Temperature Dependent RNA Landscapes
A Study Based on Partition Functions

DIPLOMARBEIT
eingereicht von
Sebastian Bonhoeffer
zur Erlangung des akademischen Grades
Magister rerum naturalium
an der Formal- und Naturwissenschaftlichen Fakultät
der Universität Wien

Januar 1992

It is a pleasure to thank all the people who helped me to do my first steps in the scientific world.

Peter Schuster gave me the possibility to join his group although I am a physicist. He introduced me into scientific work. Peter Stadler was always open for discussions. John McCaskill from Göttingen helped me to incorporate his partition function algorithm into our work. Christoph Streissler, Thomas Griesmacher, Tom Kovar and Manfred Kofranek answered my question concerning computers with never ending patience.

Christian Forst, Manfred Tacker, Pedro Tarazona, Ivo Hofacker, Subbiah Baskaran, Rick Bagley, Erich Bauer and Walter Grüner contributed to a nice working atmosphere.

Abstract

Statistical properties of RNA "folding landscapes" have been investigated extensively very recently [9]. The underlying folding algorithm used for these calculations yields the minimum free energy of a RNA molecule. John McCaskill [21] designed a new dynamic programming algorithm, which makes the partition function of RNA computable of order n^3 in the sequence length n. The partition function enables one to calculate pair binding probabilities for every possible base pair, and gives important information on the structural variability of a considered RNA molecule. Furthermore this algorithm makes the calculation of the temperature dependence of free energy and of the pair binding probabilities feasible. It is used here to compute more realistic folding landscapes and to determine the temperature dependence of statistical properties like the correlation length of the resulting landscapes. A detailed examination of how strongly mutations affect on average physical properties of RNA molecules gives an important insight into prebiotic evolution.

Deutsche Zusammenfassung

Statistische Eigenschaften von RNA "Faltungslandschaften" sind erst kürzlich Gegenstand wissenschaftlicher Untersuchungen gewesen [9]. Der diesen Untersuchungen zugrundeliegende Faltungsalgorithmus basiert auf einer Minimierung der Freien Energie eines RNA Moleüls. John McCaskill [21] entwickelte vor kurzem einen rekursiven Algorithmus, der die Bestimmung der Zustandsumme von RNA Molekülen in einer Zeit proportional zu n^3 ermöglicht, wobei n die Länge der RNA Sequenz ist. Mithilfe der Zustandsumme lassen sich die Wahrscheinlichkeiten der Bildung aller möglichen Basenpaare errechnen, woraus wichtige Informationen über die stukturelle Variabilität des betrachteten RNA Moleküls gewonnen werden können. Weiterhin gestattet dieser Algorithmus die Bestimmung der Temperaturabhängigkeit der freien Energien und der Bindungswahrscheinlichkeiten von Basenpaaren. Im Rahmen der vorliegenden Arbeit wird der Algorithmus benützt um realistischere Faltungslandschaften zu berechnen und um für die resultierenden Landschaften die Temperaturabhängigkeit statistischer Größen wie der Korrelationslänge zu bestimmen. Eine detaillierte Untersuchung, wie Mutationen im Mittel die physikalischen Eigenschaften von RNA Molekülen beeinflußen, gibt einen Einblick in den Lauf der präbiotischen Evolution.

Contents

List of Figures

List of Tables

1 Introduction

Charles Darwin in 1859 presented his famous essay on "The Origin of Species", which was one of the most important contributions to natural sciences in the last century. He was the first to recognize the principle of variation and selection as the driving force of biological evolution. Although the treatise immediately gained fame there was also a lot of criticism of Darwins theory. Apart from ideological objections, one reason, why Darwin's conclusions were rejected, was that he attributed chance an important role in his explanation of the origin of species. How could something as complex as living beings evolve from an accumulation of successive random events? No doubt, a great deal of criticism was due to a misunderstanding of Darwins principle. But still today, where the theory is well established it remains astonishing that the diversity and the complexity of life should be founded on such a simple principle.

In the beginning of this century the work of Haldane [16], Wright [33] and Fisher [8] provided a mathematical background for Darwin's theory. It became evident, that Darwinian Evolution only requires an object, which possesses the ability of selfreplication and which lives from limited resources. Eigen [4] showed in 1971, that natural selection can be extended to inanimate nature. Thereafter Eigen and Schuster introduced [5, 6, 7] the conception of the hypercycle and a theory for the origin of life was presented.

In connection with the selection Darwin coined the catchphrase "survival of the fittest", which also was cause for fatal misunderstandings and misinterpretations. The fittest may here be defined as the one having the greatest number of offsprings reaching the age of fertility. But describing the features, which determine the fitness one has to face enormous problems, so that there is a temptation, to define the fittest as the one, which survives. But the fact, that up to now it is not possible to calculate the fitness for a living being in a given surrounding is no justification for the famous objection, that the Darwinian theory reduces to the tautology "survival of the survivor".

Today evolution is often viewed as a procedure optimizing a multiparametric problem. The parameters of the fitness function which is to optimize are a chosen set of properties having an influence on the fertility of the individual. Sewall Wright in 1932 introduced the notion of fitness landscapes assigning a fitness value to every particular set of parameter values. The process of evolution can be seen as the task of finding blindfolded the highest mountain on the fitness landscape. Simplified: The strategy of evolution is to walk around randomly in the landscape with the restriction of going downhill only with low probability. Although this strategy is fairly simple, the task is very complex, since the shape of the landscape can change the problem dramatically. Thus information about the features of the landscape is substantial in order to attain a better understanding the course of prebiotic evolution.

Our interest in studying biophysical landscapes arose from investigations about evolutionary optimization and adaptation in fitness landscapes derived for RNA molecules [11]. Modern gene technology made experimental investigation of bio-

physical landscapes approachable in recent years. However, due to the vast number of nucleic acids, which have to be sequenced in order to determine complete landscapes, these techniques cannot yet be utilized to derive global properties of biophysical landscapes. Consequently computer experiments capturing the landscapes' essential features are indispensable for the characterization of evolutionary processes.

The present work is organized as follows. In section 2 we first introduce the secondary structure model for RNA molecules, which is an important prerequisite for the understanding of two subsequent subsections. We give a brief introduction in the minimal free energy folding algorithm originally designed by Zuker and Sankoff [34]. This algorithm was used for the investigation of RNA folding landscapes by Fontana et al. [9]. We will then explain in greater detail the partition function algorithm of McCaskill [21]. It permits a more realistic calculation of RNA landscapes and additionally has the advantage of introducing a temperature dependence into the computation of statistical properties of RNA landscapes.

In section 3 we will give a more precise idea of the notion of a landscape. We will discuss biologically motivated landscapes as well as landscapes resulting from combinatorial optimization problems. We then point out the intimate connection between landscapes and their corresponding optimization procedures. Finally we will return to the landscapes investigated in this work.

We derive the mathematical tools needed for the investigation of landscapes in section 4. There we will first be dealing with measures of the value distribution in the landscape, which do not take into account their spatial arrangement. Then the autocorrelation function is introduced as a measure for the ruggedness of landscapes. In the last subsection we then explain two techniques, which have been applied to explore RNA landscapes.

Our results are discussed in detail in section 5 and section 6 concludes the work.

2 Folding algorithms

2.1 The secondary structure model

A ribonucleic acid consists of a sequence of chemically linked nucleotides, synthesized from four different bases. These four bases are Adenine (**A**), Guanine (**G**), Cytosine (**C**) and Uracil (**U**). A nucleotide is a base connected to a sugar , in the case of RNA ribose, with an added phosphate group. The nucleotides are linked together by a sugar-phosphate backbone, building up a polynucleotide chain. A nucleotide has a phosphate group attached to the $5'$ position at the riboside which is connected to a phosphate group attached to the $3'$ position on the riboside of its neighbor. The backbone is held together by $5'-3'$ sugar-phosphate links. The terminal nucleotides on both sides have either a free $5'$ or $3'$ phosphate group. Therefore a RNA molecule is uniquely determined by the sequence of bases ordered from the $5'$ end to the $3'$ end of the polynucleotide chain. A string of letters chosen from the four-letter alphabet {**A**,**U**,**G**,**C**} is called the primary structure.

G and **C**, respectively **A** and **U** are complementary bases, which can form strong hydrogen bonds. A weaker bond is also possible between **G** and **U**. These interactions cause the RNA molecule to fold back upon itself and to form a complex three dimensional structure, the so called tertiary structure. The art of predicting the three dimensional structure from the linear representation of a RNA molecule seems to be still at an early stage of development. Therefore current algorithms focus on the prediction of secondary structure, i.e. what nucleotides form base pairs. Here already satisfactory results are achieved. But there is also a biological justification for for limiting our interest on the prediction of the secondary structure, because secondary structure elements are conserved in evolution [2, 20].

A RNA molecule is representable as $S = s_1, s_2, s_3, \ldots, s_n$, where $s_i \in$ {**A**,**U**,**G**,**C**} and n is the length of the polynucleotide chain. The s_i are ordered from the $5'$ end to the $3'$ end. Physically, the secondary structure is the folding in two dimensions. Fig. 1 gives an example for a secondary structure. Mathematically a secondary structure can be described by the set Φ of base pairs (s_i, s_j) with $i < j$, which are formed between complementary base pairs. Clearly, no base can be bound twice, which means if (s_i, s_j) and $(s_i, s_k) \in \Phi$ then $j = k$. Let us denote with S_{ij} the subsequence of S starting from the ith base and ending at the jth base. Φ_{ij} then represents the secondary structure of S_{ij}.

From experiments we know, that a RNA molecule cannot fold back upon itself, without leaving at least three bases unpaired. Hence, a pair $(s_i, s_j) \in \Phi$ must fulfill the condition $j - i \geq 3$. One further restriction must be imposed in order to enable dynamic programming. We allow only structures, which contain no knots. A secondary structure is knotted if (s_i, s_j) and (s_k, s_l) are base pairs and $i < k < j < l$. This constraint is crucial for the recursive algorithm, which will be discussed in detail in the forthcoming section. If (s_i, s_j) pair then every pairing base s_k with $i \leq k \leq j$ will have its partner in S_{ij}. Therefore every base pair divides the secondary structure in parts, which do not have any base pair in common. This allows to construct a

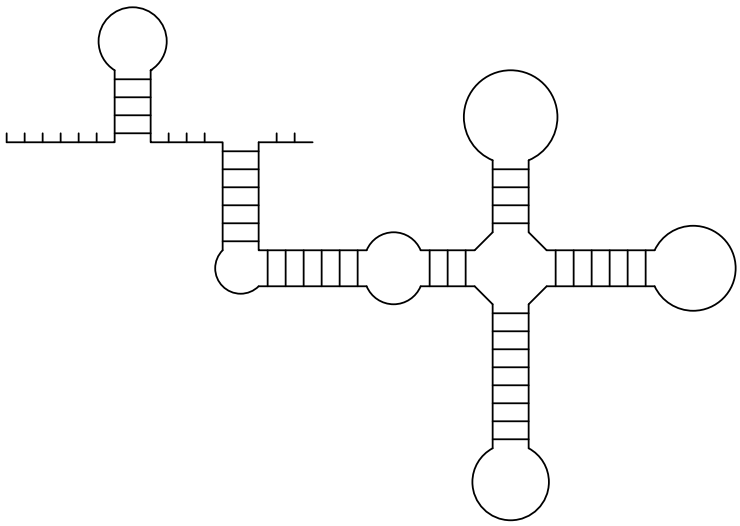

Figure 1: An example for a RNA secondary structure, with free dangling ends, stacks and loops

recursive scheme, which first looks for the secondary structure of substrings and then builds up the entire secondary structure of the considered RNA sequence.

Although in reality there are examples for secondary structures which are knotted, we have to relegate the problem of knotted structures to the tertiary structure prediction. But since there are no experimental data available on how knots affect the stability of the secondary structure, it seems justified to exclude knotted structures for the present.

Every unknotted secondary structure is built up out of several structural motifs. These are stacks, loops or bulges and external unpaired bases, like free dangling ends and joints between independent substructures. The motifs can be described more precisely in the following manner, by giving the expression loop a more general meaning.

Let (s_i, s_j) be in Φ. Then (s_i, s_j) closes a loop. Any base pair $(s_k, s_l) \in \Phi_{ij}$, with $i < k < l < j$ and no $\tilde{k} < k < l < \tilde{l}$ so that $(s_{\tilde{k}}, s_{\tilde{l}}) \in \Phi_{ij}$, is called interior to the loop closed by (s_i, s_j). (s_i, s_j) is called closing base pair, which by convention does not belong to the loop. (s_k, s_l) itself closes a loop, which is excluded from the loop closed by (s_i, s_j). If u is the number of unpaired bases and m is the number of interior base pairs in a loop, we can classify the loops by their values of u and m:

- $m = 0, u > 0$: hairpin loop
- $m = 1, u = 0$: stack
- $m = 1, u > 0$: bulge or interior loop

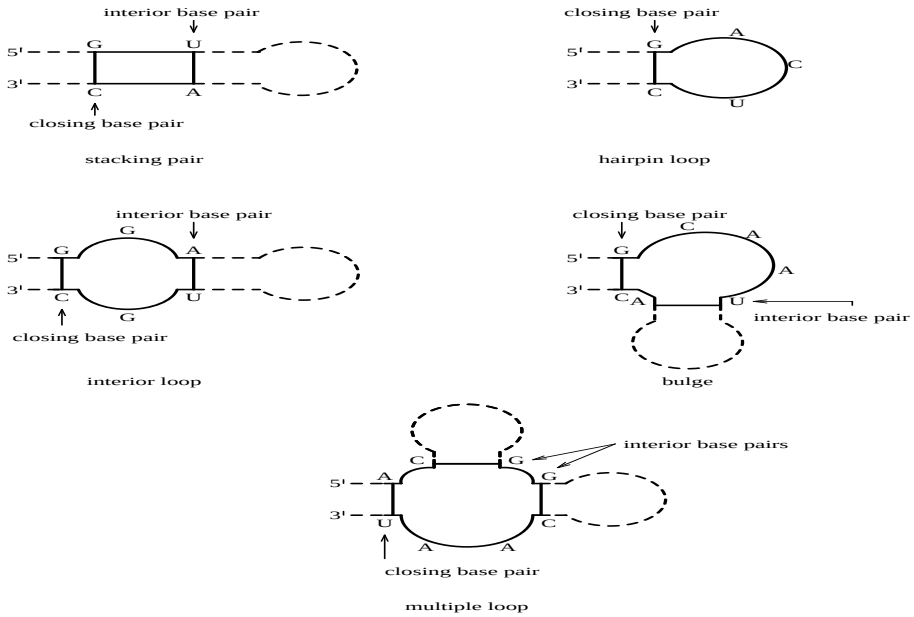

Figure 2: The classification of loops for the decomposition of RNA secondary structure

- $m > 1$: multiple loop

External are those bases, which do not belong to any loop. u is also called the length of a loop. Let $u_1 = k - i - 1$ and $u_2 = j - l - 1$, with $u = u_1 + u_2$, then one can distinguish a bulge, where $u_1 = 0$ or $u_2 = 0$, from an interior loop, where $u_1, u_2 > 0$. The loops are shown in figure 2.

From this generalization of a loop and from the assumption of unknotted secondary structure follows firstly, that if a base of a pair belongs to certain loop , its partner must as well belong to the same loop, and secondly, that each base is internal exactly to one loop or is external to all loops.

2.2 The minimal free energy folding

The classification described in the previous section makes the complete decomposition of a given secondary structure in terms of loops and external bases feasible. Biochemical data have been produced to assign energy values for all possible loops with $m \leq 1$, depending on the closing and the interior base pair, as well as the number of unpaired interior bases [14, 17, 28, 29, 23]. The data are derived from melting experiments with small RNA molecules or oligonucleotides.

Each possible stacked pair is given a negative stabilizing free energy, depending on the preceding neighbor, i. e. the closing pair, and on the strength of the interaction between the bases of the stacking pair. Internal loops and bulges have a destabilizing

effect on the secondary structure and are therefore assigned positive free energies. Apart from the involved base pairs and the length u of the loop, the value of free energy also depends on the symmetry [22], i. e. the size of $u_1 - u_2$. Unfortunately there are no experimental data available for the free energy assignment of multiple loops, but the linear ansatz

$$F_{\mathrm{ML}} = a + bm + cu \tag{1}$$

has appeared to be useful [17]. The parameters a, b, and c are adjusted through comparison of experimentally determined secondary structures with the prediction of the folding algorithm.

The free energy of each loop is assumed to contribute additively to the entire free energy of the secondary structure. The evaluation of the free energy of a given structure results in simply summing up over all loops L occurring in the secondary structure Φ.

$$F(\Phi) = \sum_{L \in \Phi} F_L$$

The task of the folding algorithm is to find out of all possible structures of a RNA molecule the one with the minimal free energy. The most simple approach would be to calculate successively the free energy for each possible structure and selecting the one with minimal free energy. But as the number of different possible structures increases exponentially with the length of the RNA molecule, such an algorithm would break down for any realistic chain length.

Fortunately there is a way out due to the additivity of the contributions of each loop. If the structure of the sequence S is optimal, which means that its free energy is minimal, than the folding of any substring S_{ij} is optimal provided that (s_i, s_j) form a pair. This enables one to make full use of dynamic programming. The minimal free energy algorithm designed by Zuker et al. [34], calculates first the optimal folding of the smallest substrings, which form a secondary structure, and then recursively constructs the optimal structure out of optimal subfragments. Zuker et al. achieved a calculation of minimal free energy and optimal structure in cubic order of the chain length n. A major advantage of the dynamic programming algorithm is its speed, which is gained by giving up the the prediction of knots and by taking into account only interactions between nearest neighbors. The additivity of the energy contributions is crucial. Incorporating more complex interactions would lead to a breakdown of the recursion scheme.

Because of the very few experimentally determined secondary structures, it is difficult to estimate the predictive power of the folding algorithm [30]. Only transfer RNA structures are known in detail, but secondary structures of other RNAs have been deduced by phylogenetic comparison, i. e. by comparison of RNA molecules of identical function in different organisms. The folding of a sequence by minimizing the energy of all possible secondary structures predicts around 80% of the either experimentally determined or phylogenetically deduced secondary structure for short RNA molecules of chain length $n \approx 100$.

2.3 Dynamic programming of the partition function

The above described algorithm yields only one structure with minimal free energy. No information is gained, whether there are other structures with the same free energy or whether there are solutions in close vicinity to the optimal folding. But the distribution of suboptimal solutions in respect to their energy value gives important information about the structural variability of a RNA sequence. In reality RNA molecules do not take on only the optimal structure, but seem to change more or less rapidly conformation between closely related structures. Waterman [31] proposed a modification of the minimal free energy algorithm, which allows to compute all structures lying within a certain energy range. A few years ago McCaskill [21] presented a completely new approach, which made it possible to determine the partition function of RNA molecules. From the partition function all thermodynamic quantities of interest can be derived. The connection between partition function and free energy is

$$F = -kT \ln Q \tag{2}$$

where Q is the partition function, T is the temperature and k is the Boltzmann factor. The important point is here, that we get now a temperature dependence of the free energy.

The partition function algorithm and the minimal free energy algorithm are closely related. Both base on the above described assumptions, both recursively scan through all possible secondary structures and in principle the calculation of the partition function has also been possible since experimental data were available for all different kinds of loops. The partition function of a given RNA molecule is

$$Q = \sum_{\Phi \in \mathcal{M}} e^{-F(\Phi)/kT} \tag{3}$$

where $\mathcal{M}$ is the set of all possible secondary structures Φ of the nucleotide sequence. Also here arises the problem, that the computational expense increases exponentially with sequence length. Therefore one major problem of the calculation of the partition function is to refine the algorithm to a computation in cubic order of sequence length.

The additivity of free energy contribution of the various loops implies a multiplicativity in the partition function.

$$Q = \sum_{\Phi \in \mathcal{M}} \prod_{L \in \Phi} e^{-F_L/kT} \tag{4}$$

Note, that the product comprises a exponentially increasing number of factors.

Let Q^b_{ij} be the partition function of the segment S_{ij} given that s_i and s_j pair, i. e. that $(s_i, s_j) \in \Phi_{ij}$. Q^b_{ij} can then be written as a recursive formula

$$Q^b_{ij} = \sum_{L} e^{-F_L/kT} \prod_{\substack{(h,l) \in L \\ i<h<l<j}} Q^b_{hl} \tag{5}$$

where the sum goes over all possible loops closed by (s_i, s_j). If L is a hairpin loop there is no pair $(h,l) \in L$, if L is an interior loop or a bulge there is exactly one pair $(h,l) \in L$, but if L is multiloop then there are m pairs $(h,l) \in l$ with $i < h_1 < l_1 < \cdots < h_m < l_m < j$. Clearly no base can pair with itself, therefore the initial condition of the above recursion formula is $Q^b_{ii} = 0$. Formation of unallowed secondary structures as for example hairpin loops with length less than 3 is penalized with infinitely high energy, which results in infinitely small contribution to the partition function.

