Sequence Redundancy in Biopolymers

A Study on RNA and Protein Structures

PETER SCHUSTER[a,b,*] AND PETER F. STADLER[a,b]

[a]Institut für Theoretische Chemie, Universität Wien, A-1090 Wien, Austria

[b]The Santa Fe Institute, Santa Fe, NM 87501, U.S.A.

[*]Mailing Address: Professor Peter Schuster
Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 40480 669 Fax: **43 1 40480 660
E-Mail: `pks@tbi.univie.ac.at` or `pks@santafe.edu`

Abstract

Mapping sequences onto biopolymer structures is characterized by redundancy since the numbers of sequences exceed the numbers of structures. The degree of Redundancy depends on the notion of structure. Two classes of biopolymers, RNA molecules and proteins are considered in detail. A general feature of sequence to structure mappings is the existence of a few common and many rare structures. Consequences of redundancy and frequency distribution of RNA structures are **shape space covering** and the existence of extended **neutral networks**. Populations migrate on neutral networks by a diffusion-like mechanism. Neutral networks are of fundamental importance for evolutionary optimization since they enable populations to escape from local optima of fitness landscapes.

Key words

Molecular evolution – neutral networks – protein structures – RNA secondary structures – sequence-structure map

1. The origin of redundancy in biopolymer structures

The notion of selective neutrality appears already in Charles Darwin's "Origin of Species" [9]. Until now, nevertheless, no quantitatively satisfactory concept of sequence redundancy has been given in evolutionary theory. Motoo Kimura's neutral evolution [40, 41] is based on neutrality in the strict sense of identical fitness values. Tomoko Ohta [51] extended Kimura's concept to variable environments. In the generic case it is very unlikely that a fittest genotype is optimal under all conditions. Commonly genotypes will be slightly deleterious in most of the environments and best adapted only under very few circumstances. Ohta's theory, often addressed as "near neutral theory", makes an important contribution to the understanding of evolution: neutrality with respect to selection does not mean fitnesses that are identical to the very last digit, it rather implies a band of fitness values among which selection is unable to distinguish. Whether or not two or more genotypes produce phenotypes of indistinguishable fitness which are therefore neutral is not only a matter of properties in isolation but also a result of environmental conditions, selection constraints, and population size.

Mapping of genotypes into fitness values is a core issue of evolutionary biology. It is commonly simplified by partitioning it in two steps:

$$\textbf{\textit{Genotype}} \quad \Longrightarrow \quad \textbf{\textit{Phenotype}} \quad \Longrightarrow \quad \textbf{\textit{Fitness}}\ .$$

Exceptions are some model landscapes that assign fitness values to genotypes more or less randomly, like spin glass models [2] or the closely related n-k model of Stuart Kauffman [38, 39]. Genotype-phenotype mappings are generally to complicated to be analyzed by rigorous techniques. *In vitro* evolution of molecules, however, reduces this map to relations between polynucleotide sequences and biopolymer structures and functions. These relations are also a primary subject of molecular biophysics. Here we shall review the first step of this combined map. It deals with the formation of biomolecular structures through folding of genotypes, being RNA or DNA sequences of polynucleotides as well as amino acid sequences of proteins. Neutrality of structures implies redundancy of sequences in the sense that two or more sequences give rise to the same structure. In the puristic view of X-ray crystallography of biopolymers, sequence redundancy is non-existent:

Small as they may be there are always differences in atomic coordinates that make structures unique. The crystallographic notion of structure, however, is vastly different from biochemical and evolutionary intuitions. Protein and RNA structures are often represented by wire diagrams. Phylogenetic conservation of structure is discussed, for example, by comparison of backbone foldings.

High precision, nevertheless, is required for active sites of enzymes and ribozymes as well as for specific recognition sites of regulators biopolymers. Conserved positions and residues are only a few compared to a hundred and more monomer units that can be changed without substantially altering biopolymer function. Whether or not a change caused by a substitution is significant is not only a matter of structure but also a question of environmental fluctuations and selection constraints.

Near-neutral theory defines a band of almost optimal fitness values and thus variations in structure and properties are tolerated. This fact calls for a coarse grained notion of structure. An operational coarse graining which is suitable for modeling evolution is not available yet. There are, however, established notions of structural coarse graining that can be analyzed in detail. In section 2 and 3 we shall discuss coarse grained sequence-structure relations of RNA molecules and proteins, respectively. The consequences of redundancy in sequence structure mappings for the course of evolutionary adaptations are reviewed in section 4. The concluding section 5 compares sequence redundancy in the two different classes of biopolymers.

2. The sequence structure map of RNA

Mapping RNA genotypes onto phenotypes becomes accessible to straightforward analysis when the phenotype can be identified with the molecular structure of the RNA molecules. This is the case with *in vitro* evolution of RNA [67] and presumably also with simple RNA viruses whose life cycles are determined by the structure of the viral RNA [75]. Molecular structures of RNA molecules are complicated objects. Often they cannot be represented by a single conformation only and then a statistical description by means of the matrix of base pairing probabilities [46] is appropriate. We shall not discuss this issue here and thus assume that the minimum free energy conformation is representative for the properties of the molecule.

2.1. RNA secondary structures

Mapping RNA genotypes into phenotypes requires a solution to the structure prediction problem [61]. Current knowledge on three-dimensional structures of RNA molecules, however, is rather limited: only very few structures have been determined so far by crystallography and NMR spectroscopy. Needless to say, spatial structures of RNA molecules are also very hard to predict by computations based on minimization of potential energies and molecular dynamics simulations. The so-called secondary structure of RNA is a coarse grained version of structure that lists Watson-Crick (**GC** and **AU**) and **GU** base pairs. A secondary structure can be represented by a planar graph without knots or pseudo-knots[1]. Secondary structures are conceptionally much simpler than three-dimensional structures and allow to perform rigorous mathematical analysis [72] as well as large scale computations by means of algorithms based on dynamic programming [77] and implementation on parallel processors [30]. RNA secondary structure predictions are more reliable than those of full spatial structures. In addition, the definition of RNA secondary structures allow to find formally consistent distance measures (D) in shape space [17, 33, 42, 54]. Some statistical properties of RNA secondary structures were shown to depend very little on choices of algorithms and parameter sets [71].

RNA secondary structures provide an excellent model system for the study of global relations between genotypes and phenotypes. The conventional approach of structural biology determining structures for single sequences is extended to a general concept that considers sequence structure relations as (non-invertible) mappings from sequence space into shape space [17, 55, 60].

2.2. Common and rare RNA structures

Application of combinatorics to RNA secondary structures [72] allows to derive an asymptotic expression from a simple recursion for the numbers of (acceptable) structures that can be formed by sequences of chain length n [32, 59]

$$S_n \;\approx\; 1.4848 \times n^{3/2}\,(1.8488)^n \;. \tag{1}$$

[1]The precise definition for an acceptable secondary structure is: (i) base pairs are not allowed between neighbors in the sequences $(i,i{+}1)$ and (ii) if (i,j) and (k,ℓ) are two base pairs then (apart from permutations) only two arrangements along the sequence are acceptable: $(i{<}j{<}k{<}\ell)$ and $(i{<}k{<}\ell{<}j)$, respectively.

