STRUCTURE FORMATION OF BIOPOLYMERS IS COMPLEX, THEIR EVOLUTION MAY BE SIMPLE

ERICH BORNBERG - BAUER
Institut für Mathematik, Strudelhofg. 4 and
Institut für Theoretische Chemie, Währingerstr. 17
Universität Wien, A-1090 Wien, Austria
email: erich@tbi.univie.ac.at, url: http://www.tbi.univie.ac.at

Evolutionary strategies depend on the ability of evolving entities to conserve acquired features and to quickly adapt to new requirements as well. We use computer simulations of simplified exact biopolymers models to investigate the influence of mutations on structure formation. Our computations on large ensembles of random RNA secondary structures show that the sequence to structure mapping is ideally suited for evolutionary optimisation under point mutations: from any random structure it is not far to any target and yet most mutations will preserve the structure.
The aim of this paper is to discuss the analogies as well as some recently developed methods to apply our approach to proteins: there we use Dill's HP - model of lattice proteins and apply a novel fast and efficient folding rule. There are remarkable similarities as both landscapes are rugged and structure formation largely depends on local interactions such that it is possible to accomplish a characterisation of the mapping similar to the RNA case.

1 Introduction

The influence of mutations on the structure, and thereby on the phenotype and the functionality of biopolymers is of interest for the understanding of both natural and artificial evolution, in particular since evolutionary design (selection and optimisation starting from *random* pools of sequences) has become a frequently used technique. To understand possibilities and limitations of natural or artificial selection it is essential to establish a theory about mechanisms that enable simple molecules to exhibit apparently inconsistent features: quick adaption to new requirements and strong conservation of once established properties such as structural features. It is commonly assumed that function depends on structure and therefore structure is conserved in evolution.
Evolutionary theories coined the term of "fitness landscapes" [29,17]: adaptation in this view takes place over a space of genotypes and follows an adaptive walk towards species with better fitness values that depend on the phenotype. Instead of assigning fitness values, structural properties such as stability and similarity of elements to given targets are often viewed as the criteria of se-

lection [7,14]. To explain tolerance of fitness values towards mutations at the genotype level several theories about evolution emphasise the importance of neutrality. For the case of proteins this has been pointed out by J. M. Smith [21] although this matter is rather involved: neutral mutations occur by redundancy at the level of the genetic code and structural tolerance with respect to the folding at the phenotype level.

A system that suits perfectly well to study neutrality in sequence space by extensive computation can be found in RNA secondary structures [20,9]. RNA is special since it combines the genotypic (the information encoded in the sequence) and the phenotypic level (the structure) into a single molecular object. In a given chemical environment the structure is implied directly and only by the sequence. RNA is also accessible to extensive computational studies. RNA received much interest since it appears to be the most promising candidate as a starting molecule of pre-biotic evolution. In several selection/amplification experiments RNA was proven to posses enzymatic activities or even affinity to substrates (dyes) that are unlikely to appear in a natural environment (for a review see [18] and references therein).

RNA secondary structure is commonly believed to be a useful approximation to the 3D structure and it can be predicted with reasonable accuracy and computational resources. Very little, however, is known from experiments about statistical features of biopolymers. Unlike proteins the 3D structure has been determined for only very few t-RNA molecules. We use, therefore, computer algorithms to investigate statistical properties of large ensembles of RNA secondary structures [10].

