned to each genotype.

It has been known since Eigen's pioneering work on the molecular quasi-species (Eigen, 1971; Eigen and Schuster, 1977; Eigen *et al.*, 1989) that the dynamics of evolutionary adaptation (optimization) on a landscape depends crucially on the detailed structure of the landscapes itself. Extensive computer simulations (Fontana and Schuster, 1987; Fontana *et al.*, 1989) have made it very clear that a complete understanding of the dynamics is impossible without a thorough investigation of the underlying landscape. Landscapes derived from well-known combinatorial optimization problems such as the Traveling Salesman Problem `TSP` (Lawler *et al.*, 1985), the Graph Bipartitioning Problem `GBP` (Fu and Anderson, 1986), or the Graph Matching Problem `GMP` have been investigated in some detail, see (Stadler, 1996) and the references therein. A detailed survey of a variety of model landscapes obtained by folding RNA molecules into their secondary structures has been performed during the last decade, see (Schuster and Stadler, 1994; Schuster *et al.*, 1997; Schuster, 1997a) and the references therein. While the use of (computationally simple) landscapes derived from spin-glasses or combinatorial optimization problems, or of the closely related Nk model (Kauffman, 1993) is certainly appealing, it is by no means clear that these models will capture the salient features of biochemically relevant landscapes. Indeed, we know now that landscapes derived from folding biopolymers into their spatial structures are quite different from spin-glass-like landscapes (Hordijk and Stadler, 1998).

One of the most important characteristics of a landscape is its *ruggedness*, a notion that is closely related to the hardness of the optimization problem for heuristic algorithms (Manderick *et al.*, 1991). Three distinct approaches haven been proposed to measure and quantify ruggedness and to subsequently compare different landscapes. Sorkin (1988), Eigen *et al.* (1989) and Weinberger (1990) used pair correlation functions. Kauffman and Levin (1987) proposed adaptive walks, and Palmer (1991) based his discussion on the number of meta-stable states (local optima). The relationship between correlation measures and local optima is discussed in detail by García-Pelayo and Stadler (1997). A mathematical framework for studying landscapes is developed in (Stadler, 1995; Stadler, 1996; Stadler and Happel, 1999).

Not surprisingly, landscapes based on sequence-structure maps (Figure 5) inherit their ruggedness even if the map from structures to fitness values is smooth or even linear, since shape space covering implies that a substantial fraction of point mutations lead to unrelated structures. On the other hand, a completely

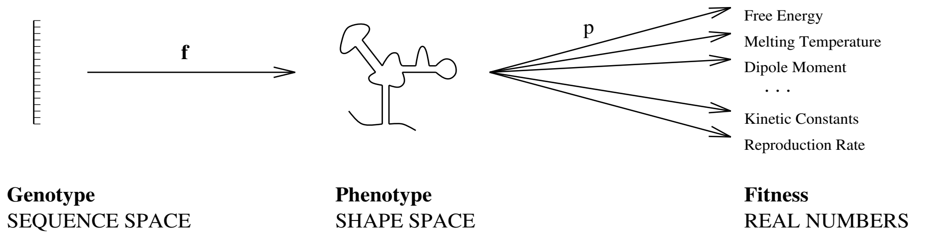

Figure 5. Landscapes based on genotype mappings can be viewed as compositions $p(f(g))$, where f : Sequence Space $\to$ Shape Space represents folding and p : Shape Space $\to \mathbb{R}$ encodes the evaluation of the structure by the environment.

random assignment of fitness values to structures cannot undo the correlation introduced by neutrality (Stadler, 1999).

Simplifying the detailed mechanisms of replication and mutation one may represent the dynamics of an infinite population by a reaction-diffusion equation of the form (Kimura, 1983; Ebeling *et al.*, 1984; Feistel and Ebeling, 1982)

$$\frac{\partial}{\partial t}\phi(x,t) = D\,\Delta\phi(x,t) + \phi(x,t)\left(F(x,\vec{\phi}) - \Phi(t)\right), \tag{4.1}$$

where $\phi(x,t)$ denotes the fraction of genotypes x at time t, $\Phi(t) = \sum_x F(x,\vec{\phi})$ is an unspecific dilution term ensuring conservation of probability, and D is a diffusion coefficient. The discrete Laplace operator Δ describes diffusion in sequence space and is a key ingredient in the mathematical theory of fitness landscapes (Stadler, 1996). In general $F(x,\vec{\phi})$ will be a non-linear function of the genotype frequencies describing the interactions between different species as well as their autonomous growth (Hofbauer and Sigmund, 1988). Within the context of this contribution $F(x,\vec{\phi}) = F(x)$ is the fitness landscape. The diffusion constant D is determined by the mutation rate p which is conveniently measured in units of mutation events per nucleotide and per generation. While this equation is not suitable for a detailed quantitative prediction of a particular model, it is a valuable qualitative heuristic for explaining some of the most important effects. One should keep in mind, however, that it is a mean field equation that does not correctly describe some important effects even in the limit of large populations. For an instructive example see (Tsimring *et al.*, 1996). In the absence of mutation, i.e., for $D = 0$ we are left with a system of coupled ordinary differential equation that in the usual way describe the population dynamics (Hofbauer and Sigmund, 1988).

Evolutionary dynamics on rugged landscapes without neutrality, such as the spin-glass like models is considered for instance in (Eigen, 1971; Ebeling *et al.*, 1984; Eigen *et al.*, 1989). For small mutation rates p a population is likely to get stuck in local optima for very long times. Populations form localized quasi-species around a "master sequence". There is a critical mutation rate p_{et} at which diffusion outweighs selection and the population begins to drift in sequence space – the genetic information is lost (Eigen, 1971; Eigen *et al.*, 1989). As an order of magnitude estimate one finds $p_{et} \approx \sigma/n$ where the "superiority" σ is a measure of the fitness advantage of the master sequence. Eigen's *error threshold* is a phenomenon that should be distinguished (Wagner and Krall, 1993) from Muller's ratchet. The latter refers to the loss of the optimal genotype in a finite population in the limit of very large genotypes. There the probability of reversing a deleterious mutation becomes zero. The error-threshold, on the other hand, appears also in a infinite population for relatively short sequence lengths.

On a flat fitness landscape, $F(x) = 1$ for all $x \in V$, the selection term disappears and we are left with a pure diffusion equation. A stochastic description can be found in (Derrida and Peliti, 1991). The situation on landscapes with a large degree of neutrality is much closer to the flat landscape than a non-neutral rugged one. There is no stationary master species surrounded by a mutant cloud, since Eigen's superiority parameter σ is so small in the presence of a large number of neutral mutants that reasonable values of p exceed the (genotypic) error-threshold by many orders of magnitude. For small values of p the neutral network of the fittest *structure*, $S(\psi)$, dominates the dynamics. Populations migrate by a diffusion-like mechanism (Derrida and Peliti, 1991; Huynen *et al.*, 1996b) on $S(\psi)$ just like on a flat landscape with the single modification that the effective diffusion constant is smaller by the factor λ, the fraction of neutral mutations.

Random drift is continued until the population reaches an area in sequence space where some fitness values are higher than that of the currently predominating neutral network. Then a period of Darwinian evolution sets in, leading to the selection of the locally fittest structure. Evolutionary adaptation thus appears as a stepwise process: phases of increasing mean fitness (transitions between different structures) are interrupted by periods of apparent stagnation with mean fitness values fluctuating around a constant (diffusion on a neutral network) (Huynen *et al.*, 1996b), Figure 6. A detailed analysis of evolutionary trajectories in terms of likely structural adaptations can be found in (Fontana and Schuster, 1998b; Fontana and Schuster, 1998a). When the fittest structure is common its neutral network extends through the entire sequence space allowing the population to eventually find the global fitness optimum. A population is not a single localized quasi-species in sequence space (Eigen *et al.*, 1989), but rather a collection of different quasi-species since population splits into well separated clusters (Huynen *et al.*, 1996b) on a single neutral network. Each cluster undergoes independent diffusion, while all share the same dominant phenotype.