Table 1. Common secondary structures of **GC**-only sequences.

	#Sequences		#Struct.	**GC***		
n	4^n	2^n	S_n	$\tilde{S}_{\mathbf{GC}}$	R_c	n_c
7	16,384	128	6	2	1	120
10	1.05×10^6	1,024	22	11	4	859
15	1.07×10^9	32 768	258	116	43	28 935
20	1.10×10^{12}	1.05×10^6	3 613	1 610	286	902 918
25	1.13×10^{15}	3.36×10^7	55 848	18 590	2 869	30 745 861
30	1.15×10^{18}	1.07×10^9	917 665	218 820	22 718	999 508 805

* The total number of minumum free energy secondary structures formed by **GC**-only sequences is denoted by $\tilde{S}_{\mathbf{GC}}$, R_c is the rank of the least frequent common structure and thus is tantamount to the number of common structures, and n_c is the number of sequences folding into common structures.

Equ.(1) is based on two assumptions: (i) the minimum stack length is two base pairs ($n_{stack} \geq 2$, i.e., isolated base pairs are excluded) and (ii) the minimal size of hairpin loops is three ($n_{loops} \geq 3$). The numbers of sequences are given by 4^n for natural RNA molecules and by 2^n for **GC**-only or **AU**-only sequences. For both classes of RNA molecules there are more sequences than secondary structures. In the evolutionary relevant cases there will be large number of sequences folding into the same secondary structure ψ. We call the set $S(\psi)$ of sequences folding into ψ the *neutral set* of structure ψ. For natural RNA molecules this general picture does not change qualitatively if non-nested pseudo-knots are allowed as well [69]. The number of acceptable structures with pseudo-knots increases asymptotically with $S_n \propto 2.35^n$ instead of 1.85^n as was derived for the structures without pseudo-knots (see equ.1). The result was derived without applying a constraint for the stereo-chemistry of pseudo-knots and thus represents rather an upper limit for the number of acceptable structures. It shows, however, that the view of RNA shape space derived from secondary structures does not change drastically when tertiary interactions are included.

Not all acceptable secondary structures are actually formed as minimum free energy structures. The numbers of stable secondary structures, $\tilde{S}_n$, were determined by exhaustive folding [26, 27] of all **GC**-only sequences with chain lengths up to $n = 30$ (table 1). The fraction of acceptable structures obtained as minimum free energy structures through folding **GC** sequences is between 20% and 50%. This fraction is decreasing with increasing chain length n. The best estimate of the exponential increase of minimum free energy structures from our data is $\tilde{S}_n \propto 1.65^n$.

Secondary structures are properly grouped into two classes, common ones and rare ones. A straightforward definition of common structures was found to be very useful:
A structure ψ is *common* if and only if

$$|S(\psi)| \geq \overline{|S|} = \kappa^n \; / \; \tilde{S}_n \,, \tag{2}$$

Here κ denotes the size of the alphabet ($\kappa = 2$ for **GC**-only or **AU**-only sequences and $\kappa = 4$ for natural RNA molecules). A structure ψ is common if it is formed by more sequences than the average structure.

The results of exhaustive folding suggest two important general properties of the above given definition of common structures [26, 27]: (i) the common structures represent only a small fraction of all structures and this fraction decreases with increasing chain length, and (ii) the fraction of sequences folding into the common structures increases with chain length and approaches unity 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 larger than computed from equ.(2) since only common structures play a role in natural evolution and in evolutionary biotechnology.

2.3. The topology of neutral sets

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 been frequently observed that seemingly unrelated protein sequences have essentially the same fold [34, 49, 50]. 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 [31, 53]. Whether these 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. Another well known example is represented by the clover leaf secondary structure of tRNAs: The sequences of different t-RNA's have very little sequence homology [16] but nevertheless fold into the same secondary structure motif.

Inverse folding can be used to determine the sequences that fold into a given structure. For RNA secondary structures an efficient inverse folding algorithm is available [30]. It was used to show that sequences folding into the same structure are (almost) randomly distributed in sequence space. On the other hand, it was noticed already in early work on RNA secondary structures [19] that a substantial fraction of point mutants are neutral in the sense that the corresponding sequences fold into the same secondary structure. Detailed data can be found in [26].

Two approaches have been applied so far to study the topology of neutral sets: a mathematical model of genotype-phenotype mapping based on random graph theory [55] and exhaustive folding of all sequences with given chain length n [27]. The mathematical model assumes that sequences forming the same structure are distributed randomly (in the space of compatible sequence, see below) and it uses the fraction λ of neutral neighbors as (the only) input parameter. If λ is large enough this model makes two rather surprising predictions:

(1) 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.

(2) The neutral sets $S(\psi)$ of all common structures form networks that percolate sequence space.

In the following two sections we shall discuss these prediction in more detail for the RNA case.

2.4. Shape space covering

It is straightforward then to compute a spherical environment (around any randomly chosen reference point in sequence space) that contains at least one sequence (on the average) for every common structure. The radius of such a sphere, called the covering radius r_{cov}, can be estimated from simple probability arguments [58]:

$$r_{cov} \quad = \quad \min\left\{h \;\middle|\; B_h \geq \tilde{S}_n\right\} \;, \tag{3}$$

Table 2. Shape space covering radius for common secondary structures.

n	Covering Radius r_{cov}				$B_{r_{cov}} / 4^{\kappa}$ [‡]
	Exhaustive Folding		Estimate from equ.(3)[⋆]		
	GC[†]	**AU**	$\kappa = 2$	$\kappa = 4$	
20	3 (3.4)	2	4	2	3.29×10^{-9}
25	4 (4.7)	2	4	3	4.96×10^{-11}
30	6 (6.1)	3	7	4	7.96×10^{-13}
50	10.7*	7*	12	6	7.32×10^{-20}
70	15.6*	11.5*	18	10	8.75×10^{-27}
100	22.9*	17.3*	26	15	4.52×10^{-37}

* Upper bounds from ref. [27], table 5.
⋆ The covering radius is estimated by means of a straightforward statistical estimate based on the assumption that sequences folding into the same structure are randomly distributed in sequence space. Equ.(1) is used as an estimate for $\tilde{S}_n$.
† Exact values derived from exhaustive folding are given in parantheses.
‡ Fraction of **AUGC** sequence space that has to be searched on the average in order to find at least one sequences for every common structure.

with B_h being the number of sequences contained in a ball of radius h. The covering radius is much smaller than the radius of sequence space. The covering sphere represents only a small connected subset of all sequences but contains, nevertheless, all common structures and forms an evolutionarily representative part of shape space.