The full partition function of the subsequence S_{ij} is the sum of the partition function of S_{ij} given that (s_i, s_j) form a pair and all configurations given that s_i and s_j do not form a pair.

$$Q_{ij} = 1 + \sum_{\substack{h,l \\ i \le h \le l \le j}} Q_{i,h-1} Q^b_{hl} \tag{6}$$

The free energy contribution for external base is assumed to be zero. Hence, in the partition function such structural elements result in a multiplication by 1. Therefore the initial conditions here are $Q_{ii} = 1$ and $Q_{i+1,i} = 1$. From these two equations the partition function can be determined recursively until one yields $Q_{1,n}$, which is the partition function of the complete sequence. But still the problem of the exponentially increasing computation time remains unsolved. While there is only one possibility to form a hairpin loop and two possibilities to form bulges between s_i and s_j, the actual difficulty is the calculation of all interior and multiple loops.

Dividing the partition function into the contribution coming from the different loop forms, equation (5) can be rewritten as

$$Q^b_{ij} = e^{-F_0(i,j)/kT} + \sum_{\substack{h,l \\ i<h<l<j}} e^{-F_1[(i,j),(h,l)]/kT} Q^b_{hl} \tag{7}$$

$$+ \sum_{\substack{h,l \\ i<h<l<j}} Q^m_{i+1,h-1} Q^b_{hl} e^{-(a+b+c(j-l-1))/kT} \tag{8}$$

where equation (1) is used and where

$$Q^m_{ij} = \sum_{h,l} (e^{-c(h-i-1)/kT} + Q^m_{i,h-1}) Q^b_{hl} e^{(-b+c(j-l-1))/kT} \tag{9}$$

with $Q^m_{ii} = 0$ and $Q^m_{i+1,i} = 0$. F_m refers to the classification of loops described in the previous section according to the value of m ($m = 0 \rightarrow$ hairpin loop, $m = 1 \rightarrow$ stack, interior loop, bulge). The third term in equation (8) represents the multiple loop contribution.

We have now a recursion for the partition function, which sums over four independent indices going from 1 to n, i. e. the algorithm is now of order n^4 in the sequence length n. In order to refine the computation of long interior loop to n^3, a maximum value u_m for the length of the loop may be introduced, whereby loops longer then u_m are regarded as prohibited. This restriction is justifiable, because

long interior loops are given rather high free energy values and therefore structures containing large loops contribute poorly to the overall partition function. Thus the calculation of interior loops results for long sequences in an algorithm of cubic order multiplied by a factor proportional to u_m^2. For short sequences with lengths close to u_m the resulting algorithm remains proportional to n^4.

A reduction of the multiple loop contribution to the partition function is achieved by introducing a new auxiliary quantity defined by

$$Q_{ij}^{\tilde{m}} = \sum_{l=i+1}^{j} Q_{il}^{b} e^{-c(j-l)/kT} \tag{10}$$

Equation (8) then has the form

$$Q_{ij}^{b} = e^{-F_0(i,j)/kT} + \sum_{\substack{h,l \\ i<h<l<j}} e^{-F_1[(i,j),(h,l)]/kT} Q_{hl}^{b} \tag{11}$$

$$+ \sum_{\substack{h \\ i<h<j}} Q_{i+1,h-1}^{m} Q_{h,j-1}^{\tilde{m}} e^{-(a+b)/kT} \tag{12}$$

and equation (9) can be written

$$Q_{ij}^{m} = \sum_{\substack{h \\ i<h\leq j}} (e^{-c(h-i)/kT} + Q_{i,h-1}) Q_{hj}^{\tilde{m}} e^{-b/kT} \tag{13}$$

If we define in addition

$$Q_{ij}^{1} = \sum_{\substack{l \\ i\leq l\leq j}} Q_{il}^{b} \tag{14}$$

with $Q_{ii}^{1} = 0$, the computation of the partition function of the subsequence S_{ij}

$$Q_{ij} = 1 + \sum_{\substack{h \\ i\leq h\leq j}} Q_{i,h-1} Q_{ij}^{1} \tag{15}$$

is achieved in cubic order.

Altogether the calculation of the full partition function Q is now possible with the above recursion formulas in cubic order for long sequences. Although the partition function algorithm is now of the same order as the minimal free energy algorithm described in the previous section, the computation of the partition function is nevertheless remarkably more time consuming, because the whole calculation has to be done with floating point numbers, whereas the computation of the minimal free energy involves only integers.

We want to stress, that the calculation of the partition function of the entire sequence yields also the full partition function of all subfragments. This will become important for the computation of the statistical properties of RNA folding landscapes later on.

Clearly the McCaskill algorithm does not predict a secondary structure of a RNA molecule, but one can calculate the base binding probabilities between all possible base pairs. The resulting base pair probability matrix contains important information on the structural variability of the molecule. It gives an impression of the uniformity of the equilibrium ensemble of the secondary structures.

A specified secondary structure Φ' is weighted in the partition function according to the Boltzmann distribution, therefore its probability of occurrence is

$$P(\Phi') = \frac{e^{-F(\Phi')/kT}}{Q} \tag{16}$$

A base pair is either a closing pair or interior to one type of loop. The probability of the formation of the base pair (s_h, s_l) may consequently be expressed as

$$\begin{aligned} P_{hl} &= \frac{Q_{1,h-1} Q^b_{hl} Q_{l+1,n}}{Q} \\ &+ \sum_{\substack{i,j \\ i<h<l<j \\ u<u_m}} \frac{P_{ij} Q^b_{hl} e^{-F_1[(i,j),(h,l)]/kT}}{Q^b_{ij}} \\ &+ \sum_{\substack{i,j \\ i<h<l<j}} \frac{P_{ij} Q^b_{hl} e^{-(a+b)/kT}}{Q^b_{ij}} (e^{-c(h-i-1)/kT} Q^m_{l+1,j-1} \\ &+ Q^m_{i+1,h-1} e^{-c(j-l-1)/kT} + Q^m_{i+1,h-1} Q^m_{l+1,j-1}) \end{aligned}$$

The first term sums over all different configurations, where (s_h, s_l) forms a closing pair. The second term takes into account all possibilities, where (s_h, s_l) is internal to a stack, a bulge or an interior loop and (s_i, s_j) closes the loop. The third term gives the contributions of multiple loop formations with (s_h, s_l) interior to the multiloop and (s_i, s_j) closing the multiloop. Again it is possible to reduce the order of the base pairing probability to n^3. Because of the restriction, that interior loops of length greater then u_m are forbidden, the only term of order n^4 for long sequence length is the multiple loop term. But also here, like in the partition function calculation, we may help ourselves by introducing

$$P^m_{il} = \sum_{\substack{j \\ j>l}} \frac{P_{ij} Q^m_{l+1,j-1}}{Q^b_{ij}} \tag{17}$$

and

$$P^{\tilde{m}}_{il} = \sum_{\substack{j \\ j>l}} \frac{P_{ij} e^{-c(j-l-1)/kT}}{Q^b_{ij}} \tag{18}$$

Equation (17) may then be written

$$P_{hl} = \frac{Q_{1,h-1} Q^b_{hl} Q_{l+1,n}}{Q} \tag{19}$$

$$+ \sum_{\substack{i,j \\ i<h<l<j \\ u<u_m}} \frac{P_{ij} Q^b_{hl} e^{-F_1[(i,j),(h,l)]/kT}}{Q^b_{ij}} \tag{20}$$

$$+ \sum_{\substack{i \\ i<h}} Q_{hl} e^{-(a+b)/kT} (P^{\tilde{m}}_{il} Q^m_{i+1,h-1} \tag{21}$$

$$+ P^m_{il} (e^{-c(h-i-1)/kT} + Q^m_{i+1,h-1})) \tag{22}$$

The resulting base pair binding probabilities can be stored in a triangular matrix. This matrix contains information on the structural variability of the considered RNA molecule.

Before going to the next section, we want to discuss briefly the difference between the usage of the experimental data in the minimal free energy algorithm and in the partition function algorithm. The data are normally determined under physiological conditions, i. e. 37° C. In order to enable the calculation of the partition function at different temperatures, one needs to extrapolate the data for stacks, loops and bulges. The data for the free energy of a stack are decomposed into enthalpy and entropy contributions. Therefore we get a temperature dependence of the stack contributions according to

$$F = H_{37} - TS_{37} \tag{23}$$

where H is the enthalpy and S is the entropy.

Unfortunately there is no decomposition of interior loops and bulges into enthalpy and entropy. But the formation of such a loop mainly decreases entropy and consequently its enthalpy is assumed to be negligible with respect to its entropy. Hence, we also have a temperature dependence for interior loops and bulges according to

$$F = -TS_{37} \tag{24}$$

The minimal free energy resulting from the Zuker algorithm accordingly does not correspond to the 0 K state of the sequence, but is the energy belonging to the most stable structure at 37° C.

3 Landscapes

3.1 The definition of a landscape

The notion of a landscape was introduced in the early thirties by Sewall Wright [33] in order to describe evolution as an adaptive walk on a fitness landscape. Nowadays landscapes appear in so different fields as in the physics of spin glasses, in the computer science of problems of combinatorial complexity, in evolution, in neural networks, in gene regulatory networks, in the maturation of immune response and in the biophysics of macromolecules.

A geographical landscape is described by the height h over all vectors $\vec{x} = (x, y)$ lying in the XY-plane. Here a landscape more generally stands for a scalar function $F(\vec{x})$, which assigns to all points $\vec{x} = (x_1, x_2, \ldots, x_n)$ of a n-dimensional space a real value. The total of all $\vec{x}$ is called the configuration space. In order to describe a landscape by its statistical properties, we need a metric in configuration space. In a geographical landscape the configuration space is two-dimensional and the components x, y of the vector $\vec{x}$ are continuous. The natural metric is the euclidian metric. We do not restrict the components of $\vec{x}$ to be continuous, in fact in all considered examples the x_i are discrete.

We can define a landscape as a triple (X, d, f) where X is a finite set, $d : X \times X \to \mathbb{R}_0^+$ is a metric and $f : X \to V$ is a function, which maps into a vector space with scalar product. This definition restricts the configuration space X to be discrete, because X has to be finite. For all landscapes discussed here, the vector space V is $\mathbb{R}$. f is often called the cost function. This expression originates from the study of fitness landscapes, where f evaluates the fitness of a species.

3.2 Fitness landscapes

Landscapes have to be seen in the context of an optimization process, which maximizes or minimizes $F(\vec{x})$. Evolution can be seen as an optimization process taking place on a fitness landscape [33]. Ever since Darwins [3] pioneering work about the origin of species evolution is understood as the interplay of mutation and selection. Mutation acts on the genotype, i. e. the genetic information, whereas selection takes place on the phenotypic level, i. e. the form emerging from the translation of the genetic information. Mutation and selection are the basis of a simple optimization procedure.

To give an example, let us consider a simplified model of molecular evolution. First of all we need a function $P(G)$, which maps the genotype in the phenotype, and a fitness landscape $f(P(G))$, which evaluates the fitness of a phenotype. Further we restrict the possibilities of mutation to point mutations, where accidentally one base is replaced by another. The probability of point mutations is assumed to be independent of the exchanged base and the position within the sequence. With the restriction to point mutations we assure that a mutation never affects sequence length. Allowing insertions or deletions would result in sequences of variable chain length and would therefore lead to a much complexer conception of configuration

space. We model the mutation-selection process by producing a random mutant of a given sequence and accepting it as the new sequence if its fitness value is not less than the fitness of the original sequence. This evolution model then is an adaptive walk on the fitness landscape.

The success of an evolutionary optimization algorithm depends strongly on the underlying fitness landscape. An adaptive walk will often fail to find the global optimum of a fitness landscape, which has many local optima, because whenever the walk reaches a local optimum it gets stuck. Clearly, looking more carefully at reality we see, that evolution is not very well described by an adaptive walk, but we need to have information on the landscape in order to estimate the power of an optimization algorithm. If we have for example a fitness landscape with one global optimum and no local optima, a gradient algorithm, which calculates the fitness of all one mutant neighbors and then accepts the one with the highest fitness, finds the optimum in fewer time steps than the adaptive walk. But the gradient algorithm will get stuck even faster, if there are local optima on the landscape.

The adaptive walk is the most simple model of evolution. Rechenberg [24] presented almost twenty years ago this and other more sophisticated optimization algorithms based on the mutation-selection principle, which have been applied to problems of mechanical engineering with great success. Developing models for evolution in nature one has to face enormous problems in finding appropriate functions, which map the genotype in the phenotype and which define a fitness for the phenotype. Although a lot of work has been done in morphogenesis, we are far away from predicting the phenotype from the genetic information. Whereas the secondary structure prediction is so far the only genotype-phenotype relationship, which reaches a level of fair reliability, we have too little knowledge on how the secondary structure of RNA molecules influences their selfreplication in order to derive a good approximation of the fitness value.

Fontana et al. [12, 11] investigated the dynamics of evolutionary optimization on RNA fitness landscapes, basing on a crude estimation of the velocity of selfreplication and the degradation rate. Fitness is here a function of the replication velocity and the degradation rate. The resulting replication and degradation landscapes have shown to be highly complex [9], which motivated our interest in the characterization of rugged landscapes.

3.3 The traveling salesman problem

A salesman, who has to visit several cities starting from his home city and returning back home after his trip, will carefully plan the order in which he visits the cities so that the total length of his tour will be as small as possible. Finding the shortest tour is the task of the traveling salesman problem (TSP). While the enormous work on the TSP problem contributed little to improve the living quality of salesmen[1],

[1] In 1832 a book appeared in Germany entitled "Der Handlungsreisende, wie er sein soll und was er zu thun hat, um Aufträge zu erhalten und eines glücklichen Erfolgs gewiss zu sein. Von einem alten Commis-Voyageur". In the last chapter we find: "By a proper choice and scheduling

many examples were found in various fields, which are closely connected with the TSP. But the importance of the TSP arises not from its direct application, but from being the most prominent example of a problem of combinatorial optimization.

If we consider a TSP with n cities, there are $(n-1)!/2$ different tours (home city and direction do not matter), where each city is visited only once. Having to do in 50 cities, the salesman had the choice between 3×10^{62} different tours. Hence, finding the shortest tour is not an easy task. The TSP belongs to the class of $\mathcal{NP}$-complete problem. $\mathcal{NP}$ complete are those problems, which can not be solved by any algorithm in polynomial time in n [15].

The TSP problem has been a playground for all kinds of optimization problems. Again we get in connection with the optimization algorithms landscapes for the traveling salesman problem [27] We can describe a tour T listing the visited cities c_i in chronological order:

$$T = \{c_1, c_2, \ldots, c_n\}$$

Hence, the configuration space is here the set of all possible tours. In order to carry out an adaptive walk on the traveling salesman landscape, we need a conception of neighboring tours and moves in configuration space. A simple example for a move in configuration space is exchanging two cities c_i and c_j from the tour T_t and generating therefrom the tour T_{t+1}. Tours, which can be mapped into each other by one single move are called neighbors. It is important that each point in configuration space is accessible from any other point by subsequent application of allowed moves.

The landscape of the TSP is

$$F(T) = \sum_{i=0}^{n} d(c_i, c_{i+1})$$

where $c_0 = c_n$ and where $d(c_i, c_j)$ is the distance between the cities c_i and c_j. Different optimization techniques, like simulated annealing, evolutionary algorithms and specially developed algorithms, have been applied to the TSP. Clearly, the landscapes corresponding to the various optimization techniques differ on which moves are allowed, i. e. which points are neighbors in configuration space. Another very often used move is the inversion, where the tour between the cities c_i and c_j is cut out and reinserted in opposite order.

The landscape of the TSP is more rugged for exchange moves than for inversions. This results from the different neighborhoods of inversion and exchange moves. Any exchange move $X_{ij}(T)$ of a tour T can be replaced by two successive inversions $I_{i+1,j-1}(I_{ij}(T))$. Accordingly, any direct neighboring tour in the X-neighborhood can be reached within two steps in the I-neighborhood. On the other hand direct neighbors in the I-neighborhood need not be close to each other in the X-neighborhood. Therefore it is plausible, that the X-landscape is more complex than the I-landscape and optimization with X-moves is the harder task.

---

of the tour, one can often gain so much time that we have to make some suggestions... The most important aspect is to cover as many locations as possible without visiting one twice". From [19]

In order to compare landscapes and their features we will have to find an appropriate measure for ruggedness. In section 4 we present tools known from statistics to characterize landscapes, but before that we have to introduce distance measures on RNA landscapes and in their configuration space.

3.4 The RNA configuration space and the free energy landscape

The configuration space of RNA molecules is the total of all possible sequences of a given length. For molecules consisting of β bases, the number of points in the configuration space is β^n, where n is the length of the chain. The configuration space is often called sequence space.

In a geographical landscape $|\vec{x} - \vec{y}|$ is a measure of the distance of two configurations. In the sequence space the distance between two molecules is given by the number of positions in which their bases differ. This distance measure is well known in information theory and is commonly called Hamming distance. Clearly, the maximal Hamming distance in the configuration space of dimension n is n. The number of sequences consisting only of **G** and **C** with length n and with Hamming distance h is

$$k = \binom{n}{h}$$

For four different base pairs this formula can be extended to

$$k = 3^h \binom{n}{h}$$

These functions have a very sharp maximum at $h = n/2$ for sequence length larger then 20. Consequently two randomly chosen sequences will most likely have Hamming distance $d = n/2$.

The configuration space for the four-letter alphabet molecules of chain length $n = 2$ is shown in figure 3. Lines are drawn between sequences of Hamming distance $h = 1$. The configuration space of **GC**-sequences of length n (see figure 4) is a hypercube of dimension n.

The Hamming distance is a distance measure in the sequence space, but we also need a distance measure of the values of the landscape. Again, in a geographical landscape this measure obviously is the difference of heights $|h_i - h_j|$. For the free energy landscape we can consequently define a distance:

$$d_e(S^i, S^j) = |F(S^i) - F(S^j)|$$

This distance measure and the Hamming distance in sequence space together with an algorithm, which determines the free energy of RNA molecules, defines a landscape, which is called the free energy landscape. For each length n and for each choice of the base set we get a different free energy landscape. Landscapes derived from the free energy computation according to the Zuker algorithm have been investigated recently [9, 13]. Here we focus on free energy landscapes derived

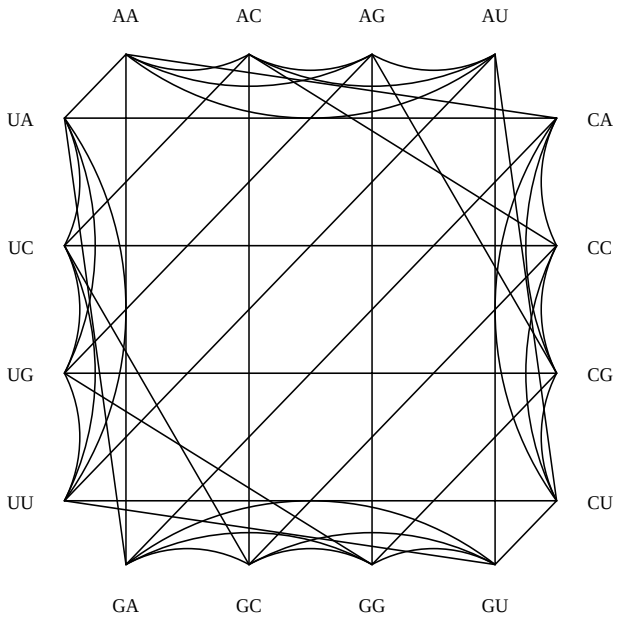

Figure 3: Configuration space of **GCAU**-sequences of chain length $n = 2$. The lines are drawn between sequence which differ in one position.

from the partition function algorithm in order to study the temperature dependence of the properties of the free energy landscape.

3.5 The landscape of the structure ensemble

Replication and degradation rates of RNA molecules depend on secondary structure motifs. For this reason it would be interesting to know, how much the secondary structure of two sequences with Hamming distance $h = 1$ differ on average. Recently, several methods have been proposed to measure distance between secondary structures [25, 18, 26]. Investigations of structure landscapes based on minimal free energy folding and tree distance of RNA secondary structures [26] are currently in work. With the partition function algorithm we can examine RNA landscapes at different temperatures. But as we described in section 2.3 the partition function algorithm yields instead of a secondary structure a base binding probability matrix. Consequently, we cannot directly apply any of the above mentioned algorithms to process the structural information contained in the base binding probability matrix. However, it is possible to modify the algorithm of Konings et al. [18], so that we compute also temperature dependence for structure landscapes.

We want to describe briefly the secondary structure comparison proposed by Konings et al. In section 2.1 we defined the secondary structure simply as the set of base pairs which are formed, when the polynucleotide chain folds back upon itself.

