Algorithm Independent Properties of RNA Secondary Structure Predictions

By

MANFRED TACKER[d], PETER F. STADLER[a,b],

ERICH G. BORNBERG-BAUER[e], IVO L. HOFACKER[f],

AND PETER SCHUSTER[a,b,c,*]

[a] Institut für Theoretische Chemie der Universität Wien
Währingerstraße 17, A-1090 Wien, Austria

[b] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

[c] Abteilung Molekulare Evolutionsbiologie, Institut für Molekulare Biotechnologie e.V.
Beutenbergstrae 11, Postfach 100813, D-07708 Jena, Germany

[d] Current Address: Österreichisches Verpackungsinstitut für Lebens- und Genußmittel
Franz-Grill-Straße 5, A-1030 Wien, Austria

[e] Current Address: Abt. Theoretische Bioinformatik, Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

[f] Current Address: Beckman Institute, 495 N Mathews, Urbana, Il 61801, USA

* Mailing Address:

Institut für Theoretische Chemie der Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: +43 1 40480 669   Fax: +43 1 40480 660
Email: pks@tbi.univie.ac.at

Abstract

Algorithms predicting RNA secondary structures based on different folding criteria – minimum free energies (mfe), kinetic folding (kin), maximum matching (mm) – and different parameter sets are studied systematically. Two base pairing alphabets were used: the binary **GC** and the natural four-letter **AUGC** alphabet. Computed structures and free energies depend strongly on both algorithms and parameter sets. Statistical properties, such as mean numbers of base pairs, mean numbers of stacks, mean loop sizes, etc., are much less sensitive to the choices of parameter sets and even algorithms. Some features of RNA secondary structures, like structure correlation functions, shape space covering and neutral networks, seem to depend only on the base pairing logic (**GC** or **AUGC** alphabet).

Key Words

Kinetic folding – minimum free energy structures – RNA secondary structures – sequence structure relations

Introduction

Recently new methods have been developed that exploit the mechanism of Darwinian selection and the features of large ensembles of randomly generated biopolymers for an evolutionary design of nucleic acid and protein molecules with predefined properties. Evolutionary biotechnology is intended to complement and to replace in part the rational design of biomolecules. The first applications have been based predominantly on RNA since it can be easily amplified in cell-free assays and it has sufficiently interesting molecular properties including specific catalytic function. Examples are the selection for RNA molecules which bind to predefined targets (proteins: Tuerk & Gold 1990; organic dyes: Ellington & Szostak 1990; ATP: Sassanfar & Szostak 1993) or discriminate between closely related molecules (theophyllin and caffeine: Jenison et al., 1994). Other studies dealt with retraining of ribozymes in order to change catalytic functions and substrate specificities (Joyce 1989, Beaudry & Joyce 1992), the search for RNA sequences that act like hammerhead ribozymess in an *in vivo* assay (Schwienhorst & Lindemann 1996), and the *de novo* design of catalytic RNA by selection techniques (Bartel & Szostak 1993). In order to be able to interpret the dynamics and to predict the outcome of these selection experiments, it is necessary to understand in detail the relations between sequences, spatial structures, and biochemical functions of the biopolymers in question.

RNA evolution experiments are simpler than those with viruses or entire organisms because RNA unites genotype (by its sequence) and phenotype (by its 3D-structure) in the same molecule. A core problem of understanding molecular evolution is thus addressed by investigating how RNA sequences fold into spatial structures (Gesteland & Atkins 1993). The spatial structure of an RNA molecule is dominated by its secondary structure which is tantamount to a list of Watson-Crick (**GC** and **AU**) and **GU** base pairs. In fact, the free energy of folding can be explained almost completely by the energy of secondary structure formation (see for example Freier *et al.* 1986, Le & Maizel 1989), and secondary structures are conserved in evolution (Sankoff *et al.* 1978, Pace *et al.* 1986, Cech 1988, Waterman 1989, Le & Zuker 1990). It is meaningful therefore to consider the folding process as consisting of two steps:

(1) folding of the sequence into a two dimensional, planar secondary structure induced by base pair formation and

(2) subsequent formation of the three dimensional structure.

Secondary structures are much easier to predict (Zuker & Stiegler 1981, Zuker & Sankoff 1984, Zuker 1989a, McCaskill 1990, Hofacker *et al.* 1994) than complete spatial structures. They can be stored in compressed forms and accordingly very large random samples can be handled on the computer (Grner *et al.* 1996a,b).

Folding RNA sequences into secondary structures is frequently used as a tool for classification and prediction of function of RNA molecules. Several prediction algorithms are in widespread use and even entire virus genomes are folded (HIV: Shapiro *et al.* 1995, Huynen *et al.* 1996a) in order to learn more about the regulation of viral life cycles in host cells. On the other hand, it is well established that all available algorithms are facing several severe problems:

(i) Predicted minimum free energy (mfe) structures do not coincide with structures derived from phylogenetic comparisons (Gutell 1992, 1993, Konings & Gutell 1995).

(ii) Energies computed for the phylogenetic structures deviate more from the mfe-values the longer the RNA sequences are (Morgan & Higgs, 1995).

(iii) The numbers of suboptimal secondary structures are extremely large and no algorithm is available that allows to compute all structures.

Tertiary interactions (non Watson-Crick base pairs, pseudoknots, base triplets, **G**-quartets, etc.) not taken into account in the computation of the energies of secondary structures may be responsible for the chain length dependent deviations mentioned in (iii). Alternatively, larger RNA molecules need not fold into mfe-structures but into those that are determined by the kinetics of the folding process. The various kinetic folding algorithms for RNA differ with respect to the folding criterium. This criterium, in essence, corresponds to the time for refolding of partial structures into more stable configurations which is alloted during the folding process. For mfe based algorithms this time is infinite.

In this contribution we shall be concerned with qualitative aspects of the mapping from sequences into secondary structures. We shall show that several properties are fairly insensitive to the choice of folding principles and parameter sets. In particular, most of the global relations between sequences and structures were found to be independent for almost all practical purposes.

Methods

Secondary Structures

Secondary structures are lists of Watson Crick and **GU** base pairs. Let $(i : j)$ denote a base pair between the nucleotides at position i and j, with $i < j$. The original definition of secondary structure (Waterman 1978) requires that

(i) Each base is involved in at most one base pair: from $(i : j)$ and $(k : l)$ follows that $i \neq k$, $i \neq l$, $j \neq k$, $j \neq l$.

(ii) The following *no knot* condition is fulfilled: For any two base pairs $(i : j)$ and $(k : l)$ with $i < k < l$ holds $i < k < l < j$.

The second condition implies that every secondary structure can be drawn as a planar graph (without crossing of strands). Both conditions are not strictly valid. Base triplets, which violate condition (i), have been found in a variety of RNA molecules such as tRNAs (Giege *et al.* 1993), and group-II introns (Michel *et al.* 1989). Pseudoknots (Pleij 1990, Westhof & Jaeger 1992), which violate condition (ii), have been identified in an increasing number of RNA molecules, e.g., in group-I introns (Michel & Westhof 1990), in viral RNA (Pleij *et al.* 1985), in ribosomal RNA (Maly & Brimacombe 1983, Gutell 1992), and even in small oligonucleotides (Puglisi *et al.* 1988).

Nevertheless the structure predictions obtained from minimum free energy structures have proven to be useful in a wealth of applications to particular sequences. For several decades they were and still are used in molecular biology in the interpretation of RNA function. Furthermore, statistical properties of secondary structures derived from large samples of random sequences have been shown to agree very well with those obtained from phylogenetic structures (Fontana *et al.* 1993b, Higgs 1993). For a recent comparison of mfe and phylogenetic structures see Konings & Gutell (1995).

Secondary structures decompose into the following structural elements which are assumed to contribute additively to mfe:

- *Stacks* are double-helical regions of RNA,
- *Loops* are unpaired regions enclosed by stacks. The degree of a loop is the number of stacks attached to it. A hairpin loop thus has degree 1, bulges and interior loops have degree 2, and all loops with degree larger than 2 are denoted as multi-loops.
- *External elements* are strains of unpaired bases that are not part of a loop. They are either free ends or joints connecting different components of the secondary structure.

Free energy of folding

The energy of a particular structure is assumed to be the sum of contributions from individual base pair stackings and loops strains. Over the last twenty years detailed sets of thermodynamic parameters for stacking energies and loop energies have been measured taking into account sequence dependences of these contributions (Prschke 1977, Fink & Crothers 1972, Gralla & Crothers 1973ab, Uhlenbeck *et al.* 1973, Borer *et al.* 1974, Sugimoto *et al.* 1987ab, Peritz *et al.* 1991, Santa Lucia *et al.* 1992).
Several energy parameter sets have been derived from these thermodynamic data (Salser 1977, Ninio 1979, Papanicolaou *et al.* 1984, Freier *et al.* 1986, Jaeger *et al.* 1989). In this paper we shall compare computations based on two different parameter sets. One has been derived in 1977 by Salser, and the other one is a recent update of the Freier *et al.* (1986) parameter set. The latter is being used nowadays by most programs for RNA folding. Apart from different numerical parameter values the Salser set is based on base pair energies whereas the recent parameter set uses stacking energies (corresponding to interactions beween base

pairs) as the source of structure stabilization. As mentioned already in the introduction all parameter sets including the most recent one are far from perfectly matching reality. For instance, there are no reliable experimental data available for the energy contributions of multiloops, and the discovery of particularly stable tetraloops (Tuerk *et al.* 1988, Santa Lucia *et al.* 1992) shows that the assumption of sequence independent loop contributions is not justified in general.