Numerical values of covering radii are presented in table 2. In the case of natural sequences of chain length $n = 100$ a covering radius of $r_{cov} = 15$ implies that the number of sequences that have to be searched in order to find all common structures is about 4×10^{24}. Although 10^{24} is a very large number (and exceeds the capacities of all currently available polynucleotide libraries), it is negligibly small compared to the size of the entire sequence space that contains 1.6×10^{60} sequences. Exhaustive folding allows to test the estimates derived from simple statistics [27]. The agreement for **GC**-only sequences of short chain lengths is surprisingly good. The covering radius increases linearly with chain length with a factor around 1/4. The fraction of sequence space that is required to cover shape space thus decreases exponentially with increasing size of RNA molecules (table 2).

We remark that, nevertheless, the absolute numbers of sequences contained in the covering sphere increase (also exponentially) with the chain length.

Every common structure is formed by a large number of sequences and hence it is highly important to know how sequences folding into the same structure are organized in sequence space. Forming the same structure is understood as neutrality with respect to genotype-phenotype mapping.

2.5. Neutral networks

In order to understand the geometric structure of a single neutral set $S(\psi)$ we shall need the *space of compatible sequences* of a secondary structure ψ. A sequence is *compatible* with a structure when it can, in principle, fold into this structure. The sequence requires complementary bases in all pairs of positions forming a base pair in the structure. When a sequence is compatible with a structure then the latter is necessarily among the foldings, minimum free energy or suboptimal, of the RNA molecule; a compatible sequence x might, but need not, form the structure ψ under minimum free energy conditions.

The neutral network of a structure is the subset of its compatible sequences that actually form the structure under the minimum free energy conditions. In the mathematical approach [55] neutral networks are modelled by random graphs in sequence space. The analysis is simplified through partitioning of sequence space into a subspace of unpaired bases and a subspace of base pairs. Neutral neighbors in both subspaces are chosen at random and connected to yield the edges of the random graph that is representative for the neutral network. The parameter λ measures the mean fraction of neutral neighbors in sequence space. The statistics of random graphs is studied as a function of λ. The connectivity of networks, for example, changes drastically threshold when λ passes a threshold value:

$$\lambda_{cr}(\kappa) \;=\; 1 \;-\; \sqrt[\kappa-1]{\frac{1}{\kappa}}\;. \tag{4}$$

The quantity κ in this equation represents the size of the alphabet. As shown in figure 0 we have $\kappa = 4$ (**A**,**U**,**G**,**C**) for bases in single stranded regions of RNA molecules and $\kappa = 6$ (**AU**,**UA**,**UG**,**GU**,**GC**,**CG**) for base pairs. Depending on the particular structure considered the fraction of neutral neighbors is commonly

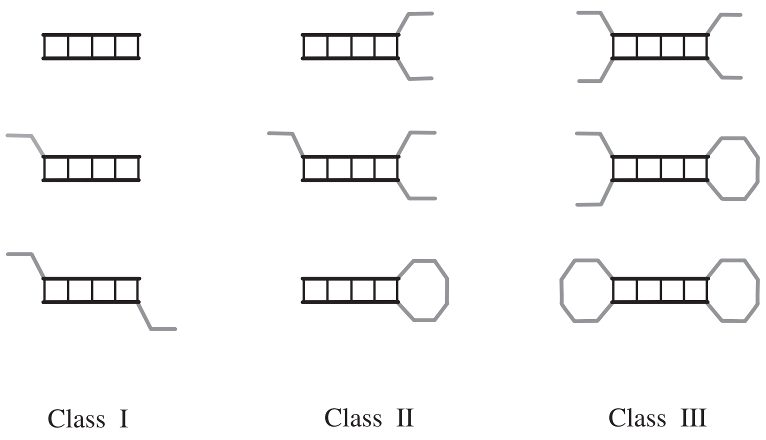


Figure 1: Three classes of RNA secondary structures forming different types of neutral networks. Structures of class I contain no mobile elements (free ends, large loops or joints) or have only mobile elements that cannot form additional base pairs. The mobile elements of structures of class II allow the extension of stacks by additional base pairs at one position. Stacks in class III structures can be extended in two positions. In principle, there are also structures that allow extensions of stacks in more than two ways but they play no role for short chain length ($n \leq 30$).

different in the two subspaces of unpaired and paired bases and we are dealing with two different parameter values, λ_u and λ_p, respectively. Neutral networks consist of a single component that spans whole 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.

Exhaustive folding allows to check the predictions of random graph theory and reveals further details of neutral networks. The typical series of components for neutral networks (either a connected network spanning whole sequence space or a very large component accompanied by several small ones) is indeed found with many common structures. There are, however, also numerous networks whose series of components are significantly different. We find networks with two as well as four equal sized large components, and three components with an approximate size ratio of 1:2:1. Differences between the predictions of random graph theory and the results of exhaustive folding were readily explained in terms of special properties of RNA secondary structures [27], see Figure 1.

Random graph theory, in essence, predicts that sequences forming the same structure should be randomly distributed in sequence space. Deviations from such an ideal neutral network can be identified as structural features that are not accounted for by non-specific base pairing logics. All structures that cannot form additional base pairs when sequence requirements are fulfilled behave perfectly normal (class I structures in figure 1). There are, however, structures that can form additional base pairs (and will generally do so under the minimum free energy criterion) whenever the sequences carry complementary bases at the corresponding positions. Class II structures (figure 1), for example, are least likely to be formed when the overall base composition is 50% **G** and 50% **C**, because the probability for forming an additional base pair and folding into another structure is largest then. If there is an excess of **G** ($\{50{+}\delta\}\%$) it is much more likely that such a structure will actually be formed. The same is true for an excess of **C** and this is precisely reflected by the neutral networks of class II structures with two (major) components: the maximum probabilities for forming class II structures are **G**:**C**=$(50{+}\delta)$:$(50{-}\delta)$ for one component and **G**:**C**=$(50+\delta)$:$(50-\delta)$ for the second one. By the same token structures of class III have two (independent) possibilities to form an additional base pair and thus they have the highest probability to be formed if the sequences have excess δ and ε. If no additional information is available we can assume $\varepsilon \approx \delta$. Independent superposition yields then four equal sized components with **G**:**C** compositions of (50+2δ):(50-2δ), 2×(50:50), and (50-2δ):(50+2δ) precisely as it is observed indeed with four component neutral networks. Three component networks are *de facto* four component networks in which the two central (50:50) components have merged to a single one. Neutral networks are thus described well by the random graph model: The assumption that sequences folding into the same structure are randomly distributed in the space of compatible sequences is justified unless special structural features lead to systematic biases.

3. The sequence-structure map of proteins

The number of possible protein sequences is enormous. For $n = 100$ residues there are 20^{100} sequences. On the other hand, the repertoire of stable native folds seems to be highly restricted or even vanishingly small [8]. For example, it has been frequently observed that seemingly unrelated sequences have essentially the same fold [34, 49, 50]. It seems, therefore, that RNA and proteins, despite their different chemistry, share fundamental properties of their sequence-structure maps.