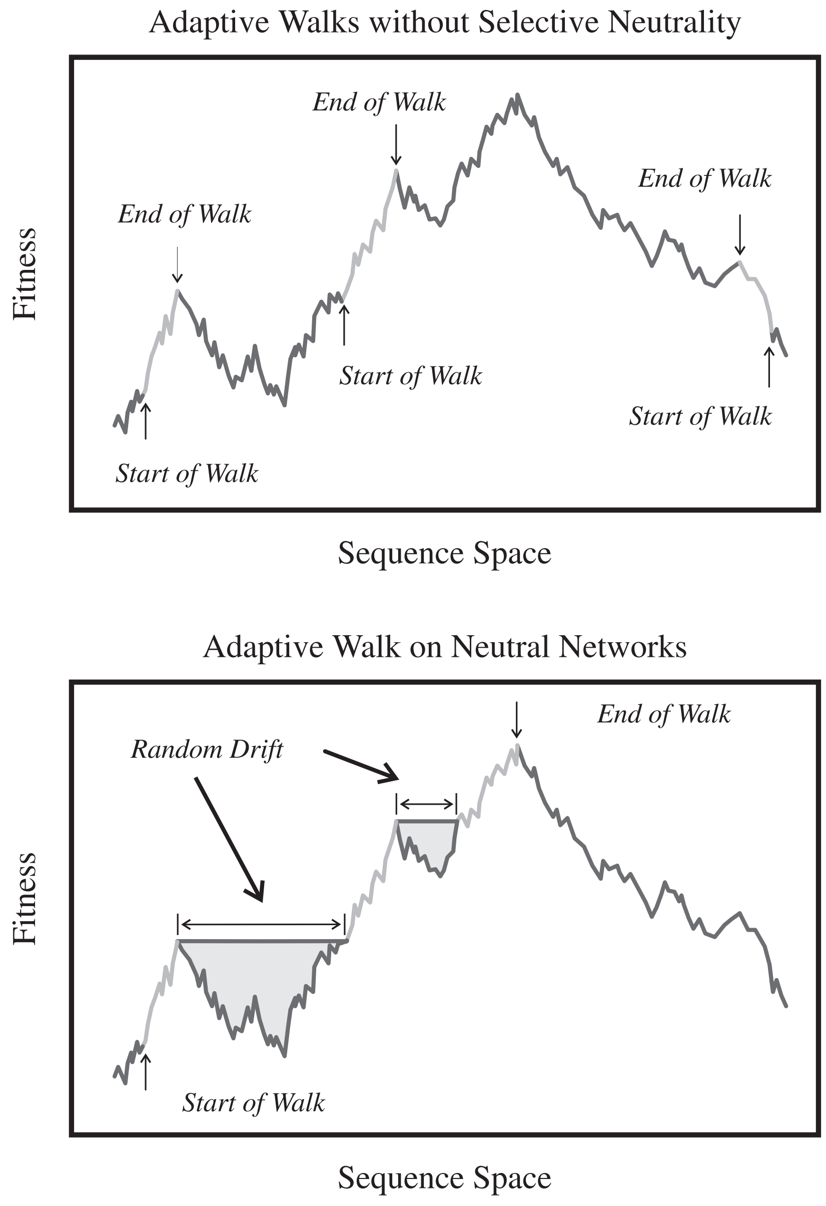

Figure 6. The role of neutral networks in evolution (Schuster, 1997b). Optimization occurs through adaptive walks and random drift. Adaptive walks allow to choose the next step arbitrarily from all directions where fitness is (locally) non-decreasing. Populations can bridge over narrow valleys with widths of a few point mutations. In the absence of selective neutrality (spin-glass-like landscape, above) they are, however, unable to span larger Hamming distances and thus will approach only the next major fitness peak. Populations on rugged landscapes with extended neutral networks evolve along the networks by a combination of adaptive walks and random drift at constant fitness (below). In this manner, populations bridge over large valleys and may eventually reach the global maximum of the fitness landscape.

It is not surprising hence that there are abundant examples of both RNA and protein structures that have been conserved over evolutionary time scales while the underlying sequences have lost (almost) all homology.

For larger mutation rates p the diffusion term dominates the dynamics and the population is not confined to the neutral network any more. The *phenotypic error threshold* (Forst *et al.*, 1995; Huynen *et al.*, 1996b; Reidys *et al.*, 1999) is the mutation rate at which the dominating phenotype is lost.

Diffusion in sequence space, the existence of phenotypic error threshold, and a close connection (Huynen *et al.*, 1996b) with Kimura's neutral theory (Kimura, 1983) which we have not discussed here, are consequences of the existence of neutral networks. Shape space covering implies a constant rate of innovation (Huynen, 1996): While diffusing along a neutral network, a population constantly produces non-neutral mutants folding into different structures. Shape space covering implies that almost all structures can be found somewhere near the current neutral network.

Hence the population keeps discovering structures that it has never encountered. When a superior structure is produced, Darwinian selection becomes the dominating effect and the population "jumps" onto the neutral network of the novel structure while the old network is abandoned. Figure 6 sketches the difference between evolutionary adaptation on spin-glass-like landscapes and on the highly neutral landscapes arising from biopolymer structures.

Neutral evolution, arising as a consequence of the high degree of neutrality observed in genotype-phenotype mappings of biopolymers, therefore, is not a dispensable addendum to evolutionary theory (as it has often been suggested). On the contrary, neutral networks, provide a powerful mechanism through which evolution can become truly efficient.

The evolution of sequences on neutral networks can be observed very clearly in RNA viruses. Our simulations show that sequence differences of as little as 10% lead almost surely to unrelated structures if the mutated sequence positions are chosen randomly (Fontana *et al.*, 1993b). The presence of conserved secondary structure elements such as the TAR or RRE region in HIV, the IRES region of picorna viruses, or the stem loop structure at the 3' terminus of flavivirus genomes, which show a significant sequence variation between different virus strains (only about 80% average pairwise sequence identity), must therefore be regarded as the result of stabilizing selection acting on the secondary structure. This effect can be used to design an algorithm that reliably detects conserved, and therefore most likely functional, RNA secondary structure elements in viral genomes based on a combination of secondary structure prediction and comparative sequence analysis (Hofacker *et al.*, 1998; Hofacker and Stadler, 1999). Evolution on neutral networks leads to an increased level of robustness against mutation since a diffusing population prefers the denser regions of the neutral network (van Nimwegen *et al.*, 1999). The effect can be observed by comparing conserved and non-conserved sub-structures in the rapidly evolving genomic sequences of RNA and retroviruses (Wagner and Stadler, 1999).

5. Energy Landscapes and Folding Kinetics

The *energy landscape* of an RNA molecule is, for our purposes, defined on the set of all secondary structures that are compatible with its sequence. Conceptually, this energy landscape is closely related to the *potential energy surfaces* (PES) which constitute one of the most important issues of theoretical chemistry (Mezey, 1987; Heidrich *et al.*, 1991). As a consequence of the validity of the Born-Oppenheimer approximation, the PES provides the potential energy $U(\vec{R})$ of a molecule as a function of its nuclear geometry $\vec{R}$. PES are therefore defined on a high-dimensional *continuous* space and they are assumed to be smooth (usually twice continuously differentiable almost everywhere). The (global) analysis of PES thus makes extensive use of differential topology. In contrast, our notion of energy landscapes is discrete. Their analysis is therefore similar to the analysis of fitness landscapes.

A crucial ingredient for the simulation of RNA folding kinetics is the choice of a "move set" for inter-converting secondary structures. This move-set defines the topology of the energy landscape by defining which secondary structures are neighbors of each other and encodes the set of structural changes that RNAs can undergo with moderate activation energies. It is the basis of all *kinetic* algorithms for RNA folding.