Folding can also be viewed as a mapping from one abstract metric space of

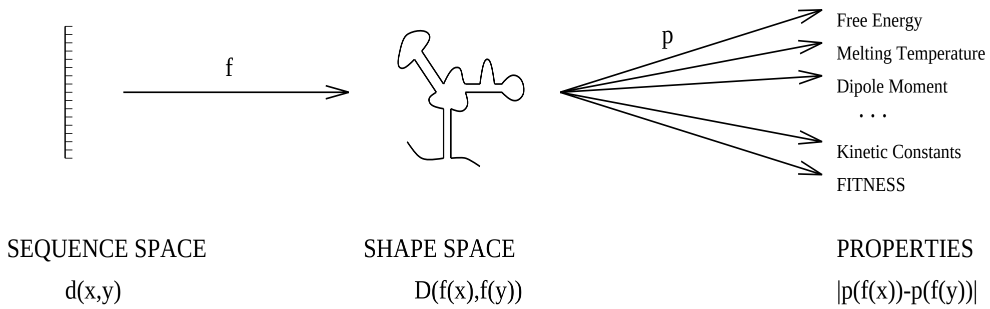

Figure 1: A generalised scheme of biopolymers folding: the sequence space consists of all κ^n sequences (when κ is the size of the alphabet the chain of length n is composed of), the shape space is the space of all possible structures.

combinatorial complexity, the sequence space to another, the shape space (see fig. 1). Size of both spaces and the isotropy of the mapping can be assumed to be of importance for the success and efficiency of adaptational processes.

Recent analytical calculations proved the size of the sequence space to be much larger, even for a (non-realistic) 2-letter alphabet [20]. It is therefore desirable to characterise the distribution, vicinity and connectivity of related structures in sequence space.

Little is known about how biopolymers actually attain their native state and it seems to be unclear if this is really the global minimum or a very deep kinetic trap where the route is encoded in the sequence [4,24]. Lattice proteins of the HP type have received much interest as they are the simplest known model to simulate protein structure formation, in particular since it has become clear that an accurate structure prediction from basic principles is still far out of reach and the pattern of hydrophobic and polar residues largely determines the geometry of the backbone [24] (for an extensive review on lattice proteins see [5], and references therein.). Furthermore there is growing evidence that at least early steps of the protein folding process that lead to the correct "area" in shape space are largely determined by kinetic effects (from this, so called "molten globe state" the final state then is accessed). There have been considerations that proteins might be selected in the course of evolution not only for structural features but also with respect to their ability to fold into a sufficiently deep minimum by a kinetic folding route [16]. Another reason why one wishes to avoid the difficulties of finding the ground-state is that this task is NP-hard even for the simple models of lattice proteins [11,28].

In the following chapter we introduce some computational methods for evaluating properties of importance to evolutionary optimisation. We will only give a short summary of the concept of RNA secondary structures; a detailed description of the properties and differences of various folding algorithms will be reported elsewhere. We will discuss to what extent the techniques and conclusions from RNA can be applied to lattice proteins.

2 Methods

2.1 RNA Secondary Structures

RNA secondary structures are defined as the list of Watson-Crick and of GU base-pairs that minimise the free energy and can be conceptually decomposed into:

1. Double-helical **S**tacks, the only elements with stabilising contribution to the overall energy.
2. **H**airpin loops, **M**ultiloops, **B**ulge loops, **I**nterior loops and
3. External elements such as **J**oints and **F**ree ends.

The secondary structure concept is a useful theoretical construct: the process of RNA secondary structure formation covers the major part of the folding energy and RNA secondary structural elements were shown to be highly conserved through evolution. Statistical properties of large ensembles of secondary structures have been shown to agree with phylogenetic data [10]. Important spatial interactions violating the above definition like pseudo-knots, triple base pairs, base stacking or non-Watson Crick base-pairs are, however, neglected [12]. The notion of secondary structure is but one of a spectrum of possible levels of resolution that can be used to define shape. Atomic coordinates as well as relative spatial orientation of the structural elements are discarded, only the topology of structure elements is retained. Folding in general can be conceptually partitioned in the two steps of formation of the secondary structure and the spatial structure.
Secondary structure graphs can be mapped one to one to strings and trees. In this contribution we only use one level of structural representation [9,19] that is an easy and efficient representations for string comparison and sorting by computational means: *fine graining* is obtained by denoting each unpaired base by a dot, each upstream paired base by an open parenthesis, and each downstream paired base by a closing parenthesis. *Coarse graining* is a symbolic notation where all distinguishable elements of secondary structure from (2) and (3) (or (2) only) are symbolised in hierarchical order. It is therefore possible to focus on the major structural features of the RNA molecule. Structure comparison and alignment can now be achieved in a similar manner as for sequences.