Computational studies similar to the explorations of RNA world reported above are precluded by the notorious complexity of the protein folding problem, and by the fact that there is no biophysically meaningful and computationally simple coarse resolution of protein structures (The term secondary structure refers in the protein world to local features that may or may not be present but do not capture the global organization of the molecule). Hence we have to resort to a less ambitious approach [3] based on inverse folding only.

3.1. Inverse folding using knowledge based potentials

In order to characterize the topology of neutral sets $S(\psi)$ we need a technique for deciding whether a given sequence x is a member of $S(\psi)$, that is, whether x folds into the structure ψ. This problem is less demanding than predicting the unknown protein structure of a given amino acid sequence. It can be investigated by inverse folding techniques [14, 6]. In contrast to the RNA case, however, we cannot derive inverse folding from a solution to the protein folding problem, as the latter is still unsolved.

The starting point is a potential function $W(x, \psi)$ evaluating the energy of a sequence x when folded into a structure ψ which is defined by the spatial coordinates of its C^α and C^β atoms, respectively. Recent studies using knowledge based potentials [5, 6, 23, 24, 25, 29, 64, 65] demonstrated that the energy of the native fold (i.e., putative ground state) of a sequence x can be estimated from the distribution of the energy values of x in its conformation space. This allows the construction of an energy scale by which conformations of different sequences can be compared. As a measure for the quality of fit of sequence x and structure ψ the *z-score* [7]

$$z(x, \psi) = \frac{W(x, \psi) - \overline{W}(x)}{\sigma_W(x)} \tag{5}$$

is used. Here $\overline{W}(x)$ is the average energy of sequence x in all conformations in a database and $\sigma_W(x)$ is the standard deviation of the corresponding distribution. Empirically, native folds have z-scores in a narrow characteristic range. Furthermore, use of the *z-score* introduces a kind of negative design [11, 28] that avoids sequences with multiple stable structures [1]. The z-score correlates well with the rms deviation of alternative structures from the native structure. The computed z-score also improves with increasing resolution of the X-ray structure when computed for the same protein [3]. The `PROSA II` potentials [7, 29, 63, 64, 65] are a particularly suitable for studying neutral sets in protein space[2] [3].

[2] For the space of amino acid sequences we use the synonym *protein space* [45]

Hence one may assume that x is a member of $S(\psi)$ if the z-score of x in conformation ψ is in the native range [7]. Of course, only native structures ψ that are already in the database can be explored by this method. Formally, we have translated inverse folding into an optimization problem on the set of all sequences: we are looking for the minima x of the z-score $z(x, \psi)$. From the computational point of view, this optimization problem appears to be very easy. Indeed, it is sufficient to use the simplest heuristic, the *adaptive walk*, which repeatedly tries random mutations (exchanges of single amino acid) that are accepted if and only if the z-score decreases.

Whether the sequences predicted by this inverse folding procedure do indeed fold to the desired structure can ultimately only be answered by experiment. Independent criteria, however, would at least indicate whether the assumption is reasonable. One finds, for instance, that the `SOPM` [22] and `PHD` [56] predictions of secondary structures of inverse folded sequences agree with the target secondary structure [3] despite the fact that the `PROSA II` potentials make no explicit reference to secondary structure.

3.2. Distribution of inverse folded proteins

Sequences generated by independent adaptive walks show little or no homology to the wild-type sequence or among each other. This is consistent with the observation that a significant sequence homology is not necessary for two proteins to have a common fold [52]. Although they lie somewhat closer together than random sequences with a typical amino acid composition (taken from the `SwissProt` database), pairs with the maximum Hamming distance do occur. Tree reconstruction methods, such as neighbor joining and the split decomposition technique [4] suggest that the sequences with wild type-like z-scores are distributed essentially randomly in sequence space.

A number of groups have argued that the pattern of hydrophobic versus hydrophilic amino acids (**HP**-pattern) has a dominating influence on protein structure [13, 37, 70]. Inverse folded sequences are very flexible at the level of individual amino acids but requires a significant level of conservation of amino acid classes.

Hence it is natural to ask whether all 20 amino acids are in fact necessary to build native protein structures, or whether this can already been done with a (small) subset of different amino acids. Not surprisingly, no sequences with wild-type like structures could be found when only hydrophilic amino acids or only hydrophobic amino acids were used. Surprisingly, however, we observed substantial differences between different alphabets that all contain both hydrophilic and hydrophobic amino acids. For instance, the two-letter alphabet **AD** gives very poor results, while other combinations of just one hydrophilic and one hydrophobic amino acid, such as **LS** or **DL**, yield wild-type like z-scores. It is not surprising that **ADL** and **ADLG** yield good sequences since **DL** is already sufficient. The inverse folded sequences in these alphabets do, however, contain a substantial fraction of **A** and **G**. The alphabet **ADLG** has been proposed as a candidate for a primordial set of amino acids, before the full genetic code was developed [47]. It is reassuring to see that this alphabet allows inverse folding of a variety of present day protein structures. It is worth noting in this context, that the **QLR** alphabet used in experimental work on random polypeptides by Sauer and co-workers [10, 57] does not yield wild-type like z-scores for globular proteins. This may not be surprising since Sauer's experimental **QLR**-peptides form multimeric structures.

3.3. Neutral networks in protein space

Inverse folded sequences with z-scores below the threshold z^* were used as starting points for neutral paths. The substitution frequencies for the production of mutants were computed from the natural frequencies of the amino acids as contained in the `SwissProt` database.

Figure 2 shows the results for four different protein structures [3]. The lengths of the neutral paths $\mathcal{L}$ are roughly equal to the lengths n of the proteins, at z-score levels comparable to the wild-type sequence. Even at z-scores about six standard deviations better than the wild-type z-score, the length of the neutral paths is still greater than three quarters of the length of the protein. The average values of $\mathcal{L}$ taken over the z-score interval $z_{\mathrm{w.t.}} - 3 \le z \le z_{\mathrm{w.t.}}$ are collected in table 3. The average Hamming distances between the endpoints of unrelated neutral paths, $\langle d\rangle_{\mathrm{nn}}$, are in the range of 90 to 95 percent of the chain length, indicating that the neutral networks span essentially the entire sequence space. It is not surprising that the Hamming distances between the end points of neutral paths is somewhat larger than the average distance, $\langle d\rangle_{\mathrm{adw}}$, between the end points of adaptive walks,

