Statistics of RNA Folding

DISSERTATION
eingereicht von
Mag. rer. nat. Manfred TACKER
zur Erlangung des akademischen Grades
DOCTOR RERUM NATURALIUM
an der Formal- und Naturwissenschaftlichen Fakultät
der Universität Wien

November 1993

Diese Arbeit verdankt Ihre Entstehung vielen Personen.

Prof. Dr. Peter Schuster verdanke ich die Einführung in dieses Gebiet. Er sorgte mit unvergleichlichem Stil, daß es nie an den Resourcen, den Ideen und der Motivation fehlte.

Dr. Peter Stadler, Assistent am hiesigen Institut hatte die harte Arbeit zu übernehmen, mich auf die richtige Fährte zu bringen, wenn ich vom Weg abgekommen war und er verbrachte viel Zeit mit wichtigen (und unwichtigen) Diskussionen.

Mag. Ivo Hofacker, Mag. Erich Bornberg-Bauer, Mag. Robert Hecht, Robert Happel, Dr. Christian Forst, Mag. Walter Grüner, Dr. Subbiah Baskaran, Thomas Griesmacher, Mag. Herbert Kratky, Bärbel Krakhofer und vor allem Judith Jakubetz schufen eine Arbeitsatmosphäre, die ihresgleichen sucht (*sic*).

Abstract

Abstract

RNA folding is a map assignins (secondary) structures to sequences. It is shown that the secondary structure prediction depends heavily on the parameter set and the algorithm used.

Generic features of the combinatory map of structures that could be characterized include the distribution of secondary structures, the statistical properties of large ensembles of molecules, the presence of neutral nets.

The distribution of structures as determined with different algorithms is higly non-uniform and follows Zipf's first law: there are a few abundant and many rare structures. Statistical investigation of secondary structure elements give the same qualitative results with all algorithm and parameter sets - the number of base pairs, loops, stacks and the free energies scale linearly with the chain length, mean loop and stack sizes get constant.

Although the structure autocorrelation lengths are very small indicating a very rugged combinatory map of structures there exist extensive neutral nets of nearest neighbours folding into the same structures. The extension of this networks is dependent on the alphabet and the ratio of sequences to structures. Consequently optimizations on the CM seems to be rather easy despite the small autocorrelation lengths.

In the last two sections a new $5' \rightarrow 3'$ kinetical folding algorithm is developed and the folding of 16S rRNAs from different archae investigated. The comparison with a minimal free energy algorithm points out that no indications of kinetic folding exist but both algorithms predict only $\frac{1}{3}$ of all phylogenetic base pairs correctly. Furthermore the energy values of the predicted structures are very different from the energy values obtained from the phylogenetic structures. This suggests an inaccuracy and incompleteness of the parameter set.

Deutsche Zusammenfassung

RNA Faltung ordnet jeder Sequenz eine Sekundär Structur zu. In dieser Arbeit wird gezeigt, daß die Vorhersage einer Sekundär Structur sehr stark vom verwendeten Parameter Satz abhängt.

Generische Eigenschaften der Sequenz Strukturabbildung müssen deshalb unabhängig vom verwendeten Algorithmus und Parameter Satz sein. Die Verteilung der Strukturen ist robust und folgt Zipf's law, was bedeutet, daß einige sehr häufige und sehr viele seltene Strukturen auftreten.

Die Statistik von Sekundär Struktur Elementen ergibt, daß die statistischen Eigenschaften qualitativ unabhängig vom Algorithmus und vom Parameter Satz sind. Die Anzahl der Basenpaare, der Loops, der Stacks und die freie Energie der Strukturen wachsen linear mit der Kettenlänge während die mittleren Loop- und Stacklängen konstant werden.

Obwohl die Struktur Autokorrelationslängen sehr klein sind, was bedeutet, daß die Strukturlandschaften sehr "rugged" sind, konnten sehr ausgedehnte Neutrale Netze nächster Nachbarn, die in die selbe Struktur falten, gefunden werden. Die Ausdehnung dieser Netze ist abhängig vom Alphabet und vom Verhältnis der Strukturen zu den Sequenzen. Adaptation auf diesen Landschaften dürfte deshalb einfacher sein, als die Autokorrelationslängen befürchten lassen.

In den letzten Kapiteln wird ein neuer $5' \rightarrow 3'$ Faltungsalgorithmus eingeführt, der die kinetische Faltung der RNA während der Transkription simuliert. Damit wurde versucht, festzustellen, wie wichtig kinetische Prozesse bei der Faltung sind. Verglichen wurden die Struktur Voraussagen dem MFE(F) Algorithmus, der die thermodynamischen Strukturen liefert. Es konnte bei der Faltung von 16S rRNAs keine Verbesserung der Voraussagegenauigkeit gegenüber des MFE(F) algorithmus festgestellt werden. Die mit beiden Algorithmen erhaltenen Sekundär Strukturen sagen nur etwa $\frac{1}{3}$ der phylogenetisch bestimmten Basenpaare richtig voraus. Die erhaltenen freien Energiewerte liegen ebenfalls sehr weit entfernt von den Energien der phylogenetischen Strukturen. Das bedeutet, daß der experimentell bestimmte Parameter Satz zu ungenau und unvollständig ist.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Historical Background

With the creation of a new paradigm Charles Darwin changed in 1859 the whole conception of biology as science [13]. He argued that all the wealth of organisms from the insects to man are not stable entities but they are the result of a natural process. The mechanisms of this process he identified as variation and selection. Randomly generated variations modify existing species and natural selection tests the newly introduced variants for their fitnesses. Different species evolve because in every generation there is an accumulation of the best adapted and most fertile organisms. The fittest will survive, as he put it.

Modern mathematical and scientific methods brought a deeper understanding of Darwinian evolution. In the 1920s and 1930s J.B.S. Haldane [33], S. Wright [88] and R.A. Fisher [19] developed the mathematical background of population biology. The underlying basic molecular mechanisms were discovered in the decades afterwards. Avery showed in 1944 that the genome is made of DNA and J. Watson and F. Crick discovered in 1953 the structure of the DNA α-helix.

Now the mechanisms of the generation of mutations and selection could be formulated in physico-chemical language. The replication of nucleic acids is error prone, the accuracy of replication is between 10^{-4} of RNA replicases in viruses and 10^{-12} in eucaryotic cells which possess many repair mechanisms. Thus the generation of variation is an unavoidable by-product of the life cycle. Selection is introduced by competition for limited resources.

A very useful method of thinking in evolutionary terms was introduced in the early 1930s by Sewall Wright. He coined the term fitness landscape [88]. By fitness landscape he defined a mapping which assigns to each geno-

type a fitness value. Evolution is understood in this terminology as a comlex optimization process on this fitness landscape.
If it is possible to calculate fitness values of every genotype *a priori* to the selection process then evolution reduces to a well defined mathematical optimization problem. The somewhat blurry notion of fitness as Darwin used it is then transformed into a perfect scientific term and the tautology avoided, that the fittest is the one which survives in a Darwinian evolution scenario.

The only problem which remains is that it is until now not possible to calculate the fitness values for species. A way out of this dilemma is the use of simpler model systems based on the replication of RNA molecules. Such models has been thoroughly investigated in the last years and their landscape characterised [23, 22, 20, 24, 76]. Pointwise characterization of landscapes can be done experimentally with gene technological methods, but a global characterization is feasible only by computer simulations.

Recently the notion of a fitness landscape has come under heavy attack. Realistic fitness landscapes, it is claimed, are not rigid, but are deformed in response to changes in the environment [42]. This is surely correct, but this argument just says that the calculation of a fitness value depend on the population size of other genotypes too. The notion of a fitness landscape remains an important way of thinking about evolution.

1.2 RNA and Evolution

RNA is a very suitable molecule for evolutionary models because it combines several important features:

- Genotype and Phenotype are embodied in the same molecule
- The phenotype can be computed from the genotype
- The origin of life is tightly associated with RNA

- Evolution experiments are performed with RNA in the test tube

1.2.1 Genotype and Pheotype

RNA is a very versatile molecule which can act as a device to store information (in the genotype) and as an effector (in the phenotype).
RNA is used as the genome in many viruses (retroviruses, class 1, class 2 and class 3 RNA viruses [82]) and more and more reactions are discovered which can be catalyzed by RNA. The main difference of RNA to DNA, the 2′-OH on the ribose is responsible for the higher vulnerability of RNA towards nucleases and the lower stability in solution but it is also very important for building a complex three dimensional structure [71].

The first catalytic RNA found was a self splicing intron from Tetrahymena [7]. Other catalytic RNAs which have been found include M1 RNA, the catalytic moiety of RNAse P and ribosomal RNAs [46]. For a summary of catalytic RNAs see the review of T. Cech [8]. Evolution experiments are exploited to find RNAs with novel catalytic properties (see below).

The presence of the genotype and the phenotype in the same molecule simplifies theoretical models and evolution experiments because no transformation step from the genotypic to the phenotypic level (as in the case of the DNA/protein system) is necessary.

1.2.2 Computation of the Phenotype

It is conceptionally possible to divide the folding process of RNA into two steps:

- Formation of the secondary structure
- Folding of the tertiary structure from the secondary structure

The simple representation of the RNA structure in the form of a planar graph, the secondary structure, which contains almost all base pairs (pseudoknots are excluded) is justified by the facts that secondary structure elements are conserved in the course of evolution [7] and that the main driving force of folding is the formation of the secondary structure base pairs. The folding of the tertiary structure is considered as the formation of additional tertiary interactions without changing the secondary structure.

The evolutionary conservation of secondary structure elements suggest that there exist a correlation between secondary structure (or more generally shape) and function. The fitness of a molecule in a special environment depends on its ability to perferm certain reactions. It is therefore helpful to divide the relationship of genotype and function (fitness) into two steps:

- Folding of the sequence into a shape
- A mapping from this structure to fitness

The folding of the genotype leads to a 3D structure (approximated in the case of RNA by the secondary structure). This shape then determines the function of the molecule (its fitness). The mapping from the genotype via the secondary structure to replication and degradation rates has been investigated in [23], the mapping from genotypes to association and dissociation constants in [83].

The most basic relationship between genotype and phenotype is consequently the mapping from genotype to shape. Statistical evalutations of the features of this mapping for different alphabets, parameter sets and algorithms is at the core fo this work.

For the prediction of secondary structures of RNA molecules there exist kinetical algorithms which simulate the kinetically determined folding and dynamic programming algorithms which find the minimal free energy structure (described in section 4). Secondary structures can be easily stored and compared. This is not the case with three dimensional structures.

1.2.3 The RNA World

It is the general assumption that life on earth started with chemical processes taking place in the 'soup' of organic, energy rich molecules in the primeval seas. Nucleic acids and amino acids have been synthesized in the laboratory under conditions thought to exist on the primitive earth [57, 63].
Manfred Eigen demonstrated in 1971 [15] that the only prerequisite for Darwinian evolution is an open system far from thermodynamic equilibrium with replication living on limited resources.

The formation of complex organic molecules is just the beginning of the story. Another necessary aspect for adaptation to take place is that information (embodied in the genotype) must be conserved. This conservation was made possible by template directed replication.

Self replicating RNAs are thought to be the first replicators of this type. The information is stored in the primary sequence, the arrangement of bases. Error prone copying of these molecules in an energy rich system provides all the ingredients for Darwinian evolution.
It is assumed, that RNA can be replicated by RNA itself, the search for the first RNA polymerase made entirely from RNA is very intensive [2]. Thus RNA alone is sufficient for providing the molecular basis of evolution.

Further hints that the first Darwinian evolving entities were made from RNA are that in todays cells many key processes are still catalyzed by RNA including protein and RNA synthesis [46].
The synthesis of deoxyribonucleotides proceeds through reduction of presynthesized ribonucleotides [82] indicating that RNA predates DNA. DNA can be seen as an RNA derivative that is chemically stabilized by removal of the

$2'$-OH group. Furthermore RNA is used as a primer by DNA polymerases whereas RNA polymerases do not need any primer at all. All these data suggest that RNA was the information carrier historical evolution on earth started with.

W. Fontana, P. Schuster *et. al* developed a computer model of RNA evolution in a chemical tank reactor [23, 22]. They could show that adaptation in this system is a very difficult task because the RNA landscape is an example of a rugged landscape (see 2) with many local optima and the adapting molecules very easily get trapped.

1.2.4 Evolution in the Test Tube

Sol Spiegelman was the first who demonstrated in the late 1960s Darwinian evolution in a laboratory experiment [58, 80]. He used the $Q\beta$ replicase system and serially diluted the RNA molecules in fresh medium. As evolution proceeded in the test tube, the surviving RNAs became progressively shorter because smaller molecules could be copied faster. He could increase the replication rate about 15 fold compared to wild type RNA from $Q\beta$.

In the last few years new methods has been developed which exploit the mechanisms of Darwinian evolution and the features of large ensembles of biopolymers to generate proteins and nucleic acids with novel properties. These methods could finally lead to a partial substitution of rational designing of molecules [8]. Due to the fact that selection acts on the phenotype whereas mutation generates new genotypes a straightforward application of in vitro evolution is possible only in the case of nucleic acids where genotype and phenotype are embodied in the same molecule. In the case of proteins another approach must be taken (see below)
A short outline of an in vitro experiment reads as follows:

1. Generation of an ensemble of random sequences

2. Selection of molecules with useful properties
3. Amplification of the selected molecules with the introduction of new mutations

which is in agreement with the theory of Darwinian evolution:

1. Generation of mutations
2. Evaluation of a fitness criterion
3. Selective amplification (with errors) according to the fitness criterion

Ensembles of random sequences can be generated at the DNA level by means of chemical modification [12, 79], incorporation of randomized oligodeoxynucleotides [53, 86] or copying by an errorprone polymerase [90, 89]. In laboratory experiments the number of quasi random RNAs is about 10^{13} in the start population [3]. It is possible to convert DNA into RNA with T7 RNA polymerase. The opposite reaction is performed through the use of reverse transcriptase. Thus there is interconversion between RNA and DNA possible at every step.

The ensuing selection step depends on the characteristics of the desired property. RNAs can be selected with respect to their ability to bind dyes [16] or to catalyse specific reactions, for example phosphoester transfer [3].

Amplification is performed by transcription of DNA by T7 RNA polymerase which amplifies cDNA about 1000fold [9] or by polymerase chain reaction (PCR).

The PCR reaction under mutagenic conditions has been used for the generation of mutations.

For proteins the situation is somewhat more complicated because there is a difference between the genotypic (DNA) and phenotypic level (protein) and

there is no simple practical possibility for generating the corresponding DNA from the protein.
A way out of this dilemma is the use of catalytic antibodies. It has been shown that second class antibodies (antibodies directed against the binding site of antigen binding antibodies) can mimic antigen [4] and theoretical models led to the concept of a shape space [65]. 10^8 antibodies should saturate this shape space and therefore provides a universal set of enzymes.

This work describes in section 2 the concept of a landscape and of combinarory maps. An overview about the RNA secondary structure model and RNA folding algorithms is found in section 4 where all the algorithms used are discussed. Chapter 5 is dealing with the statistical evaluation of secondary structure elements and the distribution of structures which follows Zipf's first law.

Section 6 presents autocorrelation functions, which are a measure for the ruggedness of landscapes. The concept of a neutral net is introduced in section 7. Chaper 8 characterizes the stability of structures against point mutation and develops new methods for characterizing the basin of attraction of secondary structures.

In 4.4.3 tries give an answer to the long standing question if RNA folding is kinetically determined. A new $5' \rightarrow 3'$ folding algorithm is introduced in this context.

A discussion of the results in section 10 concludes the work.

2 Landscapes

2.1 Definition of Landscapes

A landscape is mathematically defined by a triple (X, d, f) where X is a finite set, $d : X \times X \rightarrow R_o^+$ is a metric and $f : X \rightarrow V$ is a function which maps into a vector space with scalar product. The metric is needed in order to define neighbourhoods and distances on X. X together with the metric is called the configuration space. The choice of the metric is crucial for the properties of the landscape. It is not necessarily unique and some metrics can distort the statistical properties heavily.

The simplest type of configuration space can be represented as a finite undirect ed graph Γ. The canonical metric on (X, d) coincides with the minimum number of necessary moves to connect two vertices on Γ [6]. The topology of the configuration space is defined by the chosen metric. Topology in this context means just neighbourhood relations of configuration pairs.

A *complex combinatory map* [24] is a slight generalization of a landscape.

Definition: A combinatory mapi CM is a quintuple $(X, d_X; Y, d_Y; f)$ where X and Y are sets endowed with metrices d_X and d_Y respectively, and f is a map $X \rightarrow Y$.

In this definition the above defined landscape is a combinatory map with Y being the real numbers and $d_Y(a, b) = \mid a - b \mid$.

$(X, d_X$ is for polymers the configuration space (or sequence space). There exist a natural metric which is induced by point mutations. It is known as Hamming distance [34] and was introduced in information theory. The hamming distance $d_H(a, b)$ counts the minimum number of point mutations necessary to move from sequence a to b. The sequence space for the **GC** alphabet is depicted in fig. 1.

Assigning to each RNA sequence an energy value results in the RNA

Uncorrelated Landscapes:	$\rho(x) = \delta_{0x}$
Extremely rugged Landscapes:	$\rho(x)$ discontinuous for $x = 0$
Generic Landscapes:	$\rho(x) = 1 - \frac{1}{\mu}\|x\| + \dots$
Smooth Landscapes:	$\rho(x) = 1 - q\|x^2\| + \dots$
Nontrivial Fractal Landscapes:	$\rho(x) = 1 - \|x\|^{\alpha} + \dots$

Table 1: Classification of Landscapes

energy landscape the statistical focus of which have been the focus of detailled investigations [20, 24]. The landscape of melting was characterized in [83] and the landscape of replication and degradation constansts in [23, 22]

A very important complex combinatory map is generated by mapping to each RNA sequence a secondary structure. This complex map of secondary structures was characterized in [21].

Analogous landscapes appear in many different fields in physics, theoretical biology and computer science. They can be classified [24] according to their autocorrelation function $\rho(k)$ (cf. appendix B near the origin (see table 1).

In biological and physical applications most landscapes and CMs are generic (or rugged) landscapes. These are characterized by exponentially decaying autocorrelation functions.
Prominent examples of such rugged landscapes are:

- Sherrington Kirkpatrick Spinglasses [44, 56]
- The Traveling Salesman Problem [47, 81]
- N-k model [40, 41, 42]
- Energy Landscapes of Proteins/RNAs [25, 20, 24]

- Fitness Landscapes [23, 22]
- Affinity Landscapes of Antibodies [49]
- Melting Kinetic Constants Landscapes [83]

Optimization on rugged landscapes is a non trivial problem. It is in general not possible to find analytically the global minimum. The performance of optimization algorithms depend crucially on the structure of the underlying landscape.