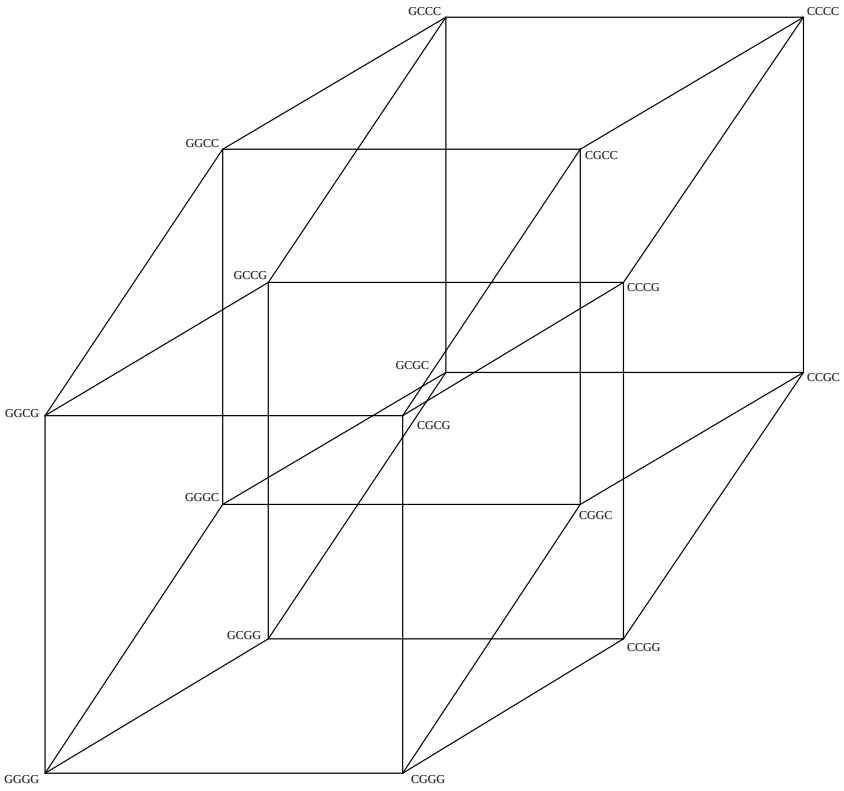

Figure 4: Configuration space of **GC**-sequences of chain length $n = 4$. The configuration space of **GC**-sequences of chain length n is a hypercube of dimension n.

A more detailed representation of the secondary structure is a linear array of the following symbols:

- $>$: if a base is upstream paired
- $<$: if a base is downstream paired
- $\circ$: if a base is in a single-stranded position

A standard maximal match alignment for linear sequences is used to compare the similarity of the so encoded secondary structure.

From the base binding probability matrix (p_{ij}) we can readily calculate the probability of a given base to be upstream, downstream or not paired:

$$p_i^{>} = \sum_{j>i} p_{ij} \tag{25}$$

$$p_i^{<} = \sum_{j<i} p_{ij} \tag{26}$$

$$p_i^{\circ} = 1 - p_i^{>} - p_j^{<} \tag{27}$$

We define in addition a similarity measure between positions in two sequences S^a and S^b as

$$\gamma = \sqrt{p_i^{>}(S^a)p_j^{>}(S^b)} + \sqrt{p_i^{<}(S^a)p_j^{<}(S^b)} + \sqrt{p_i^{\circ}(S^a)p_j^{\circ}(S^b)} \tag{28}$$

The main loop of the alignment routine is:

```
for( i=1 ; i<=sequence_length ; i++)
   for( j=1 ; j<=sequence_length ; j++)
   {
      gamma = sqrt(p_up_a[i-1] * p_up_b[j-1]) +
              sqrt(p_down_a[i-1] * p_down_b[j-1]) +
              sqrt(p_no_a[i-1] * p_no_b[j-1]);
      matrix[i][j] = maximum(matrix[i-1][j-1] + gamma,
                             matrix[i-1][j],
                             matrix[i][j-1]);
   }
dist = abs(sequence_length
           - matrix[sequence_length][sequence_length]);
```

The backtracking to find the optimal alignment can be skipped, because we are only interested in the resulting structure ensemble distance. If $S^a = S^b$ then $\gamma = 1$ for all i. Hence, $d_s(S^a, S^b) = 0$. Thus the structure ensemble distance ranges between zero and the sequence length.

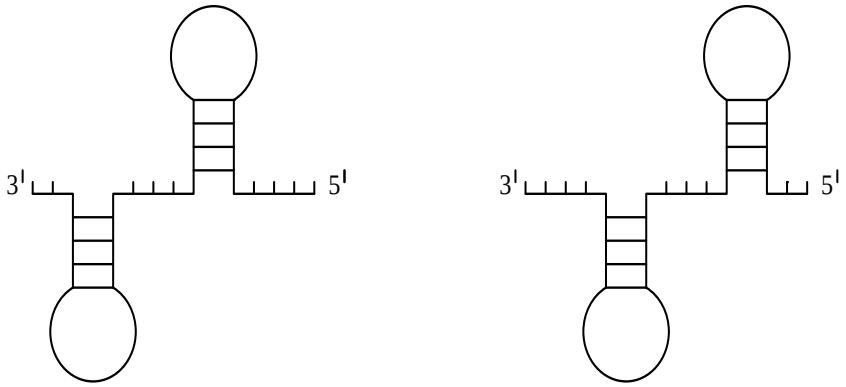

Figure 5: Two secondary structures with similar shape. The two structures differ only in the length of their free dangling ends.

Now we have a measure for the distance between the ensemble of structures of two RNA molecules. We can therefore apply all statistical methods, which will be derived in section 4 to a temperature dependent structure landscape basing on the partition function algorithm.

We want to stress here, that the above defined measure is not a measure between two structures, but measures the diversity of the secondary structure spectra of two sequences. The landscape of the structure ensemble distance is a vector landscape, i. e. each point in sequence space is assigned a vector, which contains information about the probability for each base in the sequence of being bound upstream, downstream or not being bound at all. The free energy landscape is a scalar landscape, because the free energy is a scalar quantity. Clearly we need for a landscape a measure of the distance of two values. For scalar landscapes the distance measure is obvious, but for vector landscapes the search for a convenient measure is not a simple task.

As we want to compare the similarity of two structures the length of the connecting vector of two probability vectors is not the quantity of interest. In figure 5 we see two very similar secondary structures, which differ only in the length of their free dangling ends. If we define for example the similarity measure as the number of pairs, where i is bound to j in both secondary structures, the two structures in figure 5 have a similarity of 0. Therefore we choose the above described alignment algorithm to find the maximal match between the two secondary structures.

4 Statistics

4.1 Measures of the probability distribution

Let us first consider the non-spatial value distribution of the landscape. Therefore we determine the expectation value $\langle X_i \rangle$, the variance $var(X_i)$, the skewness $skew(X_i)$ and the kurtosis $kur(X_i)$ of the probability distribution of randomly chosen value X_i in the landscape. The sample mean is defined by

$$\langle X_i \rangle = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{29}$$

The coefficient of variance is given by

$$\begin{aligned} var(X_i) &= \langle (X_i - \langle X_i \rangle)^2 \rangle \\ &= \langle X_i^2 - 2X_i \langle X_i \rangle + \langle X_i \rangle^2 \rangle \\ &= \langle X_i^2 \rangle - \langle X_i \rangle^2 \end{aligned}$$

The coefficient of skewness is

$$\begin{aligned} skew(X_i) &= \frac{\langle (X_i - \langle X_i \rangle)^3 \rangle}{(var(X_i))^{\frac{3}{2}}} \\ &= \frac{\langle X_i^3 - 3X_i^2 \langle X_i \rangle + 3X_i \langle X_i^2 \rangle - \langle X_i \rangle^3 \rangle}{(var(X_i))^{\frac{3}{2}}} \\ &= \frac{\langle X_i^3 \rangle - 3\langle X_i^2 \rangle \langle X_i \rangle + 2\langle X_i \rangle^3}{(var(X_i))^{\frac{3}{2}}} \end{aligned}$$

Finally, the kurtosis is defined by

$$\begin{aligned} kur(X_i) &= \frac{\langle (X_i - \langle X_i \rangle)^4 \rangle}{(var(X_i))^2} \\ &= \frac{\langle X_i^4 - 4X_i^3 \langle X_i \rangle + 6X_i^2 \langle X_i \rangle^2 - 4X_i \langle X_i \rangle^3 + \langle X_I \rangle^4}{(var(X_i))^2} \\ &= \frac{\langle X_i^4 \rangle - 4\langle X_i^3 \rangle \langle X_i \rangle + 6\langle X_i^2 \rangle \langle X_i \rangle^2 - 3\langle X_i \rangle^4}{(var(X_i))^2} \end{aligned}$$

Note, that the sample mean and the variance have the dimension of X_i, X_i^2 respectively, whereas the coefficients of skewness and kurtosis are dimensionless.

What the expectation value and the variance tell us about a value distribution is obvious. The skewness is a measure of the symmetry of the distribution. If the skewness has a positive sign the distribution is skewed to the right, if it has a negative sign the distribution is skewed to the left, if it is zero the distribution is symmetrical in respect to its mean value. To give an example for those, who are familiar with the geography of the alps, the Eiger seen from the Jungfraujoch is skewed to the right.

The kurtosis gives a measure of the flatness of the distribution. The smaller the kurtosis the flatter is the distribution. The Kilimandjaro consequently has smaller kurtosis than the Matterhorn. To give a mathematical example, the coefficient of skewness for the Gaussian distribution is 0 and the coefficient of kurtosis is 3.

4.2 The autocorrelation function

In order to describe the features of a landscape, we need a relationship between the distance of two configurations in configuration space and the mean change of their height. The previously introduced measures of probability distribution do not take into account the spatial structure of a landscape. An appropriate measure is the autocorrelation function defined by

$$\begin{aligned} \rho(k) &= \frac{cov(X_i, X_{i+k})}{\sqrt{var(X_i)var(X_{i+k})}} \\ &= \frac{\langle (X_i - \langle X_i \rangle)(X_{i+k} - \langle X_{i+k} \rangle) \rangle}{\sqrt{\langle (X_i - \langle X_i \rangle)^2 \rangle \langle (X_{i+k} - \langle X_{i+k} \rangle)^2 \rangle}} \end{aligned}$$

where X_i and X_{i+k} belong to the same stochastic process. $cov(XY)$ is the covariance of X_i and X_{i+k} defined as

$$cov(XY) = \langle XY \rangle - \langle X \rangle \langle Y \rangle$$

Hence, the range of the autocorrelation function is $\rho(k) \in [-1, 1]$, because

$$O \leq var(X \pm Y) = var(X) + var(Y) \pm 2cov(XY)$$

$$\rightarrow |cov(XY)| \leq var(X)$$

if X and Y are created by the same stochastic process. The formula (30) can be simplified, because $\langle X_i \rangle = \langle X_{i+k} \rangle$ and $var(X_i) = var(X_{i+k})$, to

$$\rho(k) = \frac{\langle (X_i - \langle X_i \rangle)(\langle X_{i+k} - \langle X_{i+k} \rangle) \rangle}{\langle (X_i - \langle X_i \rangle)^2 \rangle} \tag{30}$$

We can also write equation (30) as

$$\rho(k) = 1 - \frac{\langle X_i^2 \rangle - \langle X_i X_{i+k} \rangle}{\langle X_i^2 \rangle - \langle X_i \rangle^2} \tag{31}$$

If X_i represents the height over the configuration $\vec{x}_i$ and the distance on the landscape between X_i and X_{i+k} is k, then the autocorrelation is the desired relationship between distance in configuration space and distance of values, which gives us information about the ruggedness of the landscape.

If the autocorrelation function fits well to an exponential curve, with $\rho(k) \cong e^{-\lambda k}$, all information about the autocorrelation function is contained already the correlation length defined by $l = \lambda^{-1}$. A large correlation length corresponds to a slowly decaying autocorrelation function and consequently to a smooth landscape.

We will here present a derivative of equation (30), which will become important later on. If we want to correlate a function over the configuration space, which has the form $F(\vec{x}, \vec{y})$, we cannot determine the autocorrelation function according to the equations (30) and (31). It is possible to define a measure for structural similarity [26] of two secondary structures, but there is no value assigned to a single secondary structure. For this case the suitable form of the autocorrelation function is

$$\rho(k) = 1 - \frac{\langle (X_i - X_{i+k})^2 \rangle}{\langle (X_i - X_j)^2 \rangle} \tag{32}$$

An example for a function $F(\vec{x}, \vec{y})$ is the correlation of the secondary structure.

4.3 Sampling techniques for Landscapes

Sampling statistical properties of landscapes is not that simple as one might presume. We know from section 3.1, that for natural RNA landscapes there are 4^n different sequences of length n in configuration space. The probability, that two randomly chosen sequences have a Hamming distance h, is

$$p(h) = (\beta - 1)^h \binom{n}{h} \beta^{-n} \tag{33}$$

where β is the number of different bases of which the sequences consist. Hence, random points in configuration space will have almost always Hamming distance $h = n/2$ even for moderate chain lengths. Neighboring points in the configuration space for sequence of length $n = 30$ are found with probability 8×10^{-17}. For the numerical evaluation of the autocorrelation function we need to sample neighboring sequences with sufficient statistical weight. Two different techniques were applied here, which we want to discuss briefly in the following two sections.

4.3.1 The random walk technique

The random walk technique produces a time series

$$\mathcal{X} = \{X_0, X_1, X_2, \ldots\}$$

where the sequences S^i and S^{i+1} corresponding to the values X_i and X_{i+1} have hamming distance $h = 1$. The sequence S^{i+1} generated from the sequence S^i by randomly replacing one base of S^i through an other. It is important to notice, that the number of steps of the random walk does not coincide with the Hamming distance. (see fig. 6. From the time series we can directly calculate $\langle X_i \rangle$ and $\langle X_i^2 \rangle$. In order to compute the autocorrelation function according to equation (31), we have to determine $\langle X_i X_{i+j} \rangle_h$, where the subscript h denotes the Hamming distance between X_i and X_{i+j}. With this notation equation (31) is written

$$\rho(h) = 1 - \frac{\langle X_i^2 \rangle - \langle X_i X_{i+j} \rangle_h}{\langle X_i^2 \rangle - \langle X_i \rangle^2} \tag{34}$$

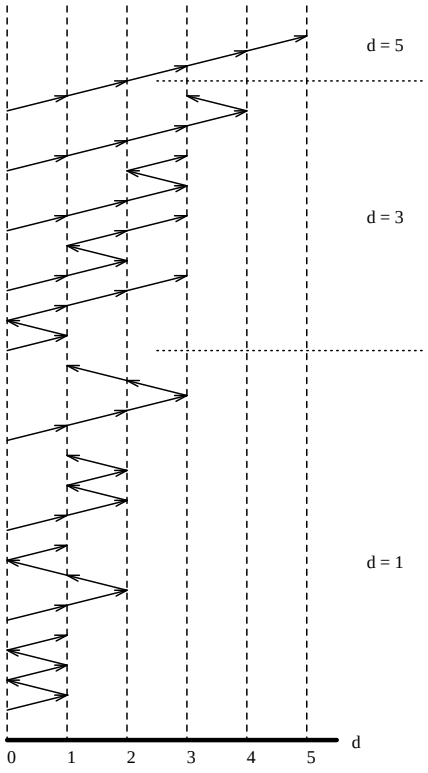

Figure 6: Possible Hamming distances for a random walk of 5 steps in the configuration space of two-letter RNA sequences

Along the time series X_i can be correlated with all preceding $X_0, \ldots, X_{i-1}$. Since hamming distances larger than $n/2$ will rarely be found because of the probability distribution (33), it is not senseful to store during the computation much more than the last $n/2$ sequences and their values. Therefore we cannot determine the full autocorrelation function with the random walk technique. But we do not need the full autocorrelation function to calculate the correlation length of the landscape.

To avoid systematic deviation due to any quasiperiodicity of the random number generator, we divide the random walk into packages of 1000 steps and start each package from a new randomly chosen sequence. The total number of points needed to assure convergence of the statistical quantities depends on the length of the sequences, but a few 100000 points normally are sufficient for chain lengths up to 50.

Let us denote with $\Phi_{hs}(n,\beta)$ the probability, that a walk of s steps has a Hamming distance h from the starting sequence, where n is the length of the sequences and β is the number of different bases contained in the sequences. Obviously

$$\Phi_{hs}(n,\beta) = 0 \text{ for } h > s \text{ or } h > n \text{ or } h < 0$$

We can derive a recursion formula to compute $\Phi_{hs}(n,\beta)$.

$$\begin{aligned}\Phi_{hs}(n,\beta) &= \Phi_{h-1,s-1}(n,\beta)\frac{n-h+1}{n} \\ &\quad +\Phi_{h,s-1}(n,\beta)\frac{h(\beta-2)}{n(\beta-1)} \\ &\quad +\Phi_{h+1,s-1}(n,\beta)\frac{h+1}{n}\frac{1}{\beta-1}\end{aligned}$$

Hence

$$\Phi_{h=s,s}(n,\beta) = \begin{cases} 1 & \text{if } s = 0,1 \\ \frac{(s-1)!}{n^{s-1}}\binom{\beta-1}{s-1} & \text{if } s > 1 \end{cases}$$

with the initial condition $\Phi_{00}(n,\beta) = 1$.

As we pointed out earlier when describing the folding algorithms, the partition function algorithm yields not only the partition function of the whole RNA sequence but also returns the partition function of subsequences. This enables one to measure all statistical quantities in one program execution also for the shorter subsequence lengths. Since the partition function algorithm requires an considerable amount of computation time, correlation of subsequences speeds up the calculation of the correlation length as a function of the chain length substantially.

The number of subsequences, which can be correlated in this manner, is limited for two reasons. One purely technical reason is due to the limitations of computer storage. The other reason is, that too short subsequences will be too seldomly affected by random mutations. Therefore we will not accumulate enough data to assure the convergence of the surveyed properties for short sequences.

The backtracking for the computation of the base binding probability matrix is also for subfragments of the whole RNA molecule of cubic order in the sublength. This follows from equation (22) on page 13. Consequently it does not pay off to correlate structural properties also for subfragments in one single program execution.

4.3.2 The neighborhood technique

For the neighborhood technique we choose a reference sequences of length n in configuration space and compute for all Hamming distances from $h = 1$ to $h = n-1$ n sequences. If we consider sequences consisting only of two different bases, there is only one sequence complementary to the reference sequence, which we also take into account in the neighborhood sample. In the case of four base sequences we choose randomly n sequences with Hamming distance n.

We compute the free energy distance between the reference sequence S^{ref} and the n sequences $S^i(h)$ with Hamming distance h according to

$$d_e = |F(S^{ref}) - F(S^i(h))|$$

and count the occurrence of a certain energy distance d_e under the condition, that the two underlying sequences have a Hamming distance h. In the same way we proceed for structural ensemble distance d_s as we defined it in section 3.5. This data are stored in twodimensional arrays $n_e(h, d_e)$, $n_s(h, d_s)$ respectively. Then we choose a new reference sequence at random and repeat the whole procedure until the surveyed properties converge.

The conditional probability to find a distance d given that two sequences have hamming distance h is

$$p(d|h) \approx \frac{n(h,d)}{\sum_{d=0}^{d_{max}} n(h,d)} \tag{35}$$

We dropped here the subscripts e and s. According to the notation of equation (34) we can now derive

$$\langle X_i - X_j \rangle_h = \sum_{d=0}^{d_{max}} d^2 p(d|h) \approx \frac{\sum_{d=0}^{d_{max}} d^2 n(h,d)}{\sum_{d=0}^{d_{max}} n(h,d)}$$

and

$$\langle X_i - X_j \rangle = \sum_{h=0}^{n} \langle X_i - X_j \rangle_h p(h) \approx \frac{\sum_{d=0}^{d_{max}} d^2 n(h,d) p(h)}{\sum_{d=0}^{d_{max}} n(h,d)}$$

where p(h) is the probability distribution (33). For the calculation of the variance $var(X) = \langle X_i - X_j \rangle$ we have to have randomly chosen sequences S_i and S_j. Because we do not choose points at random in sequence space for the neighborhood technique, we have to multiply the conditional mean square distances $\langle X_i - X_j \rangle_h$ with their probability of occurrence.

Altogether the autocorrelation function may now be expressed as

$$\rho(h) = 1 - \frac{\sum_{d=0}^{d_{max}} d^2 n(h,d)}{\sum_{h=0}^{n} \sum_{d=0}^{dmax} d^2 n(h,d) p(h)} \tag{36}$$

The neighborhood technique enables us to derive the full autocorrelation function of all hamming distances according to equation (32). The statistical properties also seem to converge faster than with the random walk technique, but correlation of subfragments is much more awkward and exhaust computer storage capability faster. The random walk technique is applied to energy landscapes, because of the ability to sample all statistic properties of interest also for sublengths. The neighborhood method is only applied to the correlation of ensemble structure, because determining the base binding probability matrix for segments of the sequence would be too time consuming.

5 Numerical results

5.1 The value distribution of free energy landscapes

The free energy of a given secondary structure is the sum over the free energy contributions of the contained loops. Neglecting tertiary interactions, the free energy of a RNA molecule is (according to equations 2 and 3)

$$F = -kT \sum_{\Phi} e^{-F(\Phi)}$$

where the free energy contributions of every possible secondary structure of the molecule are weighted exponentially for the overall free energy. From here on we will use the notion free energy always as the free energy computed with the partition function algorithm.

Clearly, the mean value of the free energy of a given secondary structures decreases linearly with increasing chain length. Consequently the overall free energy has the same dependence on chain length. Because **GC**-pairs are bound stronger than **AU**-pairs, the formation of the secondary structure for **GC**-only sequences yields on average smaller free energy than for **GCAU** sequences. Figure 7 shows the chain length dependence of the mean free energy for different temperatures. The data shown in this figure were computed with 400 independent random walks of length 1000 for each temperature. The mean value of free energy has been measured for chain lengths from 30 up to 50 with step size 2.

As the computation of the free energy of a sequence of length 50 is rather time consuming, we did not carry out the calculation for all temperatures up to chain length 50. For detailed information on computer time requirements see appendix A.