The assumptions that an RNA molecule folds into its thermodynamic ground state may well be wrong even for moderately long sequences (Morgan and Higgs, 1996). Consequently, several groups have designed kinetic folding algorithms for RNA secondary structures, mostly in an attempt to get more accurate predictions or in order to include pseudo-knots, see e.g. (Martinez, 1984; Mironov *et al.*, 1985; Abrahams *et al.*, 1990; Gultyaev, 1991; Tacker *et al.*, 1994). Only a few papers have attempted to reconstruct folding pathways (Higgs, 1995; Gultyaev *et al.*, 1995; Suvernev and Frantsuzov, 1995). These algorithms generally operate on a list of all possible helices and consequently use move-sets that destroy or form entire helices in a single move. Such a move-set can introduce large structural changes in a single move and furthermore, *ad hoc* assumptions have to made about the rates of helix formation and disruption. A more local move-set is, therefore, preferable if one hopes to observe realistic folding trajectories.

The most elementary move-set, on the level of secondary structures, consists of removal and insertion of single base pairs (while making sure that one does

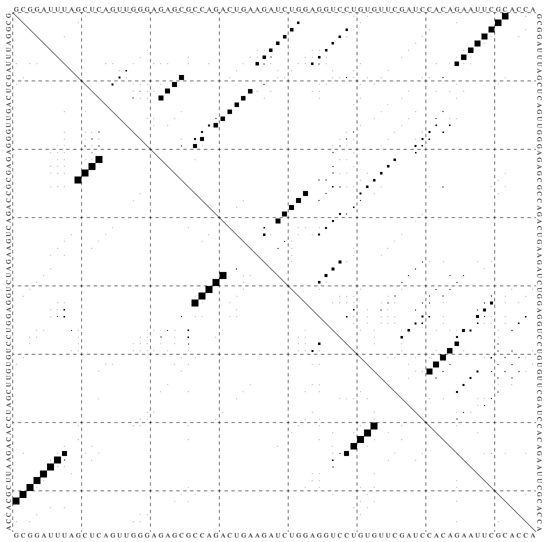

Figure 7. Base pair probabilities for an phenylalanine tRNA with and without modified bases. The equilibrium frequency p of a pair $[i, j]$ is represented by a square of area p in position i, j and j, i of the matrix. Lower left: only base pairs contained in the ground state occur with significant frequency for the sequence with modified bases. Upper right: The unmodified sequence displays a large number of base pairs from suboptimal structures, although the ground state remains unchanged.

not introduce knots or pseudo-knots). The simulations reported below are described in detail in Christoph Flamm's PhD thesis (Flamm, 1998) and (Flamm *et al.*, 1999). Either this simple move-set or, as in the data shown below, base pair insertion and deletions together with base pair "shifts" (in which a base pair $[i, j]$ is converted into a new pair $[i, k]$) are used. These shift moves facilitate sliding of the two strands of helix, which is assumed to be an important effect in dynamics of RNA molecules. The dynamics itself is simulated by an algorithm designed for stochastic chemical reactions by Gillespie (1976). The time scale is fixed using the measured hairpin formation of the oligonucleotide `AAAAAACCCCCCUUUUUU` (Pörschke, 1974). For the rates constants a symmetrical rule $k \sim \exp(-\Delta G/2kT)$ independent of the sign of ΔG has been assumed (Kawasaki, 1966) instead of the usual Metropolis rule. Additional simulations using the Metropolis rule showed qualitatively similar results.

Local minima are of particular importance for the folding dynamics, since they can trap the molecule in a misfolded state. For a given sequence the low-energy local minima can be constructed with reasonable effort: Structures within some interval of the ground state are generated through complete suboptimal folding (Wuchty *et al.*, 1999), for each structure all neighboring structures are generated

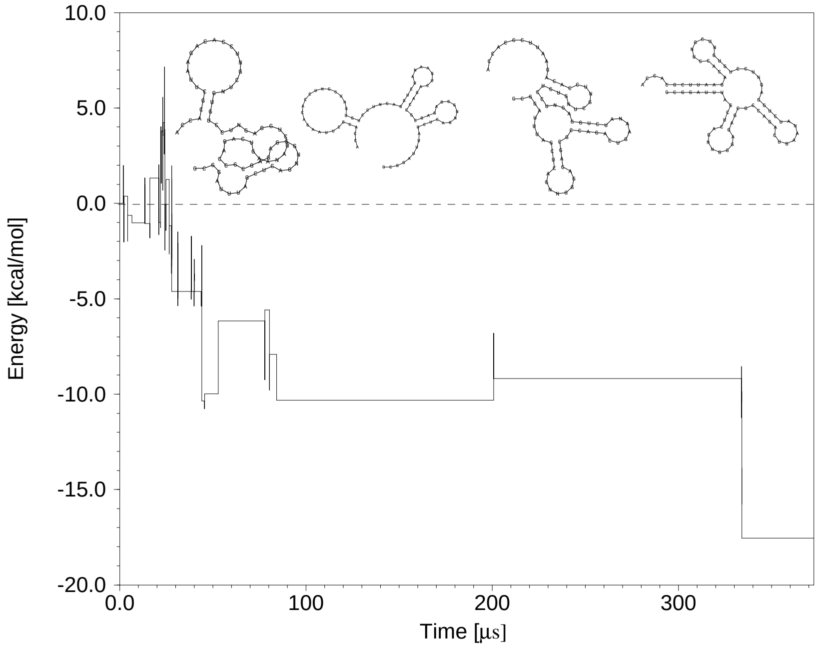

Figure 8. Energy as a function of time for a representative simulation of the modified tRNA. A few intermediate structures are shown at the top, the last one being the native clover-leaf structure. The stem closing the multi-loop forms last in most simulations.

and their energies compared to the reference structure. Figure 2 shows the density of local minima within 15 kcal/mol of the ground state for the $\mathrm{tRNA}^{\mathrm{phe}}$.

By an extension of the above procedure one can even determine the height of the energy barriers and the structures of the saddle points (transition states) connecting local minima. This analysis yields a tree with the local minima as leaves and the transition states as internal nodes; the branch lengths represent the height of the energy barriers. An example of such a tree is shown Figure 11.

Transfer RNA molecules from most organisms contain several modified bases, particularly methylations. These modified bases occur mostly in unpaired regions and often the modifications are such that base pairing is made impossible. Hence, one might speculate that the modified bases help to stabilize the correct fold.

The phenylalanine tRNA from yeast used in the following contains six modifications which prohibit base pairing its 76 nucleotides. As can be seen in Figure 7 the modifications have a strong effect on the equilibrium ensemble of structures. The frequency of the correct fold in the thermodynamic ensemble rises from 4.4% to 28% and suboptimal folding shows that the lowest six suboptimal structure are prohibited by the modifications and consequently the energy gap from the ground state to the next possible structures increases from 0.4 to 0.9 kcal/mol.