Optimization algorithms, which are universal (no information of the landscape is *a priory* necessary) are:

- Neural Networks: Originally used as models for the dynamics of the nervous system. Today they are used in certain optimization problems and in pattern recognition. [38]
- Genetic Algorithms: These algorithms apply the methods of biological evolution to optimization problems. [27]
- Simulated Annealing: This method is an improvement of Monte Carlo optimization. [44, 22]
- Grand Deluge [14]

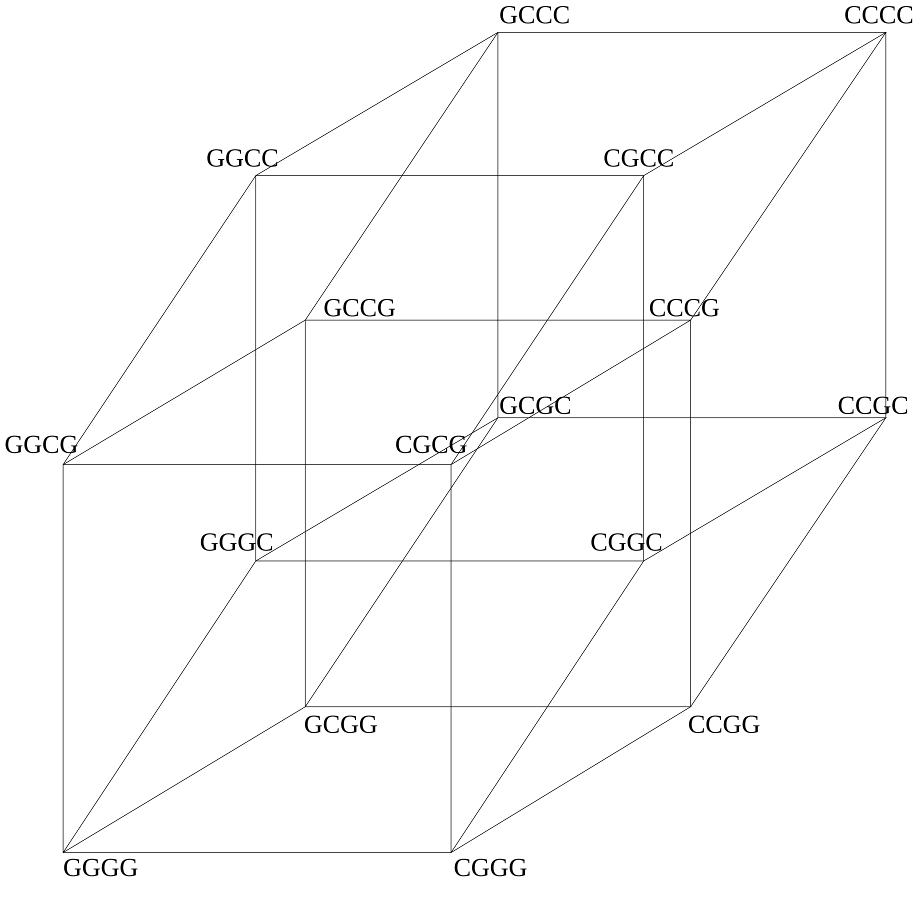


Figure 1: The configuration space of the **GC** alphabet, chain length = 4

3 RNA Secondary Structure

In the last two decades it has become easy to sequence nucleic acids [54, 72] and impressive attempts are being made nowadays to sequence the whole genome of eukaryotic organisms, even the human genome with its approximate three billion base pairs.

Apart from the deep ethical problems which are associated with this extension of the human knowledge there remains the scientific problem of exploiting the information present in the four letter code of the sequences.

It is a very simple task to translate the DNA sequence in the corresponding RNA sequence or the protein sequence. It is an enormous problem however, which is not yet solved, to deduce the tertiary structure of the protein or RNA from the primary sequences. But this three dimensional structure is an essential prerequisite to understand the processes going on in an organism - a protein or RNA is biologically active only when it assumes its correct form, incorrectly folded molecules are biologically inactive.

The usual way of a scientist coping with a problem which is too huge to be solved in one step is to break it down into smaller parts and solve the smaller problems. Fortunately this breaking down is fairly obvious in the case of RNA (and less obvious in the case of proteins):

1. Folding of the primary sequence into the secondary structure consisting of (almost) all basepairs.
2. Formation of the spatial strucure by folding the quasi-planar secondary structure into a three dimensional object.

It has been shown that the secondary structure model is not a mere theoretical construct but somehow reflects reality: secondary structure elements are conserved in evolution, for example the clover leaf structure of tRNAs class I introns [7, 48]. This conservation throughout evolution is even exploited to determine the secondary structure of RNAs by means of the phylogenetic method (see below).

Unfortunately the secondary structure model is far more difficult to apply to proteins - the concept of a secondary structure being more vague because different to RNA where a base pair is defined in a all or nothing way (either base i is paired with base j or it is not) there exist a continuum of interaction with no clear cut off value of the binding energies. Consequently for the case of proteins there exist no reliable algorithms to predict the secondary structure whereas in the case of RNA such algorithms exist (see 4.1). In the rest of the work I will consider only the folding of RNA.

The second step in the folding process, the transition from the RNA secondary structure to the three dimensional structure is more difficult to calculate. Even with the additional information from experimental data it is still more an art than science to find the correct structure (for an overview of currently used methods see [50]).

The problem is made worse by the fact, that due to experimental difficulties transfer RNAs are the only RNA molecules the three dimensional structure of which have been completely determined [43] by X-ray crystallography.

3.1 RNA Structure

An RNA molecule consists of a sequence of ribonucleotides linked together by covalent phosphate bonds. The nucleotides are composed of a sugar, ribose with phosphate groups at its 3' and 5' positions and one of four bases. Adenine (**A**) and guanine (**G**) are purines, cytosine (**C**) and uracil (**U**) are pyrimidines. The 5'-3' phosphate linkage gives the sequence of nucleotides a definite orientation. Thus RNA can be viewed as an oriented string of a definite length composed of an alphabet of 4 nucleotides, called the primary

structure.
Hydrogen bonding between **G-C** and **A-U** as well as a substantially weaker bond between **G-U** result in folding back of the RNA upon itself into the three dimensional tertiary structure. This is the biologically active form.
In naturally occuring RNAs many more bases can be found, mostly resulting from posttranscriptional modifications of the above mentioned. Transfer RNAs as well as 16S rRNAs for example contain a high percentage of unusual bases.
The folding of RNA can be conceptionally divided into two steps:

1. Folding of the primary sequence into the secondary structure consisting of (almost) all basepairs.

2. Formation of the spatial structure by folding the quasi-planar secondary structure into a three dimensional object.

The secondary structure model is not a mere theoretical construct but its biologically important role has been established by the conservation of secondary structure elements in the course of evolution, for example the clover leaf structure of tRNAs and of class I introns [7, 48].

3.2 The Secondary Structure Model

A secondary structure is defined as a collection of base pairs. In mathematical terms an RNA molecule can be represented by a string R of letters out of the chosen alphabet (**A, U, G, C** in the biophysical set) of length N. The secondary structure is then simply the set S of basepairs $i.j$ with $1 \leq i < j \leq N$ where i and j denote the i^{th} respectively j^{th} base of the RNA molecule. It is assumed that $i < j$. A base can at most bind one other base. If R_{ij} denotes a subsequence (fragment) of R e $1 \leq i < j \leq N$ where i and j

denote the i^{th} respectively j^{th} base of the RNA molecule. It is assumed that $i < j$. A base can at most bind one other base. If R_{ij} denotes a subsequence (fragment) of R extending from the i^{th} to the j^{th} nucleotide then S_{ij} is the secondary structure of R_{ij}. S_{ij} is then called a substructure of S.

Structural elements of the secondary structure:

- **stems** or **stacks**: double helical regions
- **loops** and **bulges**: internal unpaired bases
- **joints**: strings of unpaired bases joining freely mouvable substructures
- **free ends**: unpaired bases at the 3' or 5' site.

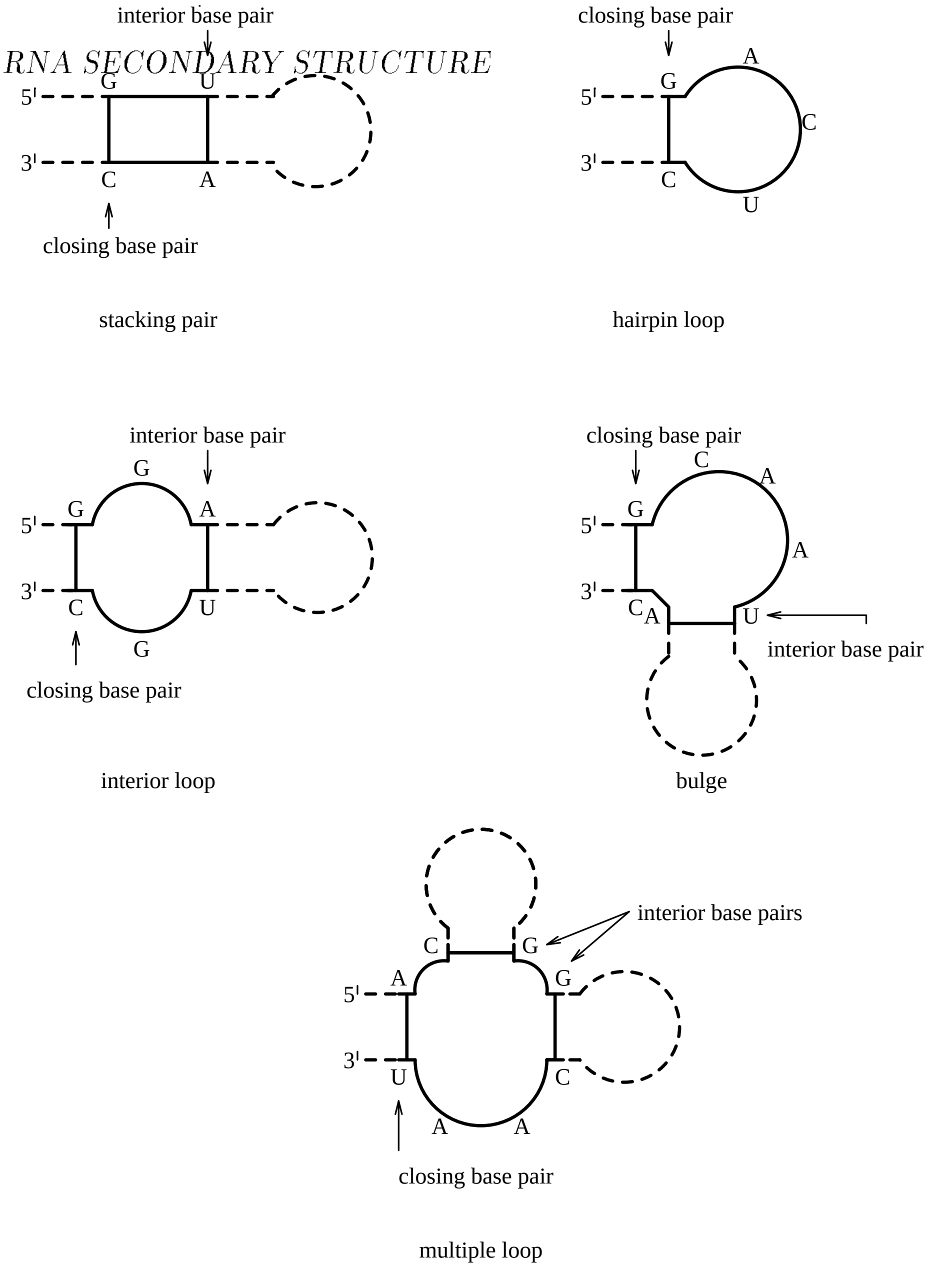
The stabilizing contributions to the secondary structure are provided by the double helical regions, the stacks. Loops and bulges destabilize the structure. Free ends do not contribute to the free energy in our implementation of the folding algorithm. Fig. 2 shows the different sorts of loops which are occuring.

The following constraints are imposed on basepairing:

1. There are at least 3 intervening bases between basepaired ribonucleotides .
2. The graph is unknotted and planar.
3. The secondary structure can be uniquely decomposed into substructures.
4. Only complementary bases can form base pairs.

The first constraint ensures that the RNA backbone cannot fold back too sharply on itself. That means that if $i.j$ form a base pair then $j > i+3$. In the folding algorithm this constraint is imposed by assigning large destabilizing energies to hairpins with less than three unpaired bases (see appendices D and E for the parameter sets).
Prohibition of pseudoknots: If $h < i < j < k$ and $i.k$ and $h.j$ form basepairs then such a structure is called a pseudoknot. Figure 3 gives a graphical representation.

Although it has been shown that pseudoknots form a common motif in RNA folding [68] they are excluded from the folding algorithm. The reason for this is that apart from slowing down the algorithm and the difficult energy assignment of the resuling loops (no experimental data is available) pseudoknots can easily be relegated to the tertiary structure because only unpaired regions not included in the secondary structure are responsible for the knot.
We assume that the secondary structure can be decomposed into substructures which contribute additively to the thermodynamic and kinetic properties of the whole molecule. That means whether substructure i is folded or not does not influence the folding of substructure j. This constraint is crucial for the MFE algorithm and will be described more in detail in the next section.



stacking pair

hairpin loop

interior loop

bulge

multiple loop

Figure 2: The classification of loops in the RNA secondary structure model

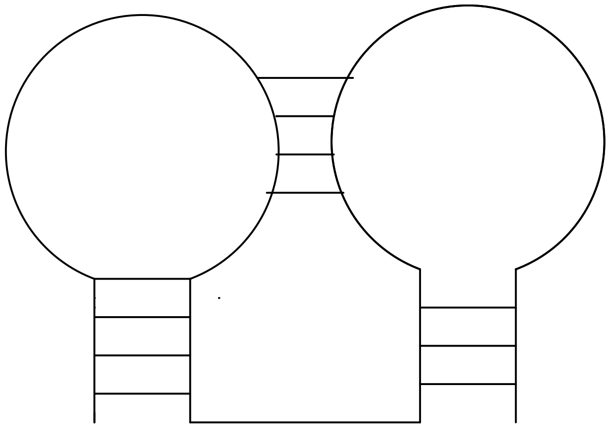


Figure 3: A secondary structure with a pseudoknot.

4 Folding Algorithms

There exist a lot of different methods to predict the secondary structure of RNA either singly or in various combinations. These methods can be divided into four major groups:

1. Phylogenetic Comparison
2. Minimum Free Energy Folding
3. Kinetical Folding
4. Prediction of Ensembles.

4.1 Phylogenetic Comparison

Phylogenetic Comparison is the method of choice for determining secondary structure provided that there are enough RNA sequences from different organisms available.

The underlying principle is that in the course of evolution mutations occur which do not alter the function of the molecule. These mutations are preserved. Mutations that destroy the function of the RNA are strongly selected against. It is assumed that function depend on structure and therefore structure should be conserved in evolution.

Based on this assumptions methods have been developed to reconstruct from covariances of nucleotides the secondary structure. The sequences are aligned first and then regions of potential base pairing are identified and searched for compensatory base pair modifications - e. g. a **AU** bp replaced by a **GC** bp. A helix is usually considered to exist if two or more covariations are found in it [64]. The sequences which are aligned have to differ by $20-40\%$ because sequences to dissimilar are very difficult to align and sequences to similar

have not enough covariations [64]. It is very difficult with this method to identify very conserved regions because they have not enough compensatory base pair changes and it is impossible to determine the regions where the secondary structure is not conserved.
The prediction capability of the method increases with the availability of homologous sequences and so the playground of phylogenetic comparison is the 16S rRNA and the 23S rRNA molecule [30, 31, 87], where hundreds of rRNAs in different species has already been sequenced. Improved methodology makes it possible to study higher order correlations and to predict pseudoknots and other tertiary interactions [32].

In general the phylogenetic comparison method can not detect all secondary structure elements in a molecule so that for the complete structure additional methods have to be used.
Novel or random RNAs can not be folded with this method because due to the lack of evolutionary conserved structures no covariance data are available.

4.2 Minimal Free Energy Folding

These algorithms find the secondary structure with the lowest free energy. These folding algorithms are the mainly used today but several objections are raised against using methods which determine a single structure with the lowest free energy. The arguments most often used are the following:

- Incomplete and inaccurate experimental data for the parameter set
- Kinetical trapping
- Distortion by tertiary interactions
- Equilibrium of several structures.

Even the best prediction algorithm is as good as the experimentally determined data available. These data were derived from melting studies on small oligonucleotides at 37° C. The only non Watson-Crick basepair considered was **G-U**.

Compilations of these parameter sets were published by Salser [73] and by Freier *et. al.* [26, 66, 35, 75]. Experimental data is available for bulges, interior loops and hairpin loops but not for multiloops.

Different parameter sets give different energy values, so it is very instructive to run the same algorithm with different parameter sets and compare the secondary structure predictions of the same RNA molecules (see below) to determine the influence of the parameter sets.

In vivo an RNA molecules starts folding during its synthesis and it is conceivable that the molecule gets trapped in a local optimum. In such a case the thermadynamically most stable structure would biologically be totally irrelevant.

Last but not least in solution there exist no single structure but an ensemble of structures in equilibrium. Most of the analytical methods (e. g. NMR) and the reaction in the cell encounter this ensemble. It is therefore desirable to develop methods which predict suboptimal foldings as well [92] or the entire equilibrium ensemble [55].

4.2.1 Combinatorial Algorithms

The easiest possibility to obtain the structure with the minimum free energy is to calculate all secondary structures and chose the best. Unfortunately the number of secondary structures grows with at least 1.8^{ν} for large ν with A,U,G,C being randomly distributed with equal probability [93].

This folding algorithm first [67, 61, 28] compiles a list of all potential helices compatible with a given sequence. The next step is then to determine all possible helix combinations. What remains to be done is to filter out the structure with the lowest free energy. This method can in principle predict pseudoknots as well as suboptimal foldings and calculate their energies. The number of possible helixes though increases with ν^2, ν being the chain length of the sequence and thus the number of possible helix combinations grow with 2^m, $m = \nu^2$. It can be shown, that a special case of the search for a maximal set of compatible regions is the well known maximal clique problem in computer science [93]. This problem is $NP - complete$, in other words no search procedure exist which can solve it in less than exponential time. It is easily seen then that this approach is bound to fail for larger ($>$ 200 nucleotides) sequences.

4.3 Dynamic Programming

The general strategy of these algorithms is to determine first the optimal structure of the smallest subfragments of a given sequence and then calculate the optimal structures of the larger fragments by adding together optimal subfragments. Crucial to these algorithms are two assumptions:

- Unique decomposition of the problem is possible
- No pseudoknots

Due to the stepwise generation of increasingly larger structures by adding together substructures it is necessary for the algorithm that a secondary structure is analytically decomposable in a unique way into a number of substructures such that each base is contained in exactly one such substructure. Pseudoknots would violate this assumptions - the inclusion of pseudoknots would lead to a breakdown of the algorithm. The same is true for the energies of the substructures. It is assumed that the main energy contribution of the secondary structure stems from the Watson Crick base pairs and that different substructures do not interact.

4.3.1 Maximum Matching (MM)

The first successful application of the dynamical programming algorithm to the prediction problem of RNA secondary structures dates back to 1978 [62]. Ruth Nussinov presented a computer method which maximized the number of base pairs in a secondary structure with the restriction of a minimal loop size of 3. In the our implementation of the programming algorithm (called MM) it is possible to tune the minimal stacklength as well.

The maximizing of base pairs is a very unphysical way of RNA folding because no loop destabilizing energies are taken into account but it is very fast and it was the starting point for the developing of more realistic algorithms. We used the MM algorithm for reasons of comparison because it represents a folding algorithm without any energy parameters at all.

4.3.2 Realistic Dynamic Programming Algorithm (MFE)

The folding algorithm used in this work is based on the dynamic programming algorithm of Zuker and Sankoff [39, 93]. This minimal free energy algorithm relies critically on the assumption that the energy of the secondary structure is the sum of the free energies of the substructures.

The strategy of the algorithm is to determine first the optimal secondary structure for each substructure. These values are stored and the optimal value of the whole molecule is obtained by adding the energies of the optimal subfragments. The runtime of the Zuker algorithm is cubic in chainlength ν.

The incorporation of more complex interactions than that of nearest neighbours would lead to a breakdown of the additivity scheme because in this case substructures need not necessarily be optimal to achieve an optimal overall folding.

The free energy data are derived from melting studies on small oligonucleotides at 37° C. There exist two compilations of parameter sets, the one by Salser (see appendix E) and a more recent version from Freier (see appendix D). The implementation of this dynamic programming algorithm with the dataset by Freier is abbreviated MFE(F), the one with the Salser dataset MFE(S).

The only non Watson-Crick basepair considered was **G-U**. Experimental data is available for bulges, interior loops and hairpin loops but not for multiloops. In this case a linear ansatz

$$\Delta G^{\circ}_{37}\mathrm{ML} = a + bn + ch \tag{1}$$

has appeared to be useful [39]. n is the number of unpaired nucleotides in the loop, h is the number of helices in the loop and a, b, c are adjustable parameters, the values of which have been determined by comparison with experimentally determined secondary structures.

4.3.3 Equilibrium Ensembles

RNA molecules are not rigid molecules which assume only one structure but they can switch between different secondary structures e.g. by opening and closing of single base pairs (zippering) or by slippage of the whole helix (both processes are very fast, see [70]). The result is that ensembles of RNA molecules show a Boltzmann distribution of secondary structures. The equilibrium partition function contains all information about this equilibrium distribution.

The full equilibrium partition function can be calculated recursively with the dynamic programming approach with ν^3 time requirement [55]. The temperature dependence of the partition function gives information about the melting temperature of the RNA [5, 83]. A matrix display of the probabilities of base pairing between different positions is very useful for assessing base pair slippage, alternate secondary structures and the stability of the dominant secondary structure. It is in general not possible, though, to reconstruct the minimal free energy structure from the base pairing probabilities.

