Eur Biophys J (1993) 22: 13–24

European Biophysics Journal

RNA Multi-structure landscapes

A study based on temperature dependent partition functions

S. Bonhoeffer [1]*, J. S. McCaskill [2], P. F. Stadler [1], P. Schuster [1, 3, 4]

[1] Institut fur Theoretische Chemie, Universität Wien, A-1090 Wien, Austria
[2] Max-Planck-Institut für Biophysikalische Chemie (Karl-Friedrich-Bonhoeffer-Institut), W-3400 Göttingen, Germany
[3] Institut für Molekulare Biotechnologie, O-6900 Jena, Germany
[4] Santa Fe Institute, Santa Fe, NM 85701, USA

Received. 5 August 1992 / Accepted: 12 January 1993

Abstract. Statistical properties of RNA folding landscapes obtained by the partition function algorithm (McCaskill 1990) are investigated in detail. The pair correlation of free energies as a function of the Hamming distance is used as a measure for the ruggedness of the landscape. The calculation of the partition function contains information about the entire ensemble of secondary structures as a function of temperature and opens the door to all quantities of thermodynamic interest, in contrast with the conventional minimal free energy approach. A metric distance of structure ensembles is introduced and pair correlations at the level of the structures themselves are computed. Just as with landscapes based on most stable secondary structure prediction, the landscapes defined on the full biophysical **GCAU** alphabet are much smoother than the landscapes restricted to pure **GC** sequences and the correlation lengths are almost constant fractions of the chain lengths. Correlation functions for multi-structure landscapes exhibit an increased correlation length, especially near the melting temperature. However, the main effect on evolution is rather an effective increase in sampling for finite populations where each sequence explores multiple structures.

Key words: Fitness landscapes – Partition function – Quasispecies – RNA secondary structures

1 Introduction

In Sewall Wright's model (Wright 1932) biological evolution is understood as a fitness optimizing process on a very complex abstract landscape. Manfred Eigen (1971) made an attempt to model the origin of biological information by placing Darwin's principle onto a physical basis. Reaction kinetics of erroneously replicating biopolymers was applied to study evolution on the molecular level and led to the concept of the (molecular) quasispecies (Eigen and Schuster 1977; Eigen et al. 1988, 1989) which may be considered as the genetic reservoir of an asexually replicating ensemble. The molecular concept enables a more complete description of the evolutionary optimization process, since it includes the relationships between sequence, structure, and function of biopolymers as the ultimate source of the reaction rate constants of the kinetic ansatz. Indeed, the fitness values are obtained as (mostly linear) combinations of these rate constants. Specific models of fitness landscapes were investigated by analytical and numerical tools (Swetina and Schuster 1982; Schuster and Swetina 1988; Nowak and Schuster 1989). Replication with decreasing replication accuracy shows an error threshold which sharpens with increasing chain lengths, thereby closely resembling cooperative transitions in biopolymers. No (finite) population can be stationary at error rates exceeding the value of the threshold which implies complete loss of genetic information. This approach was complemented by a study of the generic behavior of evolution on molecular fitness landscapes (McCaskill 1984a). Landscapes with finite correlation lengths for fitness values generically show error thresholds at copying fidelities whose locations depend on the distribution of fitness values. Correlation lengths of fitness landscapes are related to the numbers of local optima, and thus also to the speed of optimization. Besides the distribution of fitness values their autocorrelation functions and the correlation lengths derived from them, provide the most important characteristic for evolution on fitness landscapes (Eigen et al. 1989: p. 217).

RNA is now recognized as having a special role amongst the information containing molecules of molecular biology, since it does not only represent the template in replication and translation but also acts as catalyst in RNA ligation and cleavage (Cech 1990) as well as aminoacyl esterase activity (Piccirilli et al. 1992). In addition it has a predominant role in ribosomal translation: ribosomes from the eubacterium *Thermus aquaticus* retain there peptidyl transferase activity after removal of 80% of its protein (Noller et al. 1992). The secondary structure folding problem for RNA has a unique position in molecular biology as a computationally tractable, experimen-

* *Present address*: Dept. of Zoology, University of Oxford, Oxford, UK

Correspondence to. P. Schuster

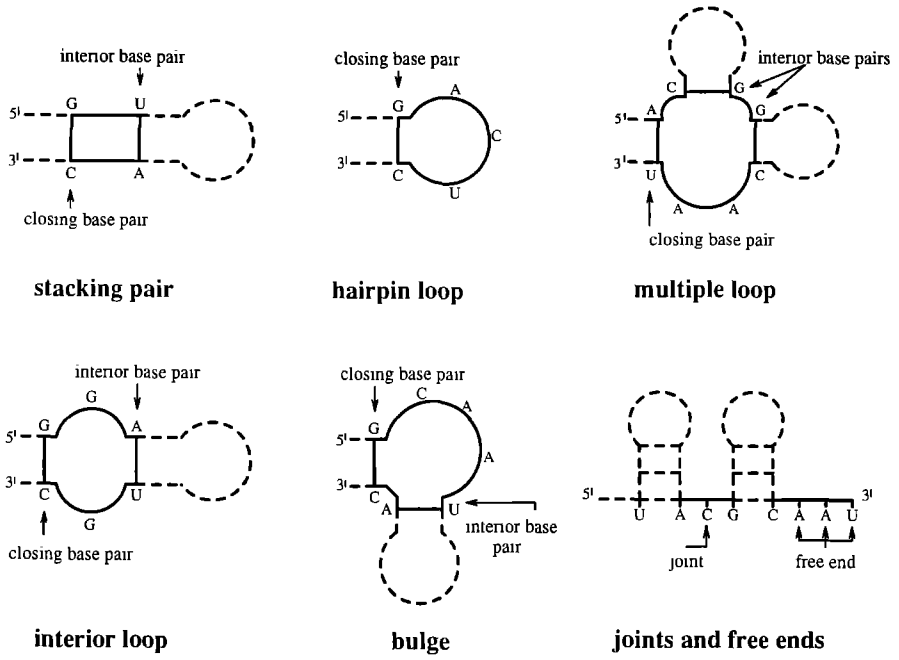


Fig. 1. Basic structure elements. Every structure within the described secondary structure model can be decomposed into such basic elements. The free energy of a secondary structure is simply the sum over all contributions of these structural elements for which detailed experimental data are available

tally well parametrized, biologically significant and clearly separable component of the sequence-function relationship (Fontana et al. 1991). The feasibility of using fitnesses based on RNA secondary structures has been verified by the computational studies of evolutionary optimization in an RNA world based on replication and mutation (Fontana and Schuster 1987, Fontana et al. 1989). In addition, the statistical properties of RNA folding landscapes and RNA structures have been investigated extensively in recent time (Fontana et al. 1991; 1992a, b).

These studies of RNA landscapes were based on a dynamic programming folding algorithm due to Waterman and Smith (1978) and originally implemented by Zuker and Stiegler (1981), which predicts only the most stable secondary structure of an RNA molecule according to the minimum free energy criterion. In reality, however, an RNA molecule is better described by an ensemble of secondary structures, which have free energies close to the minimum of free energy. A new algorithm has been published in McCaskill (1990), which allows one to calculate the partition function of the entire ensemble of secondary structures.

In this contribution, we present studies on RNA landscapes by means of this partition function algorithm. Having the partition function allows one to obtain the probabilities of individual structures or subclasses of structures and all thermodynamic quantities of interest. The partition function can be considered as a function of temperature or of various parameters of the model such as salt concentration.

2 Theory

2.1 RNA secondary structures

Predicting the three dimensional structure of an RNA molecule merely from the sequence is still at an early stage of development (Major et al. 1991). Therefore current algorithms focus on the prediction of the secondary structure, i.e. the "skeleton" of the 3D structure formed by the Watson-Crick base pairs $\mathbf{G}\equiv\mathbf{C}$ and $\mathbf{A}=\mathbf{U}$ as well as the "wobble" $\mathbf{G}-\mathbf{U}$ base pairs. This is both physically and biologically meaningful since, firstly, the main part of the free energy of structure formation results from these base pairing interactions and, secondly, a strong conservation of secondary structure elements in evolution has been reported (Sankoff et al. 1988; Cech 1988; Le and Zuker 1990).