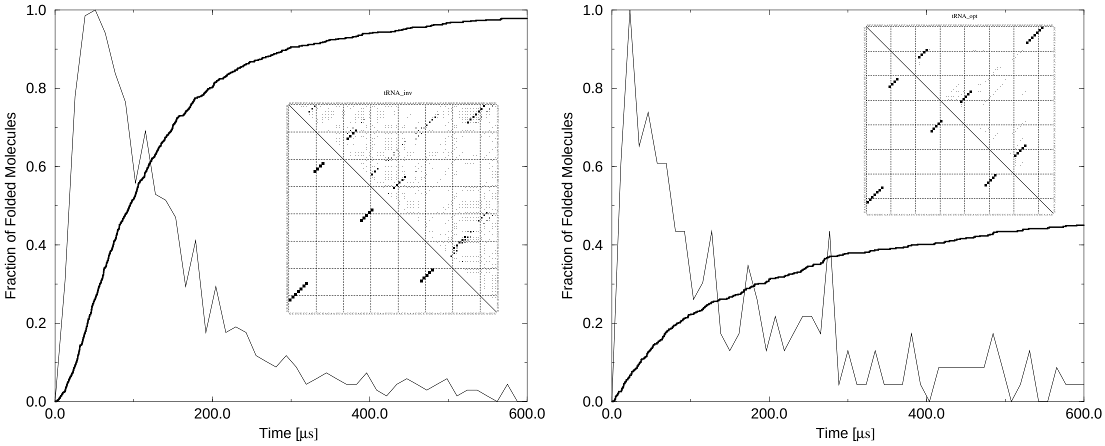

Figure 9. Thermodynamic stability and "foldability". The fraction of folded sequences $p(t)$ (thick lines) as a function of time and folding times (thin lines) for two artificial sequences designed to fold into the tRNA clover-leaf structure are derived from 1000 independent runs. Inset: dot plots showing the equilibrium base pair probabilities (upper right) as obtained from McCaskill's algorithm and the contact map of the tRNA structure (lower left). L.h.s: a randomly chosen sequence with tRNA structure shows many alternative foldings in the dot plot but nevertheless folds efficiently. R.h.s: A sequence designed to be thermodynamically stable (see inset) folds only in less than 50% of the simulations.

The modified sequence exhibits very few local minima in the low energy region, there are only 10 local minima within 5 kcal/mol of the ground state compared to 173 for the unmodified sequence (Hofacker, 1998).

To study the kinetic effect of the modifications, the folding of modified and unmodified tRNA sequence has been simulated (Flamm, 1998). The resulting trajectories were then analyzed for the existence of typical folding pathways, Figure 8. In this particular run the RNA folds somewhat slower than average, but nevertheless shows features common to all trajectories. A rapid collapse leads to a structure with almost as many base pairs as the native state but little overlap. Folding then proceeds through a series of local minima that have more and more structural elements in common with the ground state. The waiting times in the local minima increase with decreasing energy. Many trajectories visit the same low energy intermediates, in particular, the stem closing the multi-loop forms last in almost all simulations. Interestingly, the correct hairpins closest to the 5'-end are often formed first, which might support efficient folding during transcription.

As a measure of foldability we recorded the folding times, i.e., the time after which the ground state appears in the simulation for the first time. We consider in particular the fraction $p(t)$ of molecules that have reached the ground state at time t. The resulting distribution can be seen in Figure 9. For the modi-

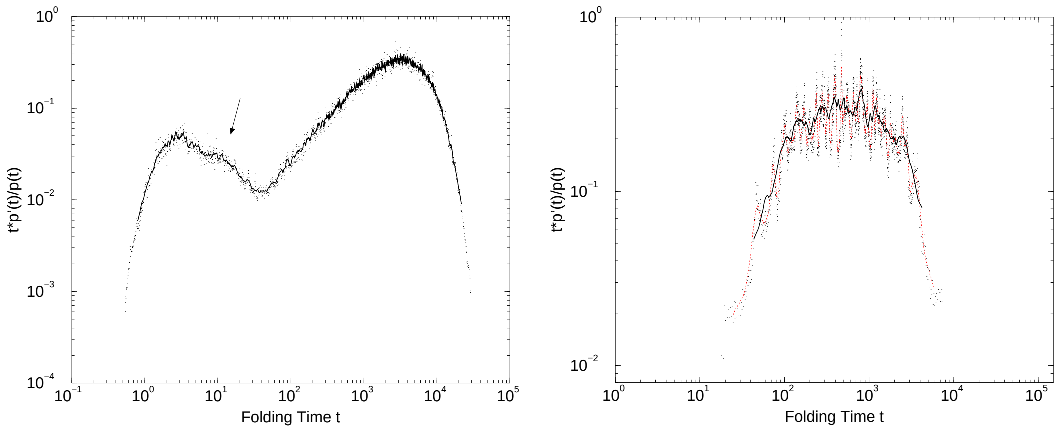

Figure 10. Folding kinetics of two different RNA molecules. L.h.s.: The (artificial) molecule whose pathway is described in detail in figure 11. The curve shows two distinct peaks corresponding to two different dominating folding pathways. A less prominent folding pathway is indicated by the shoulder on the right hand side of the first peak (indicated by an arrow). R.h.s.: The kinetic signature of the modified tRNA shows only a single peak. The time scale of folding is set by the closing of the multi-loop, see Figure 8.

fied sequence the ground state was found in all simulations. This is consistent with recent analysis by Thirumalai and Woodson (1996) of experimental data, suggesting a directed pathway to the native state for tRNAs. The unmodified sequence folds much more slowly and only 46% of runs reach the ground state within the simulation time. The fraction of folded sequences is still rising at that point and longer simulations will be needed to decide whether the curve saturates at less than unity.

In the case of phenylalanine tRNA the modified bases improved both thermodynamic stability, conferred by a large energy gap between native and misfolded states, and foldability. The same link has been claimed for lattice protein models by Šali *et al.* (1994b). To test this hypothesis we have designed two artificial sequences with the tRNA structure as ground state using the `RNAinverse` program from the `Vienna RNA Package`. The thermodynamic properties of the first sequence are typical for sequences of this size: the frequency of the ground state in the ensemble is about 7% and several alternative foldings can be seen in the base pair probability matrix, see the inset on the l.h.s. of Figure 9. The other sequence was designed to be particularly stable: its ground state dominates the ensemble with a frequency of 96% and no alternative foldings are discernible in the dot plot. We than ran 1000 folding simulations for each sequence; the results are summarized in Figure 9. Surprisingly, the thermodynamically more stable sequence folds poorly in this example.