2.2 RNA Folding Algorithms

The *minimum free energy* algorithm is based on Zuker's well known algorithm [30]. Stacks contribute additively to the overall energy, other elements have no or a destabilising effect. Various energy sets that are mainly based on experimental data from small nucleotides have been compared: from these results it is clearly seen that the parameter set is of greater importance than the chosen algorithm for the prediction of secondary structures [25].
In this contribution we present recent results for the case of fine grained structures folded with the Zuker algorithm and of length $n = 40$. A chain-length of 40 is still short but a fairly close approximation for longer *RNA*s in the sense that the average size of structural elements is already converged to the values of naturally occurring structures [10,25]. All calculations were performed using the `Vienna RNA Package` which is public domain [13].

2.3 Lattice Proteins

In the case of proteins polynomial time optimisation procedures are - to date - unknown [28,11]. In this case even the secondary structure can only be predicted by heuristics and is therefore not useful for statistical investigations. An analogy to the secondary structure notion of RNA where a clear distinction between bonded and non-bonded residues exists can not be given as a number of different forces (hydrophobic interactions, electrostatic interactions, covalent bonds, H-bonds, Van der Waals) may occur between any two residues at (nearly) any distance along the chain and without respect to any other amino acid. Ken Dill and coworkers [5] derived a strongly simplified model with fixed torsion angles and length. Chains are composed from a 2-letter alphabet: hydrophobic (**H**) residues with a stabilising interaction between each other and hydrophilic (**P**) ones without contribution to the potential. Chains are put on a lattice but each lattice site is allowed to be occupied only once (the *self-avoiding walk* criterion). To find the ground state this model is still beyond computability so that we designed a greedy chain growth algorithm (see also fig. 2b). The model is highly idealised, yet it captures several salient features of real proteins such as excluded volume and solvent interactions, implicitly regarded with the hydrophobic effect.

Our algorithm starts at one end of the unfolded chain and proceeds optimising each position out of all the next possible ones on the lattice The energy value is computed only with respect to all former residues, the positions of which have already been fixed on the lattice. "Hydrophobic" residues form a core by construction of the model. For a simple square lattice and a sample of sequences of length $n = 18$ where ground state energies are known [23] we proved that the success in finding the ground state depends exponentially on the search depth. This can be interpreted such that in an evolutionary process it is at least possible to select for sequences that can fold into the ground state by a kinetic folding process (data not shown). We used this model to investigate the influence of several parameters such as lattice dimension, alphabet size etc. on the sensibility of structure formation towards mutations.

Extensive enumeration and comparison of structures is still not as simple as for RNA but we derived several methods to characterise the structure difference. The simplest (and most restrictive) one is the base pairing analogy, i.e. the number of identical neighbourhoods of corresponding positions between two structures of equal length. Other methods such as the dynamic alignment of structures encoded as relative moves also represent a promising method. A detailed description of the folding algorithm will be given elsewhere [3]. All calculations were performed using the `Toolkit for Lattice Polymers` which is available upon email request from the author.