4.4 Kinetical Folding

These algorithms try to simulate the actual folding of RNA during transcription. The strength of these methods lies in the fact that they do not assume that the biologically active RNA structure is the minimal free energy structure.

4.4.1 Kinetical Folding I (KIN)

The earliest attempt to solve the kinetical folding problem was made by Martinez in 1984 [52]. His approach was first to calculate all possible stacks compatible with a given sequence and then generate the secondary structure by sequentially incorporating the stacks with the highest equilibrium constant.

His folding algorithm unfortunately did not account for the actual $5' \rightarrow 3'$ folding of RNA during transcription. Nevertheless some recent works eploited his approach because it is fairly easy to include the prediction of pseudo knots in the algorithm [1, 29].

Our implementation of this kinetical algorithm which in the following chapters is abbreviated KIN is constructed as follows:

- Determine all helices compatible with a given sequence and calculate the free energies

- The stem with the lowest free energy value (and therefore the largest equilibrium constant) constitutes the first stack of the secondary structure
- The next helices are added according to their equilibrium constants and the following constraints:
  - The helices must be compatible with the already incorporated stems
  - Introduction of the newly chosen helix must decrease the free energy of the secondary structure
- The secondary structure is considered to be complete when there is no helix left the incorporation of which could further decrease the free energy of the secondary structure.

It is assumed that the secondary structure obtained does not refold. Pseudoknots are excluded from the secondary structures because not enough experimental data is available for assigning reliable energy values.

4.4.2 Nondeterministic Kinetical Algorithms

The folding process is described by a Markov process which is modelled by a Monte Carlo method. It is argued that the folding process depends on the rate constant of helix formation and on the rate constant of helix decay and not on the equilibrium constant as in the Martinez approach [59].
A slight modification of this algorithm could be used for the prediction of $5' \rightarrow 3'$ folding of RNA during transcription [60].

The great advantage of this method is that it gives a Boltzmann weighted equilibrium ensemble of secondary structures. The disadvantage is that it is very timeconsuming and that many calculations have to be performed for folding a single molecule. It has therefore been applied mostly to tRNAs.

4.4.3 Deterministic $5' \rightarrow 3'$ Folding (KIN53)

Transcription of RNA proceeds from the 5' end to the 3' end. The elongation process works at a speed of about 50 nucleotids per second in E. coli [82], in eucaryots at about 15 nucleotides per second [17].
It is assumed, that freshly transcribed RNA folds immediately back onto itself and that these stacks have to be molten again to allow the structure to refold. This refolding process however can't take place when the activation energy for opening a stack is too large. In this case the stack remains closed and an energetically lower structure which requires this stack to be opened again can't be reached. So the molecule is trapped in a kinetically determined local optimum - the global optimum being inaccessible during the life time of the RNA.

The activation energy for opening a stack is proportional to the stacking energy of the bases which form the stack [69, 70]. There seems to be no substantial difference in the activation energy for melting a stack between double strand RNA and single strand RNA [70]. If one assumes a relationship between k_d, the rate of dissociation of a stack and ΔG_{stack}, the stacking energy of the helix of the form

$$k_d = A \exp\left(\frac{\Delta G_{stack}}{RT}\right) \tag{2}$$

with $A = 1.8 \cdot 10^8$ [69] then the mean dissociation time $\tau = \frac{1}{k_d}$ of a stack can be calculated (see Table 2).

It is easily seen, that stacks with stacking energies greater than ~ 13 kcal/mol can't be opened again during the typical timescale of transcription.

τ [s]	stacking energy [kcal/mol]
0.00005	5.5
0.001	7.3
0.01	8.6
0.1	10.0
1.0	11.4
10	12.8
100	14.2
1000	15.5
10000	16.9

Table 2: Dependence of τ on the stacking energy

Based on these observation the following model of the kinetical $5' \rightarrow 3'$ folding process is developed.

The Model

It is assumed that kinetical folding has to cope with the following constraints:

- Locality Constraint
- Energy Constraint
- Direction Constraint

The Locality constraint stems from the fact, that kinetically determined folding of hairpins is allowed only if the loopsize is not greater than LOOPSIZE, the value of which is 25 - 40. This assumption is justified by assuming that the folding of a hairpin has to surmount an activation energy which depends on ΔG_{loop}, the destabilizing energy of the loop which is proportional

to log(loopsize) and that the folding takes place only if there are no other stacks already present which interfere with stack formation.
Existing stacks open again with a τ as a function of the stacking energy of the helix (cf. table 2). It is thus appropriate to assume a stacking energy STACKE, which determines, that a stack is frozen if its stacking energy is greater than STACKE.
According to experimental data [70], the unzippering of helixends is very fast ($\tau \sim 1 \cdot \exp(-7)$), so it is allowed in our model.
Folding of the RNA must proceed from the $5' \rightarrow 3'$ direction. This is imposed by preferring those stacks which lie nearer to the 5' end just by assuming that if stacks A and B both fulfill the locality and the energy constraint and A and B overlap then the stack with the lower 5' value is chosen and stack B excluded.

After evaluation of the kinetically determined stacke the molecule is allowed to refold and find the thermodynamic optimum within these constraints.

The Algorithm (KIN53)

1. Calculate all stacks with stacking energies < STACKE and loop size < LOOPSIZE
2. The stacks are ordered according to their 5' positions
3. Overlapping stacks are excluded, with stacks nearer to the 5' end preferred
4. The innermost basepair of every chosen stacks is assigned a high bonus energy
5. The string is folded with a minimum free energy algorithm [93] with every structure containing the fixed basepairs getting the bonus energy. Helix ends are allowed to unzip.

The performance of the KIN53 algorithm is compared with the MFE(F) algorithm in section 9 where 16S rRNAs from different species are folded.

5 Structure Statistics

Samples of random sequences are generated with uniform distribution of nucleotides. We investigated both the natural **GCAU** and the restricted **GC** alphabet with the minimal free energy algorithm with Freier's dataset (MFE(F)) and with Salser's dataset (MFE(S)), the maximum matching algorithm (MM) and the kinetical algorithm with Freier dataset (KIN).
We sampled approximately 100000 sequences for each chain length, alphabet and algorithm.

5.1 Comparison of Predicted Structures

Most people who use folding algorithms do not want to make structure statistics and fold millions of sequences, they want just to know the secondary structure of specific sequences. Furthermore they want an algorithm, that predicts the 'correct' secondary structure.
In order to get a feeling for the reliability of the algorithms we folded the same sequences with different algorithms and compared the predictions. It is thus possible to evaluate the influence of different parameter sets and of different algorithms on the outcome of the structure prediction. This is clearly a problem of enormous practical importance because the parameterset can only be approximately correct.
Figure 4 shows how many sequences fold into the same secondary structures with different algorithms. For reasons of clarity all comparisons with MM have been omitted because these structures have nothing in common with the secondary structures predicted by the other algorithms.

A finer method of measuring the differences in structure prediction is to calculate a distance for structures of the same sequence as obtained from different algorithms. We used the tree editing distance, the base pair distance and the alignment distance for comparing secondary structures (for details concerning the structure distances see appendix C). Let $D(f(x), g(x))$ denote the distance of the secondary structures of the sequence x for two different algorithms f and g. Then we define

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

where the average in the denominator is taken over pairs of randomly chosen sequences p and q. P_{fg} measures the average deviation of the predictions of the algorithms f and g; $P_{fg} = 0$ means their predictions agree perfectly, while $P_{fg} = 1$ indicates the the predictions are completely unrelated. Fig. 5, 6 and 7 show the dependence of this parameter on the chain length. As expected, the predictions agree less and less as the chain length increases. It is easily seen that the qualitative results are not changed by the different structure distances.

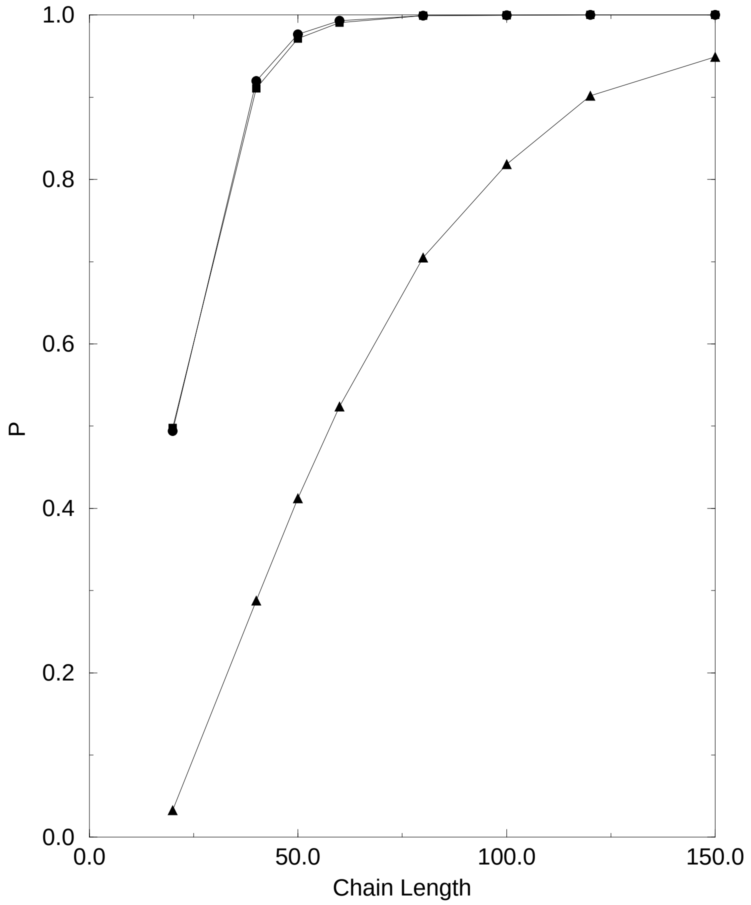


Figure 4: Fraction of sequences folded with different algorithms yielding identical structures, AUGC-alphabet

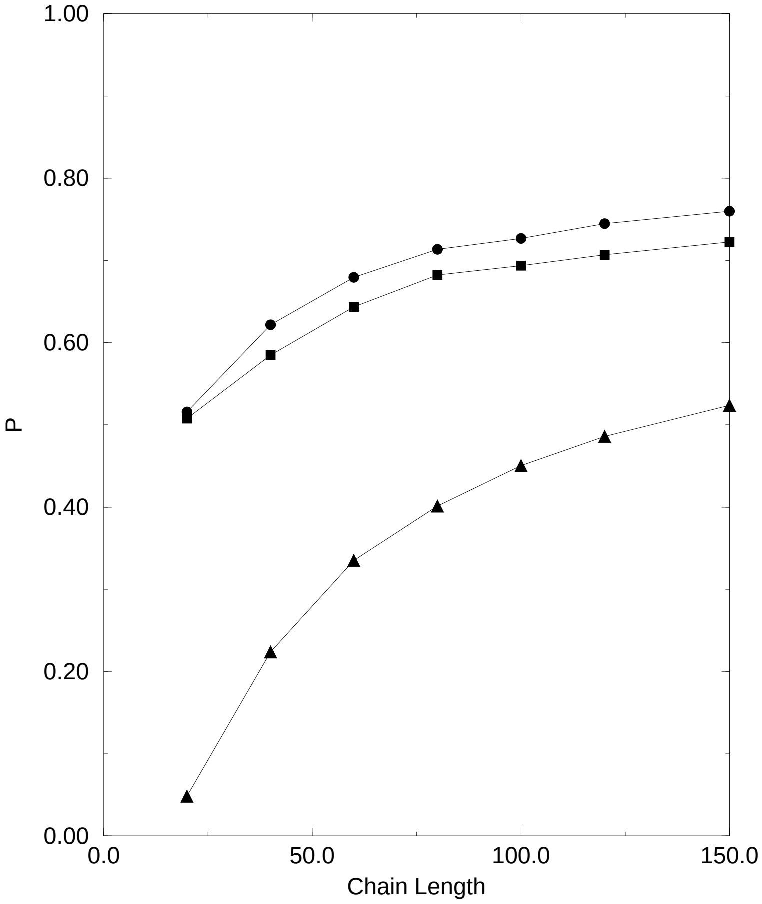


Figure 5: Mean base pair distance of sequences folded with different algorithms, AUGC-alphabet

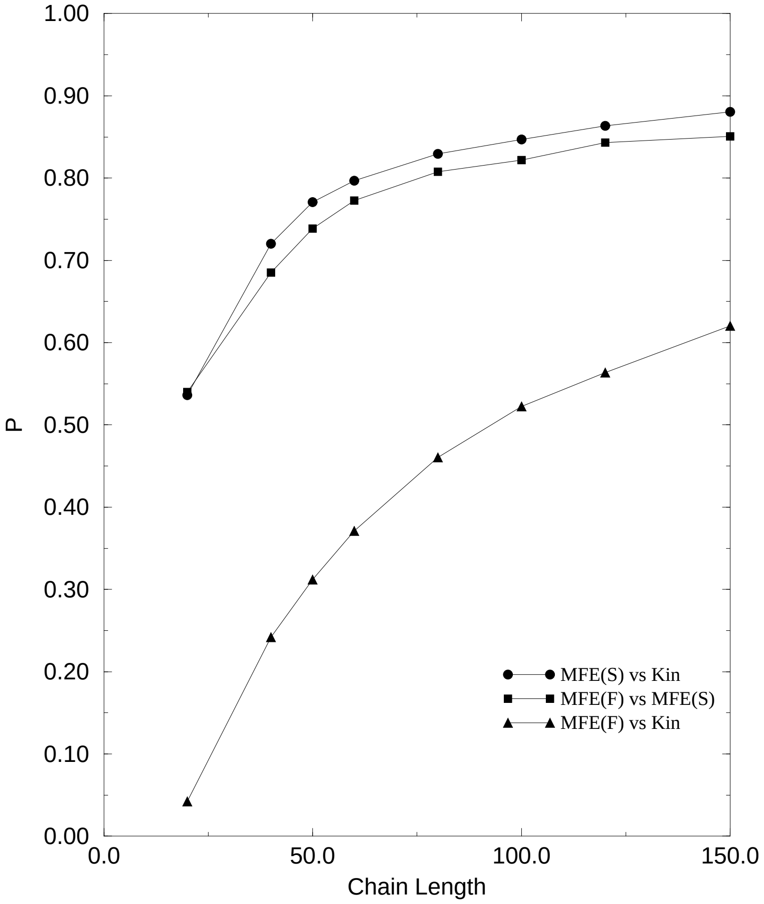


Figure 6: Mean tree edit distance of sequences folded with different algorithms, AUGC-alphabet

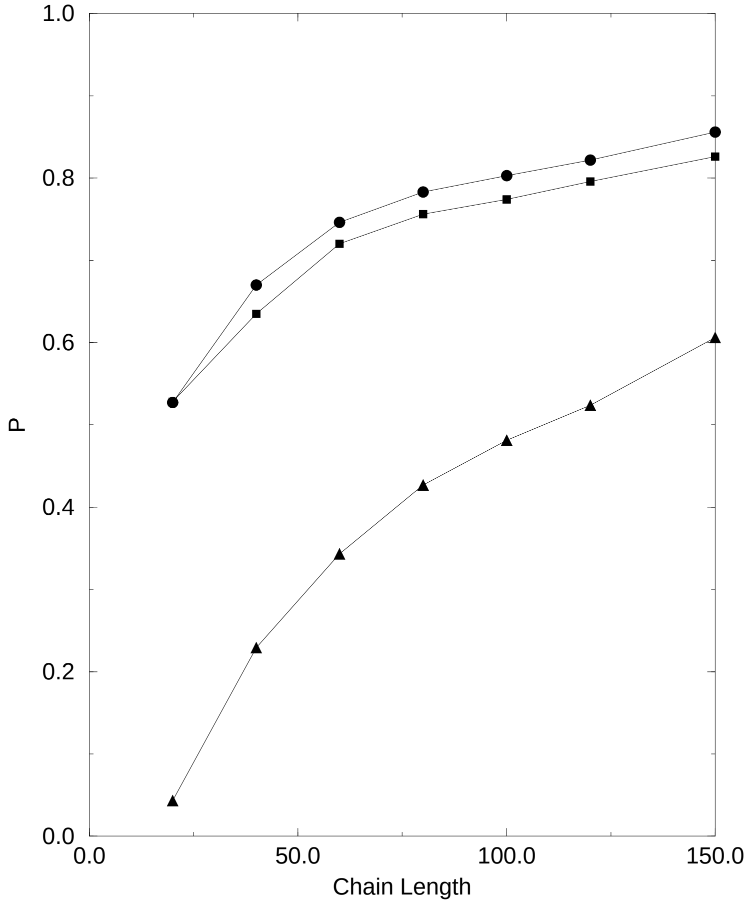


Figure 7: Mean alignment distance of sequences folded with different algorithms, AUGC-alphabet

MFE(F) vs MFE(S):
Even at chain lengths 20 only about 50% of the predicted structures are identical. For chain lengths greater than 50 the percentage of identically predicted structures is negligibly small.

MFE(F) vs KIN:
In contrast to the above comparison a much higher degree of identical structures is predicted, even sequences as long as 100 bases are folded in about 20% of the cases in the same structures.

MFE(S) vs KIN:
The percentage of identically folded sequences is a bit lower than in the first case.

MM vs all others:
Sequences folded by the maximum matching algorithm have nothing in common with the structures generated by the other algorithms.

From these results it can clearly be seen that the parameter set is much more important than the chosen algorithm for the prediction of secondary structures. Even such an algorithm as the kinetical algorithm in the most simple form predict structures with much more similarity to the minimum free energy structures when the same parameter set is used than structures which are produced by the same algorithm (`MFE`) but with a different parameter set.

A further important result is that the specific distance measure chosen does not have an influence on the qualitative results.

5.2 Distribution of Structures

For every algorithm 10000 strings of length $\nu = 100$ were folded into their secondary structures which were then ordered according to their frequencies of occurence. This ranking yields a distribution of secondary structures which follows

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

where r is the rank, and $f(r)$ is the frequency of the corresponding structure. a is a normalization constant and b can be interpreted as the number of “very frequent” structures.

This distribution follows **Zipf’s first law** which was found by Zipf in 1949 [91] when analysing literary texts. It states that ”if one takes the words making up an extended body of text and ranks them by frequency of occurrence, then the rank r multiplied by it’s frequency of occurrence $f(r)$ will be approximately a constant”. This law has attracted many researchers and a lot of papers on this subject has been published (for a list of references see [11]). The above formulation of Zipf’s first law was proposed by Mandelbrot [51].

We could find distributions following Zipf’s law for every algorithm, alphabet and chainlength. This distribution suggests that there exist a few structures which are very common and many structures which occur very rarely.

In fig. 8 and in fig. 9 the structure distribution for the **GCAU** and the **GC** alphabets are shown for the MFE(F), the MFE(S) and the KIN algorithm together with the best fit for Zipf’s first law.
The parameters obtained by curve fitting are shown in table 3.

The parameter b is a measure for the very abundant structures. The data in table 3 indicate that there are at least an order of magnitude more frequent structures for the **GC** alphabet than for the **GCAU** alphabet for all algorithms.

This is consistent with the fact that the ratio of structures to sequences is much larger for two letter alphabets. It is also interesting to note that the tail of the distribution is steeper for the kinetic algorithm than for the

	GCAU		**GC**	
	b	c	b	c
MFE(F)	70	1.72	997	2.66
MFE(S)	36	1.45	938	2.57
KIN	33	2.40	312	3.74

Table 3: Fits to Zipf's Law

minimum free energy algorithm. This indicates that many rare minimum free energy structures are not accessible at all for kinetic folding.

5.3 Statistics of Structure Elements

All statistically evaluated structure parameters for both alphabets and all algorithms are compiled in table 4 for **GCAU** sequences and in table 5 for **GC** sequences. For a recent review on secondary structure elements see [10].

The statistics of structural elements in large random samples of minimum free energy structures has been studies in great detail in [24]. A comparison of various alphabets can be found in [24, 21].

The significance of the statistical features of large ensembles of RNA molecules is highlighted by selection experiments in the test tube (see section 1.2.4) which start with as many as 10^{13} randomly generated sequences. The start distribution of RNA is therefore described by the statistical properties of random RNAs.