Folding algorithms

Three different criteria for RNA secondary structure formation are used in the algorithms compared here: maximum number of matching base pairs (mm), kinetically controlled folding (kin), and minimum free energies (mfe). In the last case it is also possible to compute base pairing probabilities derived from the partition function of secondary structures (p).

Maximum matching

For purposes of comparison we computed structures and structure statistics of RNA secondary structures with the "maximum matching" algorithm (Nussinov *et al.* 1978), which derives a structure with the maximum number of base pairs irrespective of energy parameters. It can be interpreted as a minimum free energy algorithm with a highly degenerate parameter set (in arbitary energy units): $\Delta G = -1$ for each base pair, and $\Delta G = 0$ for each loop. These parameters are tantamount to a contribution of $\Delta G = -1$ for each stacking between two base pairs and a (uniform) value $\Delta G = -1$ for each loop. This parametrization is obviously unrealistic, since in this model the loops have a stabilizing instead of the common destabilizing effect. Here maximum matching was applied with the two additional constraints that (i) hairpin loops must have at least three unpaired bases, and that (ii) isolated base pairs are forbidden (implying that the minimum lenght of a stack is two base pairs). We include maximum matching (mm) in order to search for statistical properties that are (almost) completely indepent of parameter choices, indicating that they are determined by the base pairing logic.

Kinetically controlled folding

There has been a long discussion whether an RNA molecules fold into the thermodynamically most stable structure or whether they form structures that are determined by course and kinetics of the folding process. A variety of closely related algorithms have been proposed that are based on local minimization of free energies subject to kinetic constraints (Martinez 1984, Abrahams *et al.* 1990, Gultyaev 1991). In our investigations we used a variant of the algorithm derived by Martinez (1984). All possible stacks of a given sequence are evaluated and their equilibrium constants are computed. The folding process proceeds stepwise by adding successively the stacks with the highest equilibrium constant which are compatible with already incorporated stacks. The folding process is complete when no more stacks can be found whose incorporation can lower the minimal free energy of the secondary structure.

Minimum free energy

The most common approach to predict RNA structures is based on minimization of free energy (Tinoco *et al.* 1971). Finding the minimum free energy structure is a non-linear optimization problem that can be addressed by dynamic programming. In this sense RNA folding is closely related to the loop-matching problem. In fact, loop matching can be understood as a special case of RNA mfe-folding. Dynamic programming algorithms have been published for a variety of different energy models (See for example: Waterman 1978, Zuker & Stiegler 1981, Zuker & Sankoff 1984, Waterman & Smith 1986, and for a review see: Zuker 1989b). These algorithms are readily adapted to take into account new biochemical data (for example particularly stable tetraloops). The implementation used for this study is described in detail in Hofacker *et al.* (1994). Recent improvements allow dynamic programming algorithms to predict not only a single minimum free energy structures, but an assortment of structures close to the minimum energy or all structures in a certain energy band (Waterman & Byers 1985, Zuker 1989a).

Partition function algorithm

McCaskill (1990) generalized the dynamic programming approach to produce the partition function of secondary structures. Instead of a single secondary structure this algorithm produces the base pairing probability matrix. A detailed statistical analysis of the resulting "structure ensembles" has been carried out recently (Bonhoeffer *et al.* 1993). We shall use these published data for comparison of the energy and structure landscapes. Since the partition function algorithm does not produce unique secondary structures we can only compare the correlation lengths of free energies and structures with those computed by the other algorithms.

The comparison presented here is based on the most recent parameter set ("F": Freier *et al.* 1986). The different folding criteria are applied to produce the data sets to be analyzed ($\mathtt{mm}$, $\mathtt{kin}_{(F)}$, $\mathtt{mfe}_F$, $\mathtt{p}_{(f)}$). In order to see the influence of different choices of parameter sets mfe-structures were computed with a second (older) parameter set as well ($\mathtt{mfe}_S$, "S": Salser 1977).

Distances between secondary structures

A variety of distance measures for secondary structures have been proposed. Secondary structure graphs can be mapped one to one to strings and to trees. The string representation is obtained by denoting each unpaired base by a dot, each base paired upstream by an opening bracket, and each base paired downstream by a closing bracket. Hogeweg & Hesper (1984) defined the distance of two secondary structures as an alignment distance of the two string representations (see also Konings & Hogeweg 1989). It can be shown that this distance is a metric on

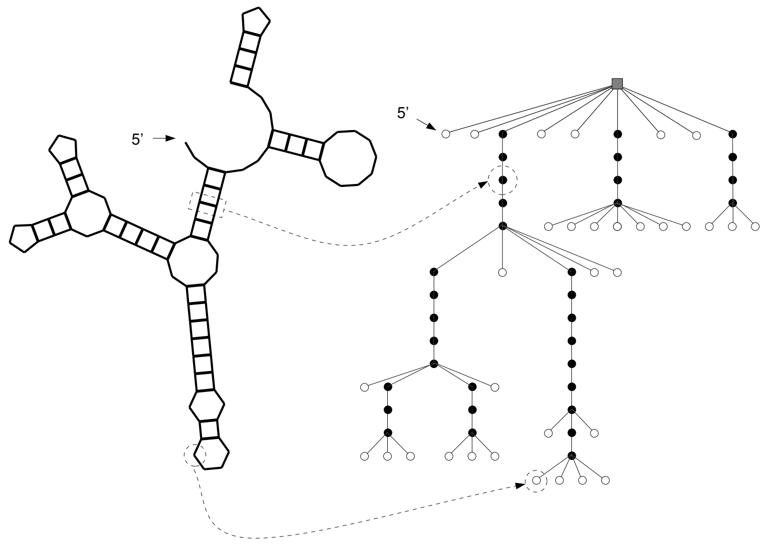


Figure 1: Representation of RNA secondary structures by means of planar trees.

the set of secondary structures.

A more sophisticated approach to a representation of secondary structure makes use of trees (Figure 1; see, for example: Shapiro 1988, Shapiro & Zhang 1990, Fontana *et al.* 1993b, Hofacker *et al.* 1994, 1996). For rooted planar trees a natural metric distance is defined by tree-editing which is a generalization of string alignments (Tai 1979, Ohmori & Tanaka 1988, Shapiro & Zhang 1990). An advantage of this distance measure is that base pairs are always aligned to base pairs while it happens frequently in string alignment that a base pair is aligned to (two) non-matching bases. Distance measures based on permutation representations of secondary structures are discussed in Reidys & Stadler (1996).

In order to compare structures belonging to a single sequence it is often useful to define distance by the number of non-identical base pairs, or even simpler, by the Hamming distance of the string representations. These measures correspond to edit distances where insertion and deletions are forbidden, and hence substitution is the only allowed edit operation.

The distance measures in shape space presented here are primarily motivated by the search for an unambiguous relation between structures. It should be mentioned, however, that tools for comparison of structures are generally context dependent and other definitions of distance in shape space might be more appropriate for comparing RNA molecules with respect to their function or evolutionary success.

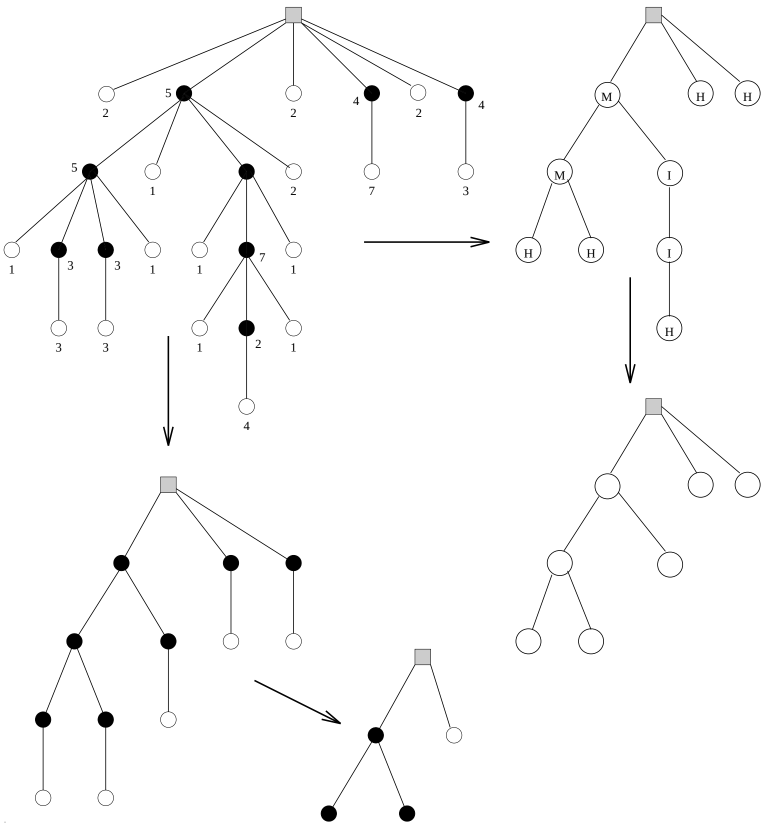


Figure 2: Coase graining of RNA secondary structures by means of planar trees.

Coarse graining of secondary structures

Tree representations in full resolution make it often difficult to focus on the essential structural features of RNA molecule because they are overloaded with details. Various coarse grained tree representations based on application dependent *ad hoc* assumptions have been invented in order to deal with this problem. Probably the most canonical approach to a coarse grained representation starts by translating the (full) tree representation into a *homeomorphically irreducible tree*, HIT, (Figure 2 and Fontana *et al.* 1993b). In a HIT stacked regions and unpaired regions, respectively, are represented by single nodes with weights indicating the number of bases or base pairs in the structural element. HITs still contain the full information on structures. Coarse graining can now be done either by omitting all nodes with weights smaller than a certain threshold (Fontana *et al.* 1993b) or by ignoring the particular arrangement of unpaired sequences in loops and labeling each stack with only the type of the loop enclosed by it (Shapiro 1988, Shapiro & Zhang 1990). An even coarser level of description is obtained by ignoring all interior loops and bulges and focusing on the branching structure of the molecule only.