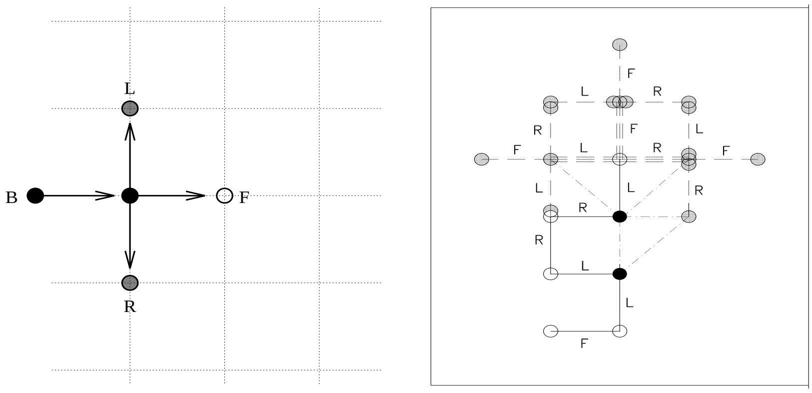

Figure 2: a) left: An example for the coding of structures by moves relative to the prior on the simple square lattice. Other lattices such as cubic, trigonal, tetraedic, face centered cubic, body centered cubic and knight's moves can be encoded in the same way and are also provided by the computer package (see text). b) right: The greedy chain growth algorithm. As an example we assume a given configuration (FLLRL) with one energy contributing neighbourhood between the two HH (dark dots) and describe the decision process for the next step (parameters are: search depth = 2, cutoff for energy contribution = 1.9). Next all possible configurations of length 2 are appended and the overall energy for all cases is evaluated. One configuration (LL) is forbidden and assigned an infinite energy contribution (∞). The "best" of all configurations (RR) with the highest energy (3.4) is selected and from there the first (R) move is appended. The resulting configuration (FLLRLR) has energy 1.7 and is used as starting configuration for the next iteration. (Full lines refer to bonds along the chain, dashed lines to possible moves, dash dotted lines to energy contributions within the cutoff distance.)

2.4 Comparison and Distances of Sequences and Structures

While the energy difference between two structures can be obtained by simple subtraction we need well defined measures to compare distances between sequences and their corresponding structures. The *hamming distance*, defined as the minimum number of point mutations to transform one sequence into another of the same length provides us with a natural metric in sequence space. *Tree alignment* provides a metric in shape space and can be applied to both structure representations of RNA as mentioned above [10].

For any scalar quantity $F(x)$ that can be assigned to a sequence x (or its structure) we can define *Local Optima*: points x such that function $F(x)$ for all neighbours (i.e. one error mutants) y one has $F(x) \geq F(y)$. Series of random searches of point mutations (prohibiting back-mutations) towards better solutions of $F(y)$ are called *adaptive walks*. The number of consecutive steps until the series is trapped at a local optimum is called the *walk length* and is an important measure of the performance of optimization processes.

Neutral neighbours are one or two error mutants that fold into the same struc-

ture as some reference sequence. Obviously it depends on where the mutation is located in regard to the structure: in a stack 2 simultaneous base exchanges are required to maintain the structure (an exception is the exchange of members from or to a GU pair), in an unpaired region a single exchange can be sufficient.
Neutral Nets are sets of connected neutral neighbours. *Neutral Paths* are subsets with monotonously increasing Hamming distances of step size 1 or 2, starting from a reference sequence.

2.5 Fitness Landscapes and Complex Combinatory Maps

The concept of *fitness landscapes* is nowadays well established and has been successfully used to describe the principles of evolutionary adaptation. Landscapes in our context are mappings from the space of genotypes (sequences) into a space of real numbers that are assigned to some phenotypic (structural) features, e.g. the minimum free energy, rate constants of structure formation or an arbitrarily chosen fitness value [9]. A suitable method to characterise landscapes is the autocorrelation function

$$\rho(h) = 1 - \frac{\langle D^2(f(x), f(y)) \rangle_{h(x,y)=h}}{\langle D^2(f(x), f(y)) \rangle_{random}} \tag{1}$$

where for example

$$D(f(x), f(y)) = \Delta G(x) - \Delta G(y) \tag{2}$$

is the difference in free energies for two sequences x and y. As analytical solutions are not available for most landscapes we use large statistical ensembles of computationally folded RNA molecules to compute $\rho(h)$. This expression can be viewed as a measure of the average similarity of energies or structures etc. as a function of the Hamming distance h of the underlying sequences.
A useful measure for the ruggedness of a landscape is the correlation length, defined as the value of h where $\rho(h) = 1/e$. It is in the order of the average distance between two local optima and therefore characteristical for the ruggedness of a landscape which in turn describes the complexity of an optimisation problem such as the performance of an evolution strategy (a recent survey on this subject was given by P. Stadler [22]). This will also work for structures since we can define a metric to compare structures.