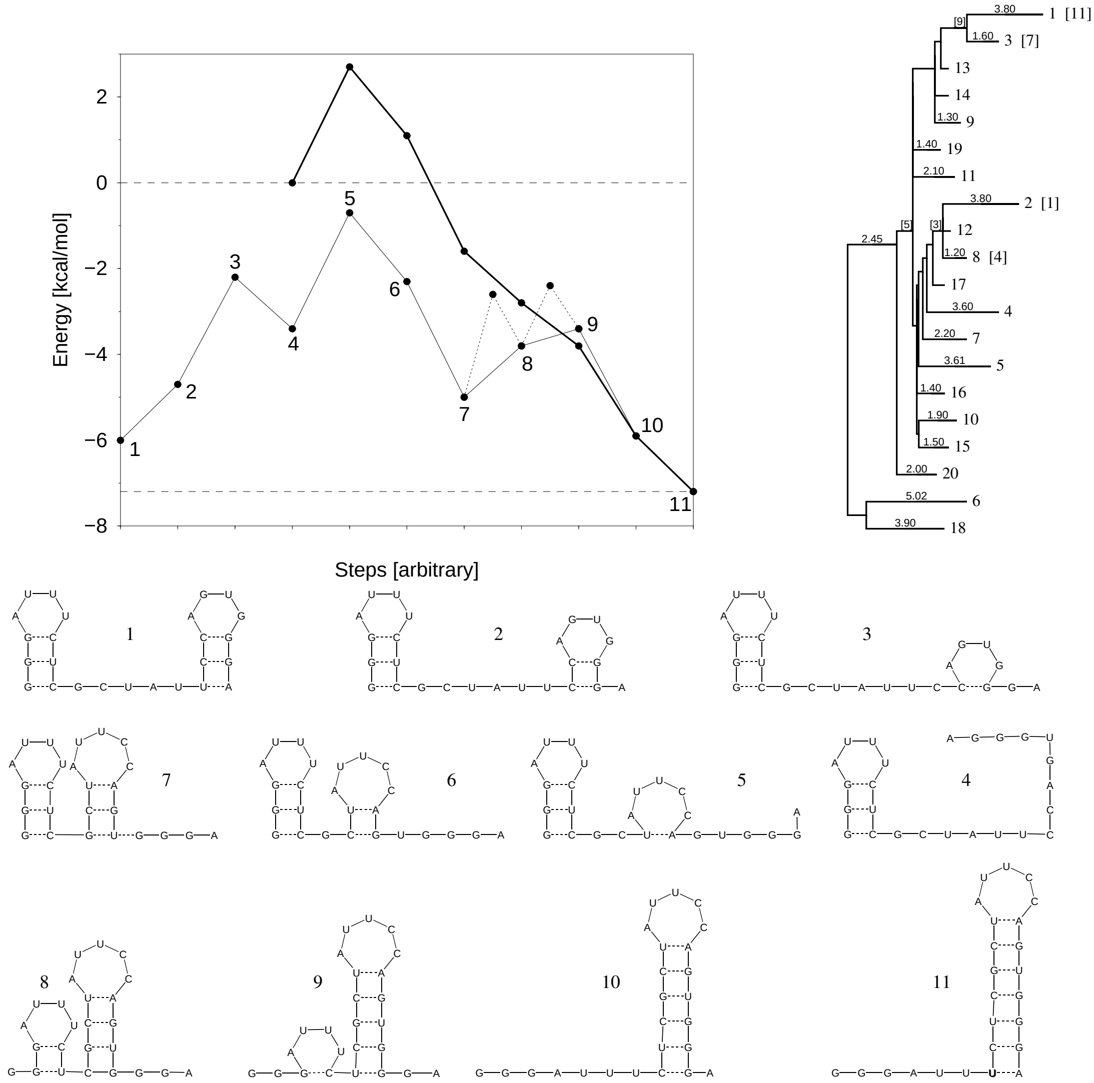

Figure 11. Trapping and escape from a local minimum. An artificial RNA molecule was designed with a low-energy mis-folded state formed by two hairpins at the 5' and 3' ends. The 3' hairpin of the misfolded state blocks the formation of the ground state, which consists of a single, much longer, 3' hairpin. The upper left plot shows the energy profile of the two most prominent folding pathways. The upper right plot shows the energy barriers between the 20 lowest local minima. The fast pathway begins with the formation of the correct hairpin at the 3' end and the rapid elongation of the stem. The only an energy barrier occurs at the first step with a height of about 2.7 kcal. The second pathway begins with the fast formation of the meta-stable structure (frame 1 in the lower plot). In order to escape from the mis-folded state the wrong stem at the 3' end has to be unfolded (steps 2 to 4). Then the correct 3' stem can be initiated (step 5). However, this stem cannot be elongated rapidly, since it is still blocked by the 5' stem. A series of shift moves (steps 8 to 10) leads to the ground state 11. Structures along this trajectory are indicated by square brackets in the tree diagram. The dashed lines in the upper part indicate the energy barriers in the absence of shifts.

Even an isolated example such as this one shows that it is easy to construct cases where the kinetics cannot be predicted from thermodynamic properties. More test cases will be needed in order to decide if and how strongly thermodynamic stability and foldability correlate on average.

Some information about folding pathways can be inferred directly from the kinetics of the folding process, in particular from $p(t)$ curves. Thirumalai and Woodson (1996) noted that meta-stable states cause dents in log-log plots of $p(t)$ versus t. We found that plotting $tp'(t)/p(t)$ as a function of $\log t$ yields a more detailed picture, see Figure 10.

Some molecules have folding pathways with very different time scales. In general, these are determined by local minima with large basins of attraction on the energy landscapes. Such local minima act as traps for the folding process. In some cases these meta-stable states are long-lived enough for experimental detection (Loss *et al.*, 1991; Biebricher and Luce, 1992; Rosenbaum *et al.*, 1993). Here we consider an artificial RNA of 25 nucleotides that was designed in such a way, that it can form either of two overlapping hairpins. This molecule is small enough that we can readily see the escape from a meta-stable state within the simulation time. Figure 11 shows a detailed analysis of the two most prominent folding pathways. The left peak in Figure 10 corresponds to "direct folding", i.e., running down the correct "funnel" of the energy landscape. Once the hairpin loop is formed, a smooth "zipper" closes the base pairs of the stack and the ground-state is reached very rapidly. The right peak corresponds to trapping in a meta-stable state. An intricate pathway, detailed in Figure 11, allows the molecule to escape by partially unfolding the meta-stable structure. The shoulder in Figure 10, finally, corresponds to shallow meta-stable states that are unfolded completely before the folding process follows the "funnel".

6. Discussion

Both folding and evolution of biopolymers can be formulated in terms of landscapes, that is, mappings from a configuration space (sequence space or shape space) into the real numbers (energy or fitness). Fitness landscapes inherit their properties from the underlying sequence-structure map. The latter is well understood in the case of RNA secondary structures because the folding problem is easily solved within this model.

The dynamics of evolutionary adaptation is determined by the interplay of the large fraction of neutral mutations and the high degree of ruggedness. These properties imply a diffusive motion along neutral network of a dominating structure punctuated by fast transitions to different structures. It seems that RNA and proteins behave very similar in this respect.

RNA secondary structures provide an ideal model system to study both structure formation and evolution. The secondary structure model is simple enough to allow efficient algorithms to compute (almost) any thermodynamic quantity of interest, yet it is still close enough to reality to address problems of practical interest. Furthermore, it is relatively easy to explore the energy landscape of a particular sequence. RNA secondary structures are thus an elegant model to address questions about foldability. In the following we very briefly point out the main differences between RNA and protein models, emphasizing the ways in which RNA presents itself as the more tractable system.

Protein folding has remained (almost) intractable for a good biophysical reason despite the efforts of many groups. Protein structure is stabilized by hydrophobic interactions and hydrogen bonds that depend on a meticulous packing of amino acid side chains. Hence the contribution of an amino acid to the overall structure is determined by the details of its entire spatial neighborhood rather than the simple specific interaction with a single pairing partner that is characteristic for nucleic acids. As a consequence, protein secondary structure is neither a particularly good description of the spatial structure nor the single most important folding intermediate.

The crucial dependence on side chain packing, which is not an important issue in RNA, has far-reaching consequences on protein folding: Not all amino acid sequences even reach a stable "native-like" structure. Instead they are stuck in a flexible, partially folded *molten globule* state. It is worth noting that not even the fraction of amino acid sequences that fold into a native-like ground state is known with any certainty. As a many-point interaction, side-chain packing is also very hard to incorporate into *knowledge based potentials of mean force* (Bauer and Beyer, 1994; Bowie *et al.*, 1991; Godzik *et al.*, 1992; Goldstein *et al.*, 1992; Grossman *et al.*, 1995; Hendlich *et al.*, 1990; Sippl, 1993a; Sippl, 1993b). Such potentials describe the effective interactions between amino acid residues and can be regarded as a natural analog of the standard energy model for nucleic acids. While such potential functions are very effective for identifying sequences that fold into a given native protein structure (the inverse folding problem) or to identify incorrectly folded proteins (or sections of proteins), they cannot be used for folding a particular sequence into its ground state structure. Inverse folding based on knowledge based potentials can be used to partially explore the sequence-structure relationships. We found neutral networks and strong indications of shape space covering (Babajide *et al.*, 1997; Babajide *et al.*, 1999), suggesting that the global properties of protein space do not differ very much from the RNA case.

The overwhelming part of theoretical investigations into protein folding are aimed at understanding the principles of the folding process rather than folding individual sequences. We may distinguish two main approaches: computer simulations based on simplified lattice models, and statistical mechanics papers.