Computer programs

All calculations presented in this study have been performed using the `Vienna RNA Package` (Hofacker *et al.* 1994). The programs are written in `ANSI-C` and are intended for UNIX systems. The package including the complete source code is available free of charge via anonymous `ftp` from `ftp.tbi.univie.ac.at` or via world wide web from `www.tbi.univie.ac.at`.

Comparison of Structures

Samples of random sequences are generated with uniform distributions of nucleotides. We studied both the natural **GCAU** and the two-letter **GC** alphabet. Statistics refer to sample size of approximately 100 000 sequences for each chain length, alphabet, and algorithm.

Reliability of structure prediction

In order to test the reliability of algoriths for the prediction of structures we have folded the same sequences with $\texttt{mfe}_F$, $\texttt{mfe}_S$, `kin` and `mm`. Reliability is clearly a question of great practical importance for the application of RNA folding algorithms, since they have been designed for the prediction of the structure of a particular sequence. Figure 3A shows the fraction of sequences that are folded into non-identical structures in pairwise comparisons of the three algorithms `kin`, $\texttt{mfe}_S$ and $\texttt{mfe}_F$. This fraction approaches one with increasing chain length implying that the three algorithms predict different structures for practically all sequences. Structure prediction seems to be more sensitive to the choice of parameters than to the folding criterium.

A more subtle method of measuring the differences in structure prediction is to calculate a distance between the structures derived by different algorithms from the same sequence. We used the tree editing distance for comparing secondary structure graphs (Figure 1). Let $D\big(f(x), g(x)\big)$ denote the distance of the secondary structures of the sequence x for two different algorithms f and g. Then we define

$$P_{fg} = \frac{\langle D\big(f(x), g(x)\big)\rangle}{\langle D\big(f(p), g(q)\big)_{\text{random}}\rangle} \tag{1}$$

where the average in the denominator is taken over pairs of randomly chosen sequences p and q. P_{fg} measures the average deviation of the predictions of the algorithms f and g: $P_{fg} = 0$ means that the predictions agree perfectly, while $P_{fg} = 1$ indicates that the predictions are completely unrelated. Figure 3B shows the dependence of P_{fg} on the chain length. As expected, the predictions agree less and less well as the chain length increases.

Individual comparisons yielded:

- $\texttt{mfe}_F$ vs. $\texttt{mfe}_s$: Even at chain lengths 20 only about 50% of the predicted structures are identical. For chain lengths greater than 50 the percentage of identically predicted structures is negligibly small.

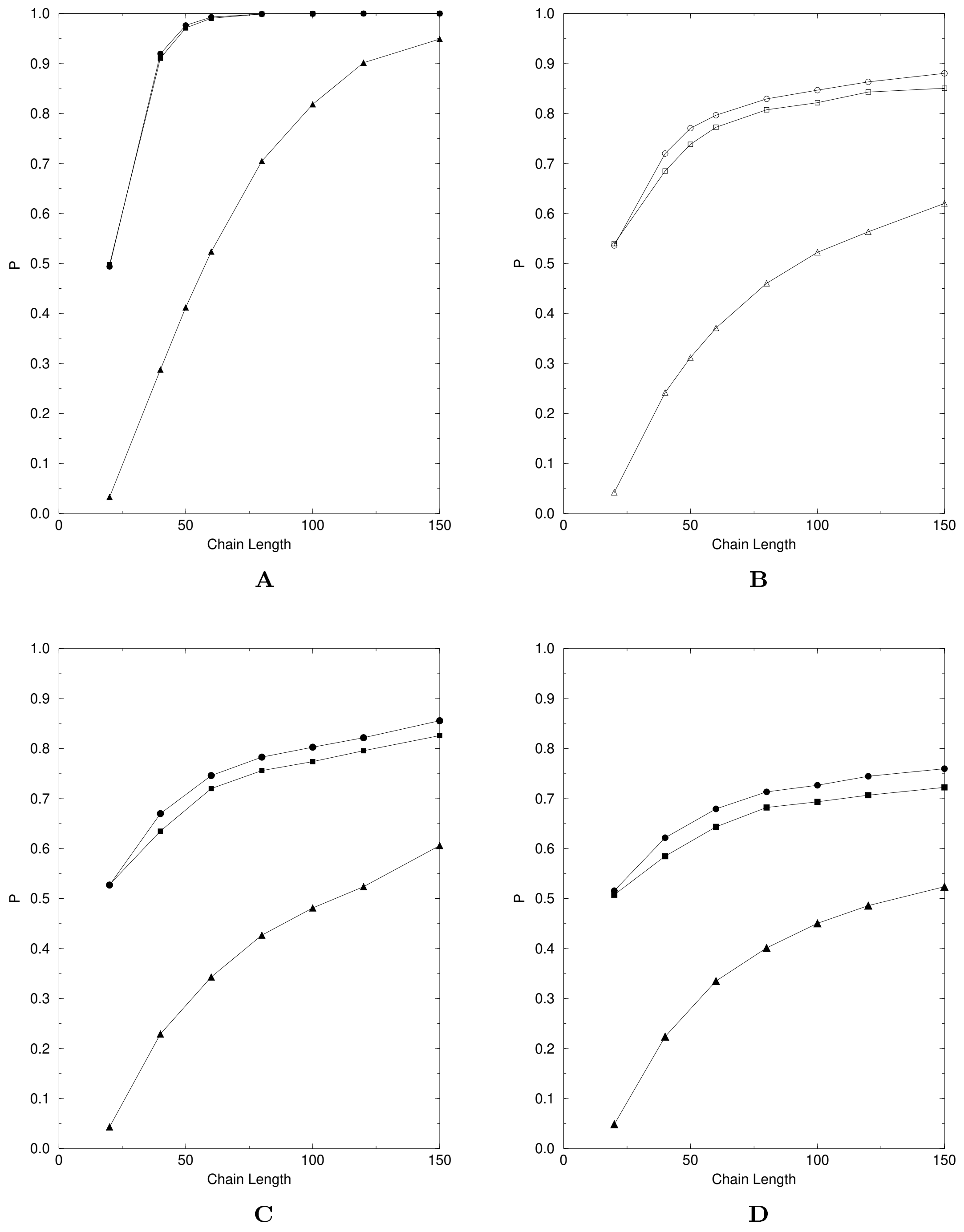


Figure 3: Average distances between structures computed from identical sequences with different algorithms and parameter sets. We show P_{fg} as a function of the chain length for the comparisons $\mathtt{mfe}_F$ versus $\mathtt{mfe}_S$: $\square$, $\mathtt{mfe}_F$ versus kin: $\triangle$, and $\mathtt{mfe}_S$ versus kin: $\circ$. Four distance measures for secondary structures are used: **A** fraction of sequences yielding non-identical structures, **B** tree edit distance, **C** alignment distance, **D** base pair distance.

- `mfe`$_F$ vs. `kin`: In contrast to the above comparison a remarkably higher degree of identical structures is predicted. Even sequences as long as 100 bases are folded in about 20% into identical structures.
- `mfe`$_S$ vs. `kin`: The percentage of identically folded sequences is even somewhat lower than in the first case.
- `mm` vs. all other algorithms: Sequences folded by the maximum matching algorithm have very little in common with the structures generated by the other algorithms.

From these results it can be clearly seen that the choice of a parameter set is more important than the folding criterium underlying the prediction algorithm for secondary structures. Even an algorithm differing substantially from the mfe-criterion as does the kinetical algorithm predicts structures with much more similarity when the same parameter set is used than those produced by the same algorithm (`mfe`) with a different parameter set.

Distribution of structures

For every algorithm 100 000 randomly generated strings of length $n = 100$ were folded into secondary structures. The structures were then ranked according to their frequencies. The ranking yields a distribution which follows a generalized Zipf's law

$$f(r) = a(r + b)^{-c} \tag{2}$$

where r and $f(r)$ are the rank and the frequency of the corresponding structure, respectively. a is a normalization constant, b can be interpreted as the number of "very frequent" structures. We found distributions following this form of generalized Zipf's law for all algorithms, parameter sets, and alphabets. This finding implies that there exist few common and many rare structures independently of the detailed base paring logics, the folding alghortihm and the energetic parameters.