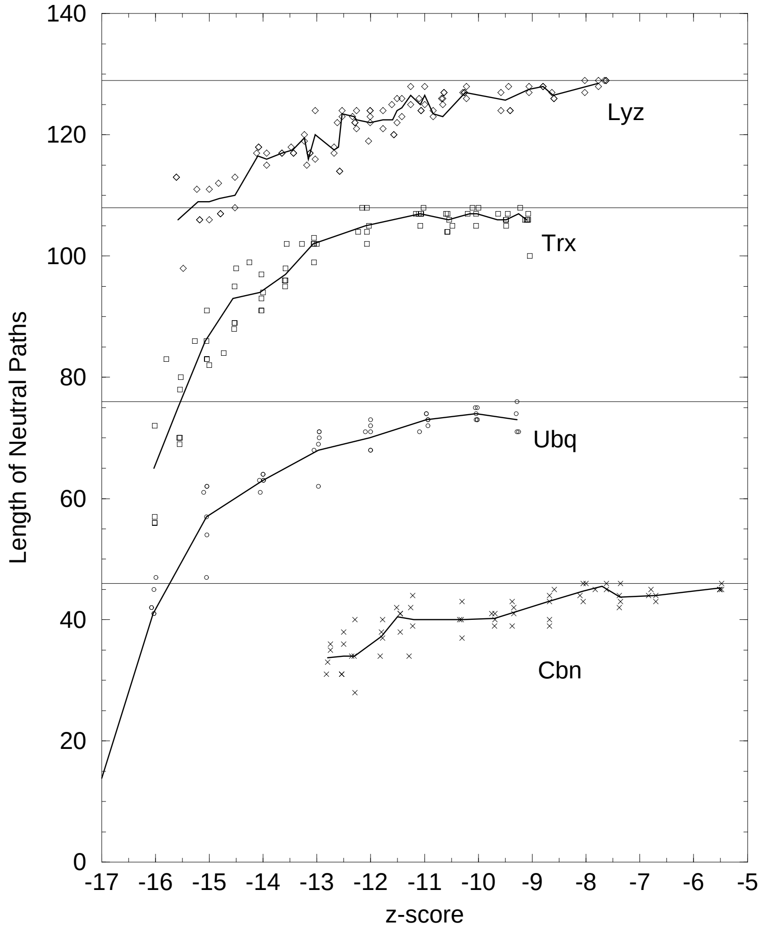


Figure 2: Length of neutral path as a function of the threshold z-score. The solid lines indicate the averages over the available data for each protein. The chain lengths of the four proteins 1cbn, 1ubq, 2trxA, and 1lyz are indicated by thin solid lines for comparison. The rightmost data points for each structure correspond to the wild-type z-scores.

since a neutral walk has a built-in bias towards sequences with a more uniform distribution of amino acids.

Extensive studies on neutral paths in restricted protein alphabets have been reported only for the "primordial alphabet" **ADLG** [3]. The average length of neutral paths in this alphabet is 76% to 87% of the sequence length. Note that the distance between random sequence is 75% of the chain length in a four letter alphabet. The neutral paths thus extends well beyond the mean distance of random sequences even in some highly restricted amino-acid alphabets.

Table 3. Characteristics of protein neutral networks.

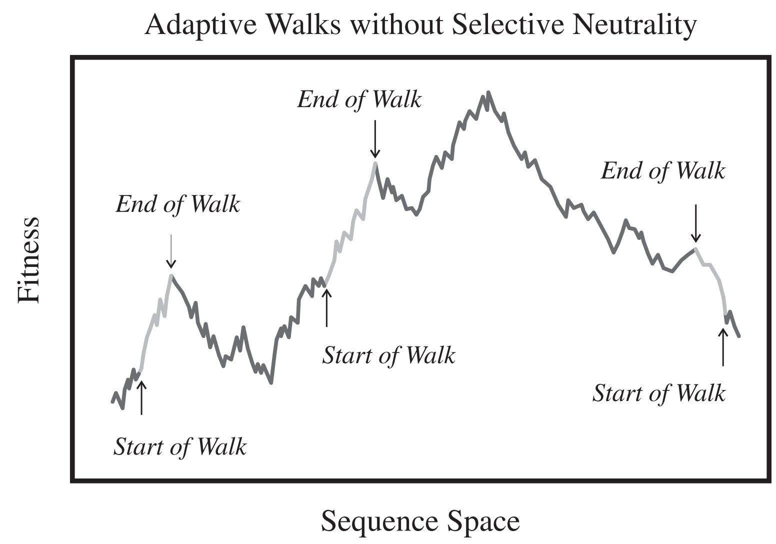
Protein	n	z	$\mathcal{L}^*$	L_{adap}	$\langle d\rangle_{\text{adw}}$	$\langle d\rangle_{\text{nn}}$
1cbn	46	−5.50	44.6	17.7	38.7	42.3
1ubq	76	−9.30	72.5	61.9	61.1	66.3
2trxA	108	−9.22	106.3	71.7	87.7	97.5
1lyz	129	−7.70	126.2	58.2	106.0	118.7

* The length of neutral path is averaged over all data with z-scores between wild type and three standard deviations better than wild type.

4. Adaptation dynamics on redundant landscapes

Landscapes assign fitness values to genotypes, i.e., to individual polynucleotide sequences in sequence space. Based on (point) mutation (and recombination) being the evolutionary adequate move set and distance measure, fitness landscapes are highly rugged in the sense that they contain a high number of local optima on all scales [21, 68, 74]. Populations optimize mean fitness by migration through sequence space. In absence of neutrality populations climb on landscapes until they reach one of the minor peaks where migration ends because all surrounding genotypes have lower fitness (figure 3). Neutral networks of RNA molecules play an important role in evolutionary optimization, as they enable populations to escape from evolutionary traps in the form of local fitness optima.

On neutral networks and, likewise, on flat landscapes populations migrate by a diffusion-like mechanism [12, 36]. Whenever adaptive migration ends on a local fitness optimum, the populations starts to drift on the neutral network belonging to the structure that corresponds to this optimum. Random drift is continued until the population reaches an area in sequence space where some fitness values are higher than that of the network. Then another period of adaptive evolution sets in. A complete optimization run thus appears as a stepwise process: phases of increasing mean fitness are interrupted by “static” periods with mean fitness values fluctuating around a constant value (figure 4). When the network belongs to a common structure its extension through whole sequence space may eventually allow the population to find the global optimum.



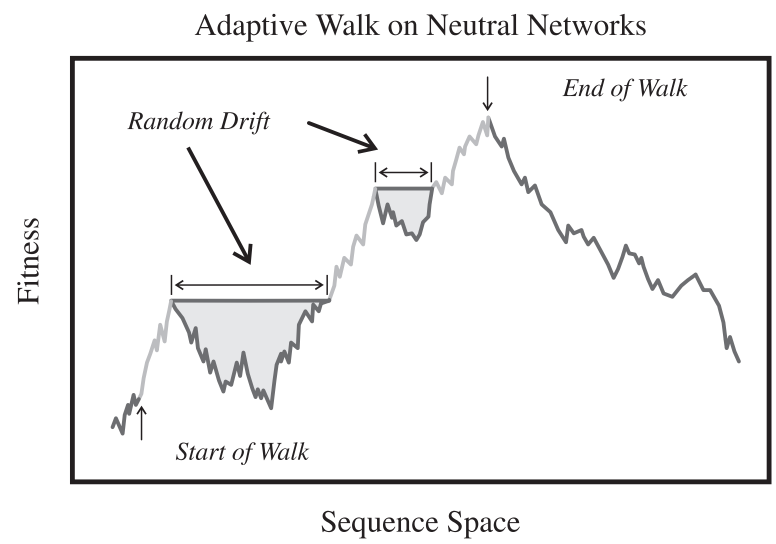


Figure 3: The role of neutral networks in evolution. Optimization occurs through adaptive walks and random drift. Adaptive walks allow to choose the next step arbitrarily from all directions where fitness is (locally) non-decreasing. Populations can bridge over narrow valleys with widths of a few point mutations. In the absence of selective neutrality (upper part) 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 (lower part). In this manner, populations bridge over large valleys and may eventually reach the global maximum of the fitness landscape.