Lattice models (Lau and Dill, 1990; Chan and Dill, 1991; Crippen, 1991; Lipman and Wilbur, 1991; Camacho and Thirumalai, 1993; Šali *et al.*, 1994b; Dill *et al.*, 1995; Chan and Dill, 1996; Li *et al.*, 1996; Bornberg-Bauer, 1997; Hart and Istrail, 1997a) provide a coarse grained view on protein structure not unlike the approximation of RNA structure by secondary structures. Unfortunately, the lattice protein folding problem is NP hard (Ngo and Marks, 1992; Unger and Moult, 1993; Hart and Istrail, 1997b). Thus most computational studies are limited to fairly short molecules ($n \ll 30$ in most work on the **HP** model), or strongly constrained sets of structures (such as 27-mers that fill a $3 \times 3 \times 3$ cube). These models allow to study the hydrophobic collapse. Furthermore they admit an intrinsic distinction between folding and non-folding sequences (a sequences folds into a native structure if the lowest-energy structure is unique); it is not clear how well this approach will generalize to more complex potential functions and larger alphabets which will lead to non-degenerate ground states for most sequences (Buchler and Goldstein, 1999). In addition, some results, such as the clustering of $S(\psi)$ and the relatively small extent of neutral networks observed in some lattice models (Bornberg-Bauer, 1997) are not very well compatible with simulations based on knowledge based potentials. This discrepancy might be explained by the short chains $n < 30$ and the two-letter **HP** alphabet used in these models. While native-like proteins can be designed from reduced alphabets, recent experiments (Davidson *et al.*, 1995; Plaxco *et al.*, 1998) as well as computer simulations (Babajide *et al.*, 1997) suggest that two letters are not sufficient.

The concept of a folding *funnel* was introduced based on an analysis of the random energy model (Bryngelson and Wolynes, 1987) and has since inspired many studies of protein folding, e.g. (Šali *et al.*, 1994a; Shrivastava *et al.*, 1995; Dill and Chan, 1997; Onuchic *et al.*, 1997). In this description the folding process is determined entirely by the density of states while the topology of the folding landscape is disregarded. The foldability of a sequence is then related to the energy gap between the ground state and the first excited state or an ensemble of mis-folded states. In the case of RNA, however, one can easily design counterexamples of sequences that fold poorly in spite of high thermodynamic stability, see Figure 9. Similar results for the protein case were presented recently by Crippen and Ohkubo (1998).

The RNA secondary structure model does not suffer from all the shortcomings and/or technical difficulties of the various protein folding models. On the other hand it deals only with a coarse grained description, which disregards both the overall three-dimensional shape and the detailed arrangement of the chemical groups that are oftentimes essential for the molecules functionality. Despite these shortcomings and the limit accuracy of the standard energy model, it is the only case that allows a complete treatment of all the various aspects, from the folding kinetics of a single molecule to the long term evolution of a population of RNA molecules *in vitro*, within a single consistent computational framework.

Acknowledgments

The research on RNA folding and evolution is an on-going joint effort with Peter Schuster and Walter Fontana at the Department of Theoretical Chemistry of the University of Vienna. Partial financial support by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* Proj. 12591-INF, the *Jubiläumsfond der Österreichischen Nationalbank* Proj. 6792, and by the European Commission in the framework of the Biotechnology Program (BIO-4-98-0189) is gratefully acknowledged.

References

Abrahams, J. P., van den Berg, M., van Batenburg, E. and Pleij, C. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.* **18**, 3035–3044 (1990).

Babajide, A., Farber, R., Hofacker, I. L., Inman, J., Lapedes, A. S. and Stadler, P. F. Exploring protein sequence space using knowledge based potentials. *J. Comp. Biol.* (1999). Submitted, Santa Fe Institute preprint 98-11-103.

Babajide, A., Hofacker, I. L., Sippl, M. J. and Stadler, P. F. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design* **2**, 261–269 (1997).

Bacher, J. M. and Ellington, A. D. Nucleic acid selection as a tool for drug discovery. *Drug Discovery Today* **3**, 265–273 (1998).

Bauer, A. and Beyer, A. An improved pair potential to recognize native protein folds. *Proteins* **18**, 254–261 (1994).

Biebricher, C. K. and Luce, R. In vitro recombination and terminal elongation of RNA by Qβ replicase. *EMBO J.* **11**, 5129–5135 (1992).

Bornberg-Bauer, E. G. How are model protein structures distributed in sequence space? *Biophys. J.* **73**, 2393–2403 (1997).

Bowie, J. U., Luthy, R. and Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170 (1991).

Bryngelson, J. D. and Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528 (1987).

Buchler, N. E. G. and Goldstein, R. A. The effect of alphabet size and foldability requirements on protein structure designability. *Proteins* (1999). In press.

Camacho, C. J. and Thirumalai, D. Minimum energy compact structures of random sequences of heteropolymers. *Phys. Lett.* **71**, 2505–2508 (1993).

Chan, H. S. and Dill, K. A. Sequence space soup. *J. Chem. Phys.* **95**, 3775–3787 (1991).

Chan, H. S. and Dill, K. A. Comparing folding codes for proteins and polymers. *Proteins* **24**, 335–344 (1996).

Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543–544 (1992).

Crippen, G. M. Prediction of protein folding from amino acid sequences of discrete conformation spaces. *Biochemistry* **30**, 4232–4237 (1991).

Crippen, G. M. and Ohkubo, Y. Z. Statistical mechanics of protein folding by exhaustive enumeration. *Proteins* **32**, 425–437 (1998).

Cupal, J. *The Density of States of RNA Secondary Structures*. Master's thesis, University of Vienna (1997).

Cupal, J., Hofacker, I. L. and Stadler, P. F. Dynamic programming algorithm for the density of states of RNA secondary structures. In: *Computer Science and Biology 96 (Proceedings of the German Conference on Bioinformatics)* (Hofstädt, R., Lengauer, T., Löffler, M. and Schomburg, D., eds.), pp. 184–186. Leipzig (Germany): Univeristät Leipzig (1996).

Davidson, A. R., Lumb, K. J. and Sauer, R. T. Cooperatively folded proteins in random sequence libraries. *Nat. Struc. Biol.* **2**, 856–863 (1995).

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

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yeo, D. P., Thomas, P. D. and Chan, H. S. Principles of protein folding: a perspective from simple exact models. *Prot. Sci.* **4**, 561–602 (1995).

Dill, K. A. and Chan, H. S. From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10–19 (1997).

Draper, D. E. Parallel worlds. *Nature Struct. Biol.* **3**, 397–400 (1996).

Ebeling, W., Engel, A., Esser, B. and Feistel, R. Diffusion and reaction in random media and models of evolution processes. *J. Stat. Phys.* **37**, 369–384 (1984).

Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften* **10**, 465–523 (1971).

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

Eigen, M., Oswatitsch-Winkler, R. and Dress, A. Statistical geometry in sequecne space: A method of quantitative comparative sequence analysis. *Proc. Natl. Acad. Sci. USA* **85**, 5913–5917 (1988).

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

Feistel, R. and Ebeling, W. Models of Darwinian processes and evolutionary principles. *Biosystems* **15**, 291–299 (1982).

Flamm, C. *Kinetic Folding of RNA*. Ph.D. thesis, University of Vienna (1998).

Flamm, C., Fontana, W., Hofacker, I. and Schuster, P. *RNA folding kinetics at elementary step resolution*. Tech. rep., Inst. f. Theor. Chemie, Univ. Vienna (1999). In preparation.

Fontana, W., Konings, D. A. M., Stadler, P. F. and Schuster, P. Statistics of RNA secondary structures. *Biopolymers* **33**, 1389–1404 (1993a).

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

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

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

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

Fontana, W., Stadler, P. F., Bornberg-Bauer, E. G., Griesmacher, T., Hofacker, I. L., Tacker, M., Taranzona, P., Weinberger, E. D. and Schuster, P. RNA folding and combinatory landscapes. *Phys. Rev. E* **47** (3), 2083–2099 (1993b).

Forst, C. V., Reidys, C. M. and Weber, J. Evolutionary dynamics and optimization: Neutral Networks as model-landscape for RNA secondary-structure folding-landscapes. In: *Advances in Artificial Life* (Morán, F., Moreno, A., Merelo, J. and Chacón, P., eds.), vol. 929 of *Lecture Notes in Artificial Intelligence*, pp. 128–147. ECAL '95, Berlin, Heidelberg, New York: Springer (1995).