An RNA molecule of chain length v can be represented as a string $I=[s_1 s_2 s_3 \ldots s_v]$, where the letters s_i are taken here from the natural four letter alphabet $\{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$, or from a binary alphabet $\{\mathbf{G}, \mathbf{C}\}$. By IUPAC convention, numbering within the string of letters begins at the 5′-end of the RNA molecule.

A secondary structure S is defined as the set of all base pairs (s_i, s_j) with $(i<j)$ fulfilling the following two requirements (Waterman 1978):

1) each base is involved in at most one base pair,

2) there are no knots or pseudoknots, i.e. if (s_i, s_j) and (s_k, s_l) are base pairs then $i<k<l<j$ or $k<i<j<l$.

Knots, pseudoknots and triple-helices are considered as parts of the tertiary structure – they are beyond the scope of our model. There are two good reasons for these restrictions in the definition of secondary structures: firstly, they are necessary in order to allow for efficient algorithms, and secondly, no reliable experimental data are available for the free energy contributions of pseudoknots or triple helices.

The basic elements of secondary structures are shown in Fig. 1. For the free energies of these building blocks experimental data are available depending on the length of the unpaired regions as well as which interior and closing base pairs are involved (Fig. 1). These elements are assumed to contribute additively to the overall free

energy of the complete secondary structure. For multiple loops a linear assumption is used (Zuker and Sankoff 1984), free unpaired regions like joints and free ends are assumed to have vanishing free energies. The data set used here has been taken from the literature (Freier et al. 1986; Jaeger et al. 1989).

The unique decomposition of secondary structures outlined above suggests a simple string representation of structures by identifying a base pair with a pair of matching brackets and denoting an unpaired digit by a circle (upstream is understood in 5′-3′ direction in accord with the IUPAC convention; downstream refers to the opposite direction):

< upstream paired base
> downstream paired base
○ single-stranded base

The so-called *mountain representation of secondary structures* derived by Hogeweg and Hesper (1984) is closely related to this string representation. The bracket notation is coding for a tree (Fontana et al. 1991). Other tree representations have been proposed for RNA secondary structures as well (Zuker and Sankoff 1984; Shapiro 1988; Shapiro and Zhang 1990; Fontana et al. 1992a).

2.2 Partition function for RNA secondary structures

The most common folding algorithm (Waterman 1978; Zuker and Stiegler 1981) predicts only the thermodynamically most stable secondary structure. At room temperature, however, RNA molecules do not take on only the most stable structure, they seem to rapidly change their conformation between structures with similar free energies. The simplest way to account for this is to compute not only the optimal structure but all structures within a certain range of free energies (Waterman and Byers 1985).

A recent algorithm (McCaskill 1990) is capable of calculating the entire partition function

$$Q(I) = \sum_{S \in M(I)} e^{-F(S)/kT} \tag{1}$$

$M(I)$ denotes the set of all secondary structures of a particular sequence I, k is Boltzmann's constant and T is the absolute temperature.

The additivity of the free energy contributions in the secondary structure model implies a factorization of the partition function which again enables a dynamic programming scheme. The partition function algorithm and the Zuker-Stiegler algorithm have the same time complexity: the performance time increases as the third power of the sequence length v.

All thermodynamic quantities of interest can be derived from the partition function Q. Here we are mainly interested in the free energy of folding

$$F(I) = kT \log Q(I), \tag{2}$$

i.e. the free energy of the ensemble of secondary structures of a given sequence I at thermodynamic equilibrium. We remark that this is not the free energy of a single structure as referred to in the conventional minimal free energy calculation. There (Tinoco 1971) the use of free energy refers to the fact that each individual secondary structure is actually thermodynamically parametrized as an ensemble of substructures. Following this parametrization, the calculation is a minimization which neglects the entropic contributions of different related secondary structures and so, one is not actually calculating a free energy.

In order to calculate the partition function at different temperatures, one needs to extrapolate the experimental data for the secondary structure elements, which are measured mostly near 37 °C. In addition enthalpies and entropies are assumed to be temperature independent in the range we are interested in here. For stacks both enthalpies ΔH°_{37} and entropies ΔS°_{37} are experimentally accessible, so we get for the free energy of stacks

$$\text{Stacks:} \quad \Delta G^\circ(T) = \Delta H^\circ_{37} - T\,\Delta S^\circ_{37}\,. \tag{3}$$

The assumption that ΔG° for loops is purely entropic has sufficed so far (Freier et al. 1986, Sugimoto et al. 1987a, b, Turner et al. 1988, Jaeger et al. 1989, Peritz et al. 1991). It is the analogue of a conventional assumption in polymer physics. Hence we have a temperature dependence for the free energy contributions of interior loops, bulges and multiple loops of the form

$$\text{Loops:} \quad \Delta G^\circ(T) = -T\,\Delta S^\circ_{37}\,. \tag{4}$$

The minimal free energy of a single structure therefore does not coincide with the 0 K limit of the free energy of the entire ensemble, since the former calculates the most stable secondary structure at 37 °C. Extrapolation of the data far away from 37 °C is of course critical, because a linear temperature dependence is only valid in a narrow range around the reference temperature (37 °C). Nevertheless, this parametrization turned out to be sufficient to model the statistical properties of RNA folding landscapes in a certain neighborhood of the reference.

The partition function algorithm, like nature, does not yield a single secondary structure, but it allows one to compute the probability that a given secondary structure S occurs in the equilibrium ensemble:

$$\text{Prob}(S) = \frac{1}{Q} \exp\left\{-\frac{F(S)}{kT}\right\} \tag{5}$$

The most probable structure is well described by the pairing matrix $P = \{p_{ij}\}$ (McCaskill 1990)

$$p_{ij} = Prob\{i \text{ and } j \text{ form a pair}\}\,. \tag{6}$$

In addition the matrix contains information about the probability of alternative structures. It can be obtained by backtracking with cubic time complexity.

2.3 Comparing equilibrium structure ensembles

The bracket notation introduced in 2.1 allows the interpretation of a secondary structure as a string $s(S)$. A standard maximal similarity alignment algorithm (Waterman 1984) can be used to define similarity and distance measures between secondary structures. This approach

has been used by Hogeweg and coworkers (Hogeweg and Hesper 1984, Konings 1989, Konings and Hogeweg 1989):

$$d(S_1, S_2) \equiv d_{al}[s(S_1), s(S_2)] \tag{7}$$

Alternative definitions of distance are based on the tree representations of secondary structures and invoke tree editing to define a distance (Fontana et al. 1991; 1992a, b). Both structure distance definitions have shown to be statistically equivalent.

Because of limitation in the computer resources we cannot explicitly use all individual secondary structures of a given sequence. Therefore we generalize the string representation of single structures and compute for each position i in the sequence the probability to be upstream paired, downstream paired or unpaired.

$$\begin{aligned} p_i^{<} &= \sum_{j>i} p_{ij} \\ p_i^{>} &= \sum_{j<i} p_{ij} \end{aligned} \tag{8}$$

The probability that the base at position i is unpaired is $p_i^{\circ} = 1 - p_i^{>} - p_j^{<}$. Note that in contrast to the corresponding string encoding for a single secondary structure the vectors $p^{<} = (p_1^{<}, .., p_i^{<}, \ldots)$ and $p^{>} = (p_1^{>}, \ldots, p_i^{>} \ldots)$ no longer form a one-to-one encoding of the secondary structure ensemble $\tilde{S}$ since the base-pairing matrix P cannot be restored from them.