Optimization of protein structures requires translation of genes into protein. Phage display or bacterial display [62, 66], for example, provide an excellent tool for *in vitro* evolution of proteins. The existence of extensive neutral networks meets a claim raised by Maynard Smith [45] for protein spaces that are suitable for efficient evolution. Empirical evidence for a large degree of *functional* neutrality in protein space was presented recently by Wain-Hobson and co-workers [44].

The scenario described above has been studied and verified by a series of computer simulations [18, 19, 36]. A very detailed study on RNA optimization has been performed by means of a simple model of molecular evolution in which the fitness of a sequence expressed in terms of its replication rate depends on the structure. The evolution of a population of RNA sequences of fixed chain length n is simulated under conditions of a flow reactor that adjust the total population size to fluctuate around a constant capacity N. Individual sequences replicate with a structure dependent rate constant that is defined to be a function of the distance between its secondary structure and a predefined target structure. Mutation is introduced by simulating a copying mechanism that copies each base with fidelity $1-p$. In this simulation as well as in evolution *in vitro* and *in vivo* there are two sources of neutrality: one is the sequence to structure mapping discussed here, the other is the structure to replication rate (fitness) mapping.

Approximating the dynamics of the flow reactor using the Moran model [48], in which the (error-prone) replication of a randomly chosen sequence is followed by the removal of a randomly chosen sequence, the diffusion coefficient D_0 can be computed for the flat landscape. One finds

$$D_0 = \frac{6anp(1+1/N)}{3+4Np} \approx \frac{6anp}{3+4Np}\,, \tag{6}$$

where a is common replication rate [36]. For small mutation rates the diffusion coefficient, D, on the structure-dependent landscape can be approximated by $D = D_0\bar{\lambda}$, where $\bar{\lambda}$ denotes the average fraction of neutral mutants for the dominant structure. The diffusion of finite populations in sequence space is directly related to Kimura's neutral theory [41], which stresses a different aspect, namely the number of nucleotide substitutions that reach fixation per generation, k, also referred to

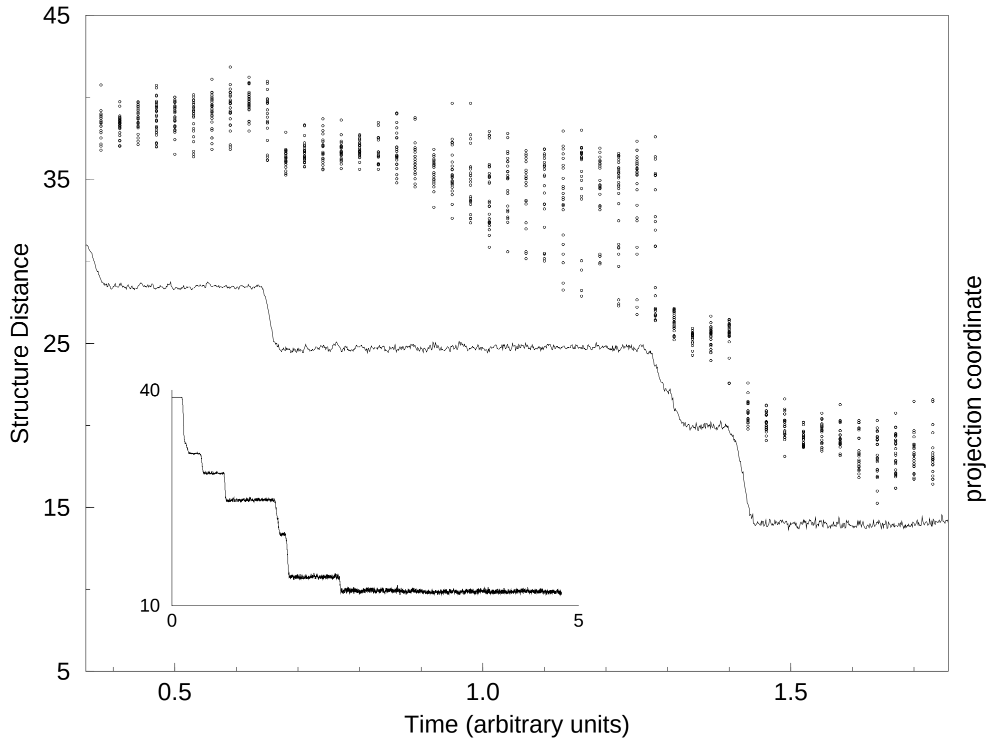


Figure 4: The stepwise course of evolutionary optimization of RNA structures [36]. A flow reactor with capacity N=1000 is initialized with that many copies of a random sequence of length n=76. The mutation rate is p=0.001 and the target secondary structure is the tRNA$^{\mathrm{Phe}}$ clover-leaf, the replication rate function is $A(D)$=1.06^{146-D} where D is the tree-edit distance to the target structure. The population average of the distance to the target is plotted against time (full line) for a specific interval of the entire run (shown in the inset). The scattered points in the upper part of the plot indicate the position of the population in sequence space as a projection to a single coordinate.
The fitness plateaus correspond to diffusion on neutral networks. The population spreads out in sequence space during these periods. Sudden jumps indicate the transition to another network [73]. Darwinian selection drastically reduces the sequence diversity at the transition points.

as the "rate of evolution". The theory yields $k = a\ p\ \nu\ \bar{\lambda}$, and hence $D = 6\,k/(3+4Np)$, for small mutation rates.

Studies of evolutionary dynamics on rugged model landscapes, which did not involve a genotype-phenotype model, showed a loss of sequence information at a critical error rate corresponding to the loss of the dominant phenotype [15]. In the presence of percolating neutrality, however, all sequence information is lost at any non-zero error rate due to diffusion and the finiteness of the population. Yet there is another threshold, the *phenotypic error threshold*, p_c, beyond which the dominant phenotype is lost as well. That is when evolutionary adaptation breaks down. The critical value p_c can be estimated from a rather complicated set of equations derived in [20].

Diffusion in sequence space, the connection with Kimura's neutral theory, and the phenotypic error threshold are consequences of the existence of neutral networks. Shape space covering implies a constant rate of innovation [35]: 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 before at a constant rate. 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 4).