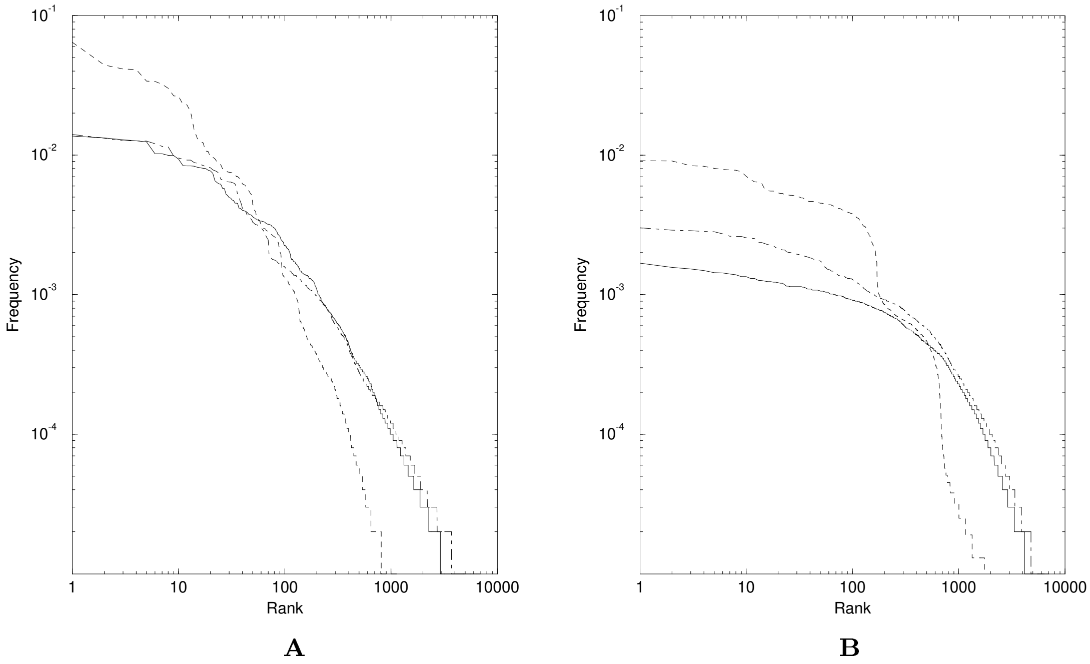


Figure 4: Zipf's law for coarse grained RNA secondary structures. $\texttt{mfe}_F$: full line, $\texttt{mfe}_S$: dash-dotted line, `kin`: dashed line. **A AUGC** alphabet, **B GC** alphabet.

Table 1. Parameters of the fits to Zipf's law for chain length $n = 100$.

	GCAU		**GC**	
	b	c	b	c
$\texttt{mfe}_F$	70	1.72	997	2.66
$\texttt{mfe}_S$	36	1.45	938	2.57
`kin`	33	2.40	312	3.74

The parameter b measures the number of very frequent structures. The data in table 1 indicate that there are at least an order of magnitude more frequent structures for the **GC** alphabet than for the **GCAU** alphabet. This is consistent with the fact that the ratio of structures to sequences is larger for two letter alphabets. It is also interesting to note that the tail of the distribution is steeper for the kinetic algorithm than for the minimum free energy algorithm. This indicates that not all of the very rare minimum free energy structures are accessible to the kinetic folding algorithm.

Statistics of Structure Elements

All statistically evaluated structure parameters for both alphabets and all algorithms are compiled in table 2. For a recent review on secondary structure elements see also Chastain & Tinoco (1991).

The statistics of structural elements in large random samples of minimum free energy structures has been studied previously in great detail (Fontana *et al.* 1993b). A comparison of various alphabets can be found in Fontana *et al.* (1993a,b). In the following we shall discuss the alphabet dependence only if it deviates from the findings for the minimum free energy structures obtained with the parameter set Freier *et al.* (1986).

The dependence of structure statistics on the chosen alphabet can be explained by the different average strength of base pairing and by the difference in stickiness **P** which is the probability that two arbitrarily chosen bases can form a base pair. In the present case we have $\mathbf{P}_{GC} = 0.5$ and $\mathbf{P}_{GCAU} = 0.375$, respectively. The strength of base pairing is larger for **GC** pairs on average, since the stacking energies for **GC.GC** and **GC.CG** are larger than for any other pairs. For both reasons we expect a larger number of base pairs in pure **GC** sequences than in **AUGC** sequences. As an immediate consequence we have smaller loop sizes and fewer external digits. Furthermore, larger stacking energies allow to stabilize more small stacks.

We compare the data from folding algorithms with a sample of random structures that fulfil the following properties: (i) all stacks have at least length 2, i.e., there are no isolated base pairs, (ii) all hairpin loops contain at least 3 unpaired bases, and (iii) they are compatible with sequences with the stickiness **P** corresponding to the **GCAU** and **GC** alphabet, respectively. A random sample of sequences with the above defined properties can be generated using the algorithm described

Table 2. Structure statistics

n	Algorithm	N_b	N_l	N_j	N_e	L_l	L_s	D_l
				GCAU				
50	mfe$_F$	10.41	2.48	1.46	15.80	5.39	4.19	1.41
	mfe$_S$	8.75	2.91	1.11	13.39	6.57	3.01	1.62
	kin	9.16	1.96	1.44	18.30	6.81	4.67	1.27
	mm	17.01	8.14	1.24	1.44	1.79	2.09	1.85
	random	9.87	4.45	1.60	6.40	5.36	2.22	1.64
100	mfe$_F$	24.40	5.57	2.37	20.62	5.49	4.38	1.58
	mfe$_S$	21.51	6.68	1.18	10.77	6.92	3.22	1.82
	kin	20.81	4.16	2.34	25.16	7.98	5.00	1.44
	mm	36.35	15.91	1.27	1.40	1.63	2.28	1.92
	random	20.70	9.31	1.85	6.70	5.57	2.22	1.80
150	mfe$_F$	38.70	8.70	3.09	24.42	5.54	4.41	1.61
	mfe$_S$	34.60	10.36	1.20	9.69	6.86	3.34	1.88
	kin	32.35	6.32	3.08	30.61	8.66	5.12	1.51
	mm	55.66	23.62	1.29	1.39	1.58	2.36	1.95
	random	31.51	14.18	1.94	6.80	5.65	2.22	1.86
200	mfe$_F$	53.67	11.94	3.67	27.09	5.49	4.49	1.69
	mfe$_S$	47.63	13.97	1.22	9.38	6.82	3.41	1.91
	kin	43.95	8.42	3.73	35.25	9.12	5.21	1.56
	mm	75.24	31.20	1.29	1.33	1.54	2.41	1.96
	random	42.31	19.04	1.99	6.91	5.69	2.22	1.90
				GC				
50	mfe$_F$	17.09	3.31	1.58	4.37	3.45	5.16	1.52
	mfe$_S$	16.99	3.51	1.24	3.27	3.64	4.85	1.65
	kin	15.47	2.62	1.58	6.42	4.82	5.90	1.40
	mm	19.01	6.39	1.19	1.12	1.70	2.98	1.81
	random	11.01	4.84	1.57	5.12	4.71	2.27	1.68
100	mfe$_F$	37.21	6.51	2.02	4.42	3.25	5.72	1.69
	mfe$_S$	37.12	6.81	1,32	2.97	3.35	5.45	1.81
	kin	33.71	5.16	2.01	6.86	4.99	6.54	1.61
	mm	40.46	11.80	1.21	1.07	1.53	3.43	1.90
	random	22.91	10.10	1.79	5.34	4.84	2.27	1.82
150	mfe$_F$	57.48	9.64	2.24	4.55	3.16	5.96	1.77
	mfe$_S$	57.46	9.99	1.39	2.93	3.22	5.75	1.86
	kin	51.89	7.64	2.29	7.30	5.10	6.79	1.70
	mm	62.04	17.04	1.22	1.06	1.46	3.64	1.93
	random	34.80	15.34	1.87	5.44	4.89	2.27	1.88
200	mfe$_F$	77.96	12.73	2.39	4.59	3.10	6.12	1.81
	mfe$_S$	77.82	13.14	1.40	2.91	3.15	5.92	1.89
	kin	70.22	10.12	2.48	7.60	5.14	6.94	1.76
	mm	83.77	22.04	1.24	1.05	1.43	3.80	1.94
	random	46.71	20.60	1.90	5.47	4.91	2.27	1.91

mfe$_F$: minimal free energy algorithm, parameter set: Freier *et. al.* 1986; mfe$_S$: minimal free energy algorithm, parameter set: Salser 1977; kin: kinetical algorithm (Martinez 1984), parameter set: Freier *et. al.* 1986; mm: maximum matching: minimum loop size is 3, minimum stack size is 2; random: a randomly chosen sample of structures created by the algorithm described in the appendix.

in the appendix. A combinatorial analysis of the statistics of random structures is described in Hofacker *et al.* (1996).

Mean free energies

Due to the higher binding energies of the **GC** pair, **GC** sequences are more stable than the **AUGC** sequences. The mean free energies scale linearly with n independent of the algorithm and the parameter set.

Mean number of base pairs N_b

All three algorithms show a linear increase of the number of base pairs with n. The $\texttt{mfe}_F$ structures have the largest average number of base pairs, the kinetic structures the smallest one (Figure 5B). Data computed from the random sample show that secondary structures as predicted by the folding algorithms (mfe or kin) do not have significantly more base pairs than random compatible **GCAU**-structures. This is, however, not true for **GC**-sequences for which the energy minimization results in a remarkable increase in the number of base pairs.

Mean number of loops and stacks N_s

The number of loops equals necessarily the number of stacks since every loop is by definition closed by a stack. For all three algorithms the mean numbers of loops (or stacks) scale linearly with n. The kinetically folded structures contain fewer loops than the mfe-structures. The number of loops and stacks is significantly smaller in the folded structures than in random compatible structures. Folding, no surprise, results in larger structural elements than random assembly.

Mean loop degree D_l

Since for every additional multiloop with degree n there must be $n-1$ additional hairpins (with degree 1), the mean loop degree of a structure with N_l loops and N_c components is $2 - N_l/N_c$. Because N_c grows only slowly we expect the average loop degree to converge to 2 in all cases. However, convergence seems to be quite slow, in particular for the kinetic algorithm.

Mean stack size L_s

The mean stack size approaches a constant value at fairly small chain lengths ($n \sim 50$). The kinetical structures have larger mean stack size than the mfe-structures. The mean stack size of structures folded with reasonable parameters is about twice as large as the mean stack size for both maximum matching and random compatible structures. This is mostly an effect of combinatorics: there are many more structures with small stacks than with large stacks. Hence they dominate the statistics if they are not suppressed by the destabilizing energy-contributions of the loops.