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. and Turner, D. H. Improved free-energy parameters for prediction of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* **83**, 9373–9377 (1986).

Fu, Y. and Anderson, P. W. Application of statistical mechanics to NP-complete problems in combinatorial optimization. *J. Phys. A* **19**, 1605–1620 (1986).

García-Pelayo, R. and Stadler, P. F. Correlation length, isotropy, and meta-stable states. *Physica D* **107**, 240–254 (1997).

Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403 (1976).

Godzik, A., Kolzinski, A. and Skolnik, J. A topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227–238 (1992).

Goldstein, R., Luthey-Schulten, Z. and Wolynes, P. Protein tertiary structure recognition using optimized hamiltonians with local interaction. *Proc. Natl. Acad. Sci. USA* **89**, 9029–9033 (1992).

Grossman, T., Farber, R. and Lapedes, A. Neural net representations of empirical protein potentials. *Ismb* **3**, 154–61 (1995).

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C. M., Weber, J., Hofacker, I. L., Stadler, P. F. and Schuster, P. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Monatsh. Chem.* **127**, 355–374 (1996a).

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

Gultyaev, A. P. The computer simulation of RNA folding involving pseudoknot formation. *Nucl. Acids Res.* **19**, 2489–2493 (1991).

Gultyaev, A. P., van Batenburg and Pleij, C. W. A. The computer simulation of RNA folding pathways using an genetic algorithm. *J. Mol. Biol.* **250**, 37–51 (1995).

Gultyaev, A. P., van Batenburg, F. H. D. and Pleij, C. W. A. *An Approximation of Loop Free Energy Values of RNA H-Pseudoknots*. Tech. rep., Gorlaeus Laboratories, Univ. Leiden (1999). Submitted.

Gutell, R. R. Evolutionary characteristics of RNA: Inferring higher-order structure from patterns of sequence variation. *Curr. Opin. Struct. Biol* **3**, 313–322 (1993).

Hart, W. E. and Istrail, S. Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than 86% of optimal. *J. Comput. Biol.* **4**, 241–259 (1997a).

Hart, W. E. and Istrail, S. Robust proofs of np-hardness for protein folding: general lattices and energy potentials. *J. Comput. Biol.* **4**, 1–22 (1997b).

Haslinger, C. and Stadler, P. F. RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bull. Math. Biol.* (1999). In press, Santa Fe Institute Preprint 97-03-030.

Heidrich, D., Kliesch, W. and Quapp, W. *Properties of Chemically Interesting Potential Energy Surfaces*, vol. 56 of *Lecture Notes in Chemistry*. Berlin: Springer-Verlag (1991).

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. and Sippl, M. J. Identification of native protein folds amongst a large number of incorrect models — the calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **216**, 167–180 (1990).

Higgs, P. G. RNA secondary structure: a comparison of real and random sequences. *J. Phys. I (France)* **3**, 43 (1993).

Higgs, P. G. Thermodynamic properties of transfer RNA: A computational study. *J. Chem. Soc. Faraday Trans.* **91** (16), 2531–2540 (1995).

Hofacker, I. L. RNA secondary structures: A tractable model of biopolymer folding. In: *Proceedings of "Monte Carlo Approach to Biopolymers an Protein Folding"* (Grassberger, P., Barkema, G. and Nadler, W., eds.). Jülich (1998).

Hofacker, I. L., Fekete, M., Flamm, C., Huynen, M. A., Rauscher, S., Stolorz, P. E. and Stadler, P. F. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.* **26**, 3825–3836 (1998).

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M. and Schuster, P. Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie* **125**, 167–188 (1994).

Hofacker, I. L., Huynen, M. A., Stadler, P. F. and Stolorz, P. E. Knowledge discovery in RNA sequence families of HIV using scalable computers. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR* (Simoudis, E., Han, J. and Fayyad, U., eds.), pp. 20–25. Menlo Park, CA: AAAI Press (1996).

Hofacker, I. L. and Stadler, P. F. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.* **23**, 401–414 (1999).

Hofbauer, J. and Sigmund, K. *Dynamical Systems and the Theory of Evolution*. Cambridge U.K.: Cambridge University Press (1988).

Holm, L. and Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.* **25**, 231–234 (1997).

Hordijk, W. and Stadler, P. F. Amplitude spectra of fitness landscapes. *J. Complex Systems* **1**, 39–66 (1998).

Huynen, M. A. Exploring phenotype space through neutral evolution. *J. Mol. Evol.* **43**, 165–169 (1996).

Huynen, M. A., Perelson, A. S., Vieira, W. A. and Stadler, P. F. Base pairing probabilities in a complete HIV-1 RNA. *J. Comp. Biol.* **3**, 253–274 (1996a).

Huynen, M. A., Stadler, P. F. and Fontana, W. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA* **93**, 397–401 (1996b).

Jaeger, J. A., Turner, D. H. and Zuker, M. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA* **86**, 7706–7710 (1989).

Kauffman, S. *The Origin of Order*. New York, Oxford: Oxford University Press (1993).

Kauffman, S. A. and Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11 (1987).

Kawasaki, K. Diffusion constants near the critical point for time-dependent Ising models. *Phys. Rev.* **145**, 224–230 (1966).

Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge University Press (1983).

Lau, K. F. and Dill, K. A. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA* **87**, 638–642 (1990).

Lawler, E. L., Lenstra, J. K., Kan, A. H. G. R. and Shmoys, D. B. *The Traveling Salesman Problem. A Guided Tour of Combinatorial Optimization*. John Wiley & Sons (1985).

Leydold, J. and Stadler, P. F. Minimal cycle basis, outerplanar graphs. *Elec. J. Comb.* **5**, R16 (1998). See `http://www.combinatorics.org`.

Li, H., Helling, R., Tang, C. and Wingreen, N. Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669 (1996).

Lipman, D. J. and Wilbur, W. J. Modelling neutral and selective evolution of protein folding. *Proc. R. Soc. London B* **245**, 7–11 (1991).

Loss, P., Schmitz, M., Steger, G. and Riesner, D. Formation of a thermodynamically metastable structure containing hairpin II is critical for infectivity of potato spindle tuber viroid RNA. *EMBO J.* **10**, 719–727 (1991).

Manderick, B., de Weger, M. and Spiessen, P. The genetic algorithm and the structure of the fitness landscape. In: *Proceedings of the 4th International Conference on Genetic Algorithms* (Belew, R. K. and Booker, L. B., eds.). Morgan Kaufmann Inc. (1991).

Mandl, C. W., Holzmann, H., Meixner, T., Rauscher, S., Stadler, P. F., Allison, S. L. and Heinz, F. X. Spontaneous and engineered deletions in the 3'-noncoding region of tick-borne encephalitis virus: Construction of highly attenuated mutants of flavivirus. *J. Virology* **72**, 2132–2140 (1998).

Martinez, H. M. An RNA folding rule. *Nucl. Acid Res.* **12**, 323–335 (1984).

Martinez, M. A., Pezo, V., Marlière, P. and Wain-Hobson, S. Exploring the functional robustness of an enzyme by *in vitro* evolution. *EMBO J.* **15**, 1203–1210 (1996).

Maynard-Smith, J. Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).

McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**, 1105–1119 (1990).

Mezey, P. G. *Potential Energy Hypersurfaces*. Amsterdam: Elsevier (1987).

Mironov, A. A., Dyakonova, L. P. and Kister, A. E. A kinetic approach to the prediction of RNA secondary structures. *Journal of Biomolecular Structure and Dynamics* **2**, 953 (1985).

Morgan, S. R. and Higgs, P. G. Evidence for kinetic effects in the folding of large RNA molecules. *J. Chem. Phys.* **105**, 7152–7157 (1996).