The temperature dependence of the free energy (see figure 8) is also easy to understand. There are two different effects, which contribute to the temperature dependence of the free energy. Beginning at low temperature, we see first a linear increase of the free energy, because the interaction between the complementary base pairs becomes weaker for increasing temperature. But as the bonds become weaker, the mean number of base pairs formed in the secondary structure will decrease. This effect obviously saturates, when the majority of structures in the ensemble are already in or at least close to the unfolded state.

In order to understand the chain length dependence of the standard deviation we want to discuss the consequences of the linear dependence of mean free energy of sequence length. Let us denote with $\langle X(l) \rangle$ the average of a random variable $X(l)$, which is a function of n. If the mean value of the random variable is linear in l, then

$$\langle X(l+m) \rangle = \langle X(l) \rangle + \langle X(m) \rangle$$

If we define $\langle X \rangle := \langle X(1) \rangle$, we get

$$\begin{aligned} \langle X(l) \rangle &= l \langle X \rangle \\ &= \frac{1}{N} \left(\sum_{i=1}^{N} X_i^1 + \ldots + \sum_{i=1}^{N} X_i^l \right) \end{aligned}$$

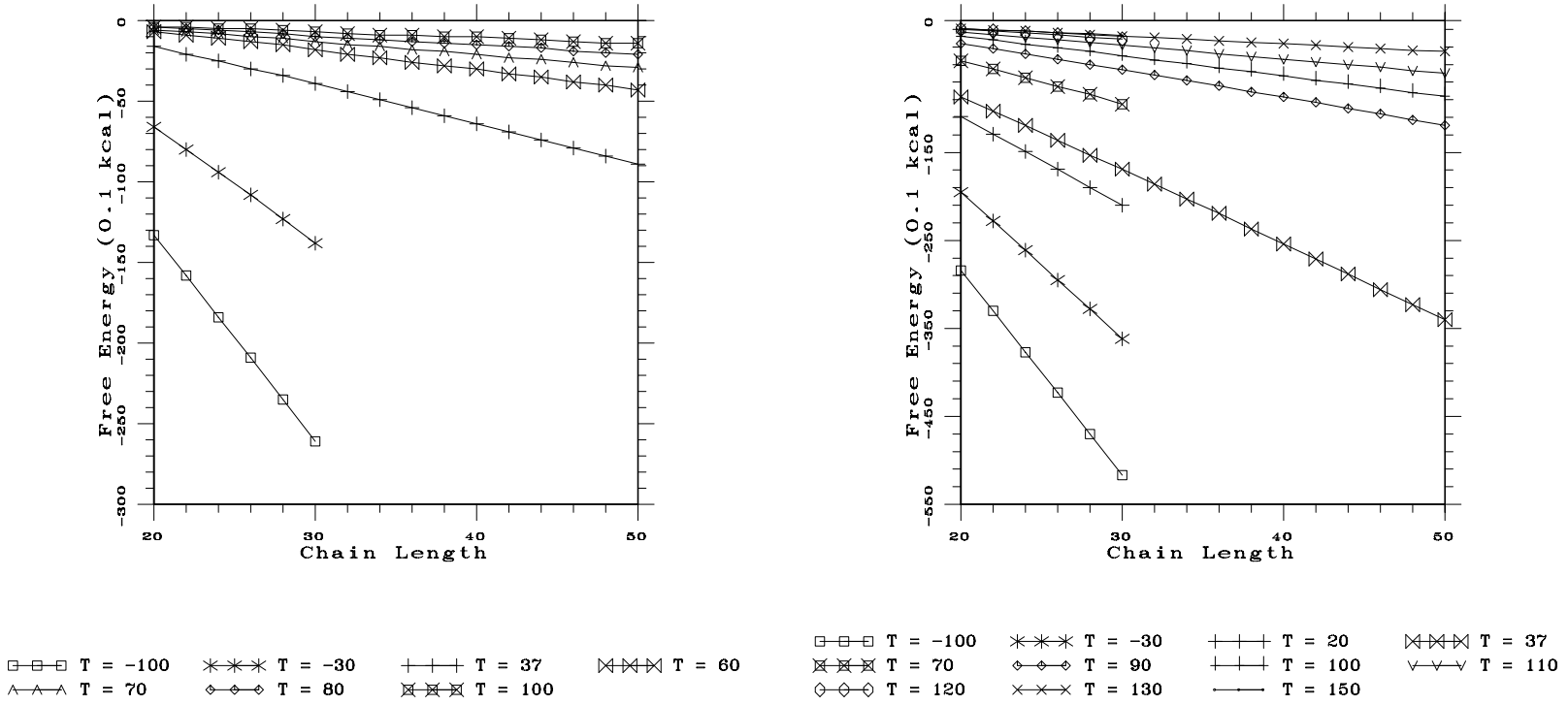

Figure 7: The mean free energy for **GCAU**-sequences (left) and for **GC**- sequences (right) versus chain length for different temperatures. Because of computer time limitations the mean value of the free energy has been sampled for some temperatures only between chain length 20 and 30.

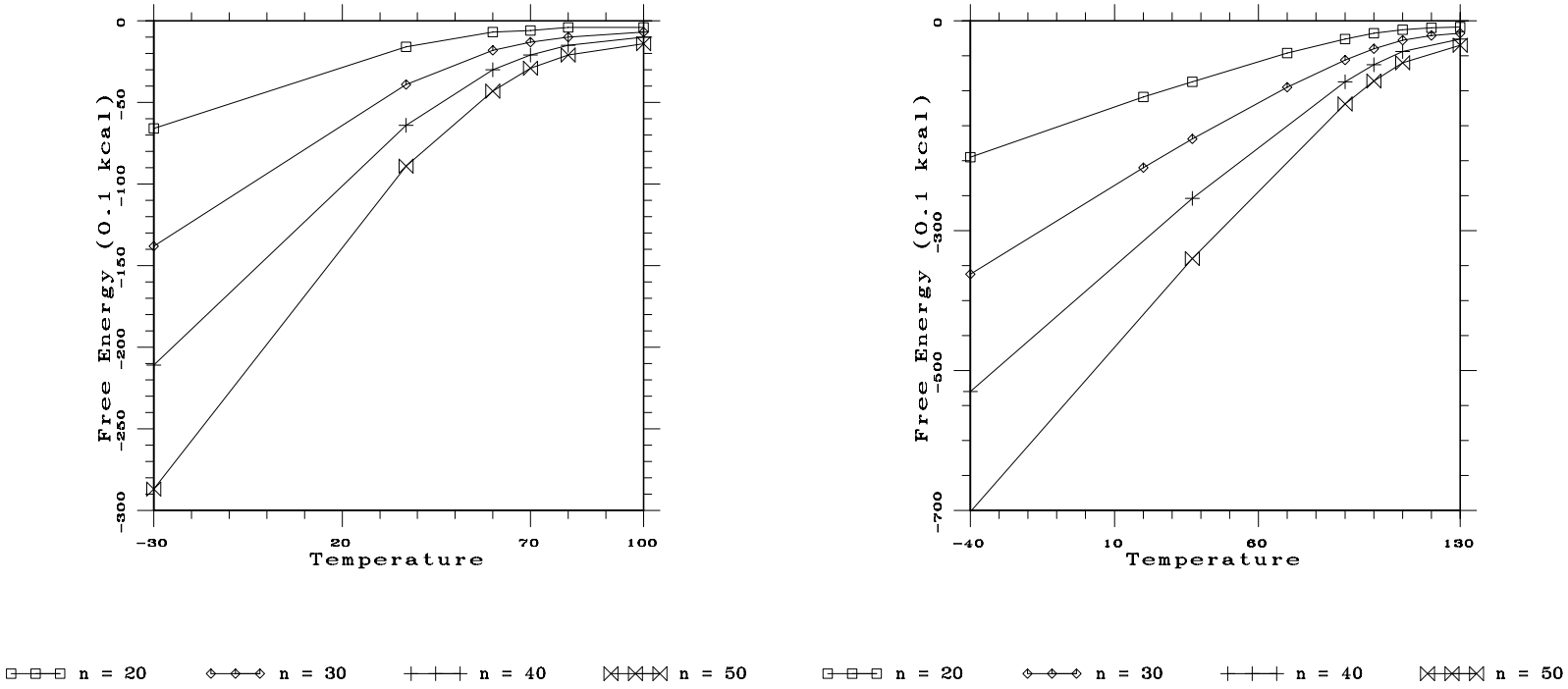

Figure 8: The mean free energy for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths between 20 and 50 in steps of ten. The **GCAU** data have been sampled at temperatures $T = -100, -30, 37, 60, 70, 80, 100°\mathrm{C}$. The **GC** data have been sampled at temperatures $T = -100, -40, 20, 37, 70, 90, 100, 110, 120, 130, 150°\mathrm{C}$.

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{l} X_i^j$$

Hence, the random variable $X(l)$ (which itself must not be linear in l), is representable as the sum over l random variables X. As the mean value of the free energy is linear in the sequence length, it can be represented as a sum over l random variables, where l does not have to match n, but has to be proportional to n.

Let us now calculate the variance of a random variable $X(l)$, which is the sum over l random variables X.

$$\begin{aligned} var(X(l)) &= \langle X(l)^2\rangle - \langle X(l)\rangle^2 \\ &= \langle \sum_{i=1}^{l}\sum_{j=1}^{l} X_i X_j\rangle - \langle \sum_{i=1}^{l} X_i\rangle \\ &= \langle \sum_{i=1}^{l} X_i^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{l} X_i X_j\rangle - \langle \sum_{i=1}^{l} X_i\rangle \end{aligned}$$

If the random variables in the sum are independent, the above equation reduces to

$$var(X(l)) = l\langle X^2\rangle - l\langle X\rangle^2$$

and the standard deviation of $X(n)$ is proportional to the square root of l.

In figure 9 the standard deviation of the free energy values is plotted versus the chain length for **GCAU** and **GC**-sequences. We see an increase of the standard deviation proportional to n^ρ, with $\rho < 1$. We expect, that in the limit of large n the standard deviation will show a square root of n behavior, because the secondary structure will then decompose in parts , which fold independently into substructures. The overall free energy will then be the sum over independent variables, and therefore its variance will increase linearly with the sequence length.

The other way around, we see how close we are to independent random contributions, if we try to fit the chain length dependence of the standard deviation to $\sqrt{n}$.

In figure 10 the dependence of the relative deviation on free energy with chain length is shown. The relative deviation is the standard deviation divided by the absolute value of the mean. Since the standard deviation of the free energy increases with chain length as n^ρ, with $\rho < 1$, the relative deviation will converge to 0 for large n. This implies, that the value distribution of the free energy becomes sharper with increasing sequence length. The reason is, that the probability to find a very stable secondary structure or to find a sequence , which does not have a secondary structure is much more likely for short sequences then for long ones. The probability to find a long sequence, which does not fold, is negligible and consequently sequences with extreme high or extreme low free energy will contribute to the distribution with very small statistical weight.

If we compare the left plot with the right plot in figure 10, we see, that the relative deviation of the free energy for **GCAU**-sequences is larger than for **GC**-sequences.

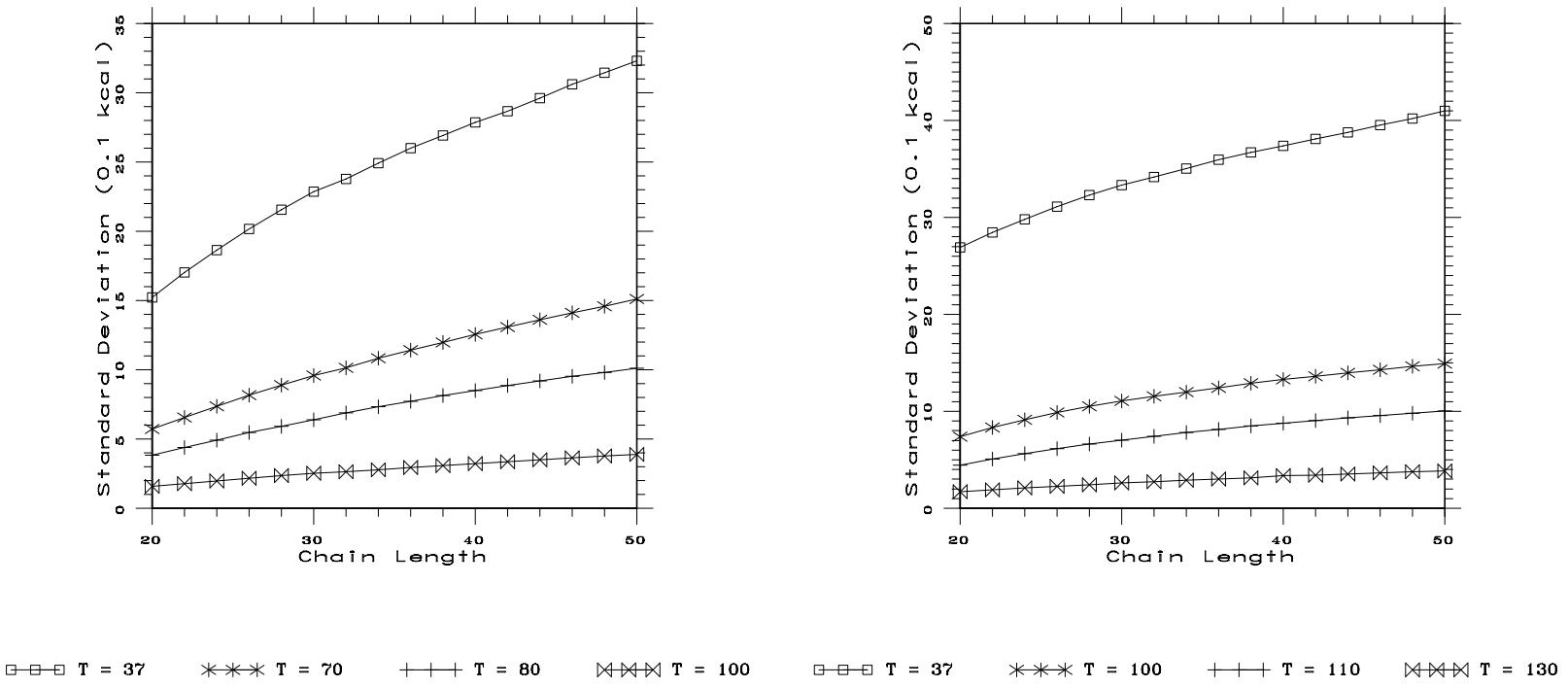

Figure 9: The standard deviation of the free energy distribution for **GCAU**-sequences (left) and **GC**-sequences (right) versus chain length.

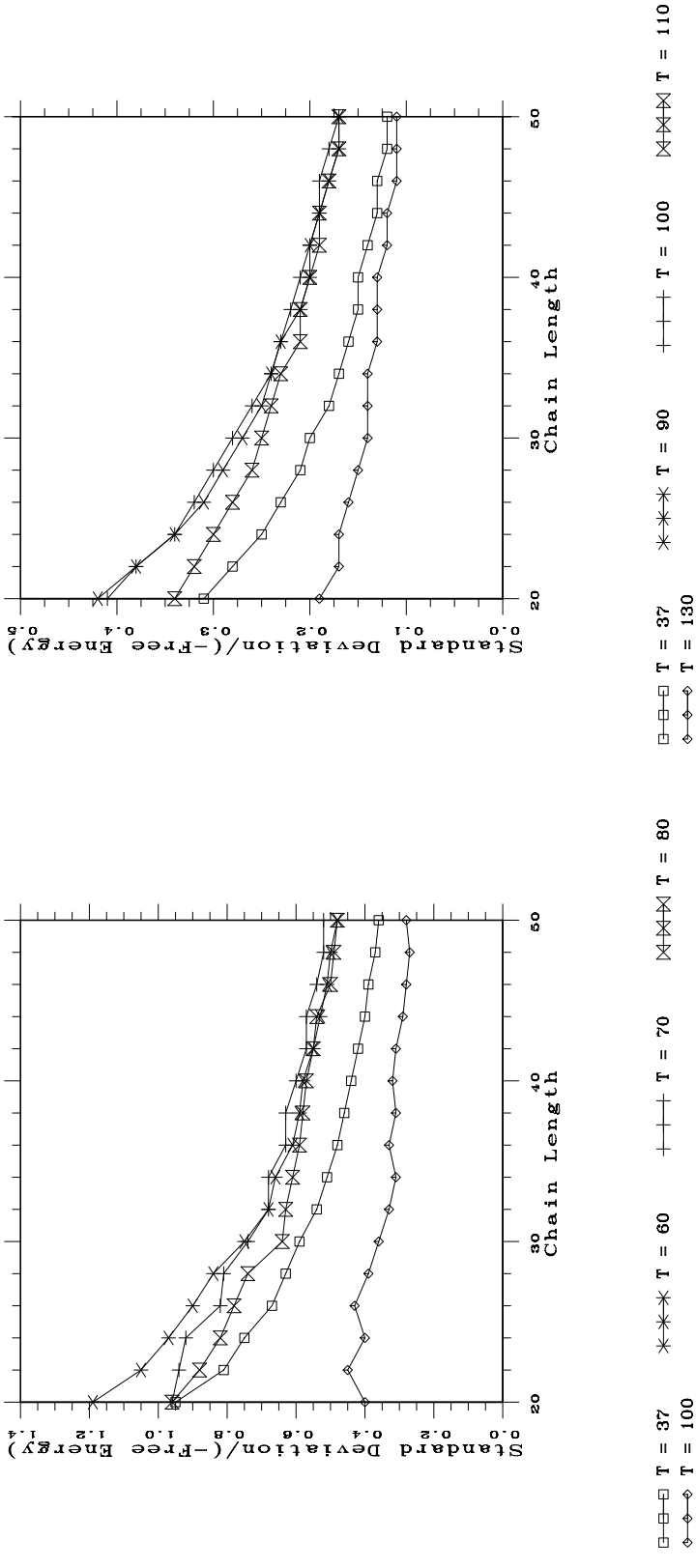

Figure 10: The relative deviation of the free energy distribution for **GCAU**-sequences (left) and **GC**-sequences (right) versus chain length.

It is clear, that the range of possible values for the free energy of **GCAU**-sequences is larger than for **GC**-sequences, because the set of all **GCAU**-sequences comprises pure **AU**-sequences, with relatively high free energy, as well as pure **GC**-sequences, with rather low free energy.

The central limit theorem of probability theory leads us to another consequence of the linearity of the mean free energy. The central limit theorem may be formulated as follows:

Central limit theorem: Let $X_1, X_2, \ldots, X_n$ be independent random variables, which have the same mean μ and the same variance σ^2. Then if $S_n = X_1 + X_2 + \cdots + X_n$,

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx$$

that is, the random variable $(S_n - n\mu)/\sigma\sqrt{n}$, which is the S_n standardized to mean 0 and variance 1, is asymptotically Gaussian distributed.

Accordingly, the same argumentation, why we expect a square root of n behavior of the standard deviation in the limit of large n, leads us to the conjecture that the distribution of free energy values becomes Gaussian for large n.

As we know, that the skewness of a Gaussian distribution is 0 and the kurtosis is 3, a look at the sequence length dependence of the skewness of the free energy distribution (see figures 11,12), tells us how close the free distribution is to the normal distribution. The data for the **GCAU**-sequences show clearly the expected convergence to the normal distribution, but for the sampled chain lengths the values of skewness and kurtosis are still far from the values for the Gaussian distribution. The lower the temperature, the closer the coefficient of skewness and the coefficient of kurtosis come to the values for the Gaussian distribution. Although the values of the skewness and the kurtosis for the **GC**-Data are closer to the Gaussian distribution than for the **GCAU**-Data, whether the skewness nor the kurtosis show a convergent behavior to the normal distribution within the range of sampled chain lengths.

The convergence of the free energy distribution to a Gaussian distribution is interesting, because under the assumption of normal distribution analytical expressions have been derived to estimate the number of local optima for not too highly correlated landscapes [27].