5. Discussion

Genotype-phenotype mappings of RNA molecules have been investigated using a prediction algorithm for secondary structures and by means of inverse folding. RNA secondary structures and protein folds are highly redundant in the sense that many sequences fold into the same shape. The most striking similarity in sequence-structure mappings of RNA molecules and proteins is the existence of few common and many rare structures. RNA secondary structures from **GC** sequences of chain length $n = 27$ yield a log-rank/log-frequency of occurrence-plot that almost coincides with the corresponding plot derived from **HP**-lattice proteins[3] (For comparison see [26, 43] and figure 5.) Such a distribution of the frequencies of structures can be characterized by a generalization of Zipf's law [76]. In the example discussed here, **GC** sequences of chain length $n = 30$, more

[3]By this expression we characterize protein models which distinguish only hydrophilic and hydrophobic residues. Sequences are folded on lattices to yield stable folds according to an energy criterion [13]. In the present case, the energy equals the number of **HH** contacts. The configurations are restricted to self-avoiding walks in a $3\times3\times3$ cube, hence n=27.

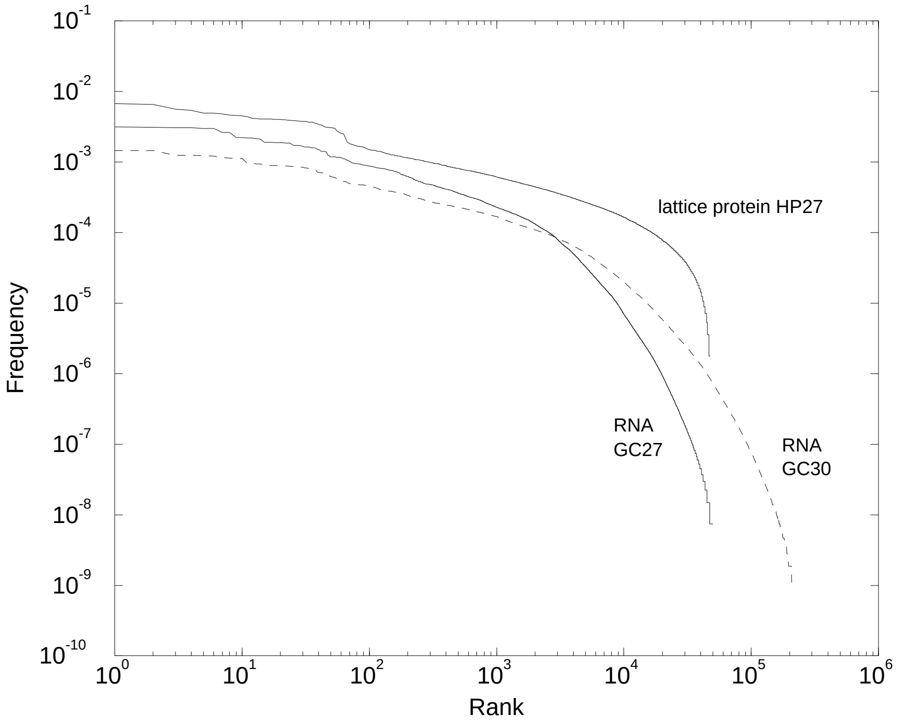


Figure 5: The distribution of frequencies of RNA secondary structures and lattice models of proteins. The diagram shows the distribution of preimage sizes of sequence structure mappings for **GC** sequences of length 27 and 30 [26] compared with the analogous plot computed for **HP** lattice proteins of chain length 27 [43].

than 93% of all sequences fold into only 10% of all structures. Extrapolation of our data to longer chains indicates an increasing percentage of sequences folding into a decreasing fraction of sequences. There are also strong indications that less coarse notions of structure give rise to very similar frequency distributions. Implications for evolutionary optimization are evident: Populations, in essence, live in a space of common structures or phenotypes. Rare phenotypes are extremely hard to find in random searches and thus play no role in evolution. Still, one important feature for understanding evolution is missing: we do not know yet the rules that determine whether a phenotype is common or rare. In other words, given the structure of a phenotype we should be able to predict the fraction of genotypes that fold into it. Investigations aiming at such a completion of the current concept in this respect are under way. First results show that the modular building principle of natural biopolymers is highly important for the probability of realization of structures.

Algorithms for folding RNA sequences into secondary structures as well as inverse folding procedures for proteins based on knowledge-based potentials predict extended connected networks of sequences with identical structure. In addition, RNA secondary structures exhibit shape space covering, that is, any common structure can be found within a small radius in sequence space. (It not yet clear whether protein space shares this property.) These observations have striking consequences for adaptation, based on a fairly realistic model of test-tube evolution. (1) Finite populations diffuse along neutral networks, where their dynamics conforms the predictions of Kimura's neutral theory. After a sufficiently long period of time (set by the diffusion coefficient) all sequence information is lost, yet the phenotype is conserved. It is the maintenance of a phenotype, not of a genotype, which defines the mutation threshold beyond which adaptation breaks down. (2) On a single neutral network the population splits into well separated clusters. A population is not a single localized quasispecies in sequence space [15], but rather a collection of different quasispecies. Each undergoes independent diffusion, while all share the same dominant phenotype. (3) Neutral networks of different structures are interwoven. While drifting on a neutral network a population produces a fraction of mutants off the network and thereby explores new phenotypes. A selection-induced transition between two structures occurs in regions of sequence space where their networks come close to one another. The independent diffusion of subpopulations increases the likelihood that a population encounters such transition regions.

Neutral evolution, therefore, is not a dispensable addendum to evolutionary theory as it has often been suggested. On the contrary, neutral networks, arising as a consequence of the the redundancy of sequence-structure (and possibly also structure-function) relationships of biopolymers, provide a powerful mechanism through which evolution can become really efficient.

Acknowledgments

Extensive discussions on the subject of this review with Drs. Walter Fontana, Christian Forst, and Martijn Huynen are gratefully acknowledged. The work on inverse protein folding is joint research with Mag. Aderonke Babajide, Dr. Ivo L. Hofacker, and Prof. Manfred Sippl. The work was supported financially by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Projects No. 10578-MAT and 11065-CHE, by the European Commission, Contract Study PSS*0884, by the Diversity Biotechnology Consortium (New Mexico) and by the Santa Fe Institute.

References

[1] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.*, 252:460–471, 1995.

[2] C. Amitrano, L. Peliti, and M. Saber. A spin-glass model of evolution. In A. S. Perelson and S. A. Kauffman, editors, *Molecular Evolution on Rugged Landscapes*, volume IX of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 27–38. Addison-Wesley Publ. Co., Redwood City, CA, 1991.

[3] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 1997. In press. Also published as Santa Fe Institute Preprint No. 96-12-085.

[4] H. J. Bandelt and A. W. M. Dress. A canonical decomposition theory for metrics on a finite set. *Adv. Math.*, 92:47–105, 1992.