We investigate the dependence of structure statistics on the alphabet, the algorithm and the parameter set.

We compare the data from folding algorithms with a sample of random secondary structures that have the following properties:

ν	algo	#bp	#loops	#joints	#external	loopsz	stacksz	loop deg
20	MFE (F)	2.36	0.64	0.59	12.32	4.64	3.68	1.08
	MFE (S)	1.23	0.45	0.38	14.99	5.71	2.74	1.16
	KIN	2.31	0.61	0.59	12.40	4.87	3.77	1.04
30	MFE (F)	4.99	1.26	0.96	13.59	5.10	3.96	1.24
	MFE (S)	3.57	1.22	0.78	15.22	6.24	2.92	1.37
	KIN	4.62	1.09	0.95	14.42	5.82	4.23	1.13
	MM	9.40	4.91	1.19	1.53	1.97	1.92	1.76
40	MFE (F)	7.65	1.87	1.23	14.80	5.28	4.08	1.34
	MFE (S)	6.20	2.09	1.00	14.19	6.42	2.97	1.52
	KIN	6.83	1.52	1.21	16.62	6.38	4.48	1.21
50	MFE (F)	10.41	2.48	1.46	15.80	5.39	4.19	1.41
	MFE (S)	8.75	2.91	1.11	13.39	6.57	3.01	1.62
	KIN	9.16	1.96	1.44	18.30	6.81	4.67	1.27
	MM	17.01	8.14	1.24	1.44	1.79	2.09	1.85
	random	9.87	4.45	1.60	6.40	5.36	2.22	1.64
60	MFE (F)	13.08	3.08	1.67	17.00	5.47	4.24	1.46
	MFE (S)	11.39	3.72	1.14	12.32	6.70	3.06	1.69
	KIN	11.41	2.40	1.64	19.92	7.20	4.76	1.31
70	MFE (F)	15.91	3.71	1.87	17.89	5.47	4.29	1.50
	MFE (S)	14.01	4.51	1.16	11.24	6.81	3.10	1.74
	KIN	13.79	2.85	1.84	21.31	7.41	4.84	1.35
	MM	24.69	11.28	1.25	1.41	1.70	2.19	1.89
80	MFE (F)	18.67	4.32	2.05	18.98	5.49	4.33	1.52
	MFE (S)	16.46	5.21	1.17	10.96	6.94	3.16	1.77
	KIN	16.09	3.28	2.02	22.86	7.60	4.90	1.39
90	MFE (F)	21.67	4.98	2.22	19.40	5.48	4.35	1.55
	MFE (S)	19.07	5.99	1.18	10.51	6.90	3.18	1.80
	KIN	18.58	3.75	2.19	23.56	7.81	4.96	1.41
100	MFE (F)	24.40	5.57	2.37	20.62	5.49	4.38	1.58
	MFE (S)	21.51	6.68	1.18	10.77	6.92	3.22	1.82
	KIN	20.81	4.16	2.34	25.16	7.98	5.00	1.44
	MM	36.35	15.91	1.27	1.40	1.63	2.28	1.92
	random	20.70	9.31	1.85	6.70	5.57	2.22	1.80
120	MFE (F)	30.22	6.85	2.67	21.84	5.51	4.41	1.61
	MFE (S)	26.80	8.18	1.20	10.09	6.89	3.28	1.85
	KIN	25.53	5.05	2.66	27.13	8.29	5.06	1.47
150	MFE (F)	38.70	8.70	3.09	24.42	5.54	4.41	1.61
	MFE (S)	34.60	10.36	1.20	9.69	6.86	3.34	1.88
	KIN	32.35	6.32	3.08	30.61	8.66	5.12	1.51
	random	31.51	14.18	1.94	6.80	5.65	2.22	1.86
200	MFE (F)	53.67	11.94	3.67	27.09	5.49	4.49	1.69
	MFE (S)	47.63	13.97	1.22	9.38	6.82	3.41	1.91
	KIN	43.95	8.42	3.73	35.25	9.12	5.21	1.56
	random	42.31	19.04	1.99	6.91	5.69	2.22	1.90

Table 4: Structure Statistic of **GCAU** sequences

ν	algo	#bp	#loops	#joints	#external	loopsz	stacksz	loop deg
20	MFE (F)	5.73	2.00	1.25	4.24	3.61	4.62	1.37
	MFE (S)	5.21	1.33	1.00	3.62	4.48	3.91	1.25
	KIN	4.89	1.04	1.01	5.40	4.65	4.70	1.03
	MM	6.39	2.82	1.09	1.29	2.11	2.27	1.61
30	MFE (F)	9.28	2.00	1.25	4.24	3.61	4.62	1.37
	MFE (S)	9.08	2.09	1.11	3.57	3.96	4.35	1.47
	KIN	8.32	1.58	1.23	5.97	4.68	5.28	1.22
	MM	10.55	4.08	1.15	1.17	1.90	2.58	1.72
40	MFE (F)	13.16	2.66	1.43	4.32	3.52	4.95	1.46
	MFE (S)	13.07	2.81	1.19	3.35	3.74	4.66	1.58
	KIN	11.90	2.11	1.44	6.18	4.75	5.65	1.32
50	MFE (F)	17.09	3.31	1.58	4.37	3.45	5.16	1.52
	MFE (S)	16.99	3.51	1.24	3.27	3.64	4.85	1.65
	KIN	15.47	2.62	1.58	6.42	4.82	5.90	1.40
	MM	19.01	6.39	1.19	1.12	1.70	2.98	1.81
	random	11.01	4.84	1.57	5.12	4.71	2.27	1.68
60	MFE (F)	21.10	3.96	1.69	4.38	3.39	5.33	1.57
	MFE (S)	21.01	4.16	1.26	3.21	3.55	5.05	1.70
	KIN	19.13	3.14	1.70	6.51	4.85	6.10	1.46
70	MFE (F)	25.08	4.60	1.78	4.41	3.35	5.45	1.61
	MFE (S)	24.99	4.83	1.28	3.11	3.50	5.17	1.73
	KIN	22.73	3.64	1.78	6.64	4.92	6.25	1.51
	MM	27.55	8.62	1.20	1.07	1.60	3.20	1.86
80	MFE (F)	29.11	5.25	1.88	4.38	3.32	5.54	1.64
	MFE (S)	29.07	5.49	1.29	3.06	3.43	5.30	1.76
	KIN	26.36	4.15	1.88	6.77	4.94	6.35	1.55
	MM	31.88	9.65	1,21	1.07	1.57	3.30	1.87
90	MFE (F)	33.11	5.88	1.96	4.44	3.29	5.63	1.67
	MFE (S)	33.13	6.15	1.31	3.00	3.37	5.39	1.79
	KIN	30.00	4.66	1.96	6.85	4.97	6.44	1.58
100	MFE (F)	37.21	6.51	2.02	4.42	3.25	5.72	1.69
	MFE (S)	37.12	6.81	1,32	2.97	3.35	5.45	1.81
	KIN	33.71	5.16	2.01	6.86	4.99	6.54	1.61
	MM	40.46	11.80	1.21	1.07	1.53	3.43	1.90
	random	22.91	10.10	1.79	5.34	4.84	2.27	1.82
120	MFE (F)	45.30	7.77	2.11	4.44	3.21	5.83	1.73
	KIN	40.91	6.13	2.15	7.19	5.05	6.67	1.65
	MM	49.10	13.88	1.22	1.06	1.49	3.54	1.91
150	MFE (F)	57.48	9.64	2.24	4.55	3.16	5.96	1.77
	MFE (S)	57.46	9.99	1.39	2.93	3.22	5.75	1.86
	KIN	51.89	7.64	2.29	7.30	5.10	6.79	1.70
	MM	62.04	17.04	1.22	1.06	1.46	3.64	1.93
	random	34.80	15.34	1.87	5.44	4.89	2.27	1.88
200	MFE (F)	77.96	12.73	2.39	4.59	3.10	6.12	1.81
	MFE (S)	77.82	13.14	1.40	2.91	3.15	5.92	1.89
	KIN	70.22	10.12	2.48	7.60	5.14	6.94	1.76
	MM	83.77	22.04	1.24	1.05	1.43	3.80	1.94
	random	46.71	20.60	1.90	5.47	4.91	2.27	1.91

Table 5: Structure Statistic of **GC** sequences

- All stacks have length of at least two, that means there are no isolated base pairs
- All hairpin loops contain at least three unpaired bases
- They have stickiness **P** corresponding to **GCAU** or **GC** alphabets.

A discussion of the statistical properties of secondary structure graphs without the restriction of being a minimum free energy or a maximum matching structure can be found in [36].

Mean Free Energies

Due to the higher binding energies of the **GC** pair, **GC** sequences are more stable than the **AUGC** sequences. The mean free energies scale linearly with ν independent of the algorithm and the parameter set (see fig. 10).

Mean Number of Base Pairs

All algorithms and parameter sets how a linear increase of the number of base pairs with ν, the MFE(F) structures having most base pairs, the kinetical structures the least (see fig. 11 and fig. 12). The data from the random sample show (see tables 4 and 5) that secondary structures as predicted from folding algorithms do not have much more base pairs than a random compatible structure in case of the **GCAU** alphabet. For **GC** sequences however, the energy minimization forces a drastic increase in the number of base pairs.

Mean Number of Loops and Stacks

The number of loops must equal the number of stacks because every loop must be closed by a stack. For all three algorithms the number of loops and stacks scale linearly with ν (see fig. 13 and 14). The kinetical structures contain less loops than the MFE structures. The number of loops and stacks is significantly smaller in folded structures than in random compatible structures, this indicates that folding favours larger structure elements.

Mean Loop Size

AUGC alphabet: For the MFE structures the average mean loop size converges even at moderate chain lengths (60) rapidly to a constant value. The kinetical structures show greater loop sizes and even at a chain length of 200 no convergence is reached (fig. 15). **GC** alphabet: For the kinetic structures a slight increase of the mean loop size is observed whereas for the MFE structures the mean loop size even decreases with chain length ν (fig. 16).

Mean Loop Degree

A hairpinloop has degree 1, an internal loop and a bulge degree 2 and a multiloop a degree greater than 2. The actual value of the mean loop degree can assume values between 1 and 2. This is due to the fact, that the lower bound is provided by structures consisting solely of hairpins, the mean loop degree of which is 1. On the other hand a structure consisting just of one big multiloop must contain at least $(n-1)$ hairpins (n is the degree of the multiloop and the number of loops as well). In this case the mean loop degree is $\frac{n+n-1}{n}$, with a limes of 2 with $n \to \infty$ The average loop degree converges to a constant value with increasing ν, the loop degree being highest for `MFE(S)` and lowest for the kinetic structures (fig. 17 and fig. 18).

Mean Stack Size

The mean stack size approaches a constant value at fairly small chainlengths ($\nu \sim 50$). The kinetical structures have larger mean stack size than the MFE structures (see fig. 19 and fig. 20).

Number of Joints and Components

A secondary structure consists of one, two or more components, which are connected by joints; hence $N_j = N_C - 1$. The number of joints is nearly indistinguishable for both alphabets for the MFE(F) and KIN. Both of them show an increase with ν which is slower than linear. It is not known yet, whether the number of joints becomes constant for very long chains ($\nu >$

1000). For Salser's parameter set we find that the number of joints rapidly settles down to a constant value (see fig. 21 and fig. 22).

Number of External Digits

Nucleotides in joints and free ends, that are all bases, which are neither in a stack or in a loop, are termed external digits. In this case the situation is similar to the number of joints. MFE(F) and the kinetic algorithm scale slightly less than linear with chain length ν for both alphabets, whereas MFE(S) decreases at small chain lengths and converges then to a constant value (fig. 23 and fig. 24).

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

In general KIN gives structures which are less stable, with larger loop size and larger stack size than MFE(F).

For the Salser data set the destabilizing effect of loops is less pronounced giving structures with less base pairs, more loops (which are larger and more branched as in the Freier data set), smaller mean stack size and the number of external digits is minimized.

Nevertheless the qualitative features of the structure statistics of secondary structure are the same for all algorithms and dataset, except for possibly the mean loop sizes. Even the quantitive difference between the algorithms are in most cases strikingly small.

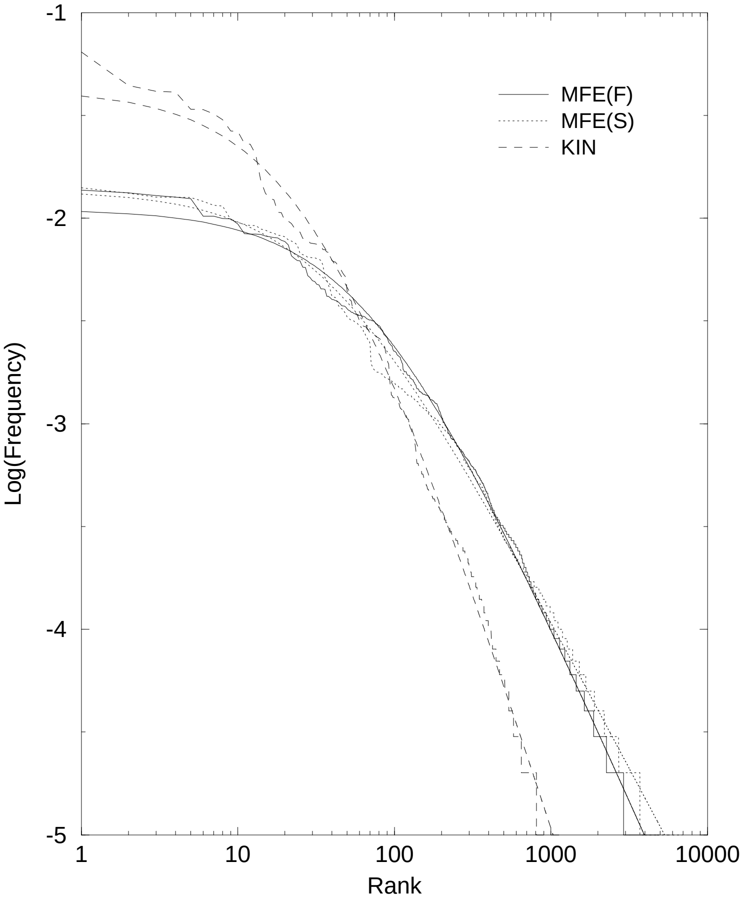


Figure 8: Zipf's first law for coarse grained (loop) RNA structures, **GCAU** sequences, $\nu = 100$

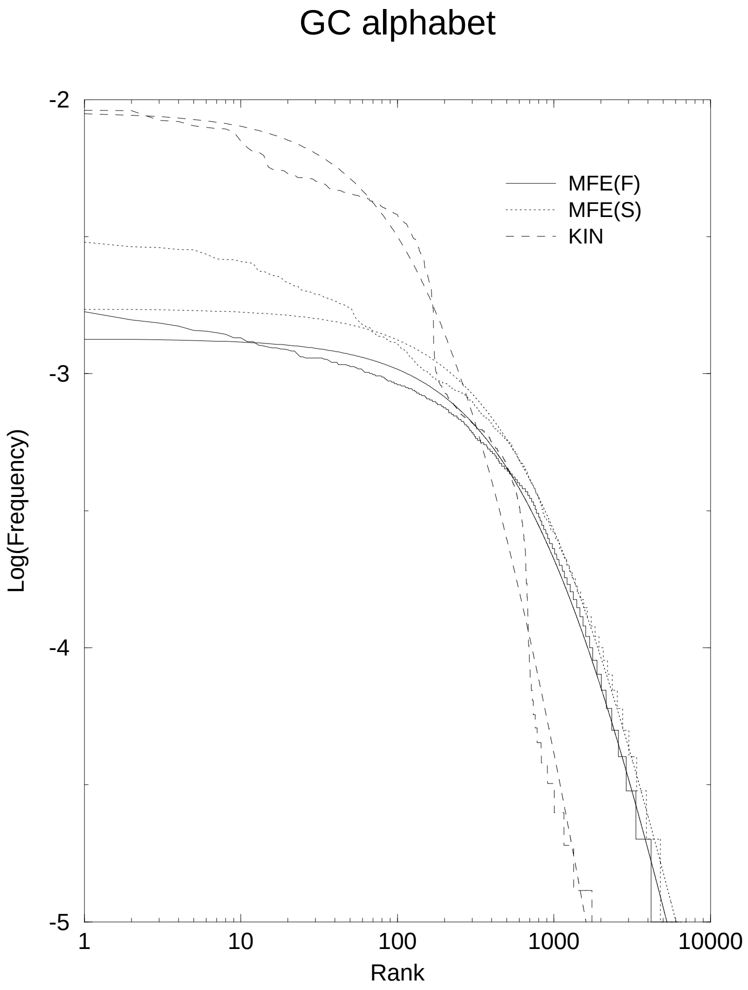


Figure 9: Zipf's first law for coarse grained (loop) RNA structures **GC** sequences, $\nu = 100$

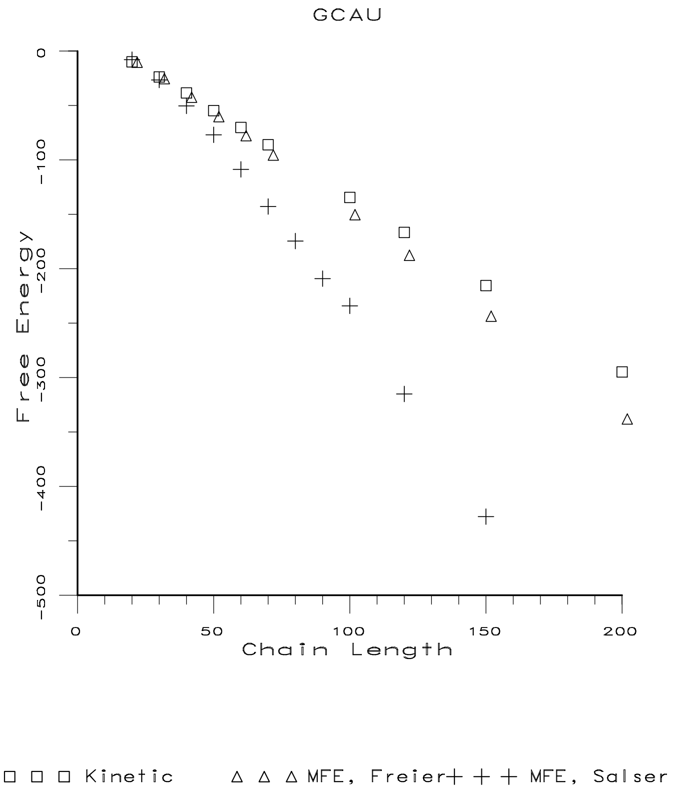


Figure 10: The mean free energies as a function of chain length for **GCAU**

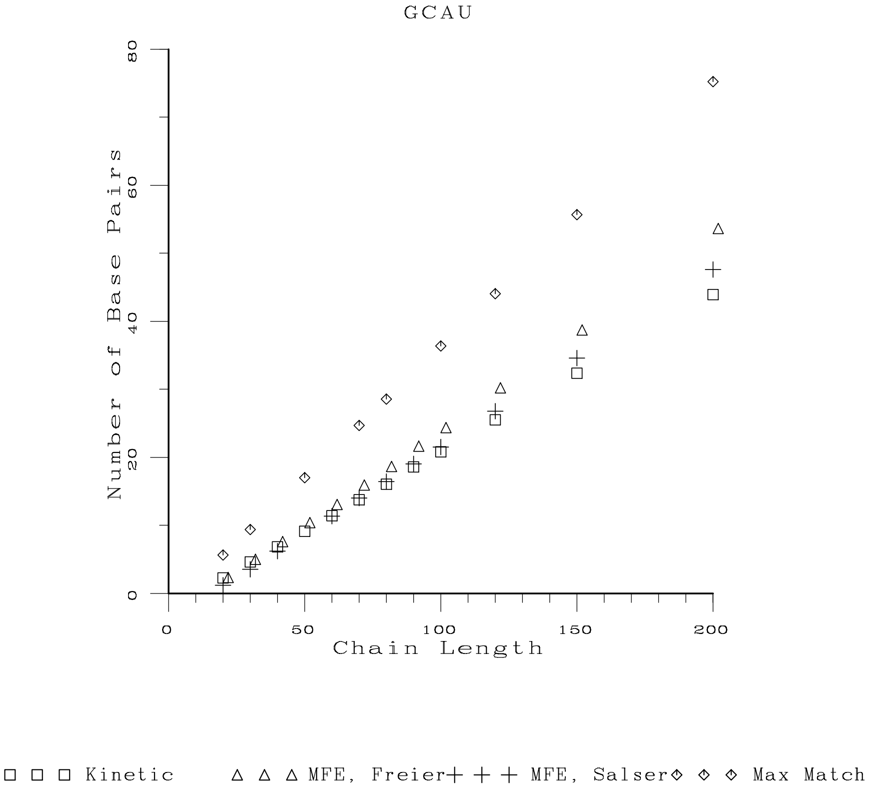


Figure 11: The mean number of base pairs as a function of chain length for **GCAU**