A reasonable definition for the distance of two such vectors, $p(\tilde{S}_1)$ and $p(\tilde{S}_2)$, uses again an alignment procedure at the level of the vectors $p^{<}$, $p^{>}$ and p°. We then define the similarity measure for an aligned position (i, j) by

$$\begin{aligned} \gamma(i,j) = &\sqrt{p_i^{>}(\tilde{S}_1)\, p_j^{>}(\tilde{S}_2)} + \sqrt{p_i^{<}(\tilde{S}_1)\, p_j^{<}(\tilde{S}_2)} \\ &+ \sqrt{p_i^{\circ}(\tilde{S}_1)\, p_j^{\circ}(\tilde{S}_2)}\,. \end{aligned} \tag{9}$$

Instead of the geometric mean in the above definitions we could for example use a logarithmic mean. The similarity measure $\hat{\gamma}(\boldsymbol{i}, \boldsymbol{j})$ of a particular alignment $(\boldsymbol{i}, \boldsymbol{j})$ of the two structure ensembles $\tilde{S}_1$ and $\tilde{S}_2$ is then given by the following sum over all aligned positions

$$\hat{\gamma}(\boldsymbol{i}, \boldsymbol{j}) = \sum_{i \text{ aligned } j} \gamma(i,j)\,. \tag{10}$$

The similarity measure of two structure ensembles is defined by the optimal alignment

$$\mathrm{sim}(\tilde{S}_1, \tilde{S}_2) = \max_{(\boldsymbol{i}, \boldsymbol{j})} \hat{\gamma}(\boldsymbol{i}, \boldsymbol{j}) \tag{11}$$

As an immediate consequence we find

$$0 \le \mathrm{sim}(\tilde{S}_1, \tilde{S}_2) \le \min(\nu_1, \nu_2) \tag{12}$$

where ν_1 and ν_2 are the chain lengths of the two molecules.

Finally, a distance measure of secondary structure ensembles may be defined by

$$\delta(\tilde{S}_1, \tilde{S}_2) = \tfrac{1}{2}(\nu_1 + \nu_2) - \max_{(\boldsymbol{i}, \boldsymbol{j})} \gamma(\boldsymbol{i}, \boldsymbol{j}) \tag{13}$$

which is metric and fulfils

$$0 \le \delta(\tilde{S}_1, \tilde{S}_2) \le \frac{\nu_1 + \nu_2}{2}\,. \tag{14}$$

In addition we define a general multistructure distance

$$\delta_m(x, y) = \sum_{S, S'} p_x(S)\, p_y(S')\, \delta(S, S') \tag{15}$$

where $p_x(S)$ is the equilibrium probability of structure S in the ensemble of secondary structures for sequence x and $p_y(S')$ the equilibrium probability of structure S' in the ensemble of secondary structures for sequence y.

2.4 Complex combinatory maps

In order to study the statistical properties of RNA molecules with respect to their genetic relation we first need a few formal definitions: the set of all sequences with given length ν composed from an alphabet A of size $\varkappa$ together with the Hamming distance (Hamming 1986) form the so-called sequence space (Maynard-Smith 1970; Eigen 1971; Swetina and Schuster 1982) (A^{ν}, d). In this contribution we deal with the two alphabets $\{\mathbf{G}, \mathbf{C}\}$ and $\{\mathbf{G}, \mathbf{C}, \mathbf{A}, \mathbf{U}\}$. A landscape is obtained by assigning a value $f(x)$, e.g. a fitness or an energy, to each point (i.e. to each sequence) x in sequence space. Recently the concept of landscapes has been generalized to complex combinatory maps where the distribution of structures rather than that of values $f(x)$ assigned to sequences is considered (Fontana et al. 1991, 1992a). What we are dealing with than is a mapping from sequence space with the Hamming distance as metric into shape space, the set of all structures (M). The tree distance δ mentioned in sect. 2.3 and obtained by tree editing (Fontana et al. 1991, 1992a) induces a metric on this set and defines a "shape space" (M, δ). The notion of shape space is due to Perelson and Oster (1979). In order to point at the distinction from landscapes we shall speak of structure and structure ensemble mappings (in contrast to fitness and other scalar properties, structures are essentially non-scalar).

Landscapes are qualitatively characterized by their ruggedness (Kauffman and Levin 1987; Kauffman et al. 1988; Macken and Perelson 1989; Eigen et al. 1989; Weinberger 1990, 1991; Fontana et al. 1991, 1992a, b; Weinberger and Stadler 1992) which can be measured conveniently by means of empirical correlation functions

$$\varrho(d) = \frac{\langle f(x) f(y) \rangle_{d(x,y)=d} - \langle f \rangle^2}{\langle f^2 \rangle - \langle f \rangle^2}\,. \tag{16}$$

The averages $\langle . \rangle_{d(x,y)=d}$ refer to pairs of sequences with prescribed Hamming distance d in sequence space and $\langle . \rangle_{\text{random}}$ refers to a pair of sequences which are chosen independently at random. As shown in Fontana et al. (1992a) (16) can be rewritten as

$$\varrho(d) = 1 - \frac{\langle (f(x) - f(y))^2 \rangle_{d(x,y)=d}}{\langle (f(p) - f(q))^2 \rangle_{\text{random}}} \tag{17}$$

In this form ϱ depends only on the differences $|f(x) - f(y)|$ of the fitness values. Fontana et al. (1992a) suggested to

generalize (17) by replacing the squared differences of values by squared distances $\delta^2(f(x), f(y))$ in shape space (M, δ):

$$\varrho(d) = 1 - \frac{\langle \delta^2(f(x), f(y)) \rangle_{d(x,y)=d}}{\langle \delta^2(f(p), f(q)) \rangle_{\text{random}}} \tag{18}$$

For any complex combinatory map one may then define the joint probability density

$$\text{Prob}\{d(x,y) = d, \delta(f(x), f(y)) = \delta\} = \wp(\delta, d) \tag{19}$$

where we assume that all pairs (x, y) are picked with equal probability. Since random pairs of sequences will almost always have a Hamming distance close to $\nu(\varkappa-1)/\varkappa$, it is more convenient for numerical purposes to compute the conditional probability density

$$\text{Prob}\{\delta(f(x), f(y)) = \delta \text{ given } d(x,y)=d\} = \wp(\delta|d) \tag{20}$$

On a sequence space the two densities are related by

$$\wp(\delta|d) = p(d) \cdot \wp(\delta, d) \tag{21}$$

where

$$p(d) = \varkappa^{-\nu} \cdot (\varkappa-1)^d \binom{\nu}{d} \tag{22}$$

is the probability that two randomly chosen points have Hamming distance d with $\varkappa$ being the number of letters in the alphabet. The density function $\wp(\delta|d)$ can easily be estimated numerically by sampling.

It has proved to be useful (Fontana et al. 1991) to characterize the autocorrelation function by a "mean correlation length" l which is defined by

$$\varrho(l) = 1/e \tag{23}$$

although $\varrho(d)$ is usually not a single decaying exponential.

In order to calculate the autocorrelation function $\varrho(d)$ from the density surface $\wp(\delta|d)$ we simply use the following identities and (18)

$$\begin{aligned} \langle \delta^2(x,y) \rangle_{d(x,y)=d} &= \sum_{\delta=0}^{\nu} \delta^2 \wp(\delta|d) \\ \langle \delta^2(x,y) \rangle_{\text{random}} &= \sum_{\delta=0}^{\nu} \sum_{d=0}^{\nu} \delta^2 \wp(\delta, d) \end{aligned} \tag{24}$$

For the present work the alignment-type distance $\delta(S_1, S_2)$ defined in the previous section has been used as metric in the conformational space consisting of the structure ensembles.

Instead of calculating $\wp(\delta|d)$ from large numbers of pairs of sequences with prescribed Hamming distance, random walks (Weinberger 1990, Fontana et al. 1991) can be used effectively as well. A random walk on sequence space is a series of sequences x_t generated from iterated point mutations applied to a initial sequence x_0. The series $\{x_t\}$ gives rise to a "time series" $\{f(x_t)\}$ of free energies with autocorrelation function

$$r(s) = \frac{\langle f(x_t) f(x_{t+s}) \rangle - \langle f \rangle^2}{\langle f^2 \rangle - \langle f \rangle^2} . \tag{25}$$

The relation of $r(s)$ and $\varrho(d)$ are determined by the geometric relaxation of the random walk in sequence space; explicit formulae can be found in Fontana et al. (1992b). This method has two disadvantages: firstly the convergence is relatively slow, because many of sampled pairs of strings are highly correlated with each other, and secondly, pairs with a Hamming distance larger than $\nu(\varkappa-1)/\varkappa$ are found only very rarely.

2.5 Fitness landscapes based on structures

On the basis of tertiary structure and reaction rates, we expect the fitness of a structure to decay more than linearly with the structural distance from a given functionally active structure, for example exponentially. This corresponds to a cooperative effect where substructures contribute multiplicatively to fitness so that a structure must be very nearly correct to have any function. In the extreme case, only the target structure will have a significant fitness. In both cases, the structural distance correlation function defined above will not correctly describe correlations of fitness in such a landscape. In fact, the multistructure landscapes may then have very different properties from those based on single structures. To see this, we consider two new distance measures, without regard to numerical tractability. The fitness oriented single-structure difference of two sequences, in the above multiplicative case, may be written

$$\delta_e(x,y) = 1 - \exp(-\alpha\, \delta(f(x), f(y))) \tag{26}$$

where α is a constant parameter describing the decay of fitness with structural mismatch.