In figure 13 we see the temperature dependence of the standard deviation of free energy for **GCAU** and **GC** sequences. If we compare these plots with the temperature dependence of the mean free energy in figure 8 we see at first glance a rather similar behavior. But looking more carefully at the temperature dependence of the standard deviation we detect, that the curvature of the plots changes the

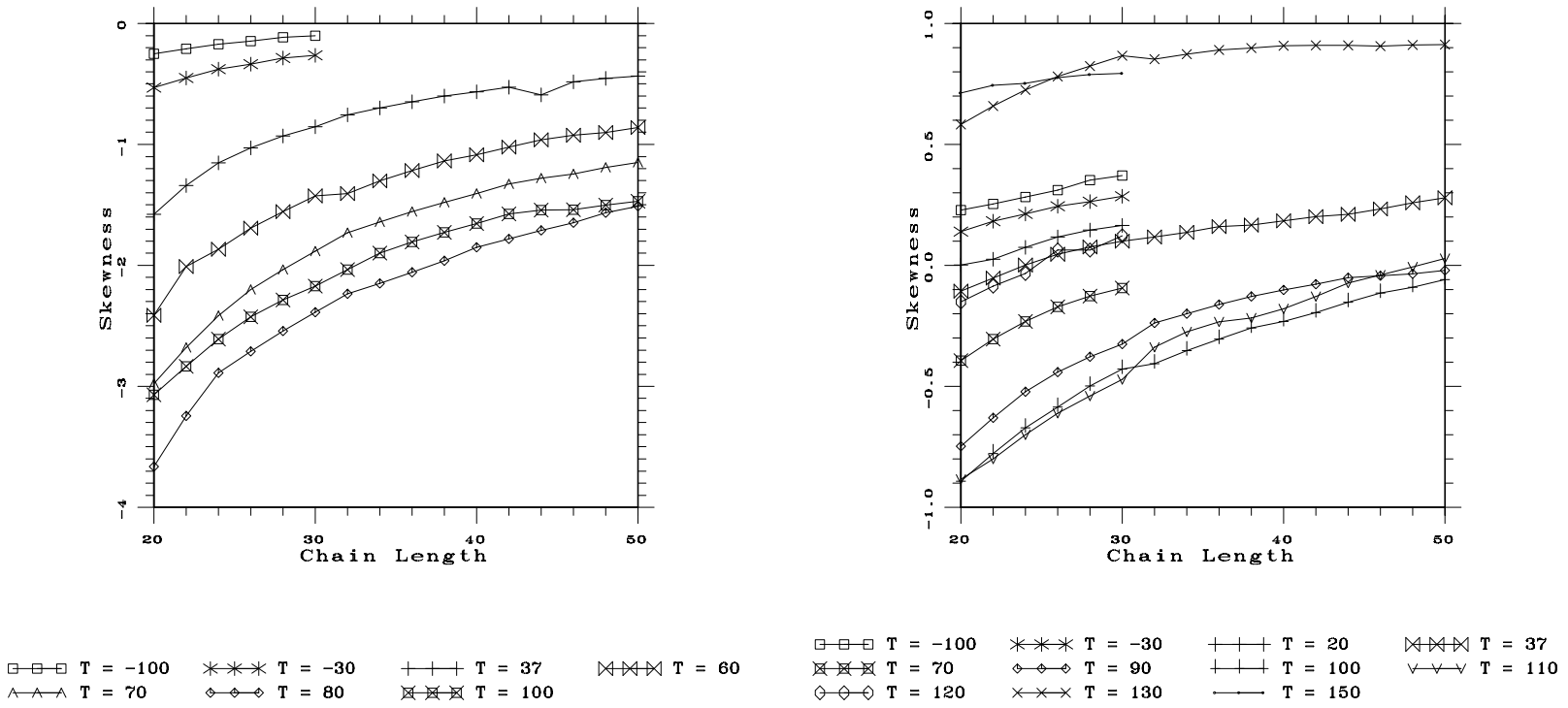

Figure 11: The skewness of the free energy distribution for **GCAU**-sequences (left) and **GC**-sequences (right) versus chain length.

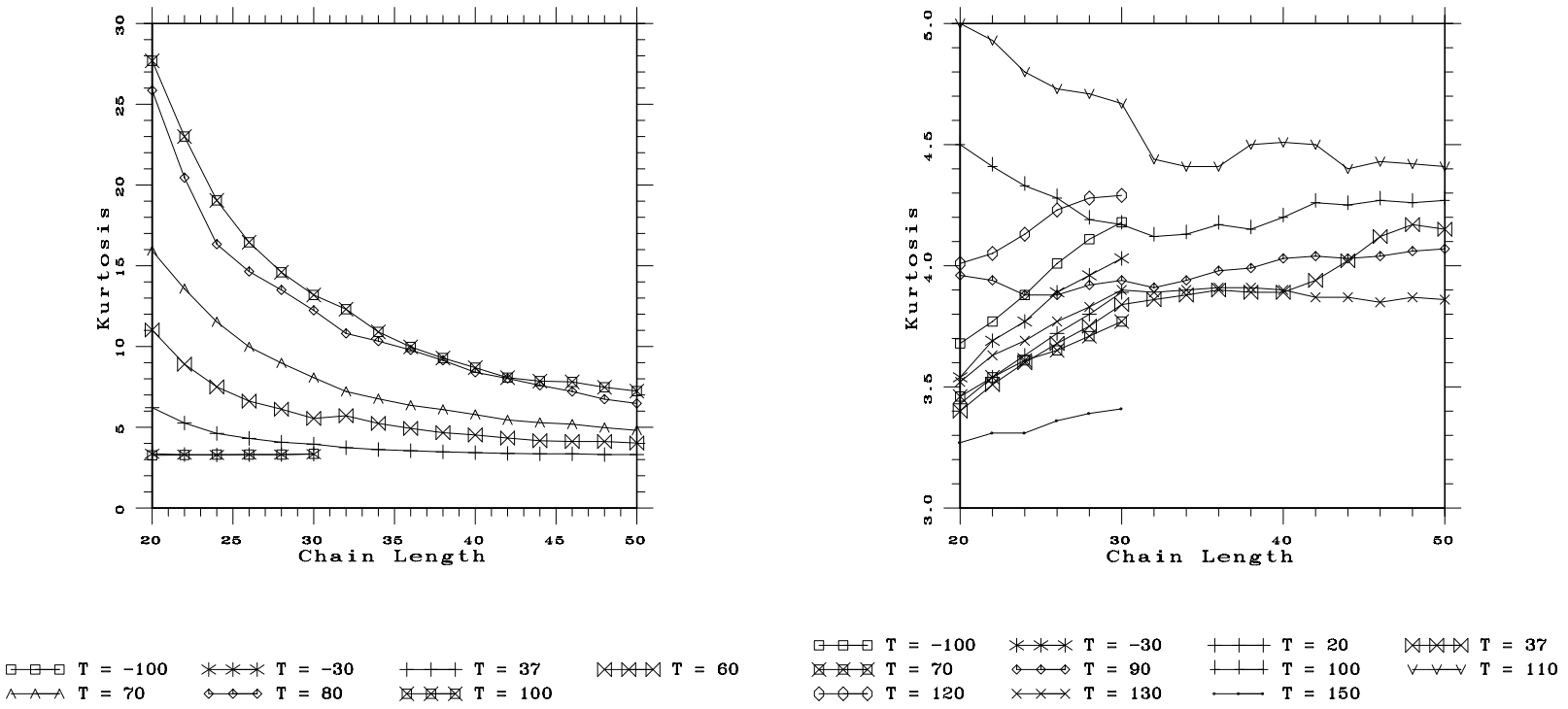

Figure 12: The kurtosis of the free energy distribution for **GCAU**-sequences (left) and **GC**-sequences (right) versus chain length.

sign. The different behavior of the standard deviation and the mean free energy with temperature becomes more evident, if we plot the relative deviation versus temperature (see figure 14).

We see that the relative deviation has a maximum. This maximum does not only depend on which base set we choose, but also depends on chain length. The maximum value of the relative deviation shifts for longer chain length to higher temperatures. We seem to observe a mean melting point of the secondary structures. The relative deviation grows with temperature, because more and more molecules remain in the unfolded state, until the point, where the majority of structures is unfolded. From this point on the relative deviation decreases, because structures with low free energy will be found more and more rarely. The mean melting point is clearly increases with chain length, because more heat is needed to melt the secondary structures.

Figure 15 shows the temperature dependence of the skewness for the free energy distribution. Approaching the melting point from low temperatures the the free energy distribution gets more skewed to the left. The energy distribution is not symmetric, because secondary structures with positive free energy are not allowed. At higher temperature the mean value of free energy and accordingly the whole distribution shifts towards zero, and therefore the plotted temperature dependence of the coefficient of skewness below the melting point seems plausible.

Beyond the melting point the behavior of the **GC** data change dramatically. Obviously the distribution changes completely. Comparing the skewness versus temperature plot with the kurtosis versus temperature plot for the **GC** data, we detect a sharp peak of the kurtosis at the temperature, where the skewness changes its behavior.

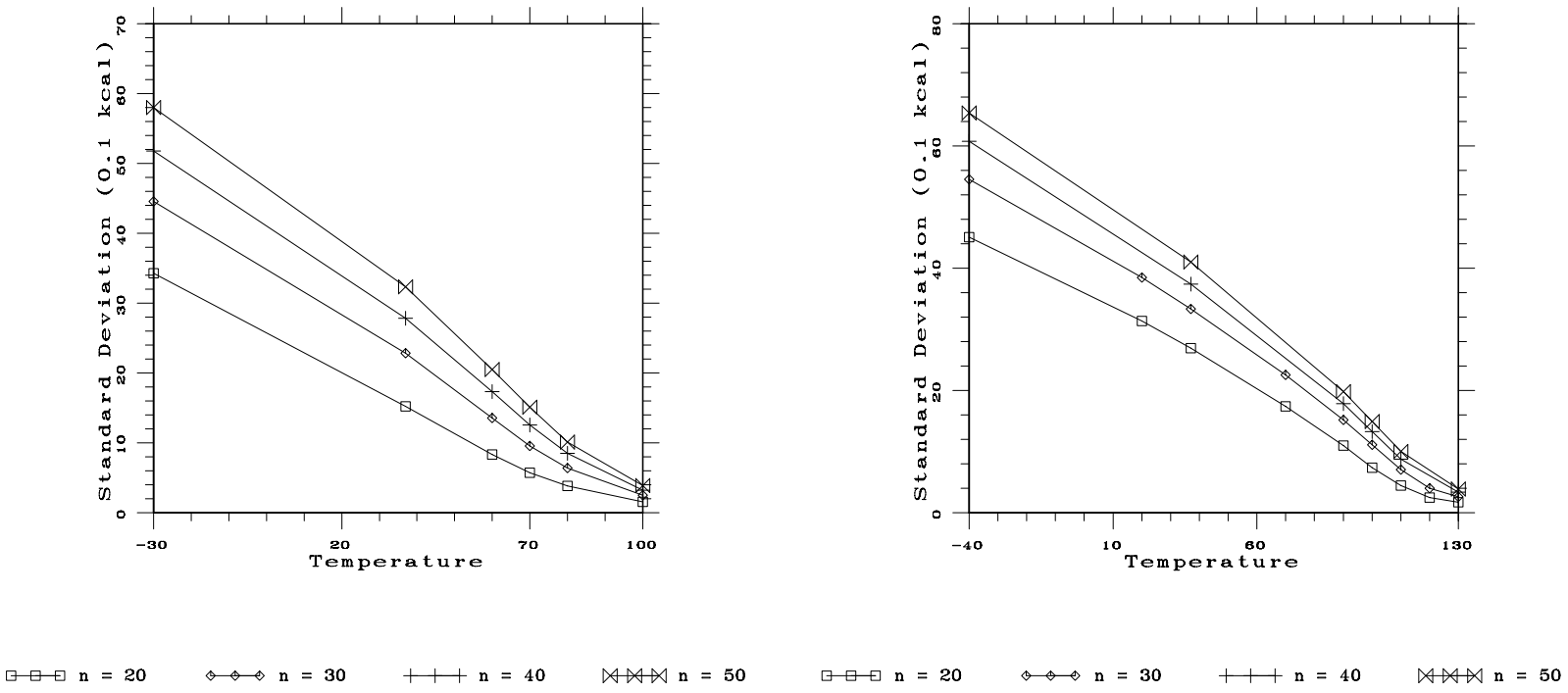

Figure 13: The standard deviation for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths between 20 and 50 in steps of ten. The **GCAU** data have been sampled at temperatures $T = -100, -30, 37, 60, 70, 80, 100°\mathrm{C}$. The **GC** data have been sampled at temperatures $T = -100, -40, 20, 37, 70, 90, 100, 110, 120, 130, 150°\mathrm{C}$.

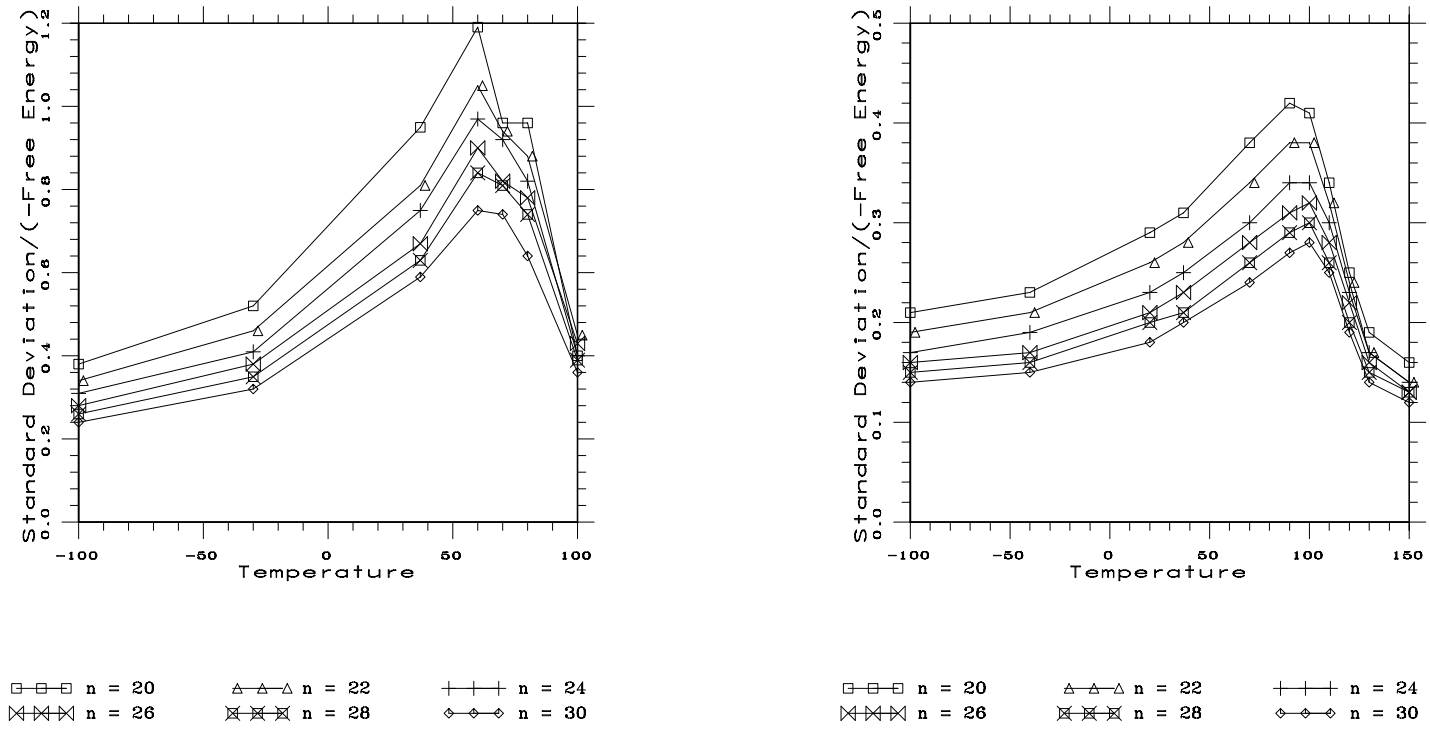

Figure 14: The relative deviation for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths between 20 and 30 in steps of 2. The **GCAU** data have been sampled at temperatures $T = -100, -30, 37, 60, 70, 80, 100°\mathrm{C}$. The **GC** data have been sampled at temperatures $T = -100, -40, 20, 37, 70, 90, 100, 110, 120, 130, 150°\mathrm{C}$.

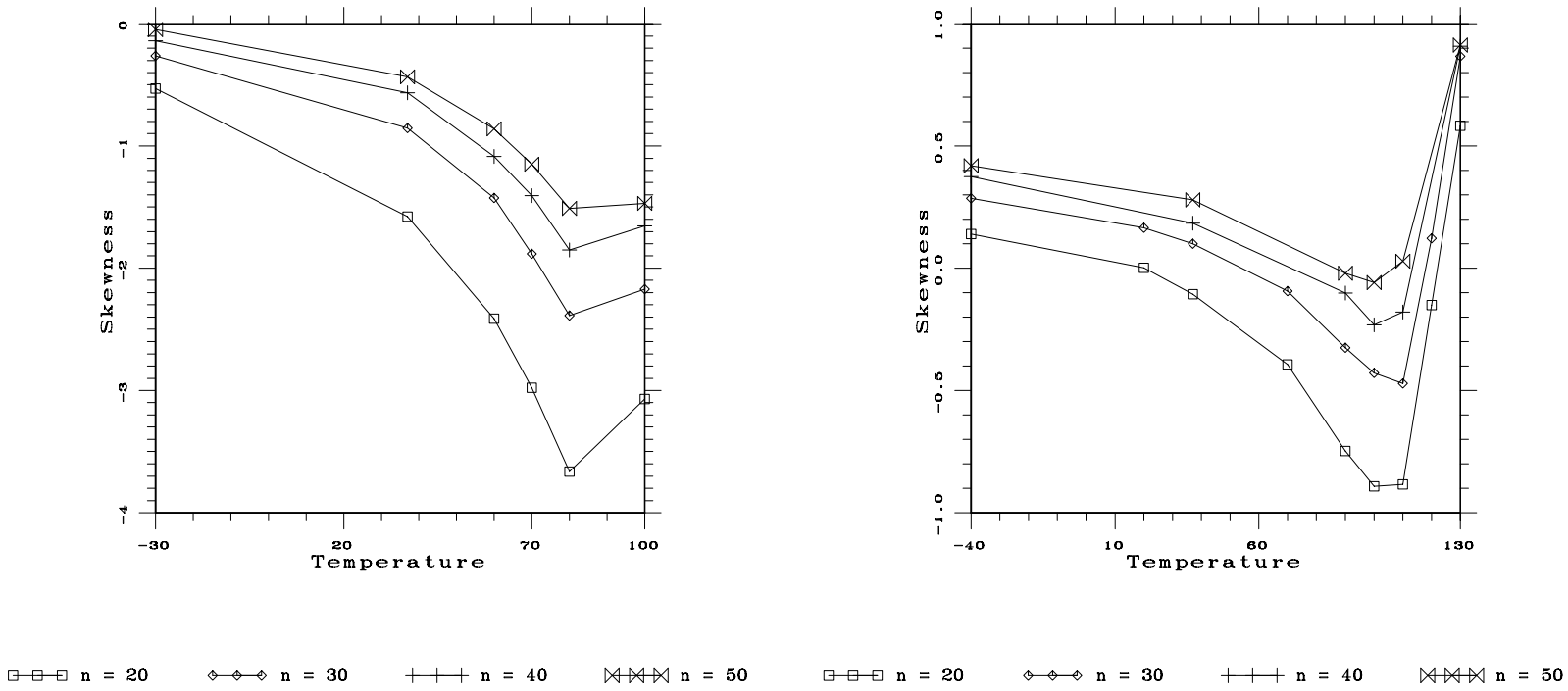

Figure 15: The skewness for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths between 20 and 50 in steps of ten. The **GCAU** data have been sampled at temperatures $T = -100, -30, 37, 60, 70, 80, 100\,^{\circ}\mathrm{C}$. The **GC** data have been sampled at temperatures $T = -100, -40, 20, 37, 70, 90, 100, 110, 120, 130, 150\,^{\circ}\mathrm{C}$.

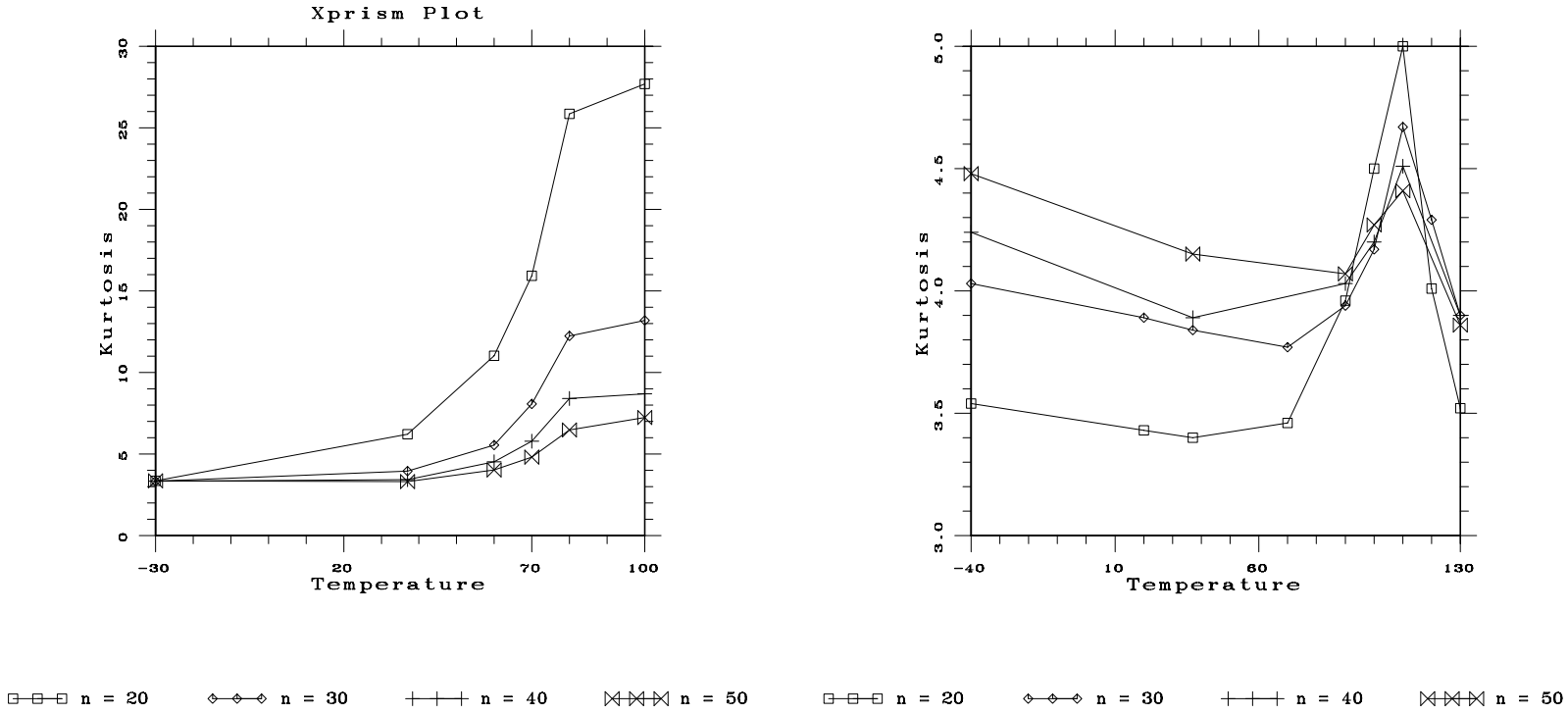

Figure 16: The kurtosis of the free energy distribution for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths between 20 and 50 in steps of ten. The **GCAU** data have been sampled at temperatures $T = -100, -30, 37, 60, 70, 80, 100^\circ$C. The **GC** data have been sampled at temperatures $T = -100, -40, 20, 37, 70, 90, 100, 110, 120, 130, 150^\circ$C.