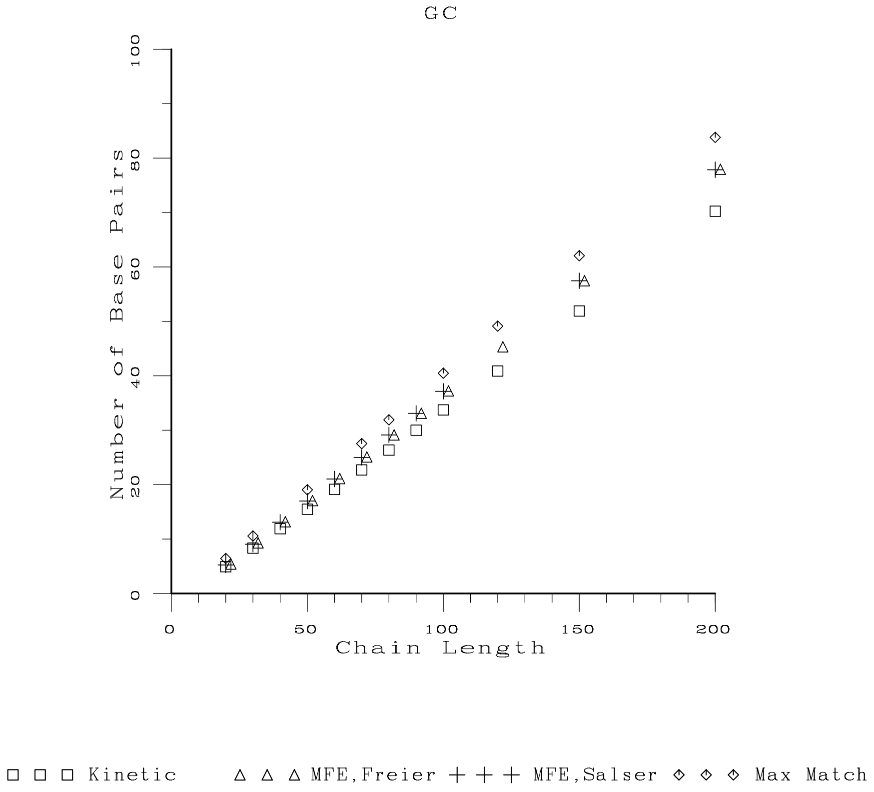


Figure 12: The mean number of base pairs as a function of chain length for **GC**

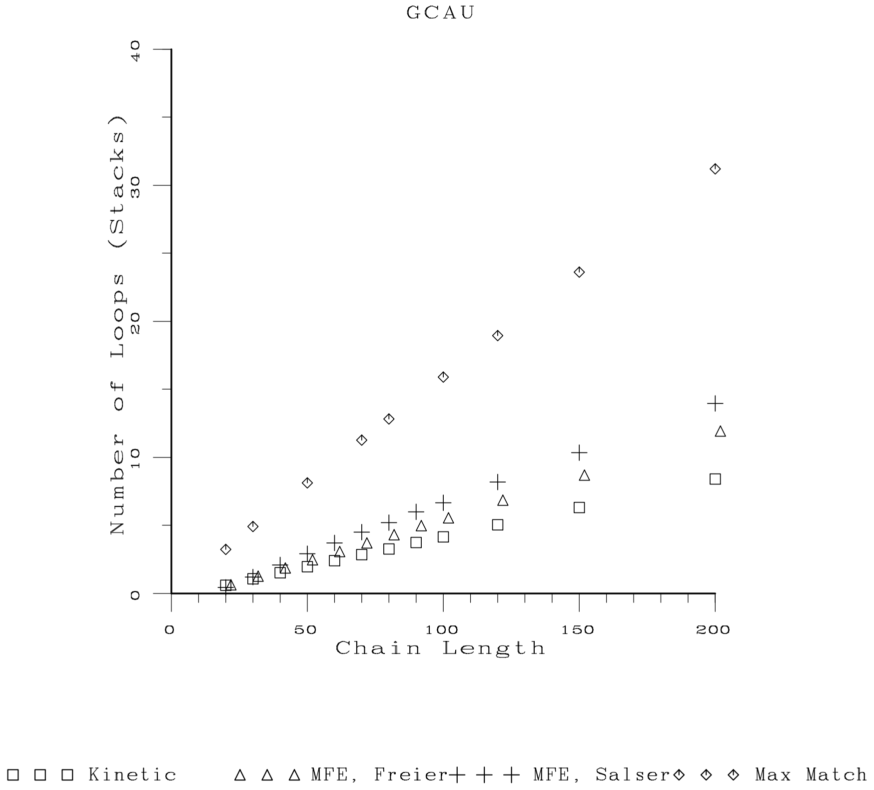


Figure 13: The mean number of loops (stacks) as a function of chain length for **GCAU**

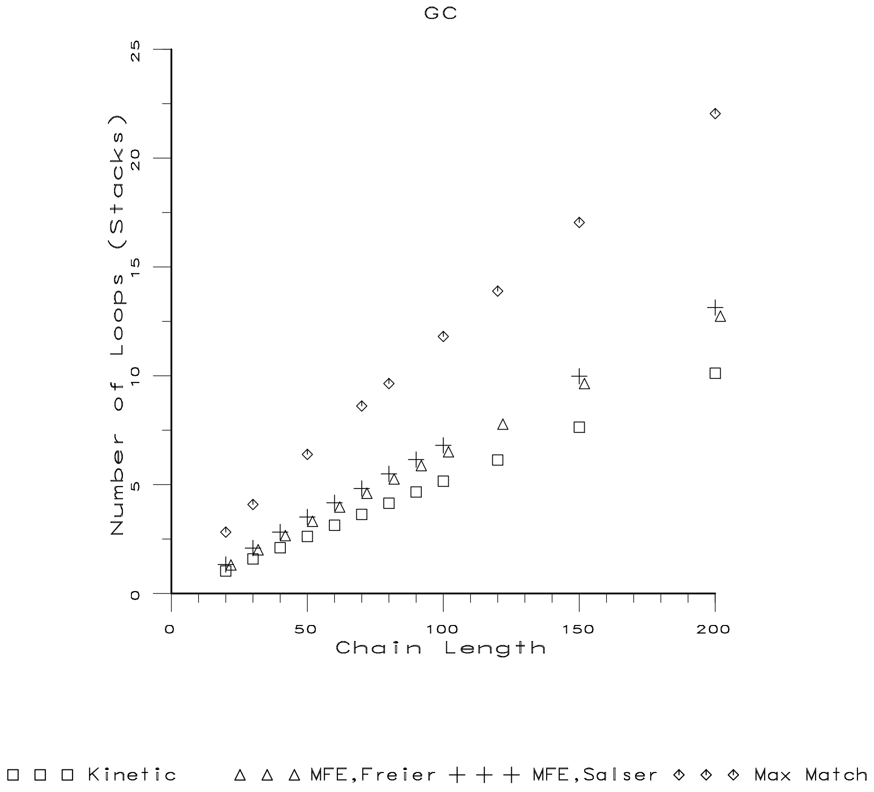


Figure 14: The mean number of loops (stacks) as a function of chain length for **GC**