3 Results

3.1 Distributions of RNA Secondary Structures:

We fold large ensembles of RNA sequences for several lengths with various algorithms and sort structures according to their frequencies. The most frequent is assigned rank 1, the next 2 and so forth. For all structure representations and algorithms [25] the ranking yields a distribution that follows a generalised Zipf law:

$$f(r) = a(r + b)^{-c} \tag{3}$$

where r is the rank and $f(r)$ is the frequency of the corresponding structure. a is a suitable normalisation constant, b can be interpreted as the number of "very frequent" structures. We are, thus, dealing with relatively few common shapes and many rare ones.

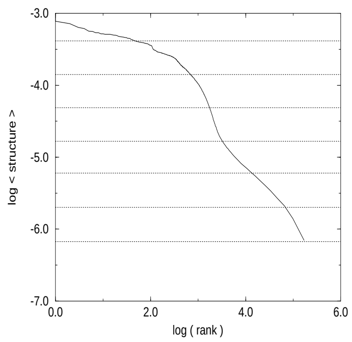

Figure 3: Zipf's law for the distribution of fine grained RNA secondary structures, a biophysical alphabet (AUGC), chain-length n=40 and a sample of 1.5×10^6 random sequences.

3.2 Neutral Mutations

Computations of neutral nets [20] showed a surprising result: neutral paths though being only a lower bound for the extension of neutral nets, have a high chance to traverse the whole sequence space. This means that starting from any random structure it is very easy to generate the sequence with the maximum hamming-distance from the underlying sequence by successive 1 and 2-error mutations without ever changing the structure. Obviously these results must depend on the probability to find a neutral neighbour at any point in sequence space which in turn depends on the frequency of the structure, i.e. the density in sequence space. Long paths were found for all alphabets and algorithms [25,2]. In the case of the "natural" 4-letter alphabet (AUGC) we found

a clear dependence of the path length on the frequency, even rare structures however have a high chance to be part of a percolating network.
A detailed comparison of neutral nets on various alphabets will be reported elsewhere [2]. Here we report the dependence of neutral mutations and paths on

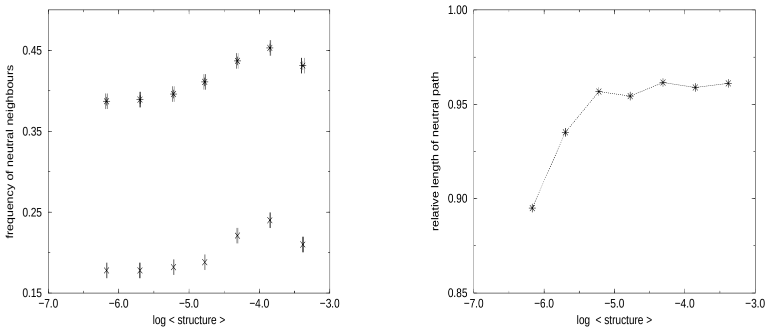

Figure 4: a) Probabilities to find neutral 1 ($\times$) and 2 ($*$) - error mutants for structures taken from samples with a certain frequency (see also fig. 3). b) the relative length (length/40) of neutral paths for the same samples.

the probability of the structure. From the large sample (see previous subsection) we picked 7 subsamples, each consisting of 250 structures with the same probability. We then calculated the number of neutral 1 and 2 - error neighbours. As there are more paired than unpaired bases in an average structure [10] and there are several ways to exchange one base-pair with another in the biophysical alphabet the frequency is much higher for 2 - error mutants. The slight decrease of neutral neighbours for very frequent structures is probably due to fact that these structures, consisting of very simple components can be easily inter-converted.
As we have a high probability to choose a very common structure, we can expect a large number of structures with many neutral neighbours and a relatively high probability to start and therefore end on a large neutral net. Consequently we expect many long neutral paths. Figure 4b shows that this is indeed the case. There appears to be a threshold value for the probability of a structure above which neutral nets become indeed connected in sequence space. For very frequent structures it is very likely that the underlying sequence can be inverted by subsequent steps of point mutations.
We also investigated the near neighbourhood of lattice proteins. Due to the lack of a well-defined and reasonable notion of a coarse grained structure representation and the enormous size of the shape space we did not succeed in finding a sufficient number of neutral neighbours for random structures.