Murzin, A. G. New protein folds. *Curr. Opin. Struct. Biol.* **4**, 441–449 (1994).

Murzin, A. G. Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394 (1996).

Ngo, J. T. and Marks, J. Computational complexity of a problem in molecular structure prediction. *Protein Engineering* **5**, 313–321 (1992).

Nussinov, R. and Jacobson, A. B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA* **77** (11), 6309–6313 (1980).

Nussinov, R., Piecznik, G., Griggs, J. R. and Kleitman, D. J. Algorithms for loop matching. *SIAM J. Appl. Math.* **35** (1), 68–82 (1978).

Onuchic, J. N., Luthey-Schulten, Z. and Wolynes, P. G. Theory of protein folding: The landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 539–594 (1997).

Palmer, R. Optimization on rugged landscapes. In: *Molecular Evolution on Rugged Landscapes: Proteins, RNA, and the Immune System* (Perelson, A. S. and Kauffman, S. A., eds.), pp. 3–25. Redwood City, CA: Addison Wesley (1991).

Plaxco, K., Riddle, D., Grantcharova, V. and Baker, D. Simplified proteins: minimalist solutions to the "protein folding problem". *Curr. Opin. Struct. Biol.* **8**, 80–85 (1998).

Pörschke, D. Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix. *Biopys. Chem.* **1**, 381–386 (1974).

Rauscher, S., Flamm, C., Mandl, C., Heinz, F. X. and Stadler, P. F. Secondary structure of the 3'-non-coding region of flavivirus genomes: Comparative analysis of base pairing probabilities. *RNA* **3**, 779–791 (1997).

Reidys, C., Forst, C. and Schuster, P. Replication and mutation on neutral networks. *Bull. Math. Biol.* (1999). In press.

Reidys, C. M. Random induced subgraphs of generalized n-cubes. *Adv. Appl. Math.* **19**, 360–377 (1997).

Reidys, C. M., Stadler, P. F. and Schuster, P. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.* **59**, 339–397 (1997).

Rivas, E. and Eddy, S. R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* **285**, 2053–2068 (1999).

Rosenbaum, V., Klahn, T. U., Lundberg, E., Holmgren, E., von Gabain, A. and Riesner, D. Co-existing structures of an mRNA stability determinant. the 5' region of the escherichia coli and serratia marcescens ompA mRNA. *J. Mol. Biol.* **229** (3), 656–670 (1993).

Šali, A., Shakhnovich, E. and Karplus, M. How does a protein fold? *Nature* **369**, 248–251 (1994a).

Šali, A., Shakhnovich, E. and Karplus, M. Kinetics of protein folding. A lattice model study on the requirements for folding of native states. *J. Mol. Biol.* **253**, 1614–1636 (1994b).

Schmitz, M. and Steger, G. Base-pair probability profiles of RNA secondary structures. *Comput. Appl. Biosci.* **8**, 389–399 (1992).

Schuster, P. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnology* **41**, 239–257 (1995).

Schuster, P. Genotypes with phenotypes: Adventures in an RNA toy world. *Biophys. Chem.* **66**, 75–110 (1997a).

Schuster, P. Landscapes and molecular evolution. *Physica D* **107**, 351–365 (1997b).

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

Schuster, P. and Stadler, P. F. Landscapes: Complex optimization problems and biopolymer structures. *Computers Chem.* **18**, 295–314 (1994).

Schuster, P., Stadler, P. F. and Renner, A. RNA Structure and folding. From conventional to new issues in structure predictions. *Curr. Opinion Struct. Biol.* **7** (1997). 229-235.

Shrivastava, I., Vishveshwara, S., Cieplak, M., Maritan, A. and Banavar, J. R. Lattice model for rapidly folding protein-like heteropolymers. *Proc. Natl. Acad. Sci. USA* **92**, 9206–9209 (1995).

Sippl, M. J. Calculation of conformational ensembles from potentials of mean force — an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883 (1990).

Sippl, M. J. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *J. Computer-Aided Molec. Design* **7**, 473–501 (1993a).

Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–362 (1993b). URL: `http://lore.came.sbg.ac.at/Extern/software/Prosa/prosa.html`.

Sorkin, G. B. *Combinatorial optimization, simulated annealing, and fractals*. Tech. Rep. RC13674 (No.61253), IBM Research Report (1988).

Stadler, P. F. Towards a theory of landscapes. In: *Complex Systems and Binary Networks* (López Peña, R., Capovilla, R., García-Pelayo, R., Waelbroeck, H. and Zertuche, F., eds.), vol. 461 of *Lecture Notes in Physics*. New York: Springer-Verlag (1995).

Stadler, P. F. Landscapes and their correlation functions. *J. Math. Chem.* **20**, 1–45 (1996).

Stadler, P. F. Fitness landscapes arising from the sequence-structure maps of biopolymers. *J. Mol. Struct. (THEOCHEM)* **463**, 7–19 (1999).

Stadler, P. F. and Happel, R. Random field models for fitness landscapes. *J. Math. Biol.* (1999). In press, Santa Fe Institute preprint 95-07-069.

Suvernev, A. and Frantsuzov, P. Statistical description of nucleic acid secondary structure folding. *J. Biomolec. Struct. Dyn.* **13**, 135–144 (1995).

Tacker, M., Fontana, W., Stadler, P. F. and Schuster, P. Statistics of RNA melting kinetics. *Eur. Biophys. J.* **23**, 29–38 (1994).

Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L. and Schuster, P. Algorithm independent properties of RNA secondary structure prediction. *Eur. Biophys. J.* **25**, 115–130 (1996).

Thirumalai, D. and Woodson, S. Kinetics of folding of proteins and RNA. *Acc. Chem. Res.* **29**, 433–439 (1996).

Tsimring, L. S., Levine, H. and Kessler, D. A. RNA virus evolution via a fitness-space model. *Phys. Rev. Letters* **76**, 4440–4443 (1996).

Uhlenbeck, O. C. A coat for all sequences. *Nature Struct. Biol.* **5**, 174–176 (1998).

Unger, R. and Moult, J. Finding the lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. *Bull. Math. Biol.* **55**, 1183 – 1198 (1993).

van Nimwegen, E., Crutchfield, J. P. and Huynen, M. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA* (1999). Submitted, Santa Fe Institute preprint 99-03-021.

Wagner, A. and Stadler, P. F. Viral rna and evolved mutational robustness. *J. Exp. Zool./ MDE* (1999). In press, Santa Fe Institute preprint 99-02-010.

Wagner, G. P. and Krall, P. What is the difference between models of error thresholds and muller's ratchet? *J. Math. Biol.* **32**, 33–44 (1993).

Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Müller, P., Mathews, D. H. and Zuker, M. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of rna folding. *Proc. Natl. Acad. Sci. USA* **91**, 9218–9222 (1994).

Waterman, M. S. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.* **1**, 167 – 212 (1978).

Waterman, M. S. *Introduction to Computational Biology: Maps, Sequences, and Genomes.* London: Chapman & Hall (1995).

Waterman, M. S. and Smith, T. F. RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences* **42**, 257–266 (1978).

Weinberger, E. D. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol. Cybern.* **63**, 325–336 (1990).

Westhof, E. and Jaeger, L. RNA pseudoknots. *Current Opinion Struct. Biol.* **2**, 327–333 (1992).

Wright, S. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In: *Int. Proceedings of the Sixth International Congress on Genetics* (Jones, D. F., ed.), vol. 1, pp. 356–366 (1932).

Wuchty, S. *Suboptimal secondary structures of RNA.* Master's thesis, Univiersity of Vienna (1998).

Wuchty, S., Fontana, W., Hofacker, I. L. and Schuster, P. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49**, 145–165 (1999).

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

Zuker, M. and Sankoff, D. RNA secondary structures and their prediction. *Bull. Math. Biol.* **46**, 591–621 (1984).

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