The general multistructure distance defined in (15) is then correspondingly replaced by

$$\delta_{me}(x,y) = \sum_{S,S'} p_x(S)\, p_y(S')\,(1 - \exp(-\alpha\, \delta(S,S'))) . \tag{27}$$

If α is large enough, the width of the structural distribution, $p_x(S)$, provides a more significant smoothing of the landscape than substructure similarity.

An overall reaction rate incurred by sequence x may be written (for example appealing to the transition state theory) as

$$R(x) = \sum_S p_x(S)\, r(S) \tag{28}$$

where $r(S)$ is the rate of reaction from structure S and might include also reaction paths via fast reacting other conformations of x as intermediates. The fitness will be some function, often sigmoidal, of R: $F(x) = \sigma(R(x))$. In any case, it is seen from (29) that correlations in $p_x(S)$ and $r(S)$ are equally important in determining the correlations in landscapes based on overall reaction rates.

Results

3.1 Free energies

The average free energy of secondary structures becomes linearly more negative with increasing chain length (Fig. 2), since the mean number of base pairs increases

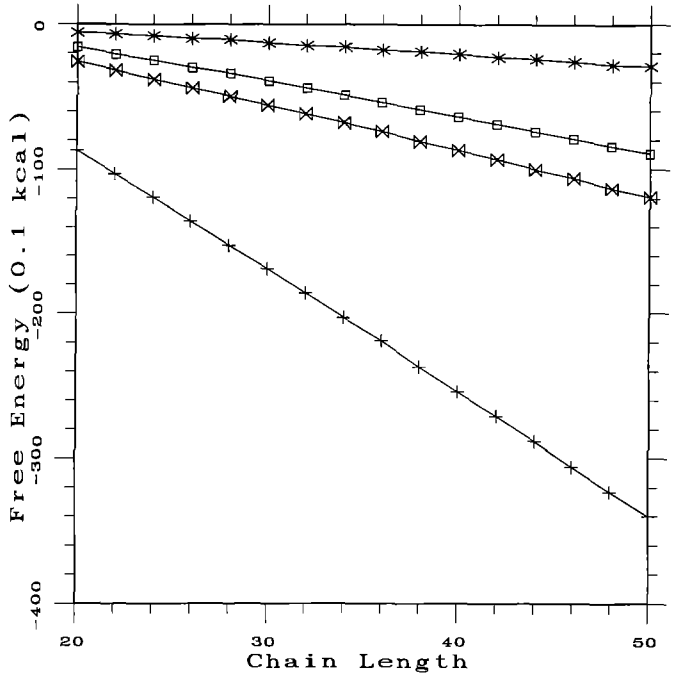


Fig. 2. Average free energies as a function of the chain length. For both the **GCAU** and the **GC** alphabet two different temperatures are shown. (**GCAU**: $T=37\,^{\circ}C$ □, $T=70\,^{\circ}C$ ★, **GC**: $T=37\,^{\circ}C$ +, $T=90\,^{\circ}C$ ⋈

linearly too (Fontana et al. 1992a). It is remarkable that the linear regime extends down to chain lengths as short as $v=20$ where one might already have expected specific effects of structural elements (see also Fontana et al. 1991). Stacking energies are higher when only **GC**-pairs are involved and therefore the formation of secondary structure from **GC**-only sequences yields a lower average free energy than secondary structures built up from the full **GCAU** alphabet. There are two different effects contributing to the temperature dependence of free energies (Fig. 3):

(*1*) In the lower temperature region there is a linear increase of the free energy, because of the negative entropies [(3) and (4)], and

(*2*) as the bonds become weaker, the mean number of base pairs formed in the secondary structure decreases too.

The second contribution obviously saturates when the majority of structures in the ensemble are already close to the unfolded state. The variances of free energies scale linearly with the chain length v because the total free energy of a secondary structure is a sum of many independent energy contributions for sufficiently large molecules. Because of the central limit theorem we expect the distribution of free energies to approach a Gaussian at sufficiently large chain lengths. Our computational data (not shown here) support this. Deviations from the Gaussian distribution increase with increasing temperature since the distribution of free energy values is essentially truncated at $F=0$ (because the totally unpaired structure has $F=0$) and becomes skew (Fig. 4).

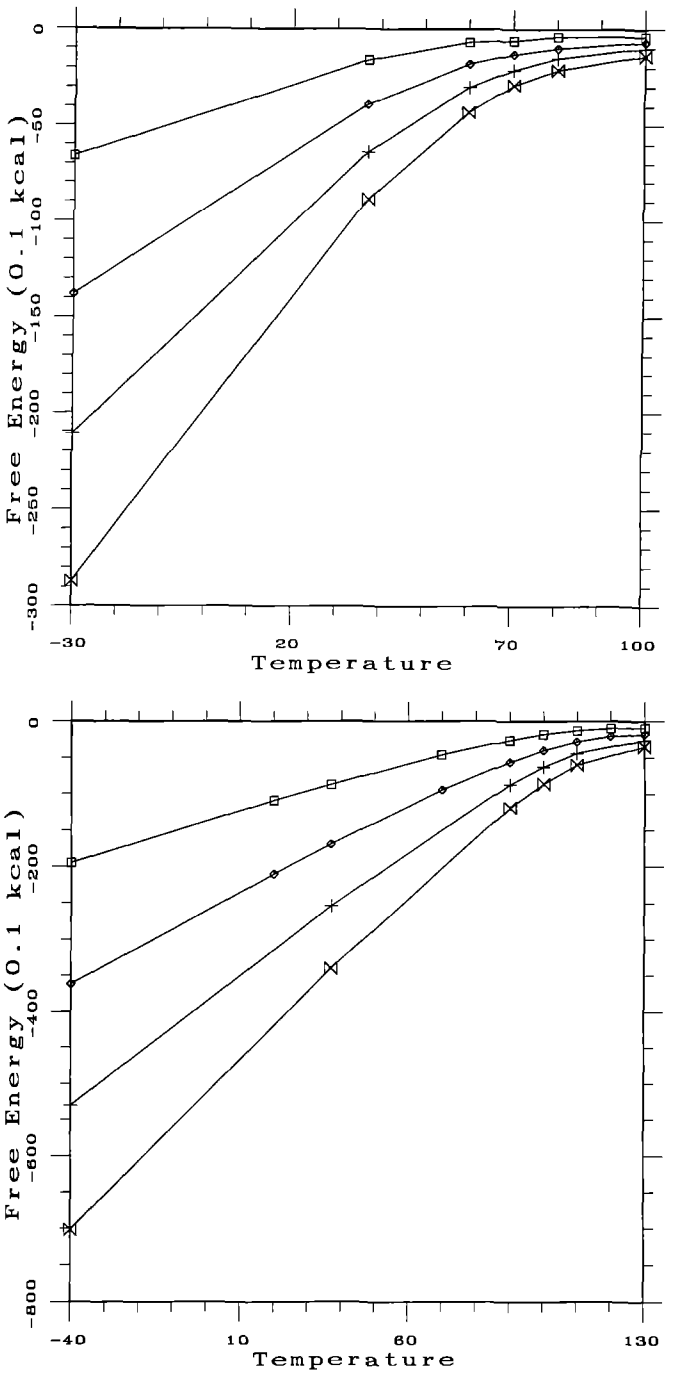


Fig. 3a, b. Average free energies as a function of temperature. **a) AUGC** alphabet (*upper*), **b) GC** alphabet (*lower*). $v=20$ □, $v=30$ ◇, $v=40$ +, $v=50$ ⋈

3.2 Correlation of energies

The correlation length l of the free energy landscape scales linearly with the chain length (Fig. 5). All correlation length reported here are obtained from a linear fit to $\log(\varrho(d))$. The correlation length can be estimated roughly from the nearest neighbor correlation (Fontana et al. 1992b)

$$l \approx \frac{2\,\mathrm{var}[f]}{\langle (f(x)-f(x'))^2 \rangle_{d(x,x')=1}} \tag{29}$$