3.3 Landscapes

Together with other statistical features of RNA the correlation lengths of free energy and structure landscapes were extensively explored in our group [1,8,10,9,26]. We find energy correlation lengths that are longer than structure correlation lengths, both are very short compared to the diameter of the sequence space. In any case we found a strong dependence on the size of the alphabet: increasing the sequence space by choosing larger alphabets smoothens the landscape. Corresponding results for the autocorrelation function of the example from the last chapter are shown in fig. 5a: the correlation length is larger for the energy landscapes than for the sequence to structure mapping. This means the distance between local optima is larger and evolutionary strategies are more likely to succeed in finding an optimal solution. This holds true for lattice polymers as well (see fig. 5b): correlation lengths are smaller for structure distances than for energy landscapes and scale linearly with the chain-length in both cases. As the structure mapping is more rugged many different structures can be tested which may have similar stability. Larger alphabets and smaller shape spaces

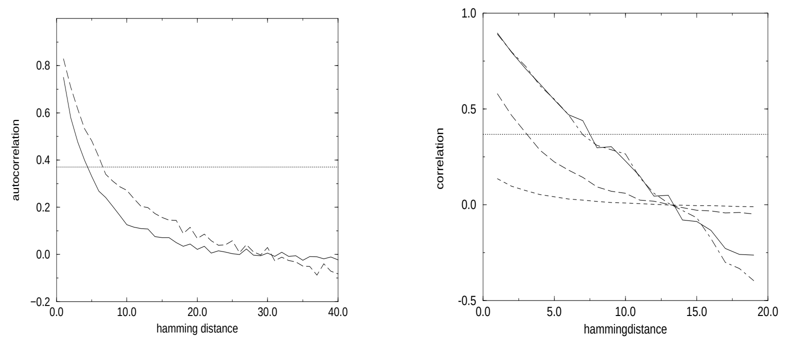

Figure 5: a) Landscape of RNA secondary structures: The autocorrelation for energy distances is larger than for tree edit distances (same parameters as in fig. 3 were used). b) Landscapes of lattice proteins: Autocorrelation functions for sequences with 67% hydrophobic residues and of length $n = 20$ folded with the greedy chain growth algorithm for free energy landscapes (solid, dot-dashed) and landscapes of contact distances (long dashed, short dashed) on the square (solid, long dashed) and the simple cubic (dot dashed, short dashed) lattice. (The horizontal line denotes $1/e$.)

(i.e. choosing lattices with a smaller number of neighbours) increase the correlation lengths. A detailed study using different measures for structure distances and lattices is in preparation [3].

4 Discussion

We presented some results from simple exact computational models to investigate possible mechanisms for evolutionary adaptation of biopolymers. Results on RNA secondary structures strengthened the perspective that it may be easy to explore shape spaces by point mutations. Adaptation of RNA should be much easier than commonly assumed: from any given structure it is not far to any desired structure and the sequences can be tried without ever changing the structure. These results depend neither on the folding algorithm, parameter sets, nor on the distance measure used for comparing the secondary structures, they are *robust properties*[25]. We suggest the number of neutral neighbours and size of sequence space and shape space and the isotropy of the mapping to be the crucial requirements for the existence of neutral nets. If one is willing to accept the secondary structure concept as a useful approximation for a realistic model of biopolymers evolution RNA appears to be an ideal playground to study also the dynamic aspects of evolutionary optimisation. Recent investigations of the evolutionary dynamics in RNA model populations have shown that neutral nets indeed play an important role under selection [15].