5.2 The free energy and the structure landscape

The aim of this work was to explore landscapes, which are somehow connected to evolutionary processes. As we already pointed out in section 3, the intrinsic structure of a landscape affects very much the power of evolution as a optimization procedure. How do we get an idea of the features of a landscape on a high dimensional support? The picture of a twodimensional landscape is helpful for the general understanding, but is often misleading.

Let us assume, we want to find the global maximum of a landscape with an adaptive walk. Will it be an easier task on a low dimensional landscape, since the total number of points in the support is comparatively small, or will the adaptive walk be more successful on a high dimensional landscape, because the number of directions, in which the adaptive walk can proceed, is larger?

Given the probability density $\rho(x)$ of the value distribution of a landscape, we can derive an approximation of the probability of local optima in the limiting case of uncorrelated landscapes.

If a is the height of a point in the landscape, the probability, that at least on direct neighboring point has a height x with $x \leq a$ is

$$\int_{-\infty}^{a} \rho(x)\, dx$$

This only is only valid if the values of neighboring points are independent. Consequently, the probability, that all direct neighbors have a height smaller than a for a n-dimensional support of the landscape, is

$$\left(\int_{-\infty}^{a} \rho(x)\, dx \right)^n$$

If we define

$$G(a) := \int_{-\infty}^{a} \rho(x)\, dx$$

we get for the probability of local maxima

$$\begin{aligned} P(max) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{a} \rho(x)\, dx \right)^n \rho(a)\, da \\ &= \int_{-\infty}^{\infty} (G(a))^n\, G'(a)\, da \\ &= \frac{1}{n+1} (G(a))^{n+1} \Big|_{-\infty}^{+\infty} \\ &= \frac{1}{n+1} \end{aligned}$$

Stadler and Schnabl [27] derived the same result for low correlated landscapes with Gaussian value distribution.

In section 4.2 we introduced the autocorrelation function and the correlation length as a measure for the ruggedness of landscapes. Numerical evaluation of the autocorrelation function showed, that it can be modeled by a single decaying exponential. Landscapes with such autocorrelation are called first order autoregressive (AR(1)) and Weinberger [32] derived a number of analytical results for them under the additional assumption of Gaussian value distribution. If a landscape is AR(1) the autocorrelation function is uniquely determined by the correlation length.

In figure 17 the correlation length of the free energy landscape is plotted versus the chain length. Due to computer time limitations the data for the correlation length larger than 30, have not been sampled with the same precision as for shorter chain length. The number of points needed to ensure convergence of the sampled property increases with chain length and as the used algorithm is of cubic order in the chain length the limits of available computer time were reached. The correlation length for sequence length between 20 and 50 shows clearly a linear dependence. Also for chain length up to 50 we see a roughly linear behavior with respect to the achieved precision of the data.

It is an interesting question, whether correlation length is linear in the sequence length also for large n. If the correlation length of a landscape does not at least increase linearly with the chain length, the relative correlation length, i. e. the correlation length divided by the chain length will vanish in the limit of large n. Hence, evolution and any other optimization algorithm will fail to find the global optimum of a landscape in finite time with probability one.

From studies of the landscape of the traveling salesman problem [27] and from studies of RNA landscapes resulting from the minimal free energy algorithm [9], we know that the correlation length gives an estimate for the mean walk length from a random point in sequence space to the next local optimum. The autocorrelation function represents the loss of information if the random walk has Hamming distance h from a chosen reference point. Accordingly there is characteristic number of steps of a random walk or a characteristic Hamming distance after which practically all information is lost. Points of this characteristic distance are statistically independent. Therefore landscapes with a relative correlation length decreasing with n, divide in more and more statistically independent and finding the global optimum in finite time will become more and more difficult as we increase the length of the RNA molecules.

Figure 18 shows the temperature dependence of the correlation length of the free energy landscape for **GC** and **GCAU** sequences.

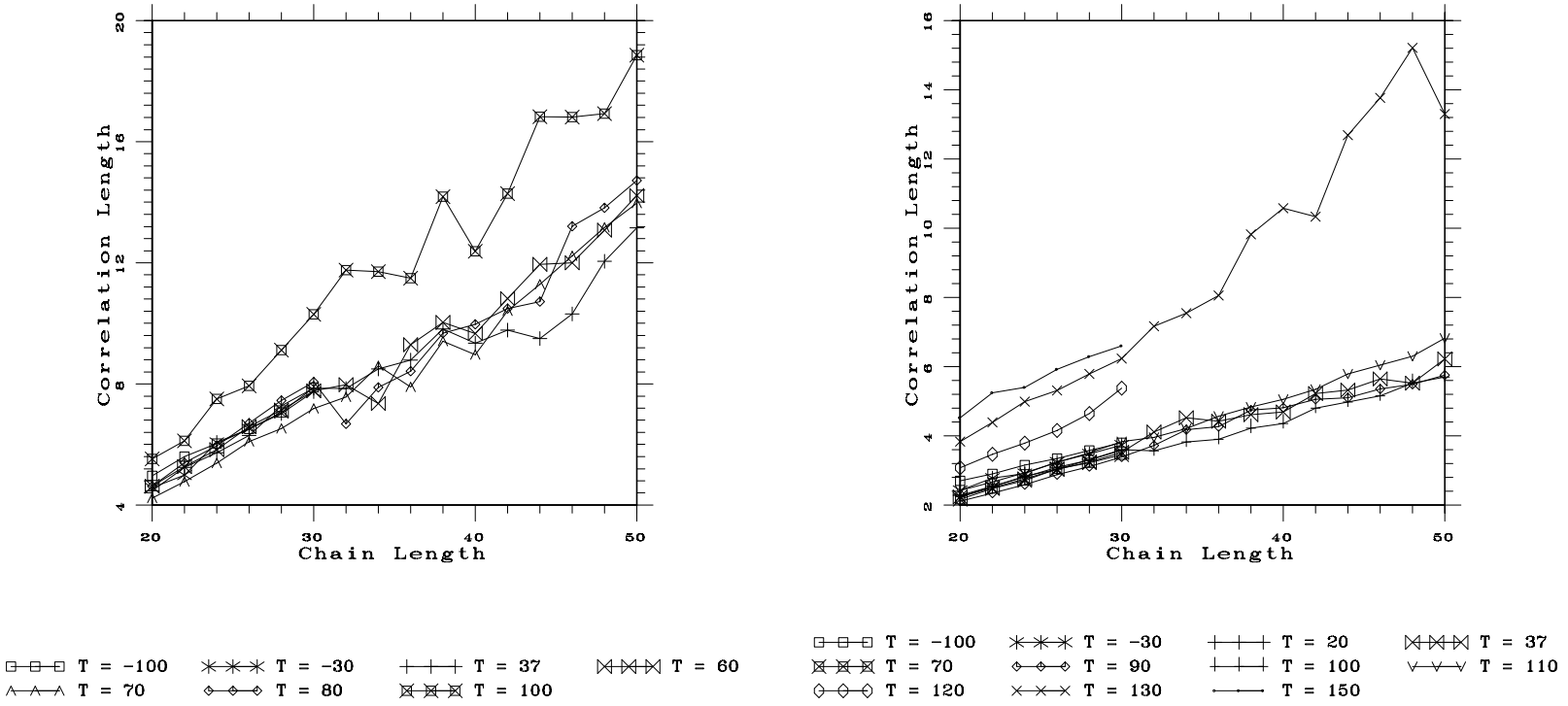

Figure 17: The correlation length for **GCAU**-sequences (left) and **GC**-sequences (right) versus chain length.

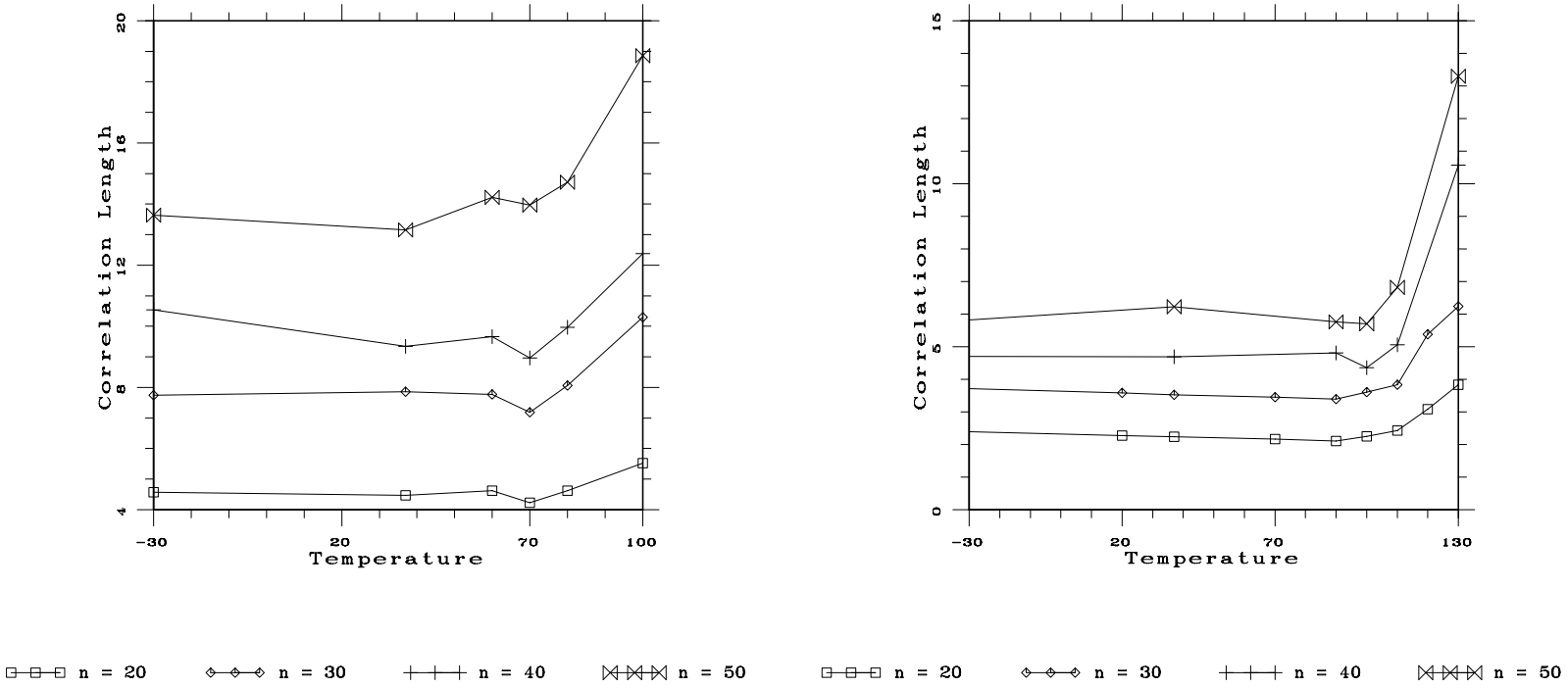

Figure 18: The correlation length for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths 20 up to 30 in steps of 2. The **GCAU** data have been sampled at temperatures $T = -100, -30, 37, 60, 70, 80, 100\,^{\circ}\mathrm{C}$. The **GC** data have been sampled at temperatures $T = -100, -40, 20, 37, 70, 90, 100, 110, 120, 130, 150\,^{\circ}\mathrm{C}$.

For both data sets the correlation length remains approximately constant up to a base set specific temperature, which can be interpreted as a mean melting temperature of the structural ensemble. The term mean reflects here, that we take the average over all sampled sequences and does not refer to all possible configurations of a specific sequence, which are already included in the partition function. We cannot claim to predict the experimentally determined melting temperature for RNA sequences, because of the assumptions which have been made in order to extrapolate a temperature dependence of the biochemical data (see end of section 2.3), but we seem to observe a general behavior of the correlation length with temperature. Beyond the mean melting temperature we see in both plots a dramatic increase.

The constant behavior of the correlation length temperatures below the melting point indicates, that the ensemble of possible secondary structures of a given RNA molecule is mainly determined by a single structure with minimal free energy, which does not change significantly with temperature. This means, that the probability for a RNA molecule to fold in the most stable structure is comparatively high and all other structures with not negligible probability of occurrence are closely related to the likeliest structure. Only at the melting point, when many stems open, we see a the correlation length becomes longer. Clearly the fewer bases pair the longer is the correlation length, because a mutation of an unpaired base will normally have a smaller effect on the change of the secondary structure, than a mutation of a base in a stack.

The values for the correlation length of free energy landscapes resulting from the Zuker algorithm lie significantly below those calculated with the partition function algorithm. For **GCAU** landscapes the values range from 70% to 80% of the partition function correlation length, whereas the values for **GC** landscapes are between 75% and 85% of the partition function correlation length. We compared the correlation length for landscapes corresponding to the minimal free energy algorithm with our data computed at 37°C. As we already pointed out, the minimal free energy algorithm determines the most stable secondary structure at 37° C, not the ground state of the RNA molecule.

Figure 19 shows the temperature dependence of the correlation length of the structural ensemble distance landscape. We see an increase of correlation length up to the mean melting temperature. Above this characteristic temperature the correlation length decreases again because of the vanishing variance at high temperatures.

In figure 20 the mean base pairing probability is plotted versus temperature for the **GCAU** sequences as well as for the **GC**-only sequences. For small chain lengths the mean base pairing probability depends on the chain length, because at least three bases must be unpaired in a region where the molecule folds back upon itself.

The next series of figures (21,22,23) shows the probability density surfaces of the free energy landscape and the ensemble structure landscape for the **GCAU** Data at three different temperatures. In these figures we plot the number of occurrence for two sequences with a given Hamming distance h and a given free energy or structure distance d. Increasing temperature obviously shifts the conditional probability

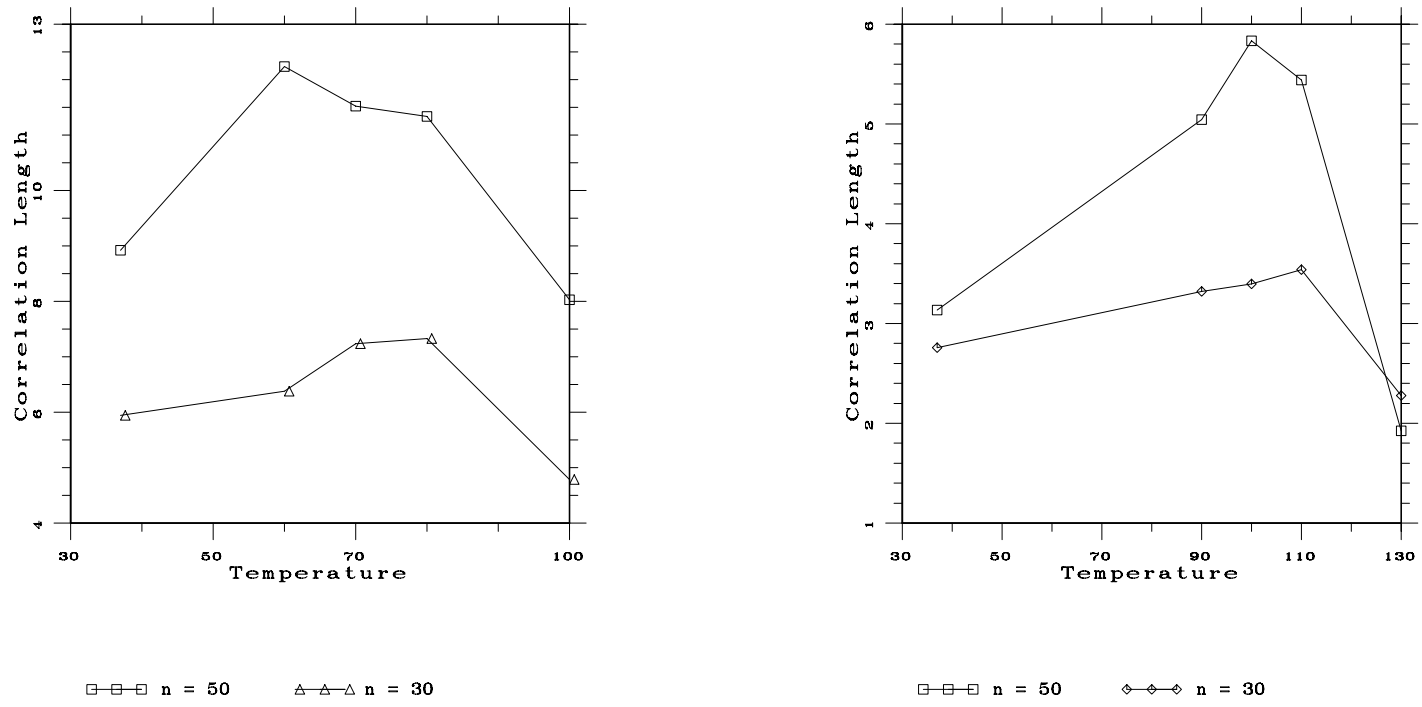

Figure 19: The correlation length of the structural ensemble distance landscape for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths $n = 30$ and $n = 50$.

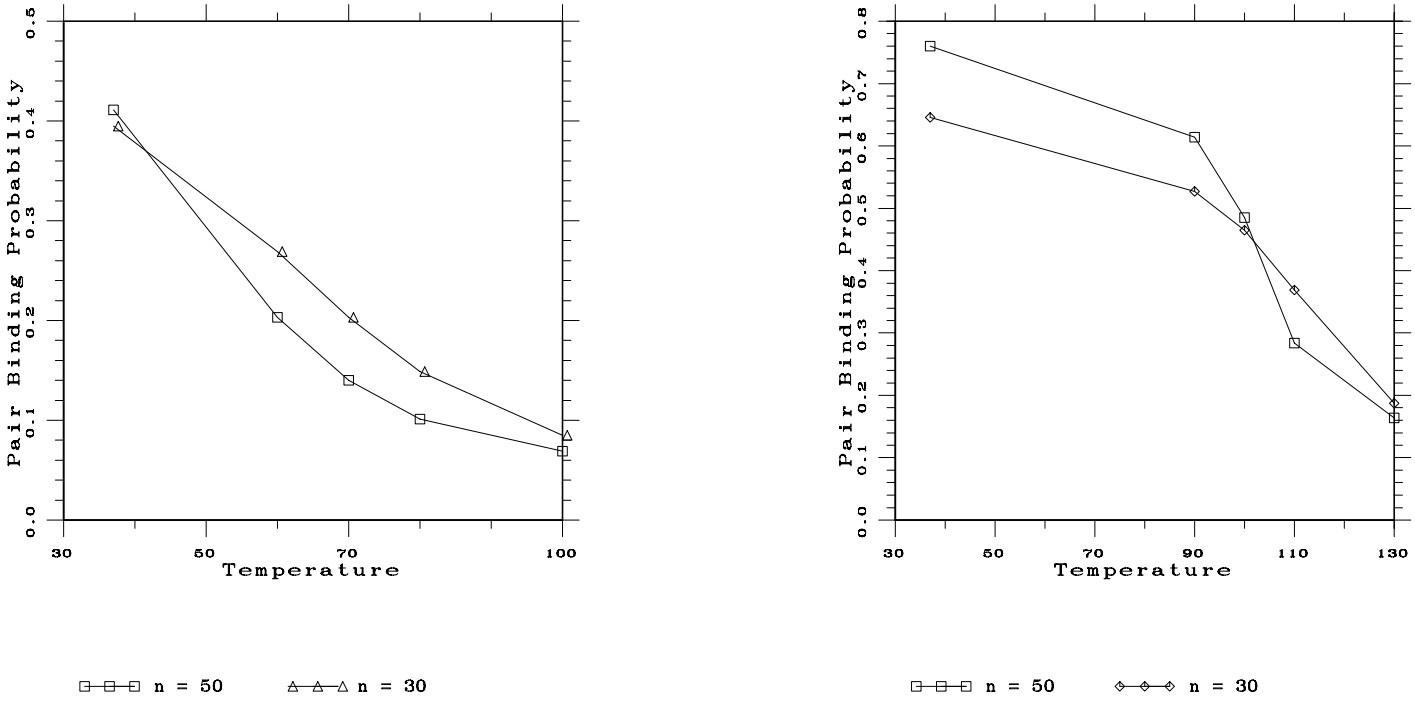

Figure 20: The mean base pairing probability for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths $n = 30$ and $n = 50$.

distribution towards vanishing structure or free energy distance. The probability surfaces at low temperature are qualitatively indistinguishable from those derived from the minimal free energy folding algorithm [10].