Mean loop size L_l

The dependence of the mean loop size on the chain length is somewhat influenced by the base pairing alphabet:

AUGC alphabet. For the mfe-structures the average mean loop size converges even at moderate chain lengths, say 60, rapidly to a constant value. The kinetically folded structures show greater loop sizes and even at a chain length of 200 no convergence is reached. The slope of the curve, however, decreases with chain length (Figure 5A) and we expect that a constant value will be reached for long chains.

GC alphabet. For the kinetic structures a slight increase of the mean loop size is observed whereas for the mfe-structures the mean loop size even decreases with chain length n.

Number of joints and components N_j, N_C

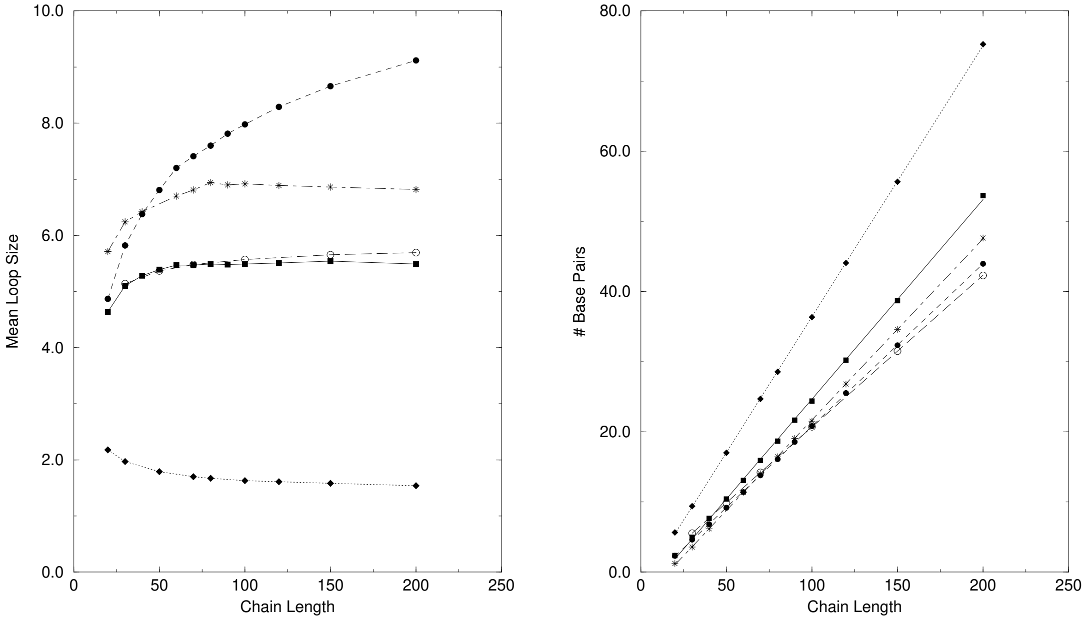
A secondary structure consists of one, two, or more components, which are connected by joints; hence $N_j = N_C - 1$. The number of joints is nearly indistinguishable for both alphabets for the $\texttt{mfe}_F$ and the kinetic algorithm. Both of them show an increase with n which is slower than linear. It is not known yet, whether the number of joints becomes constant for very long chains ($n > 1000$). For Salser's parameter set we find that the number of joints rapidly settles down to a constant value. Data from random structures are in the same ball-park as the values for structures obtained from folding mechanisms.

Number of external digits N_e
Nucleotides in joints and free ends, that are all bases, which are neither in a stack or in a loop, are termed external digits. In this case the situation is similar to the number of joints. $\mathtt{mfe}_F$ and the kinetic algorithm scale slightly less than linear with chain length n for both alphabets, whereas $\mathtt{mfe}_S$ decreases at small chain lengths and converges then to a constant value.

Except for possibly the mean loop sizes, the qualitative features of the statistics of secondary structure elements are the same for all algorithms, both parameter sets and both alphabets. Also, the alphabet-dependence of the data is qualitatively the same. Even the quantitative difference between the algorithms are in most cases strikingly small.

RNA landscapes and combinatory maps

The view of evolution as an optimization processes on complex *fitness landscape* has been introduced by Sewall Wright (1932). Since then this concept has received considerable attention (see, for example, Eigen *et al.* 1989, Kauffman 1993). A landscape is a mapping from the space of genotypes into the space of real numbers. Examples are free energies or rate constants for structure formation viewed as function of the sequence (Fontana *et al.* 1991, 1993a,b, Tacker *et al.* 1994, Bonhoeffer *et al.* 1993). In the space of genotype we have the usual Hamming distance for comparing sequences of common length.
More generally, one may consider the mapping from genotypes to phenotypes. In our application, the phenotype of an RNA molecule is defined as its secondary



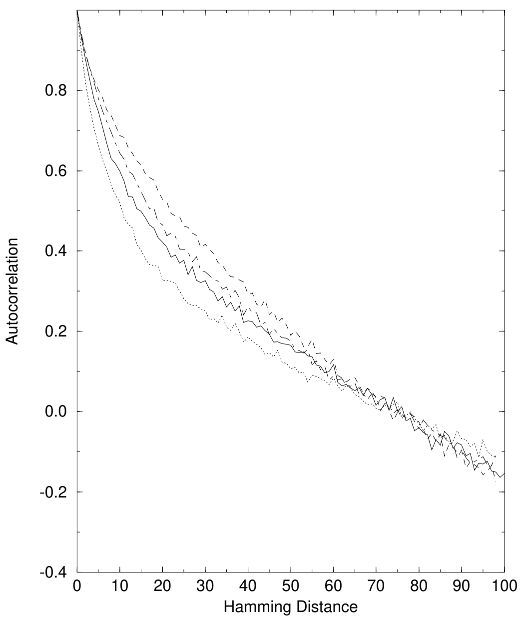
A **B**

Figure 5: Examples of structure statistics for **AUGC**-sequences. **A** Mean loops size and **B** mean number of base pairs. `mfe`$_F$ ■, `mfe`$_S$ ⋆, `kin` ●, `mm` ◇, `random` ○. More complete data are given in table 2.

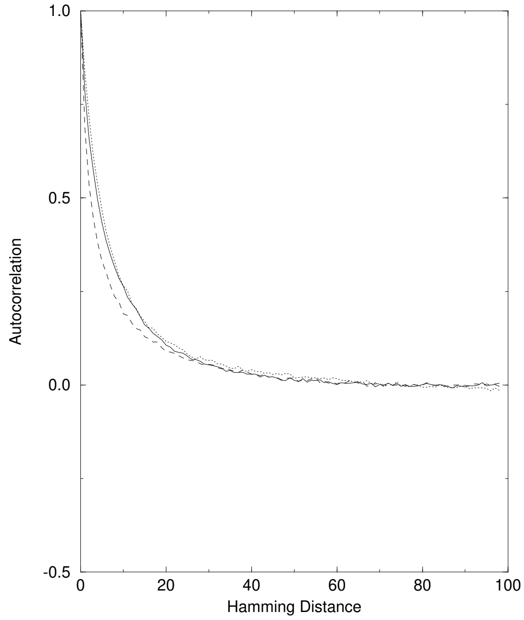
structure. The term *combinatory map* was coined for such a mapping, since the notion of a landscape is reserved for mappings into the real numbers allowing for concepts of local optima, hill climbing an the like, that have no counterparts in general genotype phenotype mappings (Fontana *et al.* 1993a,b).

Correlation

Recently, a variety of statistical measures have been proposed for characterizing (fitness) landscapes and combinatory maps. The best studied one is the autocorrelation function measuring the "ruggedness" of the mapping (Eigen *et al.* 1989, Weinberger 1990). As shown in Fontana *et al.* (1993a), it is possible to calculate autocorrelation functions for combinatory maps provided there is a metric distance



A



B

Figure 6: Autocorrelation function for **GCAU** alphabet. **A** free energy correlation, **B** secondary structure correlation calculated with tree edit distance. `mfe`$_F$: solid line; `mfe`$_S$: dash-dotted line; `kin`: dashed line; `mm`: dotted line.

measure D comparing phenotypes. The autocorrelation function is defined by:

$$\rho(d) = 1 - \frac{\langle D^2(f(x), f(y))\rangle_{d(x,y)=d}}{\langle D^2(f(x), f(y))\rangle_{\mathrm{random}}} \quad . \tag{3}$$

We may interpret $D\big(f(x), f(y)\big)$ either as the absolute difference in free energy, $D\big(f(x), f(y)\big) = |\Delta G(x) - \Delta G(y)|$, or as one of the structure distances discussed in section 3. In the following we use the tree edit distance. The autocorrelation function $\rho(d)$ measures the average similarity of structures as a function of the Hamming distance of their underlying sequences. A useful measure for the overall ruggedness of a landscape of a combinatory map is the correlation length ℓ at which $\rho(d)$ has decayed to a value of 1/e. In figure 6 we show examples of autocorrelation functions for both free energies and tree structures. Correlation lengths are compared in figure 7.