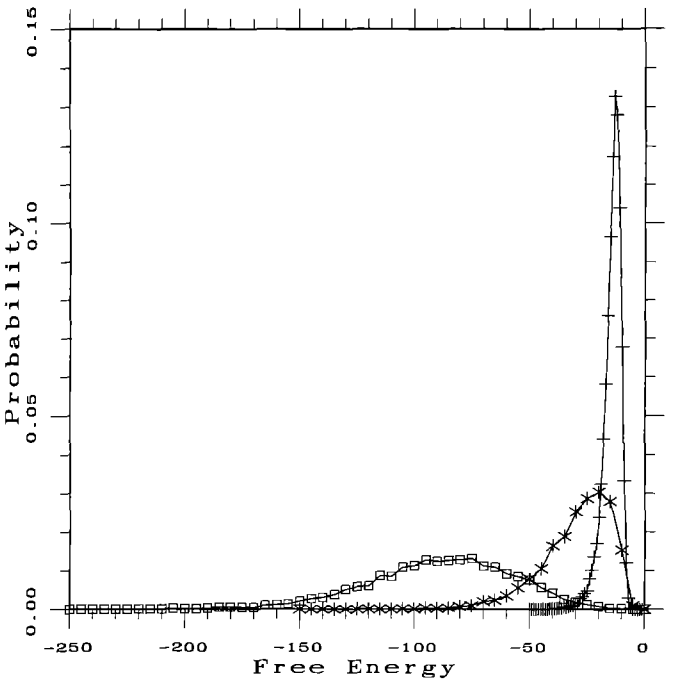
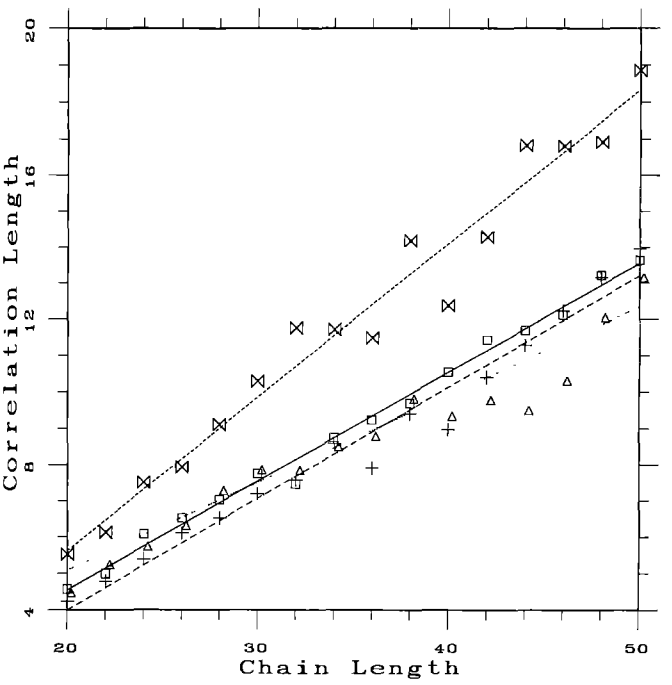


Fig. 4. Distribution of free energy values for **GCAU** sequences. $T=37\,°C$ □, $T=70\,°C$ ★, $T=100\,°C$ +, Chain length $v=40$



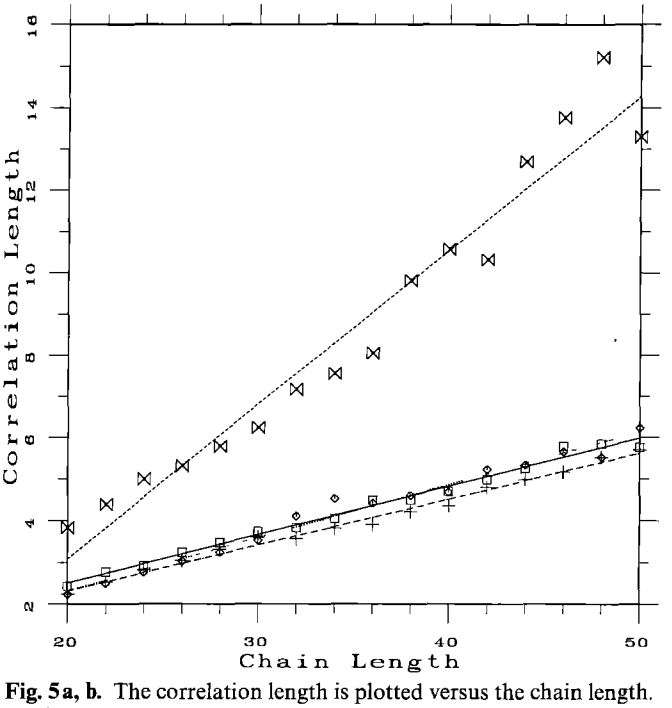


Fig. 5a, b. The correlation length is plotted versus the chain length. Straight lines are least square fits to the computed data. **a) GCAU** alphabet (*upper*): $T=-30\,°C$ □, *solid line.* $T=37\,°C$ ◇, *dotted line.* $T=70\,°C$ +, *dash-dotted line.* $T=100\,°C$ ⋈, *dashed line.* **b) GC** alphabet (*lower*): $T=-40\,°C$ □, *solid line.* $T=37\,°C$ ◇, *dotted line.* $T=100\,°C$ +, *dash-dotted line.* $T=130\,°C$ ⋈, *dashed line.* Owing to computer time limitations the data for the $v \geq 30$, have been sampled with smaller precision as for shorter chain length

It is known that the denominator is roughly constant (and bounded from above), while the enumerator scales linearly with v as shown in the previous section. The correlation length depends heavily on the alphabet: it is approximately twice as large for the natural **GCAU** landscapes than for the **GC** landscapes. Landscape with longer range correlation have fewer local optima (Weinberger 1990, Stadler and Schnabl 1992). The correlation length for both alphabets are roughly comparable with the data obtained from the minimal free energy of single structures (Fontana et al. 1991, 1992b).

Figure 6 shows the temperature dependence of the correlation length of the free energy landscape for **GC** and **GCAU** sequences. In both data sets the correlation length remains approximately constant up to a characteristic temperature T^* (which depends on the base composition). At T^* the correlation length shows a peak, indicating that the landscape becomes smoother. T^* can be identified as the average melting temperature of the structure ensemble. At T^* the average number of base pairs starts to decrease rapidly (Fig. 7). We cannot expect that our model calculations predict the experimental melting temperatures for RNA sequences reliably, since both, the ΔH° and the ΔS° values are assumed to be temperature independent. The T^* values computed, however, are not completely off the point, and relative temperature stabilities of **GC** and **GCAU** sequences are reproduced correctly.

The fact that the correlation length is independent of temperature is a consequence of the linear dependence of free energy on temperature below the melting point. We compared the correlation length for landscapes corresponding to the minimal free energy algorithm with our data computed at 37 °C. Although the values for the correlation length of free energy landscapes resulting from the Zuker algorithm are significantly smaller than the correlation lengths calculated from the partition function algorithm for $T=37\,°C$, the deviations are not tremendous. For the observed range of chain lengths they are around 25%. Best linear fits to the chain length dependence of the correlation length for both algorithms is given in Tables 1 and 2.