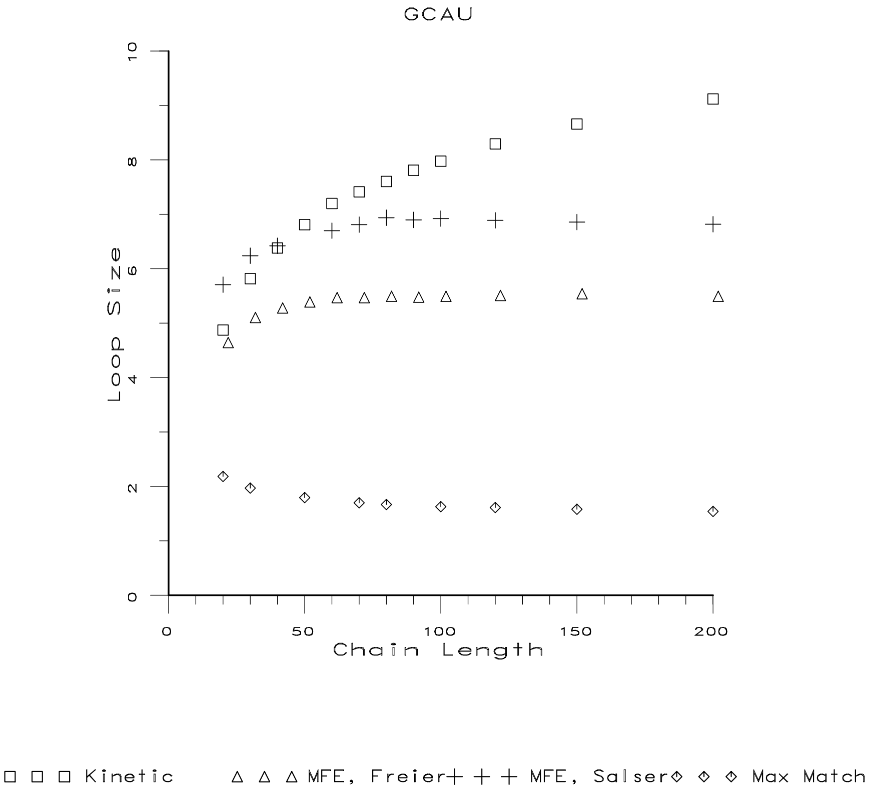


Figure 15: The mean loopsize as a function of chain length for **GCAU**

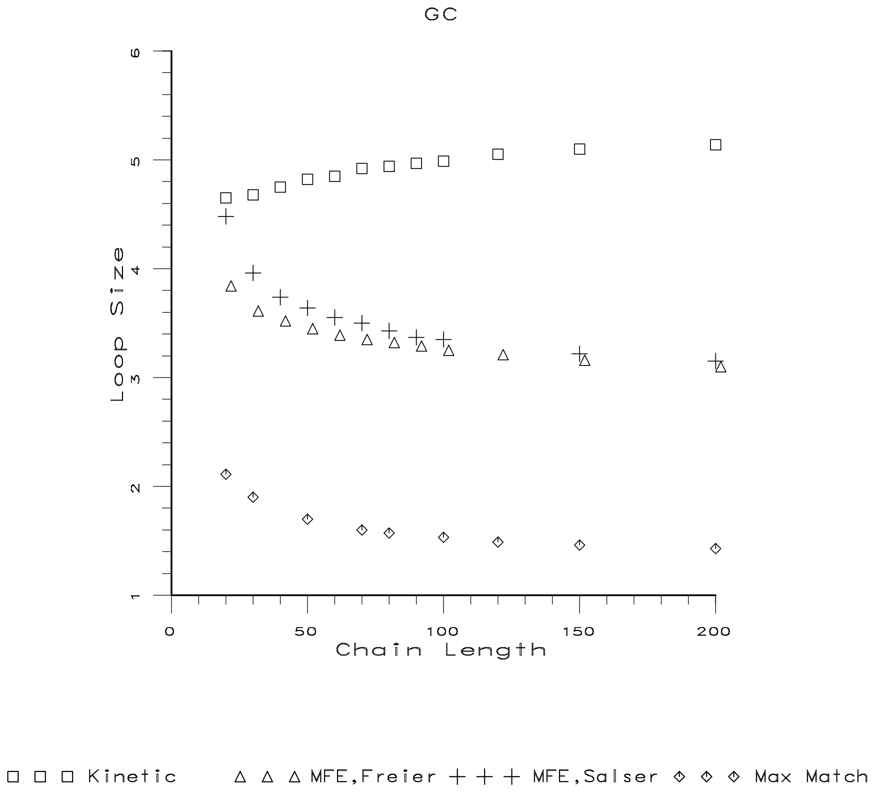


Figure 16: The mean loopsize as a function of chain length for **GC**

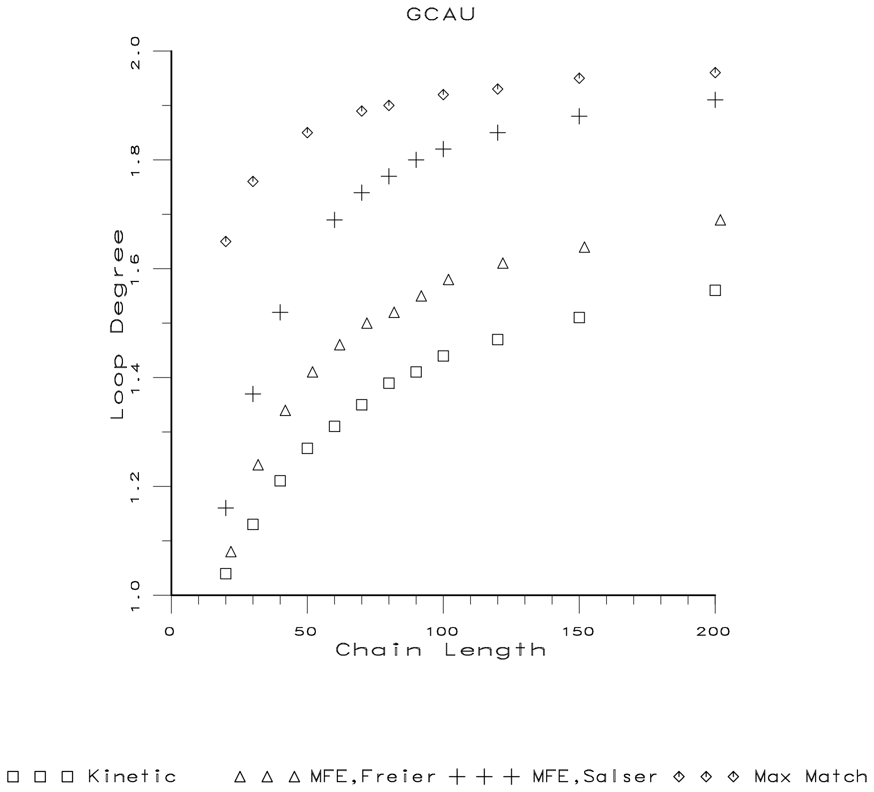


Figure 17: The mean loopdegree as a function of chain length for **GCAU**

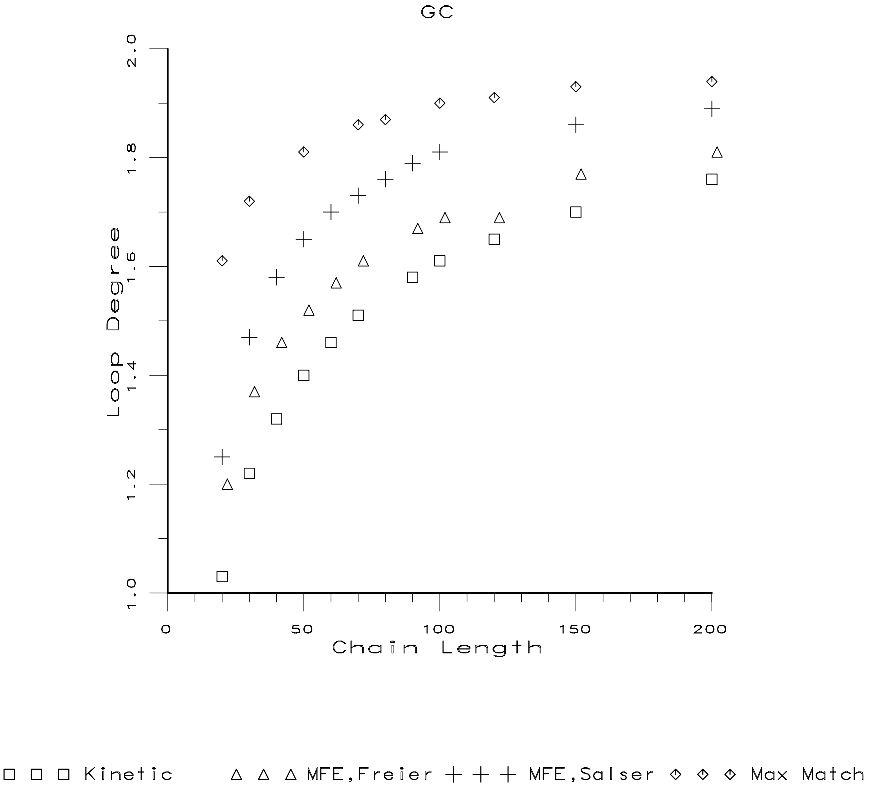


Figure 18: The mean loopdegree as a function of chain length for **GC**

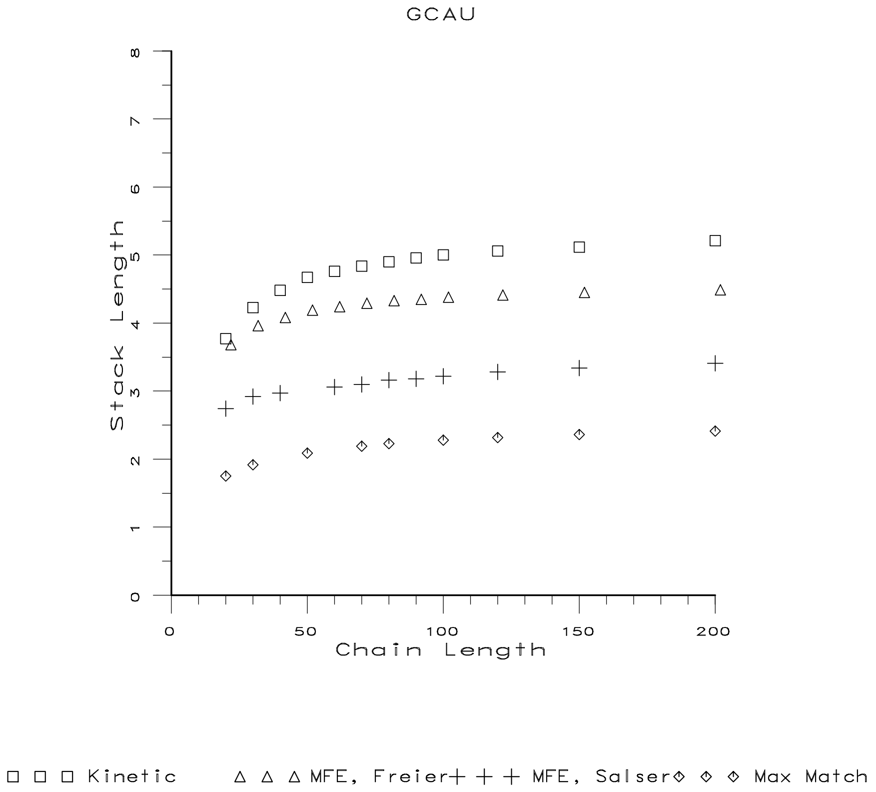


Figure 19: The mean stacksize as a function of chain length for **GCAU**

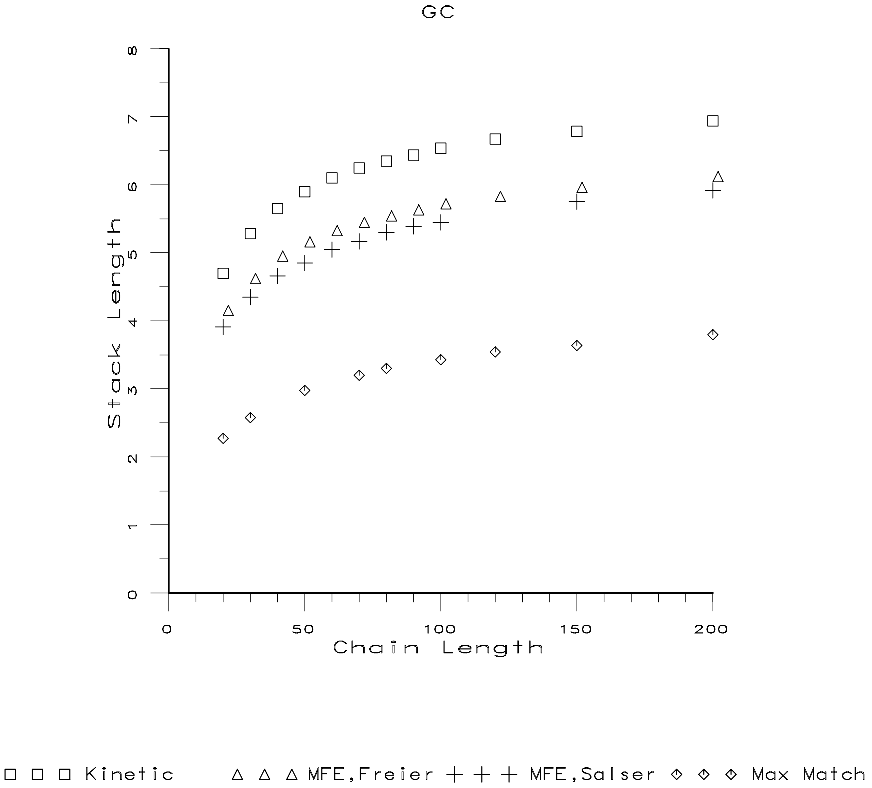


Figure 20: The mean stacksize as a function of chain length for **GC**

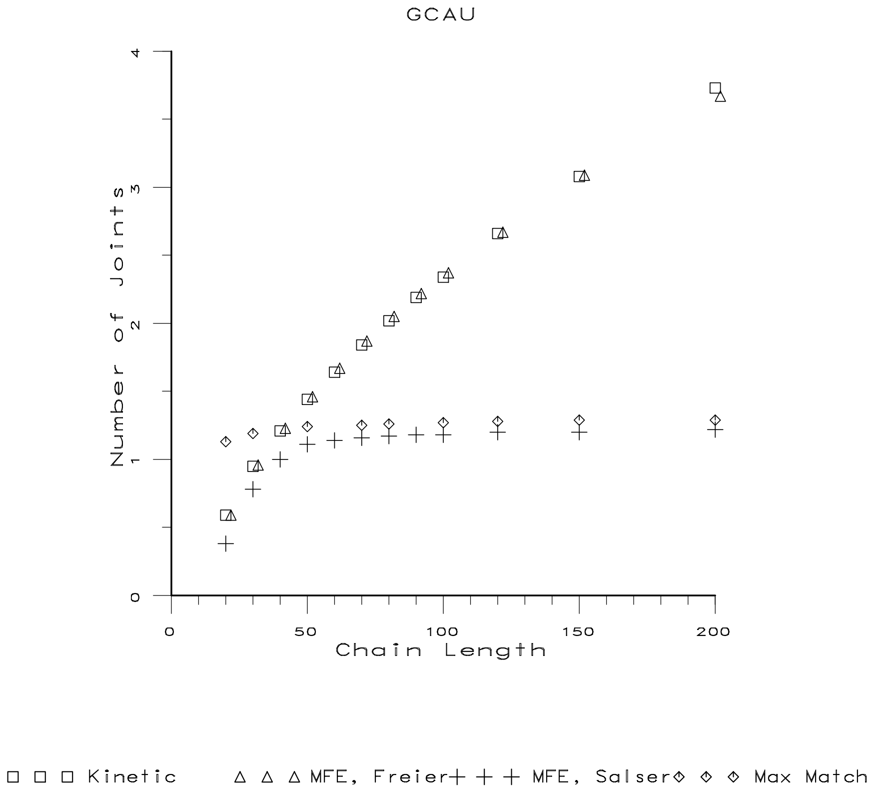


Figure 21: The mean number of joints as a function of chain length for **GCAU**

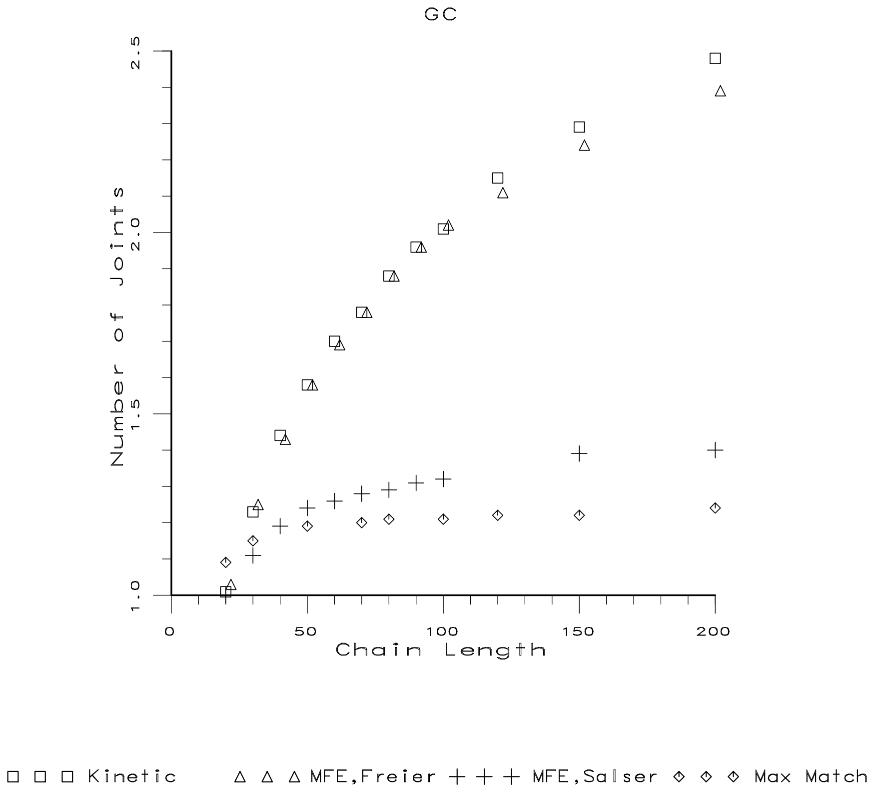


Figure 22: The mean number of joints as a function of chain length for **GC**

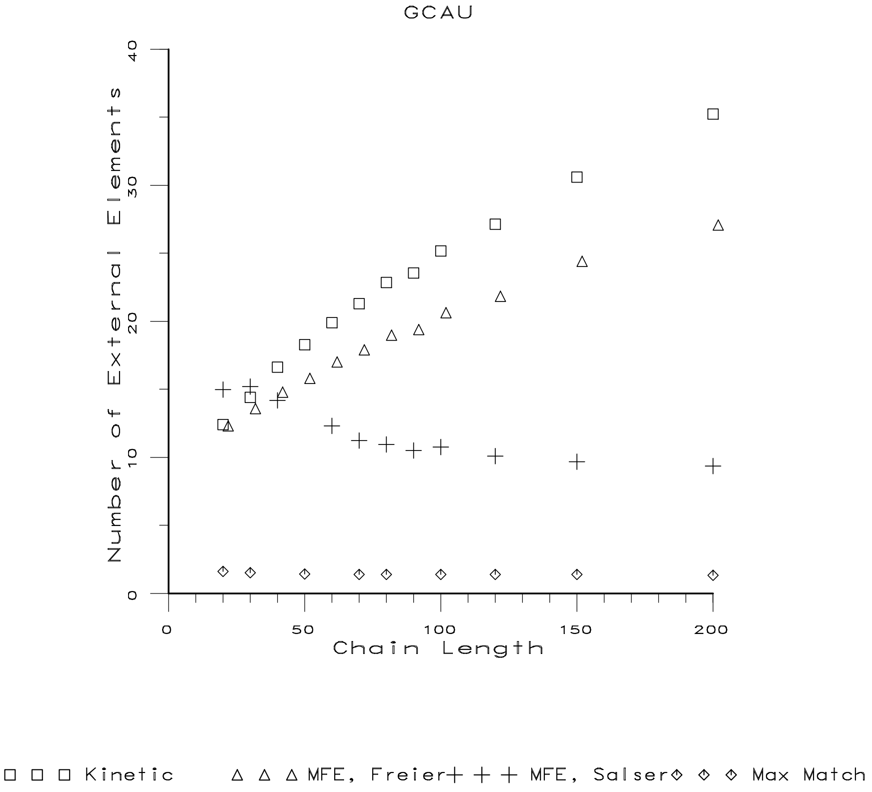


Figure 23: The mean number of external digits as a function of chain length for **GCAU**

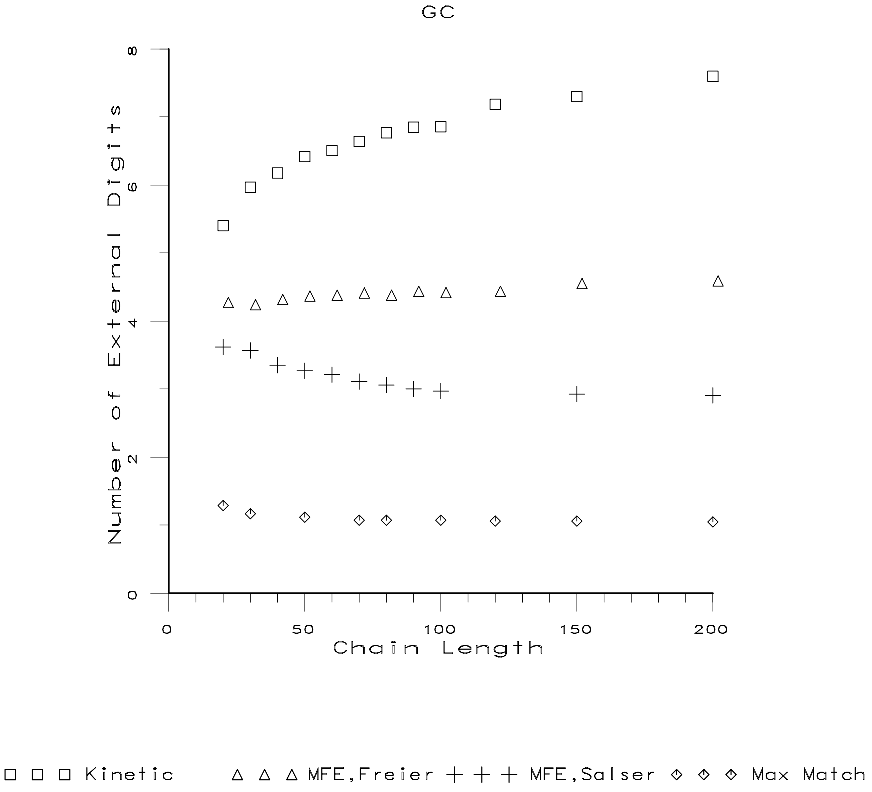


Figure 24: The mean number of external digits as a function of chain length for **GC**

6 Autocorrelation

The autocorrelation function $\rho(d)$ is a measure for the average similarity of structures as a function of the Hamming distance of their underlying sequences. For the mathematical details see appendix B.

We calculated the autocorrelation functions $\rho(d)$ on the free enery landscape and the combinatory map of secondary structure generated by the MFE(F), MFE(S), KIN and MM algorithms. Fig 25 shows as an example the autocorrelation functions of the free energy landscape, and fig. 26 shows the autocorrelation function of the CM of secondary structures for the **GCAU** alphabet.

In all cases the autocorrelation functions can be approximated by an exponential function and the correlation lengths determined as the Hamming distance at which $\rho(d)$ has decayed to a value of $\frac{1}{e}$.

Table 6 shows the energy autocorrelation lengths l_E and the structure autocorrelation lengths l_S for the **GCAU** and **GC** alphabets. For a graphical representation of the data see fig. 27 and fig. 29 for the energy autocorrelation lengths and fig. 28 and fig. 28 for the structure autocorrelation lengths. For reasons of comparison we show the correlation lengths computed with the partition function algorithm [55, 5] as well. These data are published in [5].

Larger correlation lengths imply smoother landscapes, i.e. fewer local optima. As a consequence, evolutionary optimization is easier on these landscapes [22, 24].

It is a general feature of RNA landscapes that the correlation length l_E of the energy landscape is much larger than the correlation length l_S of the corresponding combinatory map of structures. Nevertheless in both cases we observe a linear scaling of the correlation lengths with ν with the

		GCAU		**GC**	
ν	algo	l_E	l_S	l_E	l_S
20	MFE(F)	3.6	3.0	2.0	1.6
	MFE(S)	3.8	3.1	2.2	1.7
	kinetic	3.4	2.9	1.9	1.5
	mm	4.8	3.1	2.5	2.0
30	MFE(F)	5.7	4.1	3.1	1.9
	MFE(S)	6.7	4.3	3.4	2.1
	kinetic	5.1	3.8	2.5	1.8
	mm	8.3	3.7	4.1	2.5
40	MFE(F)	8.2	4.7	3.6	2.2
	MFE(S)	9.9	4.6	4.2	2.4
	kinetic	7.1	4.4	2.7	2.0
	mm	11.4	4.0	5.5	2.6
50	MFE(F)	10.6	5.3	4.8	2.4
	MFE(S)	13.8	5.1	5.6	2.5
	kinetic	8.9	5.2	3.4	2.3
	mm	15.1	4.1	6.7	2.8
60	MFE(F)	12.4	5.8	5.5	2.6
	MFE(S)	18.4	5.7	6.4	2.9
	kinetic	10.1	5.7	3.8	2.6
	mm	19.4	4.4	8.7	2.9
70	MFE(F)	16.2	6.1	6.6	2.8
	MFE(S)	20.8	5.9	7.3	3.1
	kinetic	12.5	6.3	4.2	2.7
	mm	22.5	4.7	9.6	3.2
80	MFE(F)	18.6	6.8	7.0	2.9
	MFE(S)	23.8	6.4	8.2	3.3
	kinetic	14.3	7.0	5.1	2.9
	mm	25.2	4.6	11.9	3.2
100	MFE(F)	22.3	7.3	9.2	3.3
	MFE(S)	27.9	7.1	9.6	3.8
	kinetic	17.0	7.8	5.7	3.4
	mm	33.3	4.7	13.9	3.5

Table 6: Autocorrelation Length for GCAU and GC for different algorithms

AUGC alphabet having the largest correlation lengths for both cases. The maximum matching algorithm qields the highest value of l_E, followed by MFE(S), MFE(F) and KIN. This means that for KIN the energy landscape is much more rugged than for the other algorithms - neighboring sequences are more likely to have larger differences in their energy values.

The situation is different for the correlation lengths of the combinatory map of structures. For the **GCAU** alphabet, the MFE(F), the MFE(S) and KIN show approximately the same behavior with the kinetic algorithm having the steepest slope. The maximum matching algorithm has a much smaller structure correlation length.

For the **GC** alphabet l_S is in general very small indicating a very rugged landscape for all algorithms - neighboring structures often have very different secondary structures.

For both alphabets the partition function algorithm shows higher l_S values. Ensemble averages are much more conserved than the single structures predicted by the different algorithms.

The essence of these results is that the biologically more important combinatory map of structures is rather independent of the chosen algorithm. This means that results obtained from a specific algorithm (e. g. MFE(F)) should hold for other algorithms as well.

This is especially important in the light of new results which showed that it is possible to cross through whole of sequence space without ever changing the secondary structure (see 7).

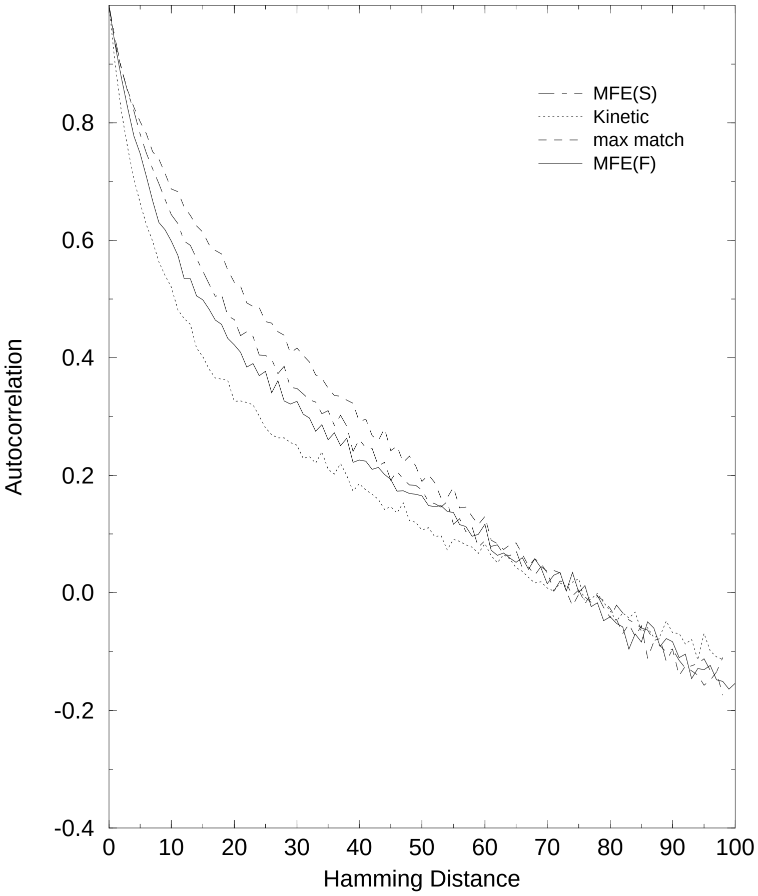


Figure 25: Energy autocorrelation functions for different algorithms, **AUGC** alphabet, $\nu = 100$

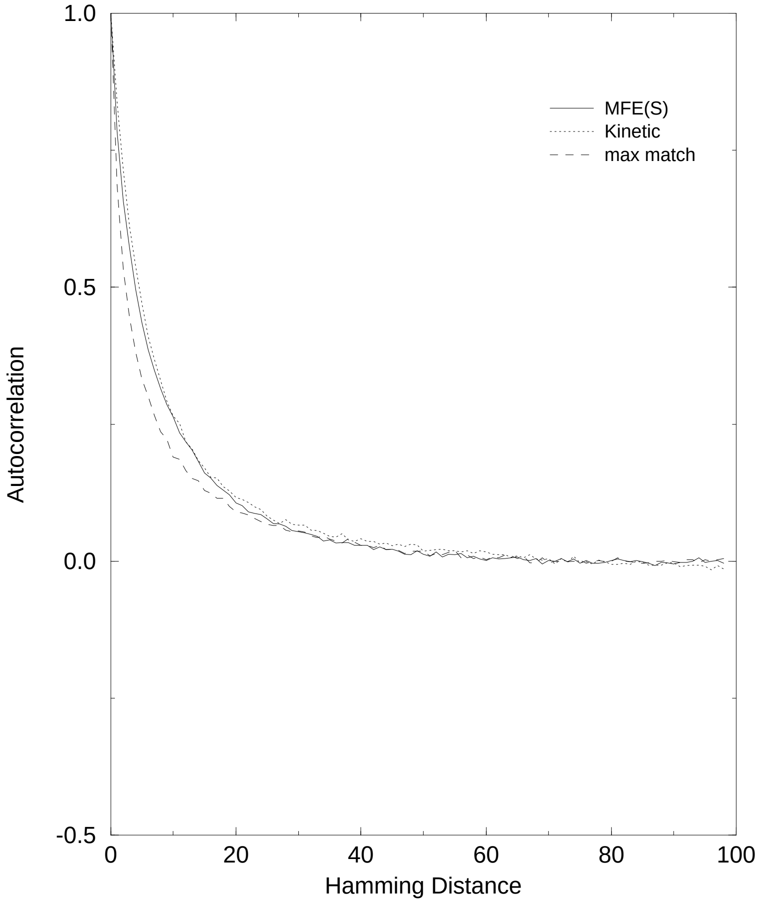


Figure 26: Autocorrelation functions for the CM of secondary structures for different algorithms, **AUGC** alphabet, $\nu = 100$