Our investigations on lattice proteins had to cope with three major problems:
1) In contrast to the RNA model the shape space is larger than the sequence space. Correlations in higher dimensional shape spaces are low. Similar to the RNA case[9] this may be partly circumvented by using larger alphabets. (Among other chemical and thermodynamic aspects, this may hint why the protein alphabet (20 amino acids) is so much larger compared to RNA (4 nucleotides). One could speculate that proteins are thus able to evolve more smoothly, i.e. to maintain a higher correlation, and to enable neutral evolution if necessary.)
2) Structure enumeration is difficult. This and the next point is subject of current investigations using a coarse grained structure representation. [3].
3) Folding the ground state is not applicable for statistical investigations. This was solved by using a fast greedy chain-growth algorithm.

Our studies revealed interesting similarities of lattice protein landscapes to the RNA folding landscapes. Unlike RNA where studies on the influence of sequence space and shape space are accomplished by larger alphabets and/or restrictions imposed on the structures to be formed we could observe that higher dimensional shape spaces have no influence on energy correlation, however heavily distort the structure correlation.

Finally we want to emphasize that in spite of their simplicity and their limited capability of solving the complex task of a comprehensive structure prediction simple models as they were used in this contribution serve well to investigate aspects of biopolymers evolution as well as to study some basic principles for structure formation.

Acknowledgments

Most of the work was done under the supervision of Prof. Peter Schuster. Walter Fontana contributed many useful concepts and hints. Programming help by Alexander Renner and stimulating discussions with Manfred Tacker, Peter Stadler and Ivo Hofacker are great-fully acknowledged. Financial support was partly provided by the FFWF project P-8526-MOB.

References

1. S. Bonhoeffer *et al.*, Eur. Biophys. J. **22**, 13 (1993)
2. E. Bornberg-Bauer *et al.*, preprint , (1995)
3. E. Bornberg-Bauer *et al.*, in preparation , (1995)
4. H. S. Chan and K. A. Dill, Physics today **2**, 24 (1993)
5. K. A. Dill *et al.*, Protein Sci. **4**, 561 (1995)
6. A. D. Ellington and J. W. Szostak, Nature **346**, 818 (1990)
7. W. Fontana and P. Schuster, Biophys. Chem. **26**, 123 (1987)
8. W. Fontana, *et al.*, Mh. Chem. **122**, 795 (1991)
9. W. Fontana *et al.*, Phys. Rev. E **47**, 2083 (1993)
10. W. Fontana *et al.*, Biopolymers **33**, 1389 (1993)
11. A. S. Fraenkel, Bull. Math. Biol. **55**, 1199 (1993)
12. R. R. Gutell, Curr. Op. Struct. Biol. **3**, 313 (1993)
13. I. L. Hofacker *et al.*, Mh. Chem. **125**, 167 (1994)
14. M. A. Huynen *et al.*, J. theor. Biol. **165**, 251 (1993)
15. M. A. Huynen *et al.*, Proc. Natl. Acad. Sci.,USA , in press (1995)
16. M. Karplus and A. Sali, Curr. Op. Struct. Biol. **5**, 58 (1995)
17. S. Kauffman and S. Levin, J. theor. Biol. **128**, 11 (1987)
18. S. Kauffman, J. theor. Biol. **157**, 1 (1992)
19. D. A. M. Konings and P. Hogeweg, J. Mol. Biol. **207**, 597 (1989)
20. P. Schuster *et al.*, Proc. Roy. Soc. (London) B **255**, 279 (1994)
21. J. M. Smith, *Nature* **255**, 563 (1970)
22. P. F. Stadler, *SFI-preprint-series* , 95-03-030 (1995)
23. P. Stolorz, *Proc. 27th Hawaii Intl. Conf. on System Sciences* , (1994)
24. L. Stryer, *Biochemistry* , Freeman (1990)
25. M. Tacker *et al.*, preprint , (1993)
26. M. Tacker *et al.*, Eur. J. Biophys. **23**, 29 (1994)
27. C. Tuerk and L. Gold. *et al.*, Science **249**, 505 (1990)
28. R. Unger and J. Moult, Bull. Math. Biol. **55**, 1183 (1993)
29. S. Wright, The roles of mutation, inbreeding, crossbreeding, and selection in evolution , Vol 1 (1932)
30. M. Zuker and D. Sankoff, *Bull. Math. Biol.* **46**, 591 (1984)