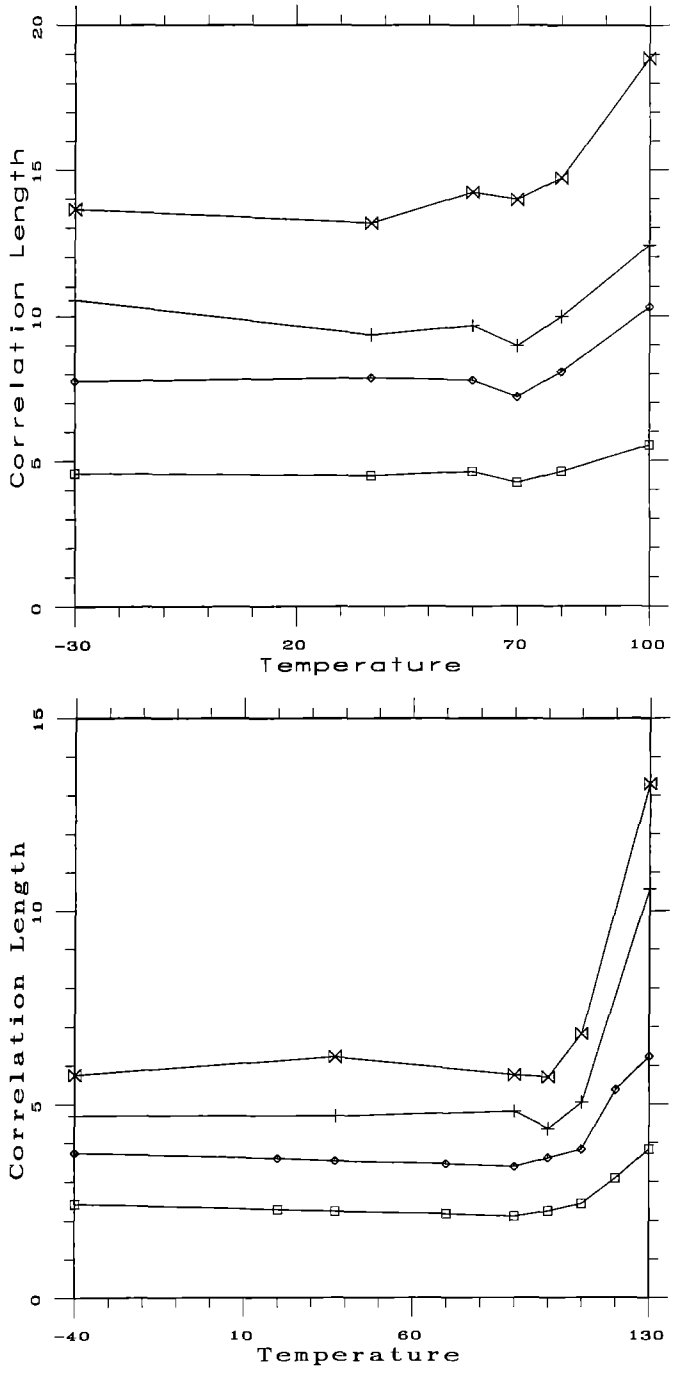


Fig. 6a, b. Temperature dependence of the free energy correlation length. **a) GCAU** alphabet (*upper*), **b) GC** alphabet (*lower*). $v=20$ □, $v=30$ ◇, $v=40$ +, $v=50$ ⋈. The critical temperatures are $T_c \approx 70\,°C$ for the **GCAU** alphabet, $T_c \approx 100\,°C$ for the **GC** alphabet, respectively

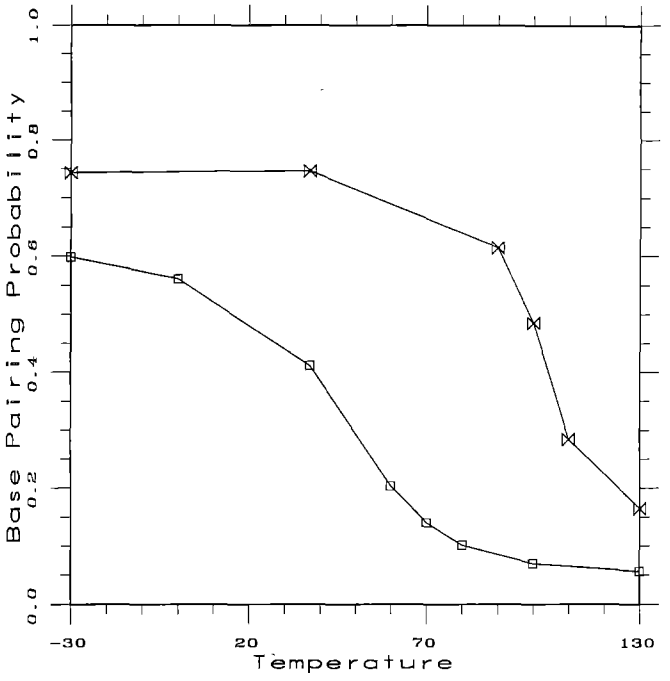


Fig. 7. Temperature dependence of the average number of base pairs for $v=50$. **GCAU** alphabet, □, **GC** alphabet, ⋈

Table 1. Correlation length l: **GCAU** alphabet, F refers to Fontana et al. 1992b

v	F	−30	37	70	100
30	6.01	7.74	7.86	7.19	10.30
40	8.14	10.54	9.35	8.97	12.37
50	11.30	13.64	13.15	13.96	18.86
Slope	0.263	0.300	0.242	0.307	0.424

Table 2. Correlation length l: **GC** alphabet, F refers to Fontana et al. 1992b

v	F	−40	37	100	130
30	2.94	3.73	3.52	3.60	6.24
40	4.00	4.70	4.68	4.36	10.57
50	4.54	5.76	6.22	5.70	13.30
Slope	0.086	0.116	0.127	0.110	0.373

3.3 Correlation of secondary structures

In Fig. 8 the correlation length of the structural ensemble mapping is shown as a function of the temperature. In contrast to the correlation length of the free energy landscape (Fig. 6) the structure correlation length does not remain constant below the mean melting temperature. We see a steady increase of the structure correlation length beginning from low temperatures up to the mean melting temperature. Apparently the structural ensemble mapping is more sensitive to temperature changes below the melting point. At temperatures above the melting point the correlation length decreases again. Naively one would expect, that the correlation length diverges in the limit of high temperatures since all RNA molecules are then in the unfolded state. Inspection of (17), however, shows that both the denominator and the numerator vanish at high temperatures. Minute remainders may cause large effects then, and it is difficult to explain therefore why the correlation length decreases again beyond the melting point.

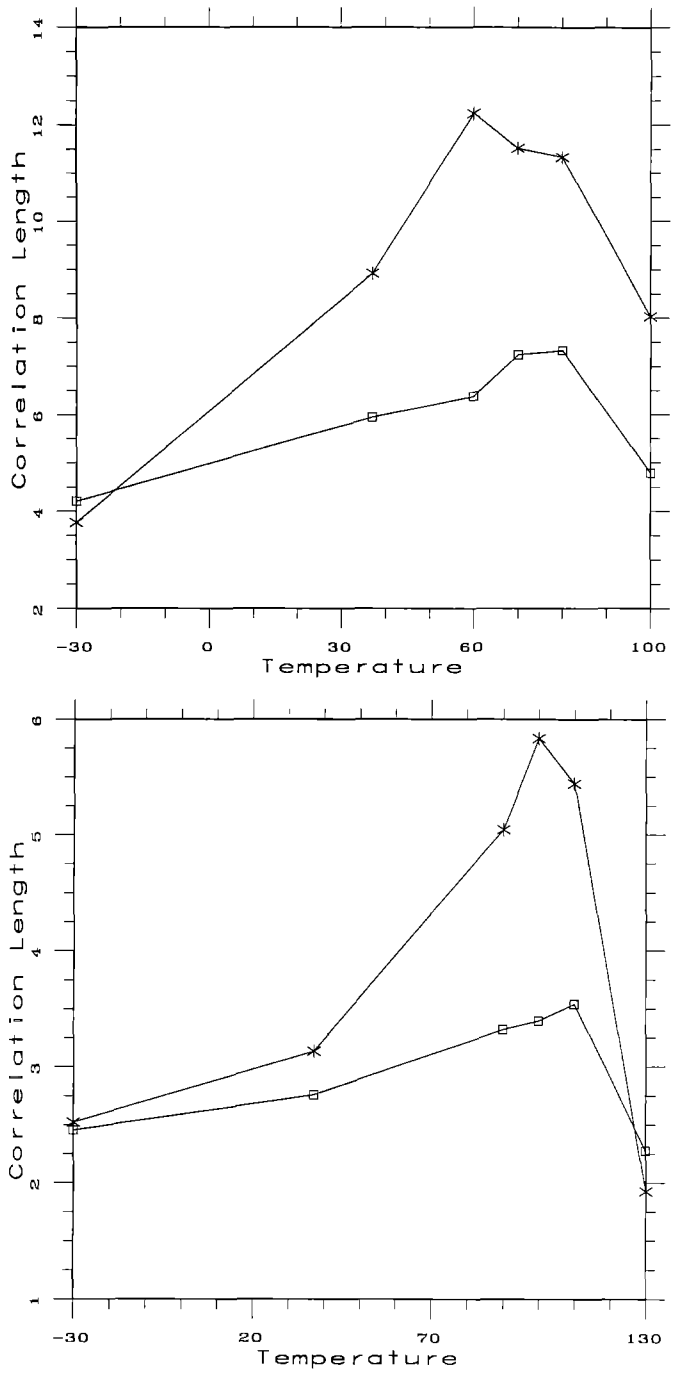


Fig. 8. Structure correlation length **a) GCAU** alphabet (*upper*), **b) GC** alphabet (*lower*). $n=30$ □, $n=50$ ★