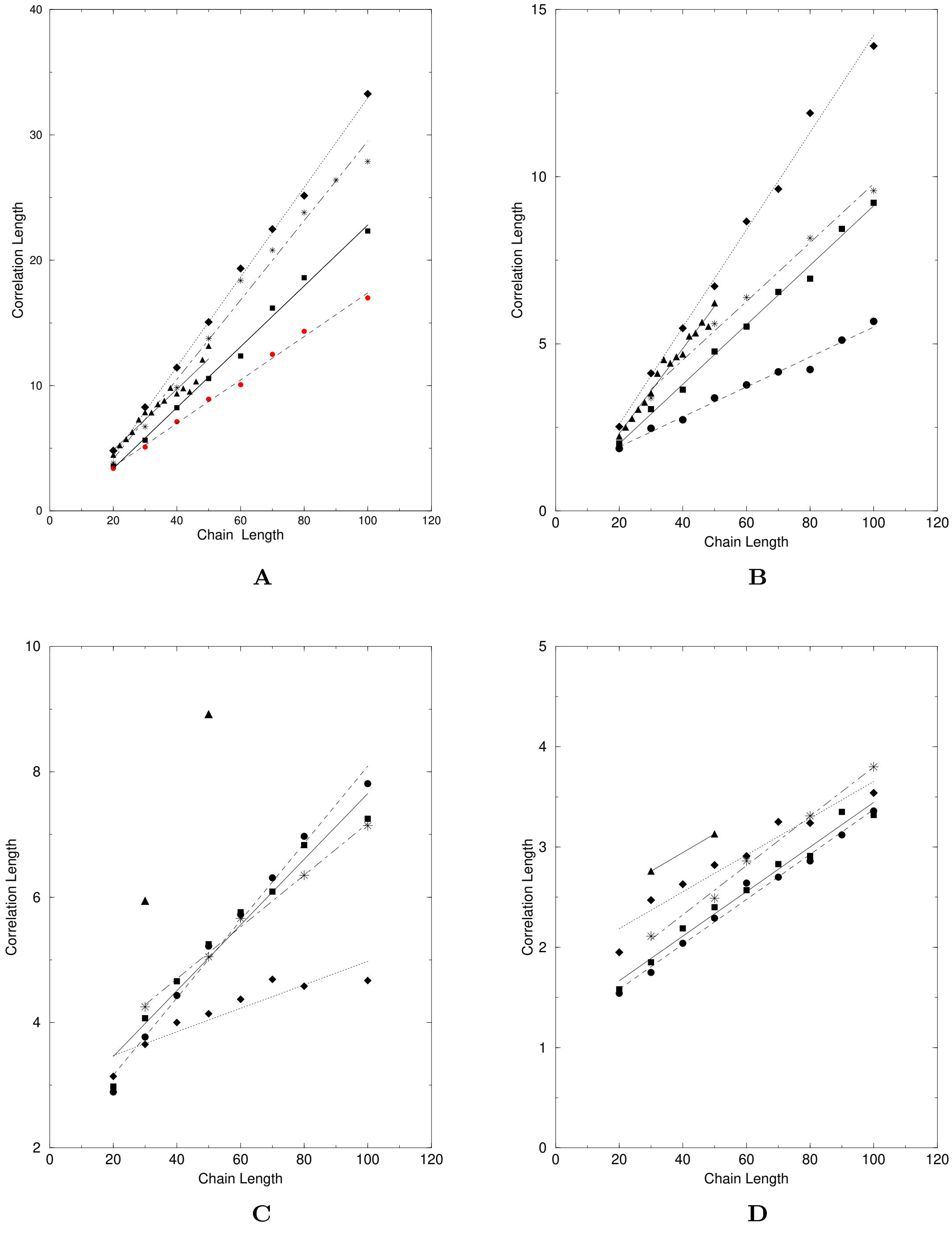


Figure 7: Energy correlation lenght for **A**: **GCAU** alphabet and **B**: **GC** alphabet; and structure correlation length calculated with the tree editing distance for **C**: **GCAU** alphabet and **D**: **GC** alphabet. Algorithms: `mfe`$_F$ ■, `mfe`$_S$ ⋆, `kin` ●, `mm` ◇, `p` △.

Larger values for correlation lengths imply smoother landscapes and, in particular, smaller numbers of local optima (Stadler & Schnabl 1992, Stadler & Krakhofer 1996). Evolutionary optimization is thus simpler on landscapes with longer correlation lenghts (For example see Fontana *et al.* 1991, 1993a, Schuster & Stadler 1995). Data on free energy landscapes and structure correlation are available in the literature for the `mfe` algorithm (Fontana *et al.* 1991, 1993ab, Schuster *et al.* 1994) and for the partition function algorithm (Bonhoeffer *et al.* 1993). Here we complement these studies by providing correlation data for two other algorithms, kinetic folding (Martinez 1984) and maximum matching.

It is a general feature of RNA landscapes that the correlation length ℓ_E of the energy landscape is much larger than the correlation length ℓ_S of the corresponding combinatory map of structures. Nevertheless, in both cases we observe a linear scaling of the correlation lengths with n with the **AUGC** alphabet having the largest correlation lengths for both cases. The auto correlation function for the combinatory map for structures is nearly identical for both algorithms, so that we can conclude that these two maps are very similar as far as ruggedness is concerned (Figure 6). The structure correlation for maximum matching on the **GCAU** mapping is particularly small since the ground state is highly degenerate, i.e., there are many more or less unrelated secondary structures with the maximum number of base pairs. We do, however, not observe the same effect for a two letter alphabet. The structure correlation as calculated from the ensemble of secondary structures (partition function algorithm) is larger than for single structures, since room temperature is already close to the average melting temperature for random **GCAU** sequences (Bonhoeffer *et al.* 1993). The situation is different though for the energy landscape. In this case ℓ_E of the landscape generated by the kinetic algorithm is smaller than the ℓ_E of the landscape generated by $\texttt{mfe}_F$. This means that the energy values of two neighbouring sequences change much faster when using the kinetic algorithm.

The essence of this result is, that the biologically more important combinatory map of structures is rather independent of the chosen algorithm. This is especially important in the light of new results which showed, that it is possible to cross the whole sequence space without ever changing the $\texttt{mfe}_F$ structure.

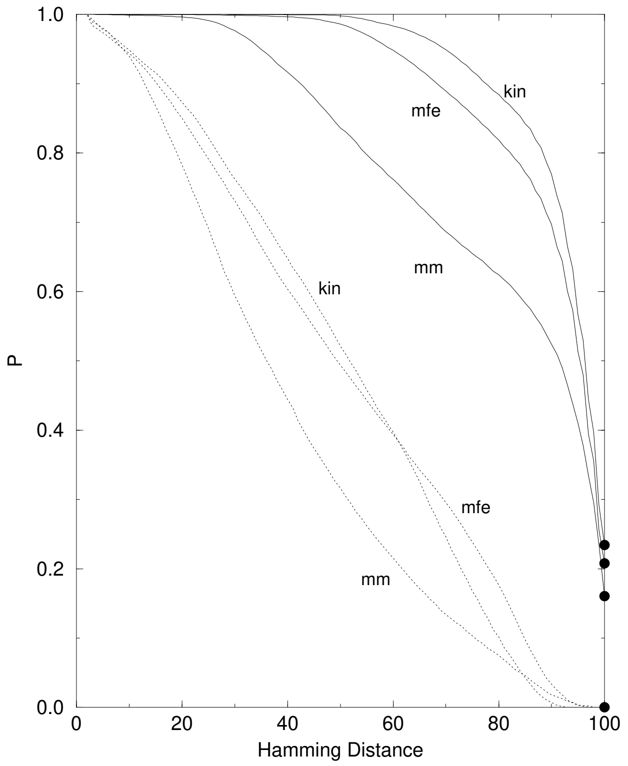


Figure 8: Probability P for a neutral path to reach at least a distance d from the starting point. Solid lines refer to **GCAU** alphabet, dotted lines belong to the **GC** alphabet. Note that the full lines do not end at $P{=}0$.

Neutral networks

A sequence x is said to be compatible with a secondary structure S if it can fold into S (irrespective of energy considerations). It is convenient to consider two sequences x and y which are both compatible with the secondary structure S as neighbors if they differ either by single base that is unpaired in S or if they differ by two bases belonging to a base pair of S. A neutral network is a subset of the set of compatible sequences consisting of those sequences that fold into the same secondary structure. It is connected in the sense of the definition of neighborhoods in the set of compatible sequences. In other words, a connected component contains all sequences which can be reached by mutating a single base or changing a single base pair of S without changing the structure into which the sequences fold. Neutral networks may consist of a single component or they may have two, three or many components (See a recent analytical approach to the problem by means of random graph theory: Reidys *et al.* 1995). It has been shown by extensive computer studies that there exist extended neutral networks percolating through the entire sequence space for the combinatory map of RNA secondary structures generated by the $\texttt{mfe}_F$ algorithm (Schuster *et al.* 1994, Grner *et al.* 1996a,b, Huynen *et al.* 1996b).

A neutral path is a path in neutral network such that the distance from the starting point increases at each step. On a neutral path of length n one can walk through all of sequence space without ever changing the secondary structure. Numerical results from the combinatory maps of all algorithms show that neutral paths are very long everywhere. Consequently we find extended neutral networks independently of the algorithm and the parameter set applied. The same is true for the phenomenon of shape space covering. It depends on essentially only two properties: there have to be many more sequences than structures and the sequences folding into the same structure have to be distributed (approximately) randomly in sequence space. Both requirements are fulfilled independently of algorithm and parameter sets.

Discussion

Results of secondary structure predictions were studied for different algorithms and parameter sets. The discussions starts with specific findings leaving the general issues for the final part.