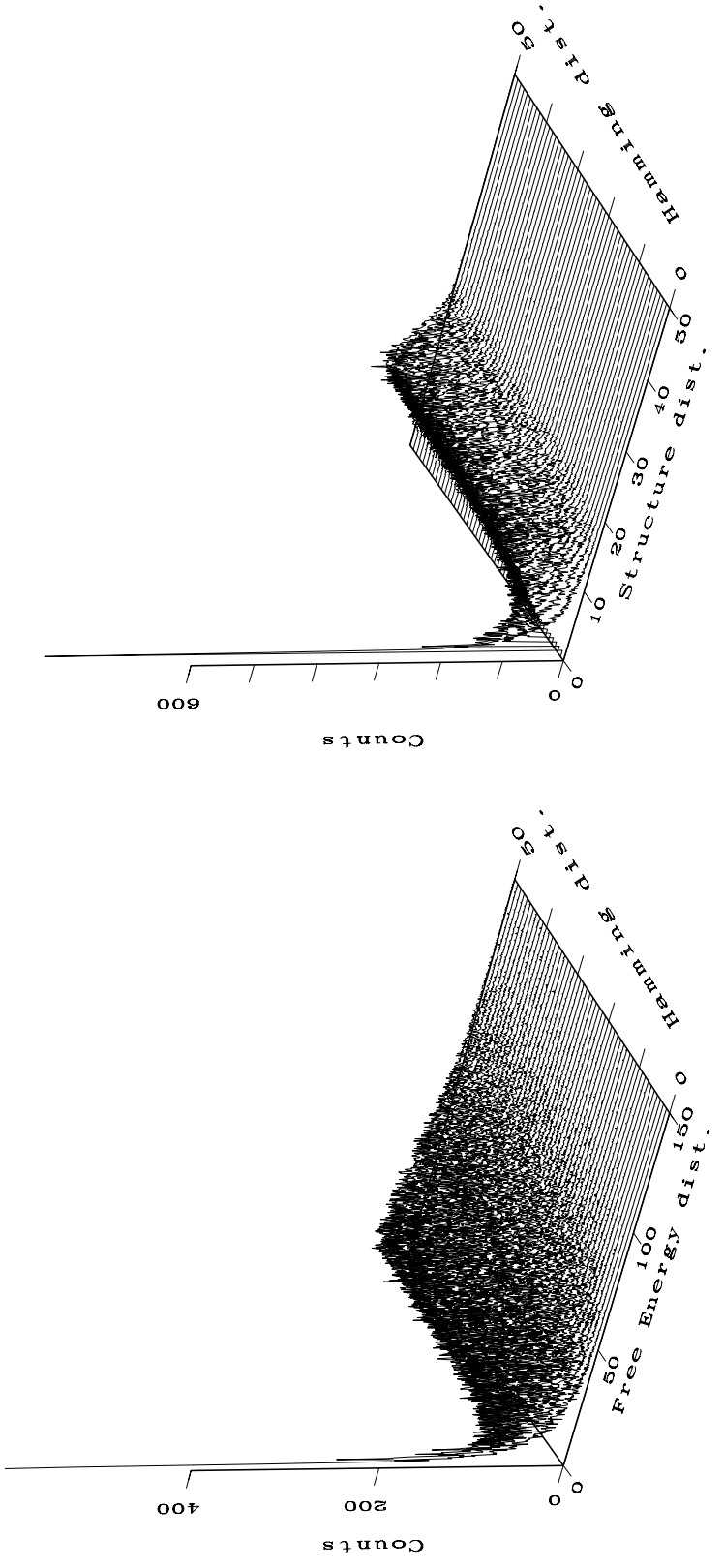

Figure 21: The probability density surfaces for the free energy landscape (left) and the ensemble structure landscape (right). The data are obtained for the **GCAU** alphabet at temperature $T = 37^{\circ}$ C.

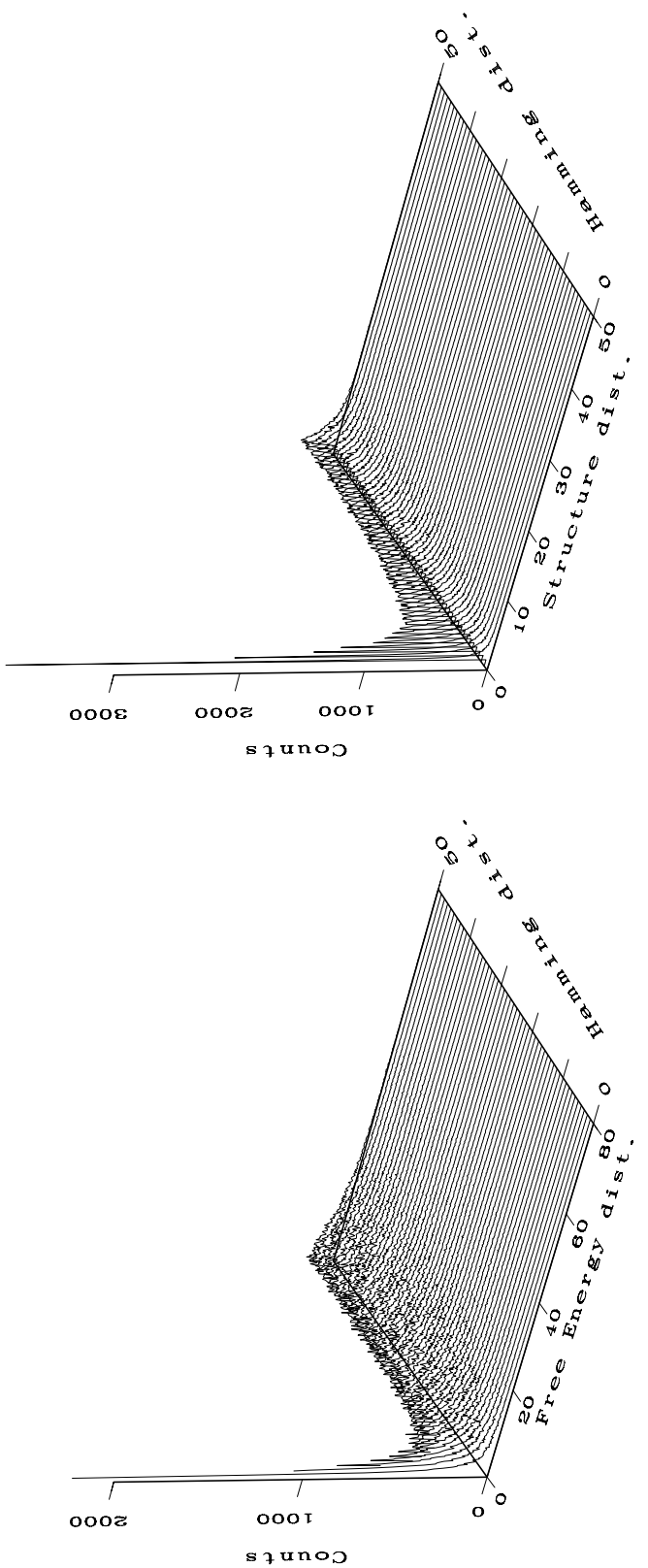

Figure 22: The mean base pairing The probability density surfaces for the free energy landscape (left) and the ensemble structure landscape (right). The data are obtained for the **GCAU** alphabet at temperature $T = 70°$ C. probability for **GCAU**-sequences (left) and **GC**-sequences (right) versus temperature for chain lengths $n = 30$ and $n = 50$.

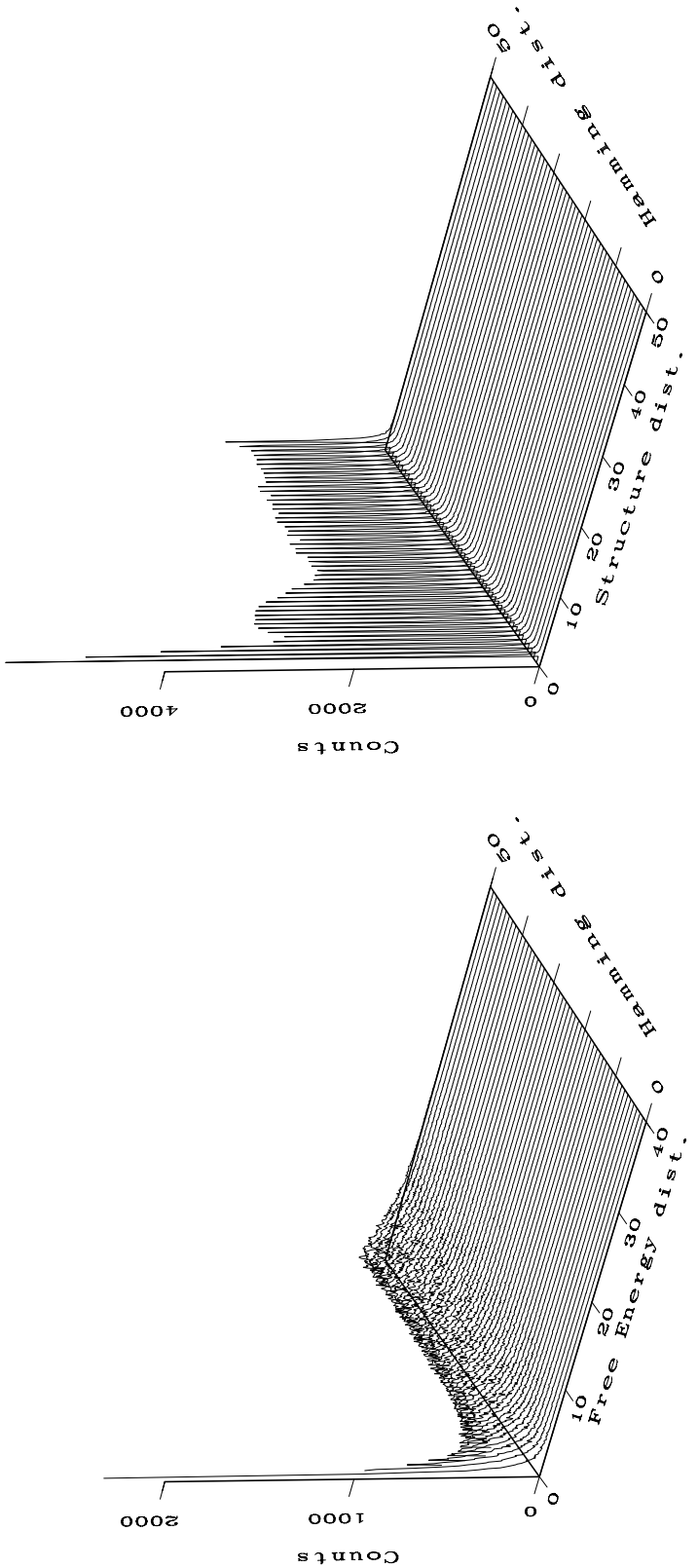

Figure 23: The probability density surfaces for the free energy landscape (left) and the ensemble structure landscape (right). The data are obtained for the **GCAU** alphabet at temperature $T = 100^{\circ}$ C.

6 Conclusion and outlook

Two different kinds of RNA landscapes, the free energy landscape and the landscape of the structural ensembles, have been analyzed in detail. A new algorithm [21], which allows the calculation of the partition function of RNA molecules, has been used to compute the value landscapes. From the partition function we derive the free energy as well as the equilibrium probabilities for the formation of all possible base pairs. In contrast to the conventional minimal free energy folding algorithms, which have previously been applied to the computation of similar landscapes, the partition function algorithm introduces temperature as a parameter into the calculation. While minimal free energy folding algorithms yield the most stable secondary structure, the partition function algorithm takes into account all possible secondary structures weighted according to the Boltzmann distribution. Consequently landscapes based on the partition function algorithm should have closer resemblance to reality, because the partition function contains important information on the structural variability. Hence, the major goal of this work was to determine, whether there is a qualitative difference between both landscapes.

The most appropriate quantity for a description of landscapes turned out to be the correlation length, which is a measure accounting for ruggedness as well an estimate for the number of local optima in the landscape.

The aim of this work then was to study the influence of temperature onto the correlation length for landscapes of short RNA molecules. Due to the enormous amount of computer time required for such simulations we investigated only landscapes corresponding to RNA molecules with chain lengths between 20 and 50 nucleotides. It is shown that the correlation length is constant below a certain characteristic temperature, which can be interpreted as an average melting temperature of the sampled RNA sequences.

We also examined the dependence of the correlation length on chain length for free energy landscapes and detected an approximately linear behavior. Although this was only shown for short RNA, it is nevertheless an interesting result, since any slower increase than the linear dependence leads in the limit of large sequence lengths to very complex landscapes. Indeed, they become so complex, that no optimization algorithm, which has a priori no information on the landscape, could ever find their global optima.

Studies of free energy landscapes [9] using the minimal free energy folding algorithms have showed, that landscapes of RNA molecules consisting of only two complementary bases are more rugged than those for real RNA consisting of four bases. This fact was also verified for partition function landscapes. Although no salient deviation from the minimal free energy results was detected, the absolute values of the correlation lengths are significantly higher for the more realistic partition function landscapes. This increase of correlation length indicates, that real RNA landscapes are less rugged, than might be expected from simulations using to the minimal free energy folding. The present work is summarized in Bonhoeffer et al. (1992) [1].

Many questions remain to be answered. Taking into consideration that most self replicating RNA molecules consist of a few hundred bases it would be interesting to expand our calculations to longer chain lengths. Unfortunately this is currently out of reach since the required amount of computation time grows proportional to n^3 (n being the sequence length). The use of the free energy algorithm, however, makes calculations up to some hundred nucleotides possible, since this algorithm is computationally less costly Therefore investigations of free energy landscapes basing on this algorithm have been performed up to chain length up to hundred bases. In these studies [9] no deviation from the linear behavior of the correlation length was detected in the range of computed sequence lengths.

It would be worthwhile, to investigate the relationship between the correlation length of a landscape and the time needed for an evolutionary algorithm to find a satisfactory solution. So far no simulations of evolutionary models have been carried out on the partition function landscapes.

The remarkable difference between the correlation length of landscapes of **GC** and **GCAU** sequences, guided our interest of our group to another point. Obviously, the introduction of a second base pair with different stacking energy results in a smoother landscape. We now want to find out for which percentage of **A**,**U**,**G** and **C** the correlation length of the corresponding landscape is maximal.

A different approach modeling the melting kinetics of RNA is currently undertaken in our group. Equilibrium constants for the melting process of RNA molecules are computed in order to derive more realistic fitness landscapes. Models of fitness landscapes do not only require a function for the evaluation of the phenotype in a given surrounding, they also need a reliable prediction of phenotype from genotype. In general the prediction of the phenotype from the genotype is still far out of reach. The only exception is the prediction of the secondary structure of RNA molecules. Here, the folding algorithms represent a fairly reliable prediction of the phenotype from the mere sequence. Hence, a model for the evaluation of the selfreproducing capability of RNA secondary structures would allow the computation of realistic fitness landscapes.

Answers to these questions will hopefully allow to gain further understanding of prebiotic evolution.

A Computer time requirements

workstation	desired result	CPU time	chain length
Sun Sparc Station 2	free energy	0.5 s	30
		2 s	50
	probability	1 s	30
		4.5 s	50
Sun Sparc Station SLC	free energy	1 s	30
		4.2 s	50
	probability	1.9 s	30
		9.1 s	50

Table 1: Computer time usage

The computations were carried out mainly on two different workstations. The calculation of the partition function, from which the free energy is derived, is of cubic order in the sequence length. As we already mentioned in section 2.3 the calculation of the base binding probability matrix from the full partition function requires an additional backtracking algorithm of cubic order. For this reason the computer time is approximately doubled for the calculation of the base binding probability matrix. The time given in the table corresponds to a single execution of the program, i. e. to find the result for a single sequence.

B Data tables

$n\backslash T$	-100	-30	37	60	70	80	100
50		-287	-89	-43	-29	-21	-14
48		-272	-84	-40	-28	-20	-14
46		-257	-79	-38	-26	-19	-13
44		-241	-74	-35	-24	-17	-12
42		-226	-69	-33	-23	-16	-11
40		-211	-64	-30	-21	-15	-10
38		-196	-59	-28	-19	-14	-10
36		-181	-54	-26	-18	-13	-9
34		-166	-49	-23	-16	-12	-9
32		-151	-44	-21	-15	-11	-8
30	-261	-138	-39	-18	-13	-10	-7
28	-235	-123	-34	-15	-11	-8	-6
26	-209	-108	-30	-13	-10	-7	-5
24	-184	-94	-25	-11	-8	-6	-5
22	-158	-80	-21	-9	-7	-5	-4
20	-133	-66	-16	-7	-6	-4	-4

Table 2: The mean free energy of **GCAU**-sequences in 0.1 kcal. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-40	20	37	70	90	100	110	120	130	150
50		-702		-340		-119	-86	-60		-35	
48		-667		-323		-113	-82	-57		-34	
46		-633		-306		-106	-77	-53		-32	
44		-599		-288		-100	-72	-50		-30	
42		-564		-271		-93	-68	-47		-28	
40		-530		-254		-87	-63	-44		-26	
38		-496		-237		-81	-58	-41		-25	
36		-462		-219		-74	-54	-38		-23	
34		-428		-203		-68	-49	-34		-21	
32		-394		-186		-62	-45	-31		-19	
30	-517	-362	-210	-169	-95	-56	-40	-28	-21	-18	-17
28	-470	-328	-190	-153	-84	-50	-35	-25	-19	-16	-15
26	-423	-295	-169	-136	-75	-44	-31	-22	-16	-14	-14
24	-377	-261	-149	-119	-65	-38	-27	-19	-14	-12	-12
22	-330	-228	-129	-103	-55	-32	-22	-16	-12	-11	-11
20	-284	-195	-109	-87	-46	-26	-18	-13	-10	-9	-9

Table 3: The mean free energy of **GC**-sequences in 0.1 kcal. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-30	37	60	70	80	100
50		57.98	32.31	20.51	15.10	10.11	3.89
48		56.80	31.44	19.93	14.57	9.81	3.78
46		55.60	30.60	19.27	14.11	9.52	3.65
44		54.31	29.62	18.65	13.60	9.20	3.50
42		52.99	28.67	18.04	13.08	8.87	3.36
40		51.75	27.86	17.34	12.55	8.50	3.23
38		50.24	26.93	16.62	11.98	8.13	3.09
36		48.80	26.01	15.91	11.41	7.73	2.94
34		47.20	24.92	15.22	10.82	7.33	2.79
32		45.60	23.77	14.37	10.14	6.89	2.65
30	63.49	44.54	22.85	13.55	9.58	6.40	2.52
28	61.30	42.67	21.55	12.65	8.89	5.93	2.35
26	58.84	40.76	20.16	11.68	8.17	5.46	2.17
24	56.35	38.80	18.64	10.62	7.37	4.92	1.98
22	53.52	36.60	17.02	9.49	6.56	4.39	1.79
20	50.11	34.28	15.22	8.31	5.73	3.84	1.58

Table 4: The standard deviation of the free energy **GCAU**-sequences in 0.1 kcal. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-40	20	37	70	90	100	110	120	130	150
50		65.41		41.00		19.82	14.94	10.03		3.86	
48		64.64		40.21		19.47	14.64	9.80		3.75	
46		63.86		39.54		19.08	14.30	9.56		3.64	
44		63.07		38.76		18.71	13.97	9.32		3.53	
42		61.94		38.09		18.32	13.63	9.06		3.41	
40		60.78		37.40		17.83	13.29	8.78		3.38	
38		59.77		36.69		17.34	12.90	8.48		3.14	
36		58.85		35.97		16.81	12.43	8.14		3.01	
34		57.87		35.05		16.31	12.00	7.80		2.88	
32		56.25		34.18		15.71	11.55	7.43		2.74	
30	70.26	54.52	38.46	33.35	22.52	15.17	11.08	7.04	3.97	2.60	2.11
28	68.29	52.89	37.28	32.29	21.72	14.51	10.51	6.62	3.72	2.44	1.98
26	66.31	51.44	35.97	31.11	20.84	13.75	9.87	6.15	3.45	2.27	1.86
24	63.97	49.44	34.52	29.79	19.81	12.91	9.13	5.63	3.15	2.09	1.70
22	61.29	47.20	33.05	28.44	18.68	12.01	8.32	5.08	2.83	1.90	1.56
20	58.30	45.08	31.38	26.92	17.38	10.96	7.39	4.46	2.50	1.70	1.40