In Fig. 9 the conditional probability density surfaces according to (19) are shown for the structure ensemble mapping and the free energy landscape at three different temperatures. Here we present only surfaces for natural **GCAU** sequences. Increasing temperature has the effect of shifting the whole distribution towards zero. The shape of the probability density surface of the free energy landscape along the free energy distance axes has a simple explanation. Since the distribution of free energy values in the landscapes is essentially Gaussian, energy differences are also Gaussianly distributed. Taking the absolute value of the differences thus amounts to a truncation

Table 3. Structure correlation length l: **GCAU** alphabet, F refers to Fontana et al. 1992a

v	F	−30	37	60	70	80	100
30	3.99	4.21	5.94	6.38	7.24	7.33	4.78
50	5.46	3.76	8.92	12.23	11.52	11.34	8.03

Table 4. Structure correlation length l: **GC** alphabet, F refers to Fontana et al. 1992a

v	F	−30	37	90	100	110	130
30	1.86	2.46	2.76	3.32	3.40	3.54	2.28
50	2.35	2.52	3.13	5.05	5.83	5.44	1.92

at zero. Obviously this cannot be true for the shape of the probability surface for the structure ensemble mapping, because structures are inherently non-scalar.

4 Discussion

The problem of dealing with the full complexity of macromolecular sequence dependent structures is truncated here at the secondary structure level. This is the limit of current computational software and hardware. True 3D spatial structures and structure-function-relations must still be explored in detail to complete the single molecule genotype-phenotype mapping. What we have shown so far is that the sequence-secondary structure relationship already has important implications for the evolution of macromolecules. How much will spatial structure formation or inaccuracies in the secondary structure model affect the conclusions? The latter point has been investigated for optimal secondary structures (rather than distributions) via a change in the thermodynamic parameters, showing that, while changes in single sequences are significant, the overall form of the secondary structure landscape remains unchanged (Fontana et al. 1992b). Tertiary interaction introduces additional base pairing such as pseudo-knots and other constraints on the flexibility of single stranded portions. The key question is whether the equilibrium distribution of spatial structures is as broad as that for secondary structures or whether for each sequence the tertiary effects can be regarded as a specific stabilization for example of one or a few of the alternative secondary structures. If tertiary effects would contribute an additional free energy stabilization of, say, 10 kcal/mol to only one particular structure at room temperature, we might expect a significant reduction in the number of alternative structures sampled. While this may be the rule for some optimized sequences, we find it hard to imagine that such specific selection of secondary structures occurs for an average sequence. Obviously, we are not yet in a position to give a definite answer to this question. At any rate, a good first approach to determine the sequence dependence of spatial structures is further biasing of the weighted secondary structure distributions calculated here.

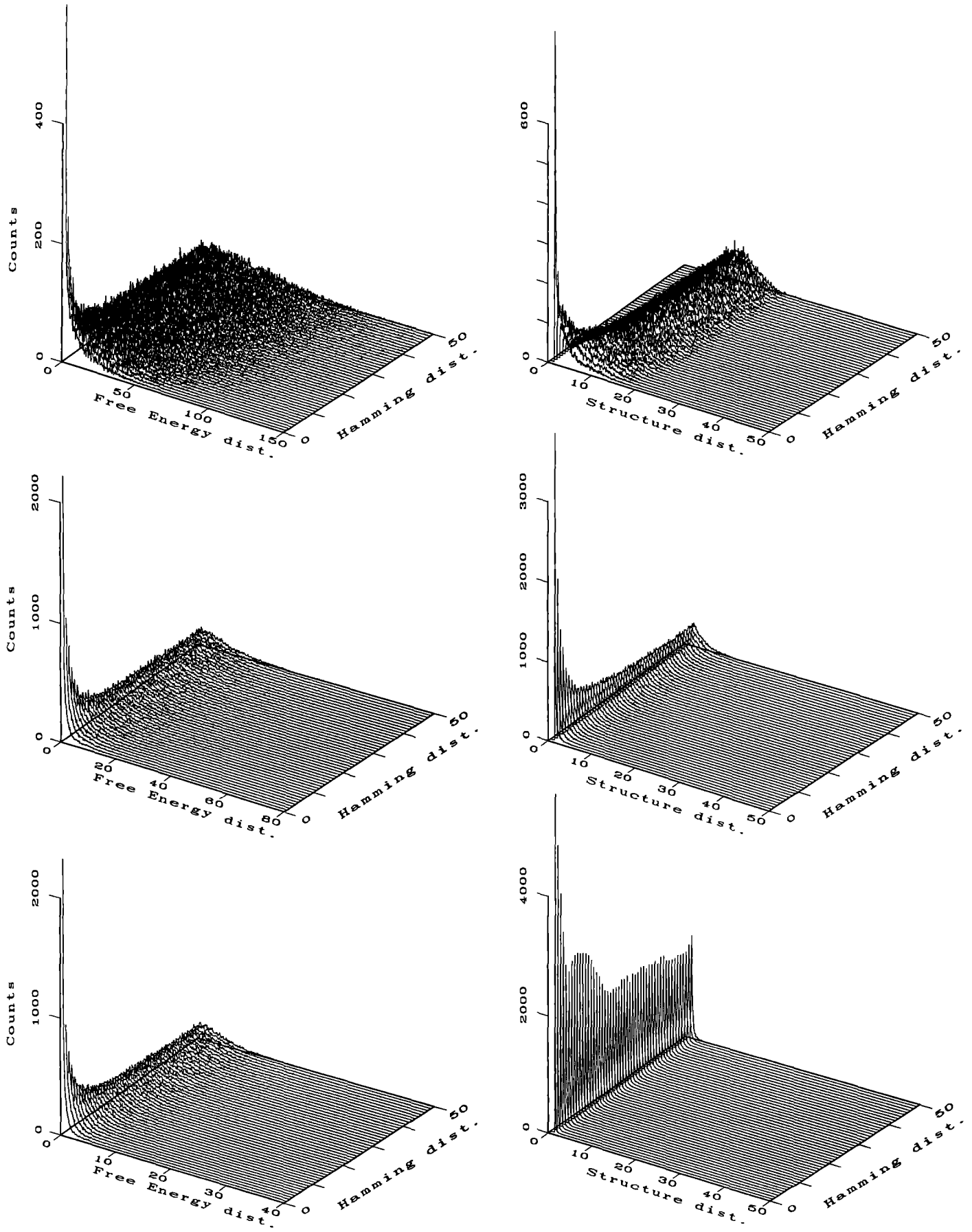


Fig. 9. Density surfaces $\wp(\delta|d)$ for **GCAU** alphabet. **a)** Energy differences $\delta = |F - F'|$ **b)** Structure distance. *top* $T = 37\,^{\circ}\mathrm{C}$, *center* $T = 70\,^{\circ}\mathrm{C}$, *bottom* $T = 100\,^{\circ}\mathrm{C}$

The free energy landscape of RNA folding has been analyzed in detail using the partition function algorithm. As a convenient measure for the ruggedness of landscapes we computed the pair correlation function in terms of the Hamming distance. In order to investigate the relation between sequences and their structure ensembles computed by the partition function algorithm, we define a conformation space of structure ensembles. A distance measure in conformation space is proposed, which is based on the base pairing probability matrix of the entire structure ensemble and a simple alignment algorithm. This measure can be readily computed and since it defines an alignment of structure ensembles it is proposed as a tool for reconstructing phylogenies. Here we used this distance measure to compute the pair correlation of mappings from sequence space into conformation space (structure ensemble mappings).