Algorithms

In order to obtain information about the specific effects an algorithm imposes on the statistical features of large ensembles of structures it is essential to consider different algorithms with the same set of energy parameters. Mfe algorithms find the global free energy minimum of the string by putting together all possible combinations of stacks and loops and selecting the one yielding the lowest free energy value. Kinetic algorithms generate the structure by successive incorporation of stacks that have the highest equilibrium constant of formation (and thus minimize stepwise free energy gains). In many cases the kinetic algorithm does not find the global minimum. This is most clearly seen in the mean free energies of structures which are somewhat lower than those of the kinetic structures. It is therefore easy to visualize that the kinetic algorithm produces structures with longer stacks which are more stable. In addition mean loop sizes are larger as well. The overall effect of folding kinetics is a constraint on base pair formation resulting in fewer pairs than would be obtained with the mfe criterion. As a direct consequence more bases must remain external. The number of joints is nearly identical for both classes of algorithms which might be accidental in the sense that it is still a finite size effect.

Dependence on parameter sets

Non unexpectedly, the choice of the parameter set has a strong influence on structure prediction. The old energy data set compiled by Salser (1977) differs from the updated version of Freier *et al.* (1986) in several points:

- the base pairing energy of the **GC** pair is higher,
- no mismatch energies for nucleotides at the ends of stacks are taken into account,
- hairpin loops have stronger destabilization effects, and
- internal loops destabilize less.

The change of these energy parameters has the effect that $\texttt{mfe}_S$ structures set have less base pairs and more loops. The stacks are shorter and the loops are larger than in the $\texttt{mfe}_F$ structures. The mean loop degree is higher in $\texttt{mfe}_S$ structures. Important qualitative differences can be seen in the number of joints where $\texttt{mfe}_S$ structures reach a constant value at fairly low chain lengths ($n \approx 50$), whereas the $\texttt{mfe}_F$ parameter set yields an increase with n which is also shown by the data derived with the kinetic algorithm. The effect observed with the numbers of external digits is a consequence of the numbers of joints. The reasons for these differences are to be seen mainly in the fact that hairpin loops are more strongly destabilized by the $\texttt{mfe}_S$ parameterization and hence more internal loops and multiloops are formed. This explains also the higher mean loop degree in a straightforward way. Shorter stacks are found because internal loop are relatively favorable and greater loop sizes are obtained since they are less destabilized than the small ones. The number of base pairs increases somewhat less than with the $\texttt{mfe}_F$ parameters. With respect to the numerical values the $\texttt{mfe}_S$ structures are most stable but this is certainly an effect of the numerical parameter values and hence not conclusive for stability arguments in general.

General conclusions

The choice of algorithms and parameter sets has a strong influence on the details of the predicted structures. Structure statistics, however, was found to be much less sensitive. The differences are readily explained in terms of differences in folding criteria and parameters. The same is true for the autocorrelation function of free energies.

As a kind of surprise we found also features of RNA secondary structures that were (almost) independent of both algorithms and parameters: these were the autocorrelation function of structures, the existence and the component structure of neutral networks as well as the phenomenon of shape space covering. Apparently, these features depend primarily on the base pairing logic and consequently results obtained with **GC**-sequences differ strongly from those derived from the natural **AUGC** alphabet.

Acknowledgements. The work reported here has been supported financially be the Austrian *Fonds zur Frderung der wissenschaftlichen Forschung* (Projects No. P-8526-CHE and P-9942-PHY). Many fruitful discussions with Dr. Walter Fontana are gratefully acknowledged.

References

Abrahams JP, van den Berg M, van Batenburg E, Pleij C (1990) Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. Nucl. Acids Res. 18:3035-3044

Bartel DP, Szostak JW (1993) Isolation of ribozymes from a large pool of random sequences. Science 261:1411-1418

Beaudry AA, Joyce GF (1992) Direct evolution of an RNA enzyme. Science 257:635-641

Bonhoeffer S, McCaskill JS, Stadler PF, Schuster P (1993) RNA multistructure landscapes. Eur.Biophys.J. 22:13-24

Borer PN, Dengler B, Tinoco Jr. I, Uhlenbeck OC (1974) Stability of ribonucleic acid double stranded helices. J.Mol.Biol. 86:843-852

Chastain M, Tinoco Jr. I (1991) Structural elements in RNA. Proc.Nucleic Acids Res. 41:131-177

Cech TR (1988) Conserved sequences and structures of group I introns: building an active site for RNA catalysis - a review. Gene 73:259-271

Eigen M, McCaskill JS, Schuster P (1989) The molecular quasispecies. Adv.Chem.Phys. 75:149-263

Ellington AD, Szostak JW (1990) *In vitro* selection of RNA molecules that bind specific ligands. Nature 346:818-824

Fink TR, Crothers DM (1972) Free energy of imperfect nucleic acid helices I. The bulge defect. J.Mol.Biol. 66:1-12

Fontana W, Griesmacher T, Schnabl W, Stadler PF, Schuster P (1991) Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. Mh.Chem. 122:795-819

Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tarazona P, Weinberger ED, Schuster P (1993a) RNA folding and combinatory landscapes. Phys.Rev.E 47:2083-2099

Fontana W, Konings DAM, Stadler PF, Schuster P (1993b) Statistics of RNA secondary structures. Biopolymers 33:1389-1404

Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH (1986) Improved free-energy parameters for predictions of RNA duplex stability. Biochemistry 83:9373-9377

Gesteland RF, Atkins JF, eds. (1993) The RNA world. Cold Spring Harbor Laboratory Press, Plainview, NY

Giegé R, J.D. Puglisi JD, and C. Florentz C (1993) tRNA structure and aminoacylation efficiency. Progr.Nucl.Ac.Res.Mol.Biol. 45:129-206

Gralla J, Crothers DM (1973a) Free energy of imperfect nucleic acid helices II. Small hairpin loops. J.Mol.Biol. 73:497-517

Gralla J, Crothers DM (1973b) Free energy of imperfect nucleic acid helices III. Small internal loops. J.Mol.Biol. 78:301-139

Grner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P (1996a) Analysis of RNA sequence structure maps by exhaustive enumeration. Part I: Neutral networks. Mh. Chem. 127:355-374

Grner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P (1996b) Analysis of RNA sequence structure maps by exhaustive enumeration. Part II: Structures of neutral networks and shape space covering. Mh. Chem. 127:375-389

Gultyaev AP (1991) The computer simulation of RNA folding involving pseudoknot formation. Nucl. Acids Res. 19:2489-2494

Gutell RR (1992) Evolutionary characteristics of 16S and 23S rRNA structures. In: Hartman H, Matsuno K (eds) The Origin and Evolution of the Cell. World Scientific, Singapore, pp.243-310

Gutell RR (1993) Comparative studies of RNA: inferring higher-order structure from patterns of sequence variation. Curr.Opin.Struct.Biol. 3:313-322

Higgs PG (1993) RNA secondary structure: a comparison of real and random sequences. J.Phys. I France 3:43-59

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. Mh.Chem. 125:167-188

Hofacker IL, Schuster P, Stadler PF (1996) Combinatorics of RNA secondary structures. Submitted to SIAM J.Disc.Math.

Hogeweg P, Hesper B (1984) Energy directed folding of RNA sequences. Nucl. Acids Res. 12:67-74

Huynen MA, Perelson AS, Vieira WA, Stadler PF (1996a) Base pairing probabilities in a complete HIV-1 RNA. J.Comp.Biol., in press

Huynen MA, Stadler PF, Fontana W (1996b) Smoothness within ruggedness: The role of neutrality in adaptation. Proc.Natl.Acad.Sci.USA 93:397-401

Jaeger JA, Turner DH, Zuker M (1989) Improved predictions of secondary structures for RNA. Biochemistry 86:7706-7710

Jenison RD, Gill SC, Pardi A, Polisky B (1994) High-resolution molecular discrimination by RNA. Science 263:1425-1429

Joyce GF (1989) Amplification, mutation, and selection of catalytic RNA. Gene 82:83-87

Kauffman SA (1993) The origins of order. Oxford University Press, Oxford UK

Konings DAM, Hogeweg P (1989) Pattern analysis of RNA secondary structure. Similarity and consensus of minimal energy folding. J.Mol.Biol. 207:597-614

Konings DAM, Gutell RR (1995) A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. RNA 1:559-574

Le SY, Maizel Jr. JV (1989) A method for assessing the statistical significance of RNA Folding. J.Theor.Biol. 138:495-510

Le SY, Zuker M (1990) Common structures of the 5' non-coding RNA in eneteroviruses and rhinoviruses. Thermodynamical stability and statistical significance. J.Mol.Biol. 216:729-741

Maly P, Brimacombe R (1983) Refined secondary structure models for the 16S and 23S ribosomal RNA of Escherichia coli. Nucl. Acids Res. 11:7363-7386

Martinez HM (1984) An RNA folding rule. Nucl. Acids Res. 12:323-334

Michel F, Umesono K, Ozeki H (1989) Comparative and functional anatomy of group II catalytic introns - a review. Gene 82:5-30

Michel F, Westhof E (1990) Modelling the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J.Mol.Biol. 214:585-610

McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structures. Biopolymers 29:1105-1119

Morgan SR, Higgs PG (1995) Thermodynamics of RNA folding. When is an RNA molecule in equilibrium? In: Morán F, Moreno A, Merelo JJ, Chacón P (eds) Advances in Artificial Life. Lecture Notes in Artificial Intelligence, Proc.Vol.929, Springer-Verlag, Berlin, pp.852-861

Ninio J (1979) Prediction of pairing schemes in RNA molecules. Loop contributions and energy of wobble and non-wobble pairs. Biochimie 61:1133-1150

Nussinov R, Peiczenik G, Griggs JR, Kleitman DJ (1978) Algorithms for loop matchings. SIAM J.Appl.Math. 35:68-82