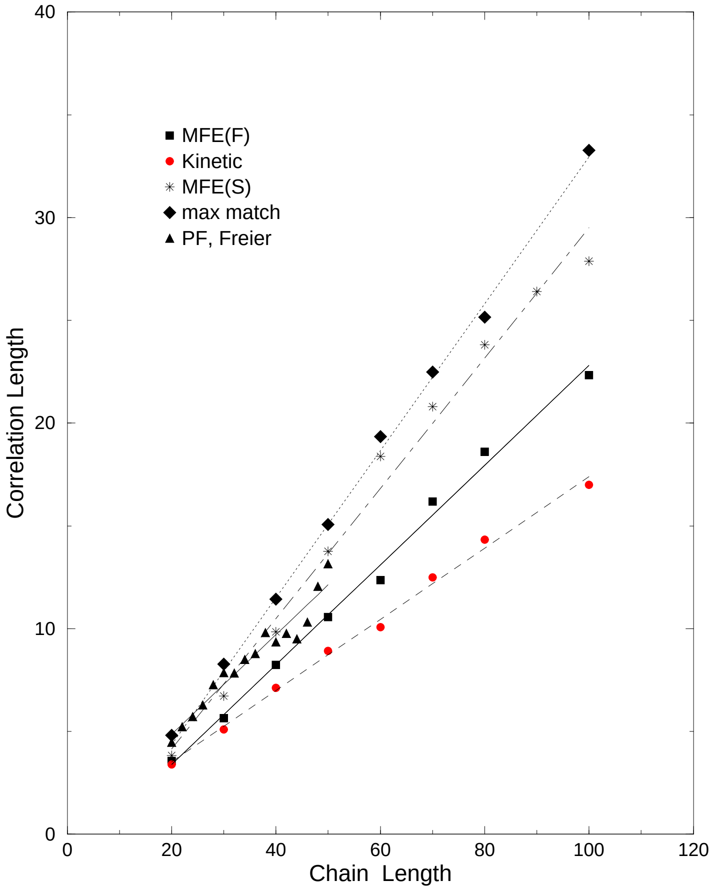


Figure 27: Energy autocorrelation lengths for different algorithms, **AUGC** alphabet

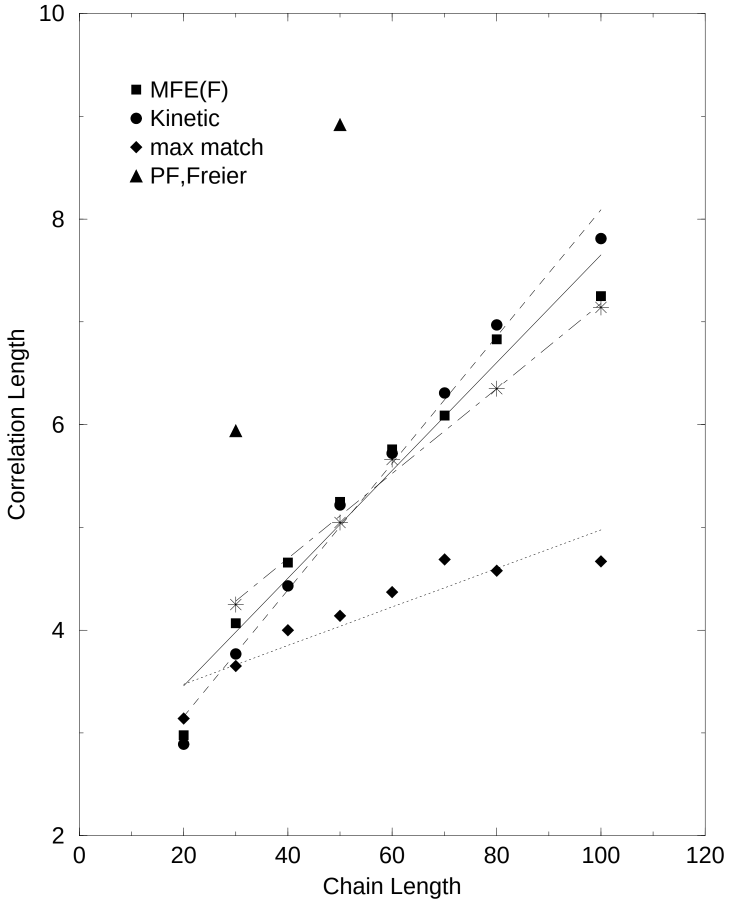


Figure 28: Structure autocorrelation lengths for different algorithms, **AUGC** alphabet

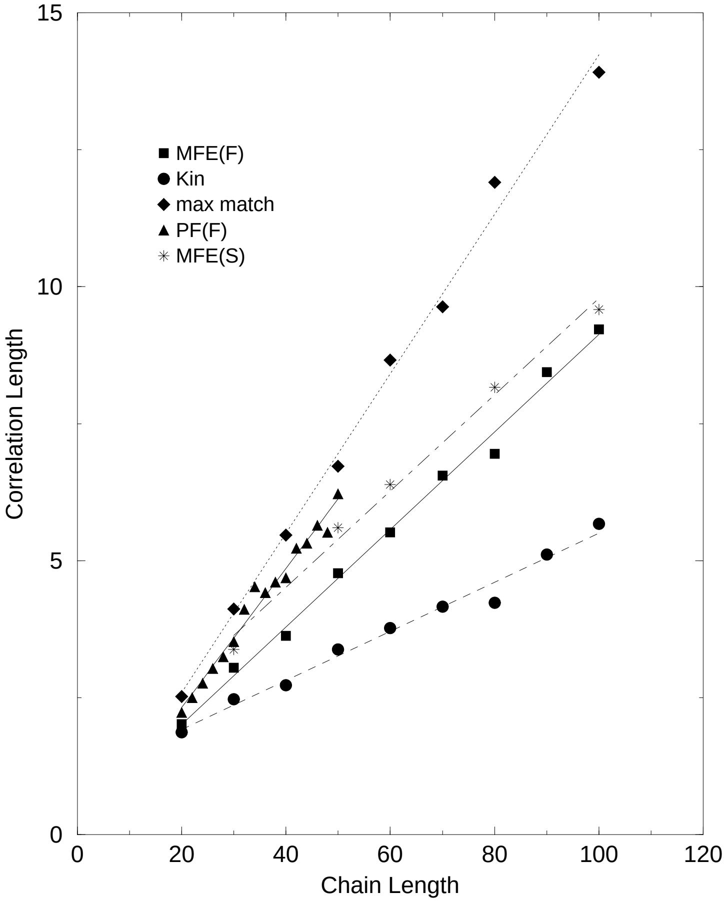


Figure 29: Energy autocorrelation lengths for different algorithms, **GC** alphabet

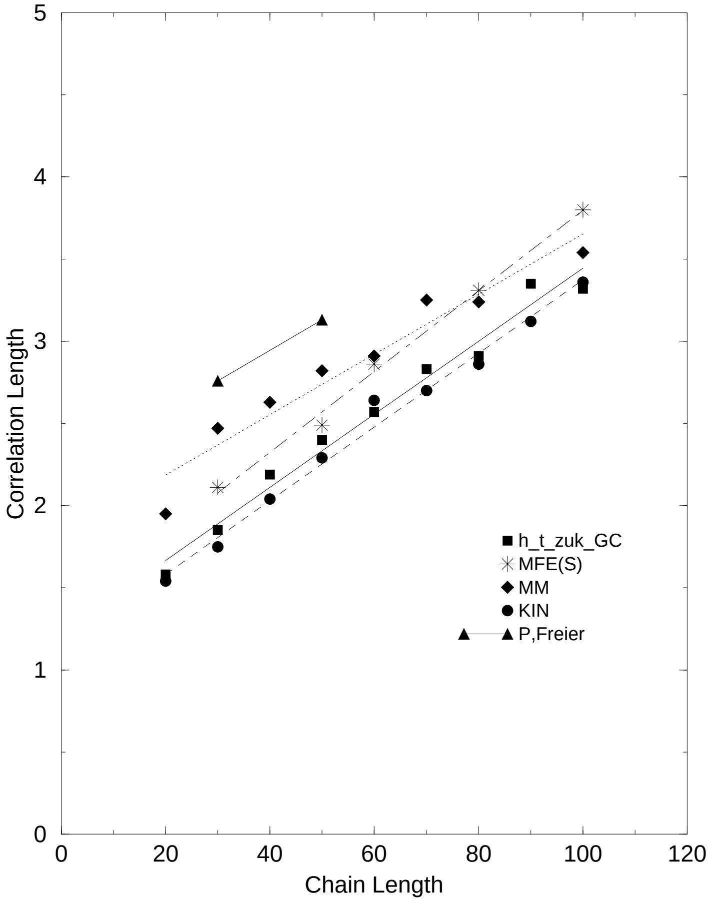


Figure 30: Structure autocorrelation lengths for different algorithms, **GC** alphabet

7 Neutral Networks

The distribution of structures obeys Zipf's first law (see section 5, indicating that there exist a few very common structures and many very rare ones for all alphabets and algorithms.

The next logical step was then to investigate how sequences which fold into the same secondary structures are distributed over sequence space. In order to do this we searched for the largest connected set of sequences which folded into the same secondary structure. A sequence is connected to another sequence if the Hamming distance h between them is at most 2 and both sequences fold into identical secondary structures.

This connected subgraph N of the sequence space $C(S)$ is called a *neutral network*.

We performed computer simulations to assess the extension of these networks. The starting point was a random initial sequence. Then this sequence was folded into its secondary structure (different algorithms were use to determine the dependence on the folding algorithm) and then the neighbourhood of the sequence was searched for neutral mutations. It was investigated how far one can walk through sequence space without changing the structure.

The results are shown in fig. 31 for chain length 100 and both alphabets. It is clearly seen that for the **GCAU** alphabet most neutral paths cross the whole sequence space with MFE(F) and KIN showing the same behaviour. Neutral networks for the maximum matching algorithm are smaller than for the other algorithms.

Neutral networks for the **GC** alphabet are smaller. This is due to the fact that combinatorial reasons allow only one sequence with Hamming distance ν from the starting sequence whereas for the **GCAU** alphabet much more (4ν) sequences with Hamming distance ν exist. In this case the KIN and MFE(F) algorithms show the same behaviour whereas the maximum matching algorithm yields smaller networks.

The dependence of the walk length on the probability of the structure is shown in fig. 33 for the case of $\nu = 30$ and the **GCAU** alphabet. The size of the neutral network is only slightly dependent on the probability of the structure in the structure distribution. Even for very rare structures (probability of 10^{-5}) most networks percolate through more than $\frac{5}{6}$ of sequence space. There is a big difference though in the number of neutral networks percolation through whole of sequence space. For very abundant structures about 60% to 70% percolate through the whole sequence space, for the very rare structures it is only 10%. Fig. 33 shows the dependence of the mean walk length on a neutral net as a function of the probability of the structure

The extension of the neutral network is dependent on the ration of sequences to structures. This is clearly seen when comparing the extension of neutral networks generated by the kinetic algorithm for different alphabets and chainlength 100. Fig. 34 shows the results for the **GCAU** alphabet, fig. 36 for the **GC** alphabet and fig. 35 for the **AU** alphabet. For the **GCAU** alphabet most neutral networks extend through more than $\frac{4}{5}$ of sequence space. For the **GC** alphabet there are much more structures compared to sequences [36]. Therefore the neutral nets are much smaller (see fig. 31 as well). The neutral nets are more extended for the **AU** alphabet. This reason for this is that many structures which are accessible for the **GC** sequences are not accessible to the **AU** sequences due to energetical reasons. This observations are in agreement with results discussed in section 8.1 where **GC** sequences are less stable against the introduction of point mutations than **AU** and **GCAU** sequences.

The same behaviour can be observed with the maximum matching algorithm. Changing the minimal allowed stack size allows a fine tuning of the number of structures accessible. The higher the value of the minimal stack size the lower the number of potential structures. Fig. 37 shows that for the four letter alphabet nearly all neutral networks extend through more than $\frac{4}{5}$ of sequence space with the higher values for the minimal stack size having larger extensions.The same is true for the two letter alphabet (shown in fig. 38) where for small values of the minimal stack size a distribution of neutral paths analogous to the **GC** alphabet folded with the kinetic algorithm is observed whereas for larger minimal stack sizes the distribution is very much alike the **AU** alphabet.

The existence of neutral nets is very important for adaptations on ladscapes. For evolutionary adaptation to take place efficiently only a rough shape (embodied by the secondary structure) has to be found. The fine tuning of the molecule is then possible by walking on this extended neutral net which allows stepwise improvement without entirely loosing functionality. Thus large portions of the sequence space are accessible via point mutations without ever changing secondary structure. The existence of neutral nets is a robust property of the combinatory map of secondary structures. It can be found with mappings generated by all folding routines used in this work, even with the primitive maximum matching algorithm.

The extension of neutral nets depend on the alphabet with **GCAU** showing larger nets than **GC** and on the folding algorithm. The probability of the secondary structure determines the size as well, but even very rare structures show extended neutral nets.

The folding of RNAs with MM with different parameters of the minimal stack size shows that the ration of structures to sequences determines the extension of neutral nets.

Evolutionary adaptation seems to be easier than earlier thought of, especially that for all these landscapes the structure correlation length l_S is very small. This suggests that l_S is not a very good measure for adaptibility. A possibility to verify this assumption is to measure the performance of optimization algorithms on landscapes as a function of the correlation length and the extension of neutral nets. This work is on the way in our group.

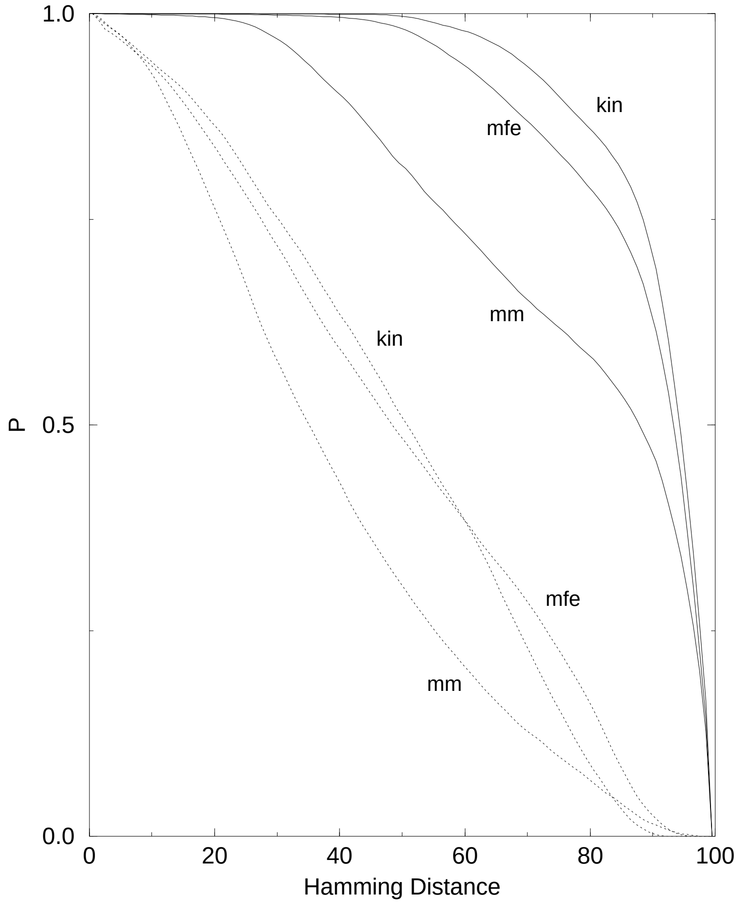


Figure 31: The mean extension of neutral nets for the **GCAU** and **GC** alphabet for the MFE(F), mm and kinetic algorithms, $\nu = 100$

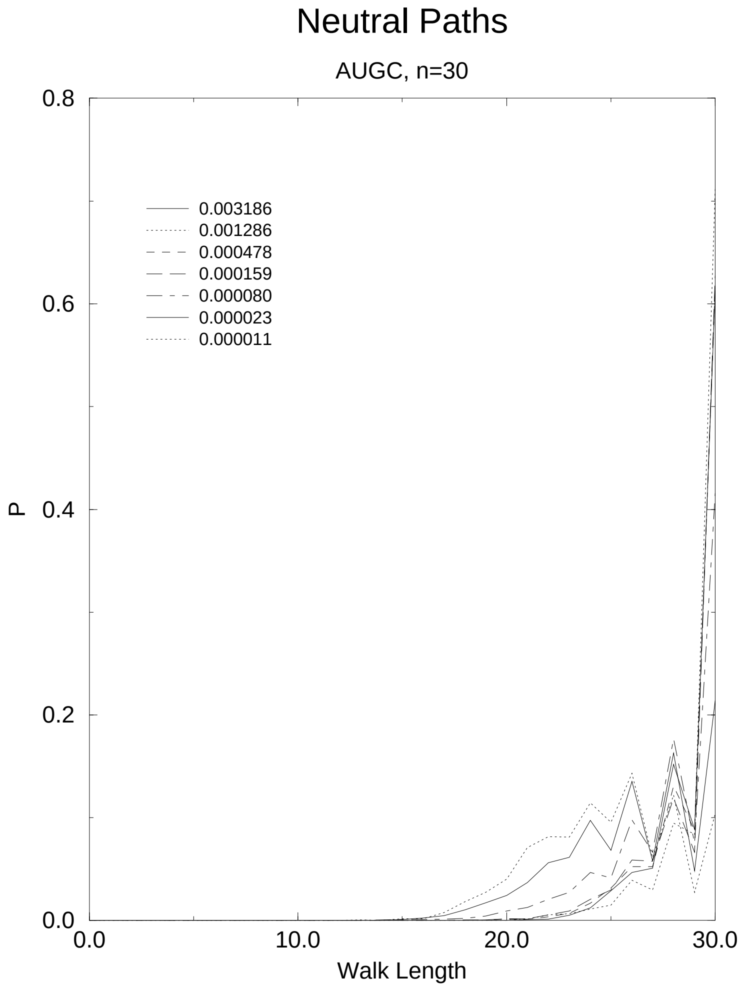


Figure 32: The dependence of the extension of neutral nets for the **GCAU** alphabet on the probability of the structure, MFE(F), $\nu = 30$

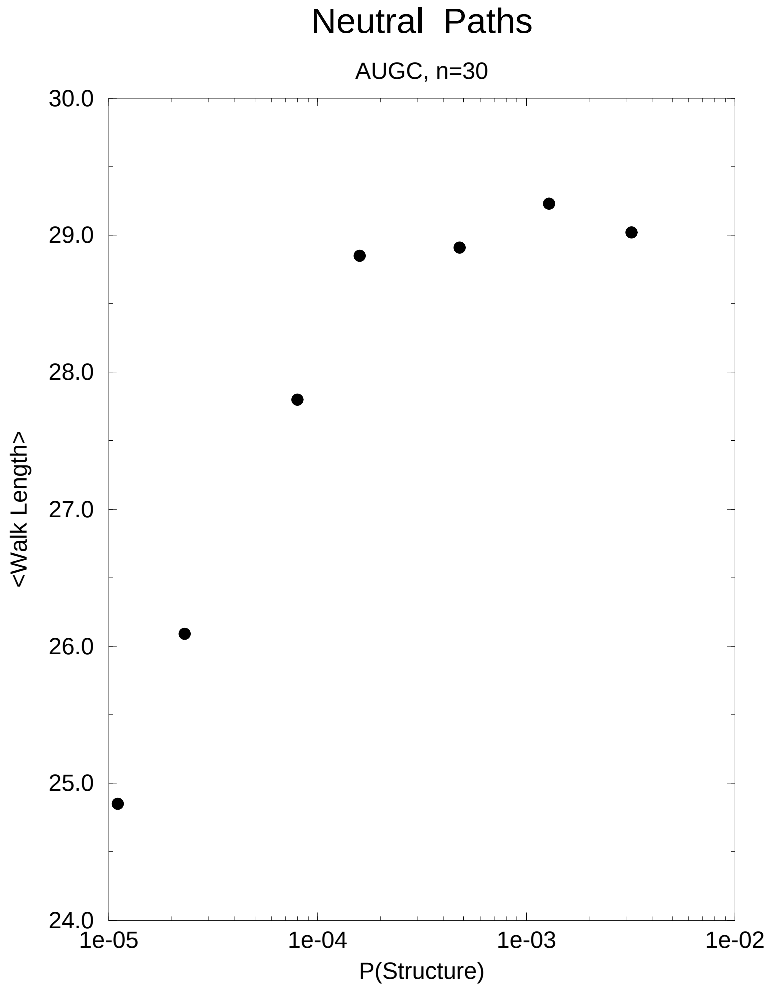


Figure 33: The dependence of the mean extension of neutral nets for the **GCAU** alphabet on the probability of the structure, MFE(F), $\nu = 30$