[5] A. Bauer and A. Beyer. An improved pair potential to recognize native protein folds. *Proteins*, 18:254–261, 1994.

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

[7] G. Casari and M. J. Sippl. Structure-derived hydrophobic potentials — hydrophobic potentials derived from X-ray structures of globular proteins is able to indentify native folds. *J. Mol. Biol.*, 224:725–732, 1992.

[8] C. Chothia. Proteins. one thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[9] C. R. Darwin. *The Origin of Species*, volume 811 of *Everyman's Library*, page 81. J. M. Dent & Sons, London, 1967.

[10] A. R. Davidson and R. T. Sauer. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci., USA*, 91:2146–2150, 1994.

[11] W. F. DeGrado, Z. R. Wasserman, and J. D. Lear. Protein design, a minimalist approach. *Science*, 243:622–628, 1989.

[12] B. Derrida and L. Peliti. Evolution in a flat fitness landscape. *Bull. Math. Biol.*, 53:355–382, 1991.

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

[14] K. E. Drexler. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci.*, 78:5275–5278, 1981.

[15] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasispecies. *Adv. Chem. Phys.*, 75:149–263, 1989.

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

[17] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[18] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaptation. *Phys. Rev. A*, 40:3301–3321, 1989.

[19] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophys. Chem.*, 26:123–147, 1987.

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

[21] R. García-Pelayo and P. F. Stadler. Correlation length, isotropy, and metastable states. *Physica D*, 1997. in press, Santa Fe Institute Preprint 96-05-034.

[22] C. Geourjon and G. Deleage. SOPM: a self optimised prediction method for protein secondary structure prediction. *Protein Engineering*, 7:157–164, 1994.

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

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

[25] T. Grossman, R. Farber, and A. Lapedes. Neural net representations of empirical protein potentials. *Ismb*, 3:154–61, 1995.

[26] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Monath. Chem.*, 127:355–374, 1996.

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

[28] M. H. Hecht, J. S. Richardson, D. C. Richardson, and R. C. Ogden. De novo design, expression, and characterization of felix: a four-helix bundle protein of native-like sequence. *Science*, 249:884–891, 1990.

[29] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. 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.

[30] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[31] I. L. Hofacker, M. A. Huynen, P. F. Stadler, and P. E. Stolorz. Knowledge discovery in rna sequence families of HIV using scalable computers. In E. Simoudis, J. Han, and U. Fayyad, editors, *Proceedings of the 2nd International*

Conference on Knowledge Discovery and Data Mining, Portland, OR, pages 20–25, Menlo Park, CA, 1996. AAAI Press.

[32] I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics of RNA secondary structures. Santa Fe Institute Preprint 94-04-026, 1996.

[33] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucl. Acids Res.*, 12:67–74, 1984.

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

[35] M. A. Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43:165–169, 1996.

[36] M. A. Huynen, P. F.Stadler, and W. Fontana. Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, 93:397–401, 1996.

[37] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht. Protein design by binary patterning of polar and nonpolar amino acids. *Science*, 262:1680–1685, 1993.

[38] S. A. Kauffman. *The Origins of Order. Self-Organization and Selection in Evolution.* Oxford University Press, Oxford, UK, 1993.

[39] S. A. Kauffman and E. D. Weinberger. The N-K model of rugged fitness landscapes and its application to maturation of the immune response. *J. Theor. Biol.*, 141:211–245, 1989.

[40] M. Kimura. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA*, 41:144–150, 1955.

[41] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK, 1983.

[42] D. A. M. Konings and P. Hogeweg. Pattern analysis of RNA secondary structure. Similarity and consensus of minimal-energy folding. *J. Mol. Biol.*, 207:597–614, 1989.

[43] H. Li, R. Helling, C. Tang, and N. Wingreen. Emergence of preferred structues in a simple model of protein folding. *Science*, 273:666–669, 1996.

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

[45] J. Maynard-Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

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

[47] S. Miller and L. Orgel. *The Origin of Life on the Earth.* Prentice Hall, Englewood Cliffs NJ, 1974.

[48] P. A. P. Moran. Random processes in genetics. *Proc. Camb. Phil. Soc.*, 54:60–71, 1958.

[49] A. G. Murzin. New protein folds. *Curr. Opin. Struct. Biol.*, 4:441–449, 1994.

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

[51] T. Ohta. Population size and rate of evolution. *J. Mol. Evol.*, 1:305–314, 1972.

[52] C. A. Orengo, D. T. Jones, and J. M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372:631–634, 1994.

[53] S. Rauscher, C. Flamm, C. Mandl, F. X. Heinz, and P. F. Stadler. Secondary structure of the 3'-non-coding region of flavivirus genomes: Comparative analysis of base pairing probabilities. *RNA*, 1997. in press, Santa Fe Institute Preprint 97-02-010.

[54] C. Reidys and P. F. Stadler. Bio-molecular shapes and algebraic structures. *Computers Chem.*, 20:85–94, 1996.

[55] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps — Neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[56] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232:584–599, 1993.

[57] R. T. Sauer. Protein folding from a combinatorial perspective. *Folding & Design*, 1:R27–R29, 1996.

[58] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnology*, 41:239–257, 1995.

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

[60] P. Schuster and P. F. Stadler. Landscapes: Complex optimization problems and biopolymer structures. *Computers & Chem.*, 18:295–314, 1994.

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

[62] J. K. Scott and G. P. Smith. Searching for peptide ligands with an epitope library. *Science*, 249:386–390, 1990.

[63] M. J. Sippl. 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.

[64] M. J. Sippl. 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, 1993.

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

[66] G. P. Smith. Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science*, 228:1315–1317, 1985.

[67] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 4:213–253, 1971.

[68] P. F. Stadler. Towards a theory of landscapes. In R. Lopéz-Peña, R. Capovilla, R. García-Pelayo, H. Waelbroeck, and F. Zertuche, editors, *Complex Systems and Binary Networks*, pages 77–163, Berlin, New York, 1995. Springer Verlag.

[69] P. F. Stadler and C. Haslinger. RNA structures with pseudoknots. *Bull. Math. Biol.*, 1997. Submitted. Also published as Santa Fe Institute Preprint No. 97-03-30.

[70] S. Sun, R. Brem, H. S. Chan, and K. A. Dill. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.*, 8:1205–1213, 1995.

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

[72] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Adv. Math. (Suppl. Studies)*, 1:167–212, 1978.

[73] J. Weber. *Dynamics of Neutral Evolution – A case study on RNA secondary structures*. PhD thesis, Friedrich Schiller Universität, Jena, Germany, 1997.

[74] E. D. Weinberger and P. F. Stadler. Why some fitness landscapes are fractal. *J. Theor. Biol.*, 163:255–275, 1993.

[75] C. Weissmann. The making of a phage. *FEBS Letters (Suppl.)*, 40:S10–S12, 1974.

[76] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Reading, MA, 1949.

[77] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.