The free energy of an equilibrium ensemble of secondary structures for a given sequence (calculated from the partition function) is a scalar measure of the mean stability of folding and an obvious first candidate for a measure of fitness. Differences in free energy for different sequences reflect different stabilities of the structural ensembles. Defining a distance between the two structure ensembles of a pair of sequences provides a finer measure of the correlations between sequence and their structure ensembles. We have investigated both correlations in this paper.

The correlation length as a function of temperature exhibits a characteristic temperature for both mappings (free energies and structures) which can be interpreted as the mean melting temperature of RNA. Both the free energy landscape and the structure ensemble mapping have been shown to be much less rugged for natural **GCAU** sequences than for pure **GC** sequences. This effect was also observed with single structure mappings (Fontana et al. 1992a). The shape of the probability density surfaces (Fig. 9) along the Hamming distance axis gives us more insight. We see that the density surfaces remain unchanged for distances larger than a characteristic Hamming distance, which is about two to three times the correlation length. Apparently sequences with distances larger than this characteristic Hamming distance are statistically independent. Work in progress addresses the related question as to whether a good approximation to any chosen structure can be found within this distance from an arbitrary sequence. If so, evolution has not to walk far through the sequence space to find solutions to predefined problems. Furthermore, this would mean that if we are interested only in satisfactory local optima, the evolutionary optimization procedure can be restricted to sub-spaces of the diameter of this characteristic Hamming distance.

The increased correlation length for the multi-structure RNA landscapes, when compared with that of the single structure landscape, provides a minor smoothing of the evolution problem of finding sequences with a given structure. However, there is a much more significant effect of the equilibrium structural variety calculated in this work. Until now, we have used implicitly the distance between the actual and a given desired structure or between the actual and desired structural ensembles as our measure of fitness. The structural correlation function characterizes such a mapping.

In order to go beyond this simple distance related structural fitness we defined new distance measures in section 2.5 [(26) and (27)] which account for most stringent functional requirements that are not well represented by a linear dependence on structure or structure ensemble distance. Depending on the choice of a tuneable parameter (α) the fitness relevant subset of structures or structure ensembles may vary from a broad distribution to a single target structure. The difficulty of optimization may again vary considerably with relevant functional relationship. (It matters indeed whether F or $\exp(-F)$ is the quantity to be optimized.)

There is a second evolutionary advantage of the structural ensemble: it increases the effective population size. Since many of the different secondary structures in an ensemble may be sampled within a given replication cycle, the effective number of sampled conformations is much larger than mere population size. In the quasispecies model and in other optimization algorithms the rate of evolution depends strongly on population or sampling size. In the limit of infinite size populations approach a global optimum exponentially in the quasispecies model since all sequences are already present. In finite populations exponential amplification of advantageous sequences can take place only after they were formed in single copies by a mutation event and this lead to major delays in the optimization process (McCaskill 1984b, Fontana and Schuster 1987, Fontana et al. 1989). The sampling of many conformations means that partially occupied structures in the distribution may serve as a basis for differential amplification even when sampled only a small fraction of the time by a single member of the population.

The consequences of conformational ensembles rather than single structures of sequences were cast into the language of reaction kinetics in (28). Here we conclude this digression into evolutionary optimization dynamics by means of an illustrative example. Consider the target structure or distribution of acceptable target structures as a point or area in sequence space. Optimization is successful if it hits the target. The genealogy of a single sequence represents a trajectory through sequence space. The quasispecies concept broadens the trajectory to a band, the concept of conformational ensembles widens this band through sequence space further. Clearly, optimization is more efficient and faster the broader this band is. Neglecting non-equilibrium structures means an underestimation of the power of Darwinian selection in molecular evolution.

In summary, modelling RNA secondary structures with the partition function algorithm provides significant insight beyond the Zuker and Stiegler (1981) approach. Direct comparison of the landscapes generated by the two different folding algorithms shows a good qualitative agreement but significant quantitative differences. At physiological temperatures the landscapes obtained by the partition function algorithm are saliently smoother. Closer consideration shows major advantages for the use of structure ensembles rather than single structures in the

attempt to understand evolution and the possibility to make a quantitative link to functional landscapes based on reaction rates.

Acknowledgements. Stimulating discussions with Walter Fontana are gratefully acknowledged. We thank Danielle A. M. Konings for supplying an updated parameter set for the folding algorithm. Part of this work has been supported financially by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj. No. 8526-MOB and by the *Stiftung Volkswagenwerk*.

References

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

Cech TR (1990) Self-splicing of group I introns. Annu Rev Biochem 59:543–568

Eigen M (1971) Self-organization of matter and the evolution of macromolecules. Naturwissenschaften 10:465–523

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

Eigen M, McCaskill JS, Schuster P (1988) The molecular quasispecies. J Phys Chem 92:6881–6891

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

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

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

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, Konings DAM, Stadler PF, Schuster P (1992a) Statistics of RNA secondary structures. (Santa Fe Institute Preprint No. 92-02-008). Biopolymers (in press)

Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tacker M, Tarazona P, Weinberger ED, Schuster P (1992b) RNA folding and combinatory landscapes. Phys Rev E (in press)

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

Hamming RW (1986) Coding and Information Theory (2nd ed.) Prentice Hall, Englewood Cliffs, NJ

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

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

Kauffman SA, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. J Theor Biol 128:11–45

Kauffman SA, Weinberger ED, Perelson AS (1988) Maturation of the immune response via adaptive walks on affinity landscapes. In: Theoretical Immunology, Part I (Santa Fe Institute Studies in the Sciences of Complexity) Perelson AS (ed). Addison-Wesley, Reading, Mass

Konings DAM (1989) Pattern analysis of RNA secondary structure. Proefschrift, Rijksuniversiteit te Utrecht

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

Le SY, Zuker M (1990) Common structures of the 5′ non-coding RNA in enteroviruses and rhinoviruses: thermodynamical stability and statistical significance. J Mol Biol 261:729–741

Macken CA, Perelson AS (1989) Protein evolution on rugged landscapes. Proc Natl Acad Sci 86:6191–6195

Major F, Turcotte M, Gautheret D, Lapalme G, Fillion E, Cedergren R (1991) The combination of symbolic and numerical computation for three-dimensional modeling of RNA. Science 253:1255–1260

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

McCaskill JS (1984a) A localization threshold for macromolecular quasispecies from continuously distributed replication rates. J Chem Phys 80:5194–5202

McCaskill JS (1984b) A stochastic theory of macromolecular evolution. Biol Cybern 50:63–73

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

Noller HF, Hoffarth V, Zimniak L (1992) Unusual resistence of peptidyl transferase to protein extraction procedures. Science 256:1416–1419

Nowak M, Schuster P (1989) Error thresholds for replication in finite populations. Mutation frequencies and the onset of Muller's ratchet. J Theor Biol 137:375–395

Perelson AS, Oster GF (1979) Theoretical studies of clonal selection: minimal antibody and reliability of self- and non-self discrimination. J Theor Biol 81:645–670

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

Piccirilli JA, McConnell TS, Zaug AJ, Noller HF, Cech TR (1992) Aminoacyl esterase activity of the tetrahymena ribozyme. Science 256:1420–1424

Sankoff D, Morin AM, Cedergren RJ (1988) The evolution of 5S RNA secondary structures. Can J Biochem 56:440–443

Schuster P, Swetina J (1988) Stationary mutant distributions and evolutionary optimization. Bull Math Biol 50:636–660

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

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

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

Swetina J, Schuster P (1982) Selfreplication with errors. A model for polynucleotide replication. Biophys Chem 16:329–353

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

Turner DH, Sugimoto N, Freier S (1988) RNA structure prediction. Ann Rev Biophys Chem 17:167–192

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

Waterman MS (1984) General methods of sequence comparison. Bull Math Biol 46:473–500

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

Waterman MS, Smith TF (1978) RNA secondary structure: A complete mathematical analysis. Math Biosc 42:257–266

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

Weinberger ED (1991) Local properties in the *N-k* model, a tuneably rugged energy landscape. Phys Rev A 44:6399–6413

Weinberger ED, Stadler PF (1992) Why some fitness landscapes are fractal. Submitted to J Theor Biol 1992

Wright S (1932) The role of mutation, inbreeding, crossbreeding, and selection in evolution. In: Proc 6th Int Congress on Genetics Vol. 1, pp. 356–366

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

Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. Nucl Acid Res 9:133–148