Nussinov R, Jacobson AB (1980) Fast algorithm for predicting the secondary structure of single stranded RNA. Proc.Natl.Acad.Sci.USA 77:6309-6313

Nussinov R, Tinoco Jr. I (1982) Small changes in free energy assignments for unpaired bases do not affect predicted secondary structures in single stranded RNA. Nucl. Acids Res. 10:341-350

Ohmori K, Tanaka E (1988) A unified view on tree metrics. In: Ferraté G (ed) Syntactic and Structural Pattern Recognition. NATO ASI Series Vol. F 45. Springer-Verlag, Berlin, pp.85-100

Pace N, Olsen G, Woese C (1986) Ribosomal RNA phylogeny and primary lines of evolutionary descent. Cell 45:325-326

Pörschke D (1977) Elementary steps in base recognition and helix-coil transitions in nucleic acids. In: Pecht I, Rigler R (eds) Chemical Relaxation in Molecular Biology. Molecular Biology, Biochemistry and Biophysics, Vol.24, Springer-Verlag, Berlin, pp.191-218

Papanicolaou C, Gouy M, Ninio J (1984) An energy model that predicts the correct folding of both tRNA and the 5S RNA molecules. Nucl. Acids Res. 12:31-44

Peritz AE, Kierzek R, Sugimoto N, Turner DH (1991) Thermodynamic study of internal loops in oligoribonucleotides: symmetric loops are more stable than asymmetric loops. Biochemistry 30:6428-6436

Pleij CWA (1990) Pseudoknots: a new motif in the RNA game. TIBS 15:143-147

Pleij CWA, Rietveld K, Bosch L (1995) A new principle of RNA folding based on pseudoknotting. Nucl. Acids Res. 13:1717-1731

Puglisi JD, Wyatt JD, Tinoco Jr. I (1988) A pseudoknotted RNA oligonucleotide. Nature 331:283-286

Reidys C, Stadler PF, Schuster P (1995) Generic properties of combinatory maps. Neutral networks of RNA secondary structures. Bull.Math.Biol., submitted

Reidys C, Stadler PF (1996) Bio-molecular shapes and algebraic structures. Computers Chem. 20:85-94

Salser W (1977) Globin messenger sequences – Analysis of base-pairing and evolutionary implications. Cold Spring Harbor Symp. Quant. Biol. 42:985-1002

Sankoff D, Morin AM, Cedergren RJ (1978) The evolution of 5S RNA secondary structure. Can.J.Biochem. 56:440-443

Santa Lucia Jr.J, Kierzek R, Turner DH (1992) Context dependence of hydrogen bond free energy revealed by substitutions in an RNA hairpin. Science 256:217-219

Sassanfar M, J.W. Szostak, JW (1993) An RNA motif that binds ATP. Nature 364:550-553

Schuster P, Stadler PF (1994) Landscapes. Complex optimization problems and biopolymer structures. Computers Chem. 18:295-324

Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc.Roy.Soc.Lond.B 255:279-284

Schwienhorst A, Lindemann BL (1995) Hammerhead-RNA from random sequences. Proc.Natl.Acad.Sci.USA, in press

Shapiro BA (1988) An algorithm for comparing multiple RNA secondary structures. CABIOS 4:381-393

Shapiro BA, Zhang K (1990) Comparing multiple RNA secondary structures using tree comparisons. CABIOS 6:309-318

Shapiro BA, Chen JH, Busse T, Navetta J, Kasprzak W, Maizel J (1995) Optimization and performance analysis of a massively parallel dynamic programming algorithm for RNA secondary structure prediction. Int.J.Supercomp.Appl. 9:29-39

Stadler PF, Schnabl W (1992) The landscape of the travelling salesman problem. Phys.Lett.A 161:337-344

Stadler PF, Krakhofer B (1996) Local minima of p-spin models. Rev.Mex.Fis., in press

Sugimoto N, Kierzek R, Turner DH (1987a) Sequence dependence for the energetics of dangling ends and terminal base pairs in ribonucleic acid. Biochemistry 26:4554-4558

Sugimoto N, Kierzek R, Turner DH (1987b) Sequence dependence for the energetics of terminal mismatches in ribooligonucleotides. Biochemistry 26:4559-4561

Tacker M, Fontana W, Stadler PF, Schuster P (1994) Statistics of RNA melting kinetics. Eur.Biophys.J 23:29-38

Tai, K (1979) The tree-to-tree correction problem. J.Ass.Comput.Mach. 26:422-433

Tuerk C, Gauss P, Thermes C, Groebe DR, Gayle M, Guild N, Stromo G, D'Aubenton-Carafa Y, Uhlenbeck OC, Tinoco Jr. I, Brody EN, Gold L (1988) CUUCGG hairpins: Extraordinarily stable RNA secondary structures associated with various biochemical processes. Proc.Natl.Acad.Sci.USA 85:1364-1368

Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249:505-510

Tinoco Jr. I, Uhlenbeck OC, Levine MD (1971) Estimation of secondary structure in nucleic acids. Nature 230:362-367

Tinoco Jr. I, Borer PN, Dengler B, Levine ND, Uhlenbeck OC, Crothers DM, Gralla J (1973) Improved estimation of secondary structure in ribonucleic acids. Nature 246:40-41

Uhlenbeck OC, Borer PN, Dengler B,Tinoco Jr. I (1973) Stability of RNA hairpin loops: $A_6-C_m-U_6$. J.Mol.Biol. 73:483-496

Waterman MS (1978) Secondary structure of single-stranded nucleic acids. Adv.Math.Suppl.Studies 1:167-212

Waterman MS, Smith TF (1978) RNA secondary structure: a complete mathematical analysis. Math.Biosci. 42:257-266

Waterman MS, Byers TH (1985) A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. Math.Biosci. 77:179-188

Waterman MS, Smith TF (1986) Rapid dynamic programming algorithms for RNA secondary structure. Adv.Appl.Math. 7:455-464

Waterman MS (1989) Consensus methods for folding single-stranded nucleic acids. In: Waterman MS (ed) Mathematical Methods for DNA Sequences. CRC Press, Boca Raton FL, pp 185-224

Weinberger ED (1990) Correlated and uncorrelated fitness landscapes and how to tell the difference. Biol.Cybern. 63:325-336

Westhof E, Jaeger L (1992) RNA pseudoknots. Curr.Opin.Struct.Biol 2:327-333

Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Jones DF (ed) Proceedings of the Sixth International Congress on Genetics, Vol.1, pp 356-366

Zuker M (1989a) On finding all suboptimal foldings of an RNA molecule. Science 244:48-52

Zuker M (1989b) The use of dynamic programming algorithms in RNA secondary structure predictions. In: Waterman MS (ed) Mathematical Methods for DNA Sequences. CRC Press, Boca Raton FL, pp 159-184

Zuker M, Stiegler P (1981) Optimal computer colding of large RNA sequences using thermodynamics and auxiliary information. Nucl. Acids Res. 9:133-148

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

Appendix: Random Structures

The number of possible secondary structures compatible with a sequence can be enumerated recursively as shown in Waterman and Smith (1978). Let $m = 3$ be the minimum size of a hairpin, and $\Pi_{\sigma_p,\sigma_q} = 1$ iff σ_p and σ_q can pair, 0 otherwise. Denote by $S_{p,q}$ the number of possible structures on the subsequence $[\sigma_p \ldots \sigma_q]$, then

$$S_{l,n+1} = S_{l,n} + \sum_{k=l}^{n-m} S_{l,k-1} S_{k+1,n} \Pi_{\sigma_k,\sigma_{n+1}}$$

with the initial conditions $S_{i,j} = 1$ for $j - i \geq m$.

For random sequences with stickiness p the expected number $\bar{S}_n$ of compatible structures is then

$$\bar{S}_{n+1} = \bar{S}_n + p \sum_{k=1}^{n-m} \bar{S}_{k-1} \bar{S}_{n-k} = \bar{S}_n + p \sum_{k=m}^{n-1} \bar{S}_k \bar{S}_{n-k-1}.$$

Extending this relation to structures with a minimum stack length l along the lines shown in (Hofacker *et al.* 1996), we get the coupled recursions

$$\begin{aligned}
\bar{\Psi}_{n+1}(l) &= \bar{\Psi}_n(l) + \sum_{k=m+2l-2}^{n-1} \bar{\Psi}^*_k(l) \bar{\Psi}_{n-k-1}(l) \\
\bar{\Psi}^*_n(l) &= p \sum_{k=l-1}^{(n-m)/2} \bar{\Psi}^{**}_{n-2k}(l) p^k \\
\bar{\Psi}^{**}_n(l) &= \bar{\Psi}_n(l) - \bar{\Psi}^*_{n-2}(l)
\end{aligned}$$

$$\bar{\Psi}_n(l) = \Psi^{**}_{n+1}(l) = 1 \quad n < m + 2l,$$

$$\bar{\Psi}^*_n(l) = 0 \quad m + 2l - 2$$

These relations can now be used to construct structures with a probability distribution identical to choosing a sequence at random from an alphabet with stickiness p and then randomly choosing a structure from the set of all structures compatible with this particular sequence.

For the simpler case of minimum stack length 1 it is easy to see how to construct such structures recursively: The first base of a structure of length n has to be either unpaired or paired to some base $k \leq n$, the probability for the first case being S_{n-1}/S_n and $S_{k-2}S_{n-k}/S_n$ for the latter. We now have reduced the problem to constructing a structure of length $n-1$ or two structures of lengths $k-2$ and $n-k$, respectively. We proceed until we reach substructures of length $m = 3$ which then must be entirely unpaired. For minimum stack lengths > 1 the procedure is completely analogous if slightly more involved.