Table 5: The standard deviation of the free energy of **GC**-sequences in 0.1 kcal. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-30	37	60	70	80	100
50			0.36	0.48	0.52	0.48	0.28
48			0.37	0.50	0.52	0.49	0.27
46			0.39	0.51	0.54	0.50	0.28
44			0.40	0.53	0.57	0.54	0.29
42			0.42	0.55	0.57	0.55	0.31
40			0.44	0.58	0.60	0.57	0.32
38			0.46	0.59	0.63	0.58	0.31
36			0.48	0.61	0.63	0.59	0.33
34			0.51	0.66	0.68	0.61	0.31
32			0.54	0.68	0.68	0.63	0.33
30	0.24	0.32	0.59	0.75	0.74	0.64	0.36
28	0.26	0.35	0.63	0.84	0.81	0.74	0.39
26	0.28	0.38	0.67	0.90	0.82	0.78	0.43
24	0.31	0.41	0.75	0.97	0.92	0.82	0.40
22	0.34	0.46	0.81	1.05	0.94	0.88	0.45
20	0.38	0.52	0.95	1.19	0.96	0.96	0.40

Table 6: The standard deviation divided by the negative free energy **GCAU**-sequences. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-40	20	37	70	90	100	110	120	130	150
50				0.12		0.17	0.17	0.17		0.11	
48				0.12		0.17	0.18	0.17		0.11	
46				0.13		0.18	0.19	0.18		0.11	
44				0.13		0.19	0.19	0.19		0.12	
42				0.14		0.20	0.20	0.19		0.12	
40				0.15		0.20	0.21	0.20		0.13	
38				0.15		0.21	0.22	0.21		0.13	
36				0.16		0.23	0.23	0.21		0.13	
34				0.17		0.24	0.24	0.23		0.14	
32				0.18		0.25	0.26	0.24		0.14	
30	0.14	0.15	0.18	0.20	0.24	0.27	0.28	0.25	0.19	0.14	0.12
28	0.15	0.16	0.20	0.21	0.26	0.29	0.30	0.26	0.20	0.15	0.13
26	0.16	0.17	0.21	0.23	0.28	0.31	0.32	0.28	0.22	0.16	0.13
24	0.17	0.19	0.23	0.25	0.30	0.34	0.34	0.30	0.23	0.17	0.14
22	0.19	0.21	0.26	0.28	0.34	0.38	0.38	0.32	0.24	0.17	0.14
20	0.21	0.23	0.29	0.31	0.38	0.42	0.41	0.34	0.25	0.19	0.16

Table 7: The standard deviation divided by the negative free energy of **GC**-sequences. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-30	37	60	70	80	100
50		-0.049	-0.434	-0.860	-1.150	-1.510	-1.470
48		-0.050	-0.454	-0.903	-1.191	-1.565	-1.501
46		-0.070	-0.481	-0.924	-1.244	-1.648	-1.540
44		-0.096	-0.592	-0.961	-1.280	-1.710	-1.544
42		-0.116	-0.528	-1.022	-1.328	-1.781	-1.575
40		-0.139	-0.566	-1.085	-1.407	-1.851	-1.654
38		-0.154	-0.601	-1.138	-1.480	-1.961	-1.727
36		-0.181	-0.647	-1.215	-1.554	-2.058	-1.806
34		-0.206	-0.697	-1.303	-1.640	-2.150	-1.897
32		-0.223	-0.754	-1.407	-1.730	-2.234	-2.036
30	-0.100	-0.263	-0.854	-1.426	-1.883	-2.388	-2.171
28	-0.115	-0.285	-0.929	-1.556	-2.035	-2.543	-2.287
26	-0.147	-0.337	-1.030	-1.691	-2.200	-2.710	-2.426
24	-0.173	-0.378	-1.153	-1.866	-2.415	-2.890	-2.608
22	-0.211	-0.451	-1.340	-2.010	-2.678	-3.245	-2.836
20	-0.252	-0.531	-1.577	-2.412	-2.978	-3.663	-3.070

Table 8: The skewness of the free energy distribution for **GCAU**-sequences. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-40	20	37	70	90	100	110	120	130	150
50		0.419		0.279		-0.020	-0.059	0.028		0.912	
48		0.405		0.258		-0.035	-0.091	-0.007		0.911	
46		0.375		0.234		-0.043	-0.115	-0.039		0.906	
44		0.372		0.211		-0.051	-0.152	-0.073		0.910	
42		0.383		0.201		-0.078	-0.196	-0.129		0.909	
40		0.375		0.184		-0.101	-0.231	-0.180		0.908	
38		0.368		0.166		-0.129	-0.258	-0.218		0.899	
36		0.358		0.160		-0.162	-0.305	-0.234		0.891	
34		0.321		0.137		-0.200	-0.353	-0.274		0.873	
32		0.312		0.118		-0.238	-0.406	-0.338		0.852	
30	0.372	0.285	0.165	0.100	-0.094	-0.325	-0.429	-0.472	0.123	0.866	0.794
28	0.352	0.264	0.146	0.076	-0.127	-0.378	-0.498	-0.539	0.064	0.824	0.789
26	0.311	0.244	0.118	0.046	-0.172	-0.441	-0.585	-0.611	0.064	0.781	0.776
24	0.283	0.213	0.075	0.001	-0.232	-0.523	-0.672	-0.700	-0.034	0.725	0.752
22	0.254	0.183	0.025	-0.052	-0.305	-0.630	-0.778	-0.800	-0.085	0.659	0.744
20	0.229	0.139	0.002	-0.107	-0.394	-0.748	-0.892	-0.884	-0.151	0.583	0.712

Table 9: The skewness of the free energy distribution for **GC**-sequences. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-30	37	60	70	80	100
50		3.35	3.30	4.03	4.81	6.47	7.23
48		3.32	3.32	4.13	4.97	6.74	7.47
46		3.29	3.35	4.11	5.18	7.22	7.80
44		3.29	3.36	4.17	5.29	7.60	7.85
42		3.34	3.38	4.33	5.45	8.02	8.08
40		3.33	3.44	4.53	5.78	8.41	8.69
38		3.30	3.47	4.67	6.10	9.14	9.28
36		3.28	3.54	4.93	6.35	9.78	9.97
34		3.29	3.63	5.24	6.76	10.34	10.90
32		3.30	3.74	5.71	7.22	10.81	12.30
30	3.36	3.33	3.96	5.54	8.07	12.23	13.20
28	3.32	3.31	4.08	6.11	8.98	13.49	14.59
26	3.31	3.31	4.30	6.63	9.98	14.64	16.45
24	3.30	3.29	4.61	7.51	11.53	16.33	19.04
22	3.29	3.32	5.27	8.90	13.56	20.46	23.01
20	3.26	3.36	6.21	11.03	15.94	25.86	27.69

Table 10: The kurtosis of the free energy distribution for **GCAU**-sequences. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-40	20	37	70	90	100	110	120	130	150
50		4.48		4.15		4.07	4.27	4.41		3.86	
48		4.44		4.17		4.06	4.26	4.42		3.87	
46		4.36		4.12		4.04	4.27	4.43		3.85	
44		4.29		4.02		4.03	4.25	4.40		3.87	
42		4.27		3.94		4.04	4.26	4.50		3.87	
40		4.24		3.89		4.03	4.20	4.51		3.90	
38		4.23		3.89		3.99	4.15	4.50		3.91	
36		4.28		3.90		3.98	4.17	4.41		3.91	
34		4.22		3.88		3.94	4.13	4.41		3.90	
32		4.12		3.86		3.91	4.12	4.44		3.89	
30	4.18	4.03	3.89	3.84	3.77	3.94	4.17	4.67	4.29	3.90	3.41
28	4.11	3.96	3.80	3.75	3.71	3.92	4.19	4.71	4.28	3.83	3.39
26	4.01	3.89	3.72	3.68	3.65	3.88	4.28	4.73	4.23	3.77	3.36
24	3.88	3.77	3.63	3.60	3.61	3.88	4.33	4.80	4.13	3.69	3.31
22	3.77	3.69	3.54	3.51	3.54	3.94	4.41	4.93	4.05	3.63	3.31
20	3.68	3.54	3.43	3.40	3.46	3.96	4.50	5.00	4.01	3.52	3.27

Table 11: The kurtosis of the free energy distribution for **GC**-sequences. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-30	37	60	70	80	100
50		13.64	13.16	14.22	13.97	14.72	18.86
48		13.22	12.06	13.08	13.16	13.82	16.93
46		12.12	10.32	12.00	12.23	13.21	16.81
44		11.69	9.50	11.96	11.29	10.71	16.83
42		11.43	9.78	10.82	10.41	10.48	14.29
40		10.54	9.35	9.66	8.97	9.97	12.38
38		9.70	9.82	10.03	9.41	9.67	14.18
36		9.23	8.79	9.29	7.90	8.41	11.49
34		8.75	8.50	7.35	8.59	7.89	11.71
32		7.46	7.84	7.97	7.57	6.68	11.76
30	7.88	7.75	7.86	7.77	7.19	8.07	10.30
28	7.06	7.03	7.28	7.14	6.53	7.45	9.117
26	6.52	6.52	6.30	6.59	6.10	6.70	7.93
24	6.03	6.07	5.73	5.83	5.38	5.96	7.51
22	5.60	4.99	5.24	5.27	4.78	5.40	6.12
20	4.96	4.57	4.47	4.62	4.23	4.62	5.52

Table 12: The correlation length of the free energy landscape of **GCAU**-sequences. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

$n\backslash T$	-100	-40	20	37	70	90	100	110	120	130	150
50		5.76		6.22		5.76	5.70	6.83		13.30	
48		5.84		5.52		5.49	5.52	6.30		15.21	
46		5.78		5.65		5.36	5.16	6.05		13.77	
44		5.24		5.32		5.10	4.98	5.79		12.69	
42		4.98		5.23		5.06	4.80	5.35		10.33	
40		4.70		4.68		4.81	4.36	5.06		10.57	
38		4.50		4.61		4.75	4.22	4.84		9.82	
36		4.48		4.42		4.27	3.90	4.56		8.05	
34		4.04		4.53		4.18	3.82	4.22		7.55	
32		3.84		4.11		3.72	3.57	3.96		7.17	
30	3.80	3.73	3.59	3.52	3.46	3.39	3.60	3.83	5.38	6.24	6.60
28	3.58	3.47	3.30	3.24	3.22	3.11	3.31	3.51	4.65	5.79	6.29
26	3.34	3.24	3.09	3.03	3.04	2.88	3.04	3.24	4.16	5.31	5.92
24	3.15	2.91	2.81	2.76	2.71	2.59	2.84	2.94	3.79	4.99	5.40
22	2.90	2.78	2.55	2.50	2.48	2.35	2.54	2.67	3.47	4.39	5.25
20	2.70	2.42	2.28	2.23	2.17	2.11	2.24	2.42	3.08	3.83	4.52

Table 13: The correlation length of the free energy landscape of **GC**-sequences. The chain length n ranges from 20 to 50 in steps of 2. The temperature T is measured in degree Celsius.

chain length $n = 30$			
T	l_e	l_s	$\overline{p}$
-30	7.21	4.21	0.54
37	5.38	5.94	0.40
60	5.73	6.38	0.27
70	7.09	7.24	0.20
80	7.79	7.33	0.15
100	8.71	4.78	0.08
chain length $n = 50$			
T	l_e	l_s	$\overline{p}$
37	13.34	8.92	0.41
60	12.63	12.23	0.20
70	11.90	11.52	0.14
80	15.13	11.33	0.10
100	17.96	8.03	0.07

Table 14: The correlation length of the free energy (l_e) and the structure (l_s) landscape for **GCAU** sequences at different temperatures. The third column shows the mean binding probability at a given temperature.

chain length $n = 30$			
T	l_e	l_s	$\overline{p}$
-40	3.6790	2.2855	0.681
37	3.6371	2.7582	0.646
90	3.76	3.32	0.53
100	3.56	3.40	0.47
110	3.87	3.54	0.37
130	6.43	2.28	0.19
chain length $n = 50$			
T	l_e	l_s	$\overline{p}$
37	5.75	3.13	0.76
90	5.57	5.05	0.61
100	5.49	5.83	0.49
110	5.57	5.44	0.28
130	12.50	1.92	0.16

Table 15: The correlation length of the free energy (l_e) and the structure (l_s) landscape for **GC** sequences at different temperatures. The third column shows the mean binding probability at a given temperature.

C The experimental data

Enthalpies for stacked pairs (0.1 kcal)						
5′/3′	CG	GC	GU	UG	AU	UA
CG	-122.0	-80.0	-77.0	-77.0	-105.0	-76.0
GC	-142.0	-122.0	-75.0	-81.0	-133.0	-102.0
GU	-81.0	-77.0	-67.0	-67.0	-67.0	-69.0
UG	-75.0	-77.0	-68.0	-67.0	-69.0	-67.0
AU	-102.0	-76.0	-67.0	-69.0	-66.0	-57.0
UA	-133.0	-105.0	-69.0	-67.0	-81.0	-66.0

Entropies for stacked pairs (0.1 kcal)						
5′/3′	CG	GC	GU	UG	AU	UA
CG	-29.7	-19.4	-20.0	-20.0	-27.8	-19.2
GC	-34.9	-29.7	-20.0	-20.0	-35.5	-26.2
GU	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0
UG	-20.0	-20.0	-20.0	-20.0	-20.0	-20.0
AU	-26.2	-19.2	-20.0	-20.0	-18.4	-15.5
UA	-35.5	-27.8	-20.0	-20.0	-22.6	-18.4

Free energies for loops[2](0.1 kcal):

hairpin AU : 9999, 9999, 45, 55, 49, 51, 52, 55, 58, 59, 60, 61, 62, 63, 64, 64, 65, 65, 66, 67, 67, 68, 68, 69, 69, 69, 70, 70, 71, 71

hairpin GC : 9999, 9999, 45, 55, 49, 51, 52, 55, 58, 59, 60, 61, 62, 63, 64, 64, 65, 65, 66, 67, 67, 68, 68, 69, 69, 69, 70, 70, 71, 71

bulge : 39, 31, 35, 42, 48, 50, 52, 53, 54, 55, 57, 57, 58, 59, 60, 61, 61, 62, 62, 63, 63, 64, 64, 65, 65, 65, 66, 67, 67, 67

internal loop AU&AU : 9999, 41, 45, 49, 53, 57, 59, 60, 61, 63, 64, 64, 65, 66, 67, 68, 68, 69, 69, 70, 71, 71, 71, 72, 72, 73, 73, 74, 74, 74

internal loop GC&GC : 9999, 41, 45, 49, 53, 57, 59, 60, 61, 63, 64, 64, 65, 66, 67, 68, 68, 69, 69, 70, 71, 71, 71, 72, 72, 73, 73, 74, 74, 74

internal loop AU&GC : 9999, 41, 45, 49, 53, 57, 59, 60, 61, 63, 64, 64, 65, 66, 67, 68, 68, 69, 69, 70, 71, 71, 71, 72, 72, 73, 73, 74, 74, 74

[2] with length $1, \ldots, 30$

Mismatch energies (0.1 kcal)																
$5'/3'$	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA				-10			-11			-19		-15	-8		-8	
AC				-7			-11			-10		-9	-7		-7	
AG				-11			-16			-19		-15	-8		-8	
AU	-8	-10	-10	-10	-7	-7	-7	-7	-8	-10	-10	-10	-8	-8	-8	-8
CA				-8			-13			-20		-14	-10		-10	
CC				-6			-6			-11		-9	-7		-7	
CG	-19	-20	-19	-19	-10	-11	-10	-8	-19	-19	-19	-19	-14	-15	-14	-12
CU				-6			-8			-15		-11	-8		-8	
GA				-11			-13			-19		-15	-10		-10	
GC	-11	-13	-13	-13	-11	-6	-6	-5	-16	-15	-14	-15	-8	-8	-8	-7
GG				-12			-14			-19		-16	-10		-10	
GU	-8	-10	-10	-10	-7	-7	-7	-7	-8	-10	-10	-10	-8	-8	-8	-8
UA	-10	-8	-11	-9	-7	-6	-3	-5	-11	-9	-12	-9	-3	-6	-3	-5
UC				-5			-5			-8		-7	-7		-7	
UG	-15	-14	-15	-14	-9	-9	-7	-7	-15	-14	-16	-14	-9	-11	-9	-9
UU				-5			-7			-12		-9	-8		-8	

References

[1] S. Bonhoeffer, J. McCaskill, P. F. Stadler, and P. Schuster. Temperature dependent RNA landscapes, a study based on partition functions. *submitted to European Biophysics Journal*, 1992.

[2] T. R. Cech. Conserved sequences and structures of group I introns: building an active site for RNA catalysis. *Gene*, 73:259–271, 1988.

[3] C. Darwin. *The Origin of Species.* reprinted in Penguin Classics, 1859.

[4] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 10:465–523, 1971.

[5] M. Eigen and P. Schuster. The hypercycle. a principle of natural selforganization. Part A: The emergence of the hypercycle. *Naturwissenschaften*, 64:541–565, 1977.

[6] M. Eigen and P. Schuster. The hypercycle. a principle of natural selforganization. Part B: The abstract hypercycle. *Naturwissenschaften*, 65:7–41, 1978.

[7] M. Eigen and P. Schuster. The hypercycle. a principle of natural selforganization. Part C: The realistic hypercycle. *Naturwissenschaften*, 65:341–369, 1978.

[8] R. Fischer. On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341, 1922.

[9] W. Fontana, T. Griessmacher, W. Schnabl, P. F. Stadler, and P. Schuster. Statistics of landscapes based on free energies. *Monatshefte der Chemie*, 122:795–819, 1991.

[10] W. Fontana, D. A. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *submitted to Biopolymers*, 1992.

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

[12] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophysical Chemistry*, 26:123–147, 1987.

[13] W. Fontana, P. F. Stadler, E. Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. Statistical properties of rna free energy landscapes. *submitted to Physical Review A*, 1992.

[14] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Biochemistry*, 83:9373–9377, 1986.

[15] M. Garey and D. Johnson. *Computers and Intractability. A Guide to the Theory of $\mathcal{NP}$ Completeness.* Freeman, San Francisco, 1979.

[16] J. Haldane. A mathematical theory of natural and artificial selection. *Transactions of the Cambridge Philosophical Society*, 23:19–41, 1924.

[17] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Biochemistry*, 86:7706–7710, 1989.

[18] D. A. M. Konings and P. Hogeweg. Pattern analysis of RNA secondary structure, similarity and consensus of minimal-energy folding. *Journal of Molecular Biology*, 207:597–614, 1989.

[19] E. Lawler, J. Lenstra, A. R. Kan, and D. Shmoys. *The Traveling Salesman Problem. A Guided Tour of Combinatorial Optimization.* John Wiley & Sons, 1985.

[20] S.-Y. Le and M. Zuker. Commmon structures of the $5'$ non-coding RNA in enteroviruses and rhinoviruses, thermodynamical stability and statistical significance. *Journal of Molecular Biology*, 216:729–741, 1990.

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

[22] C. Papanicolou, M. Gouy, and J. Ninio. An energy model that predicts the corecct folding of both the tRNA and the 5 S RNA molecules. *Nucleic Acid Research*, 12:31–44, 1984.

[23] A. E. Peritz, R. Kierzek, N. Sugimoto, and D. H. Turner. Thermodynamic study of internal loops in oligoribonucleotides: Symmetric loops are more stable than asymmetric loops. *Biochemistry*, 30:6428–6436, 1991.

[24] I. Rechenberg. *Evolutionstrategie.* problemata. Frommann-Holzboog, 1973.

[25] B. A. Shapiro. An algorithm for comparing multiple RNA secondary stuctures. *CABIOS*, 4(3):387–393, 1988.

[26] B. A. Shapiro and K. Zhang. Compating multiple RNA secondary structures using tree comparisons. *CABIOS*, 6:309–318, 1990.

[27] P. F. Stadler and W. Schnabl. The landscape of the traveling salesman problem. *Physics Letters A*, 161:337–344, 1992.

[28] N. Sugimoto, R. Kierzek, and D. H. Turner. Sequence dependence for the energetics of dangling ends and terminal base pairs in ribonucleic acid. *Biochemistry*, 26:4554–4558, 1987.

[29] N. Sugimoto, R. Kierzek, and D. H. Turner. Sequence dependence for the energetics of terminal mismatches in ribooligonuleotides. *Biochemistry*, 26:4559–4561, 1987.

[30] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 17:167–192, 1988.

[31] M. Waterman and T. Byers. *Math. Biosci.*, 77:179–188, 1985.

[32] E. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol.Cybern.*, 63:325–336, 1990.

[33] S. Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.

[34] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, 1984.

Curriculum vitae

Sebastian Bonhoeffer

$*$ 1965-10-16, Tübingen (Germany)

1971-1975	:	Grundschule, Tübingen
1975-1984	:	Gymnasium (classical education), Tübingen
1984	:	Abitur
1984-1985	:	Civil service
1984-1987	:	Studies of Music (Cello) at the Musikakademie Basel, Switzerland with Prof. Heinrich Schiff
1987.6	:	Lehrdiplom at the Musikakademie Basel
1987.11-1989.7	:	Studies of Physics at the Ludwigs-Maximilians-Universität in München, Germany
1988.6	:	Konzertdiplom mit Auszeichnung at the Musikakademie Basel
1989.3	:	Vordiplom at the Ludwigs-Maximilians-Universität in München
1989.10-1992.2	:	Studies of Physics at the Universität Wien
1991.3-1992.2	:	Diplomarbeit with Prof. Peter Schuster at the Institute of Theoretical Chemistry of the Universität Wien.