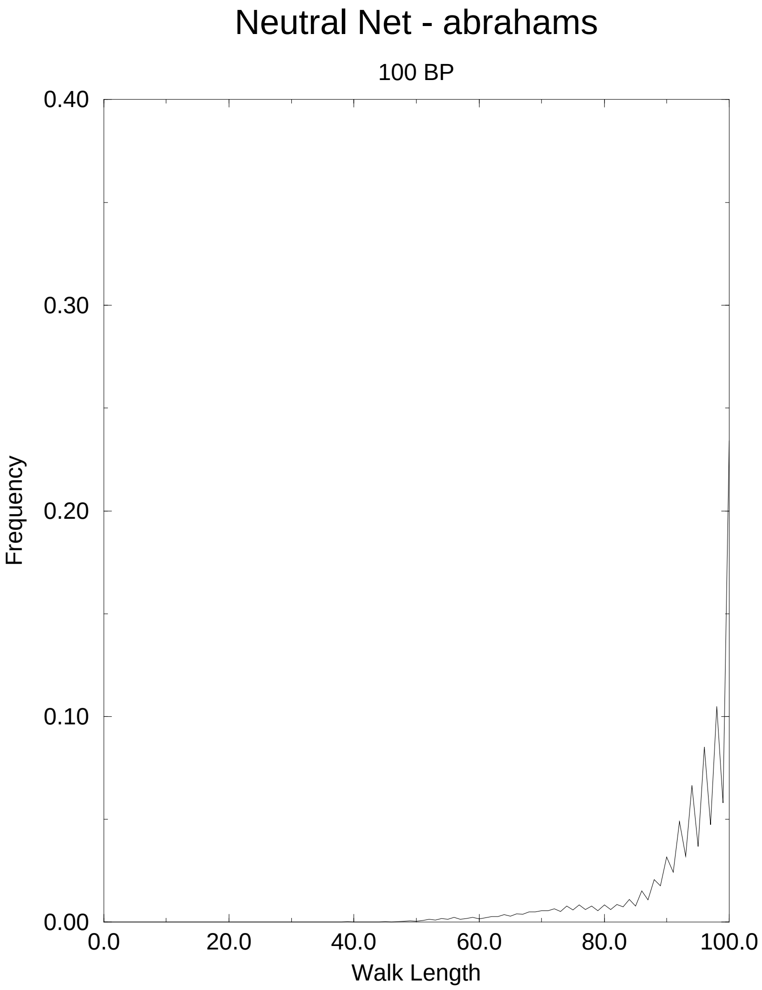


Figure 34: Neutral Network for the **GCAU** alphabet, KIN, $\nu = 100$

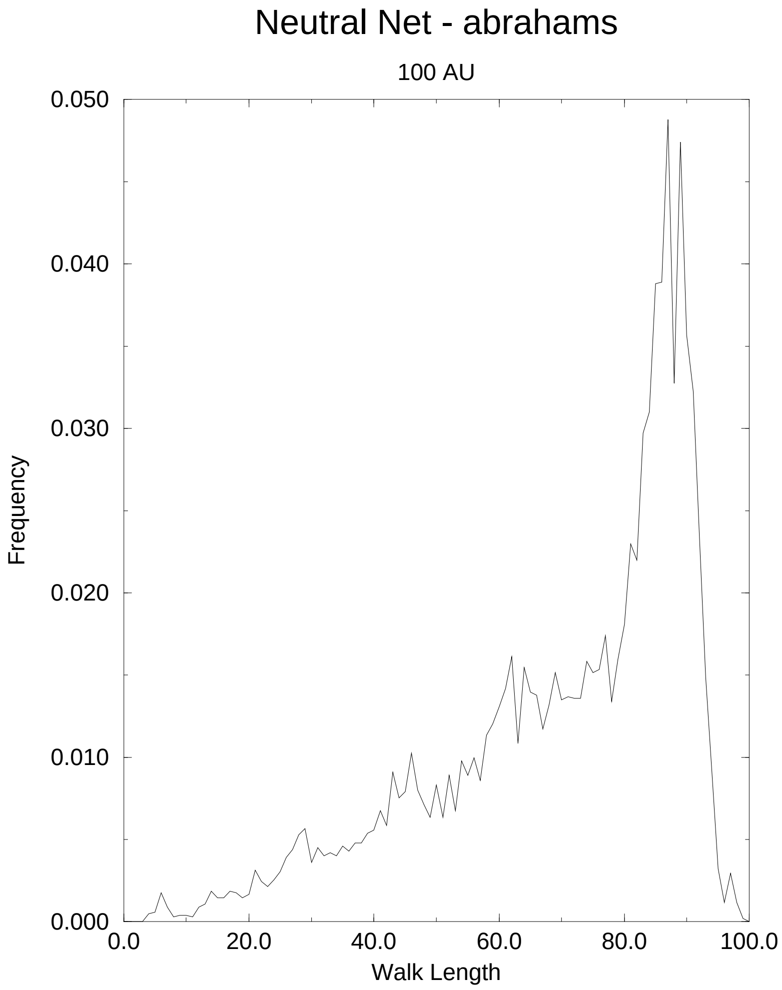


Figure 35: Neutral Network for the **AU** alphabet, KIN, $\nu = 100$

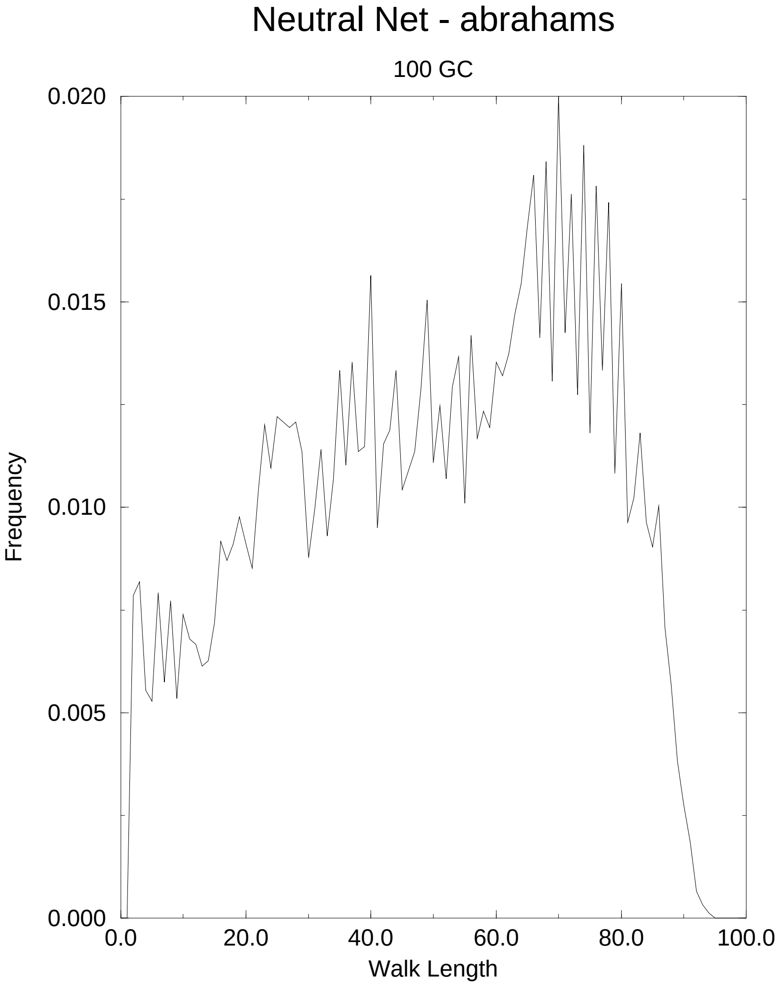


Figure 36: Neutral Network for the **GC** alphabet, KIN, $\nu = 100$

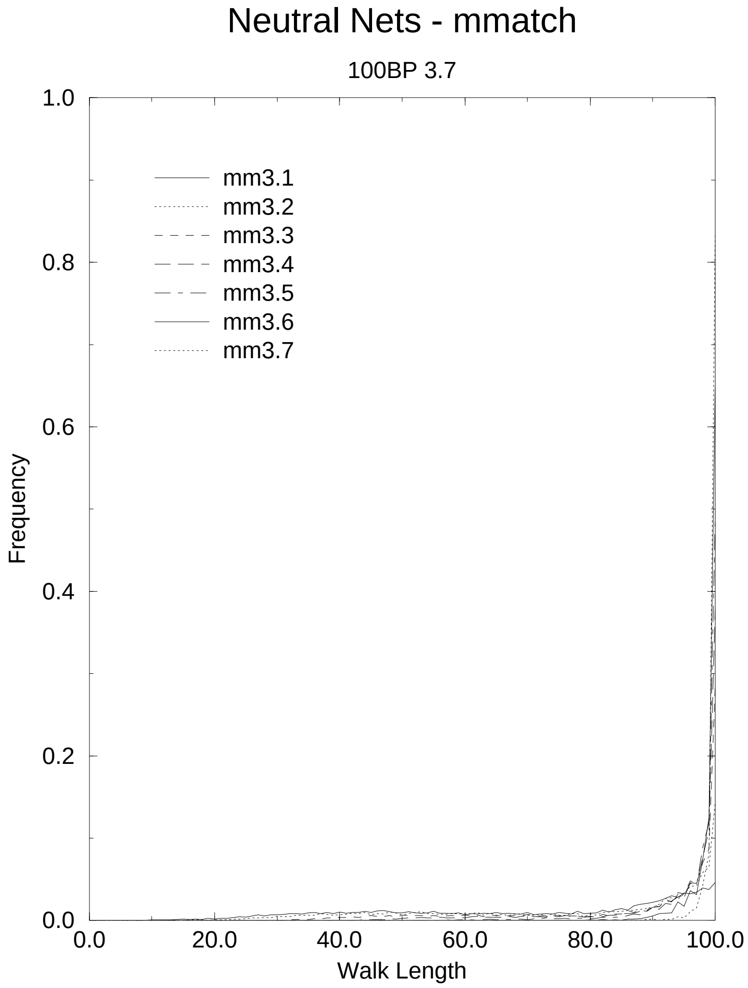


Figure 37: Neutral Network for the **GCAU** alphabet, MM with different values for the minimal stack size $\nu = 100$

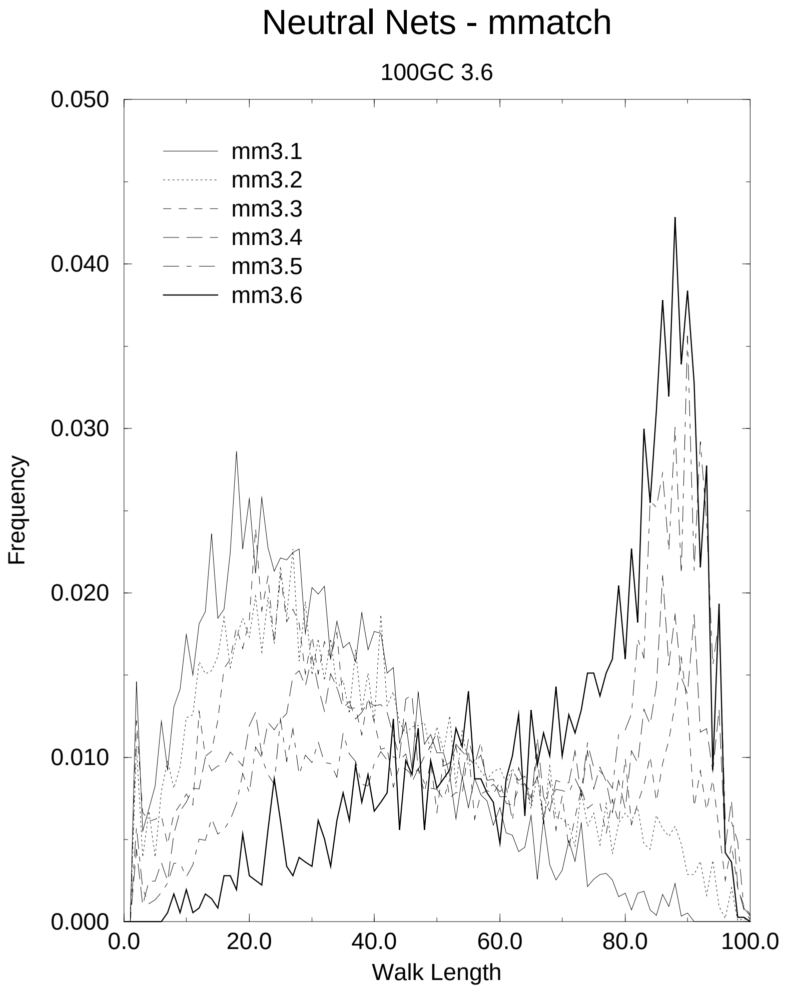


Figure 38: Neutral Network for the **GC** alphabet, MM with different values for the minimal stack size, $\nu = 100$

8 Stability of the Thermodynamic Structure

The stability of the thermodynamically most stable secondary structure is investigated under two points of view:

- Stability against point mutations
- Thermic stability

8.1 Stability against Mutations

The stability of sequences against mutations was investigated by folding sequences into its secondary structures. Then mutations were introduced, the mutated strings folded and the number of cases counted where the mutation induced a folding into a different secondary structure. We folded several thousands of RNA molecules with the MFE(F) algorithm to determine the stability against mutations as a function of the alphabet. The results are shown in fig. 39 where the number of neutral neighbours were monitored.

The four letter alphabet **GCAU** is more tolerant against mutations than the two letter alphabets. The two letter alphabet behave very differently with the **GC** strings being much more unstable against mutations than **AU**.

This result can be compared with the structure correlation lengths shown in table 6 which measure the mean change in secondary structure as a function of the error class (see appendix B). **GC** shows a smaller autocorrelation length than **GCAU**. This result is paralleled by the mean tree distance as a function of the error class (fig. 40) where for lower error classes **GC** and **AU** show nearly identical and higher mean structure distances than **GCAU**. For high error classes all alphabets behave in the same way - the distance between random structures being the high error class limit.

It is interesting to note that although the mean tree distance is nearly identical for **GC** and **AU** there is a marked difference in the number of neutral mutations which are observed (fig. 39). The **AUGC** alphabet shows again the highest number of neutral mutations but **AU** sequences have a much higher percentage of neutral neighbours than **GC** sequences. This suggests that mutations in the **GC** alphabet induce a much higher rate of refolding than in the **AU** or **GCAU** alphabets but that the refolded structures are not very different from the original structure. In the case of **AU** sequences the rarer refoldings lead to very different structures.

The next step was the investigation of the differences between subclasses of RNA: strings which fold into the same secondary structures with the MFE(F) and KIN and strings which do not. Figures 41 to 43 show the results of numerical simulations. It is easily seen that the strings which fold into the same secondary structure with both algorithms are more stable against mutations than strings which are not.

This result is not unexpected because strings which fold into different secondary structures when folded with MFE(F) and the kinetic algorithm have stable potential stacks which are not part of the thermodynamic secondary structure. These stacks are a source of instability when a perturbation is introduced by mutations - so the probability of refolding is higher.

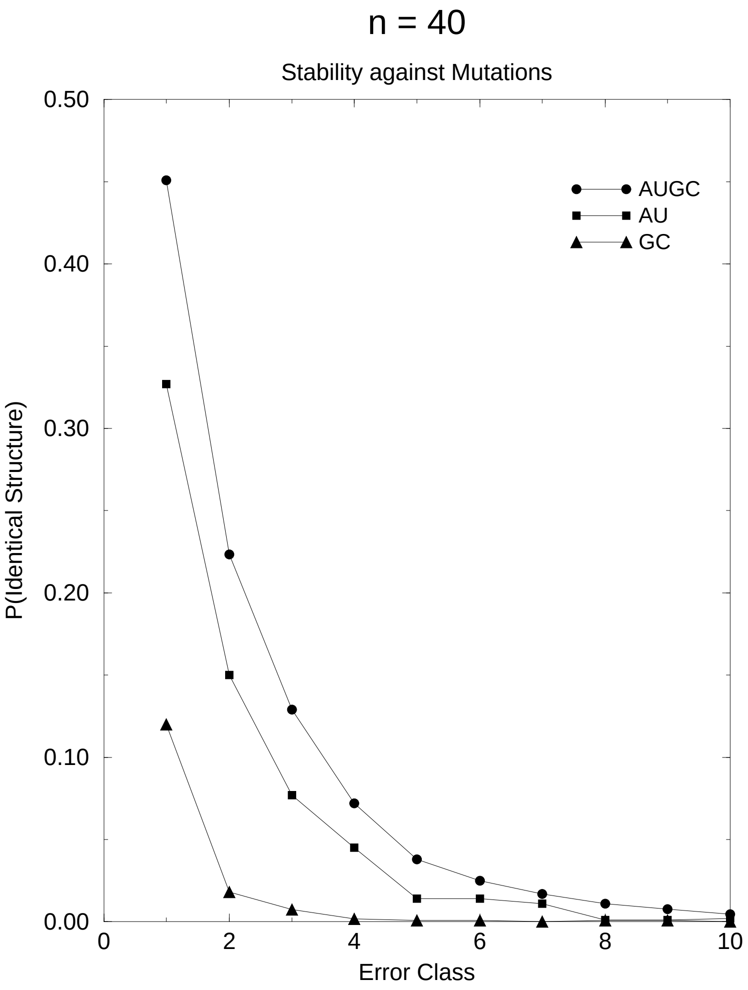


Figure 39: Stability against mutations, comparison of different alphabets, the number of neutral mutations is shown, $\nu = 40$

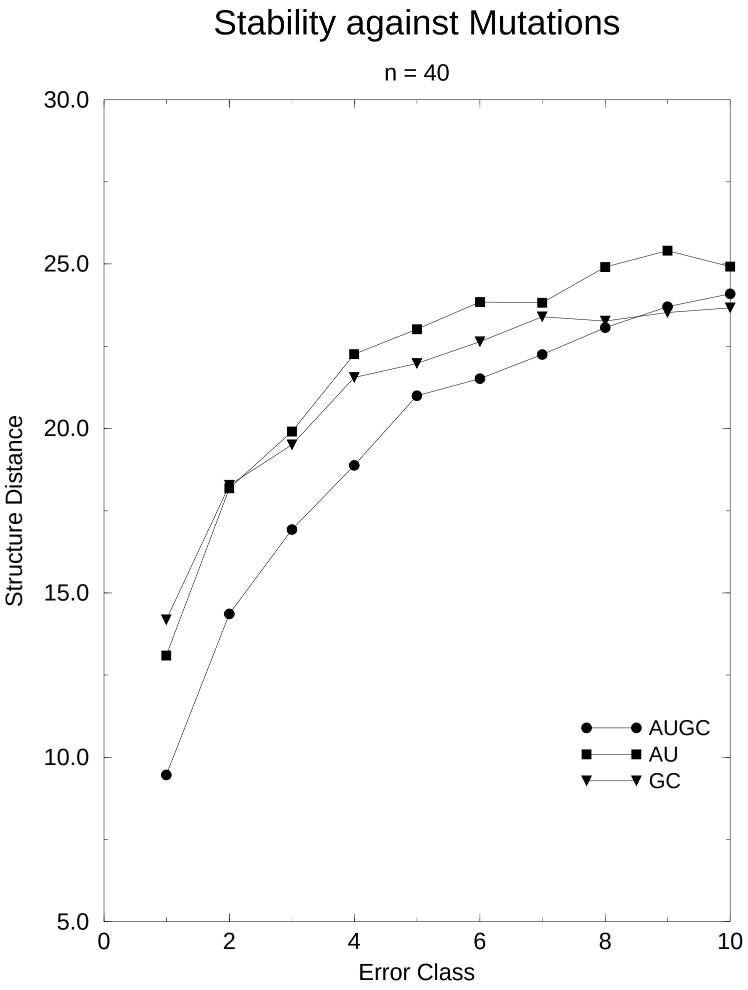


Figure 40: Stability against mutations, comparison of different alphabets, the mean tree distance of the mutations to the start sequence is shown, $\nu = 40$

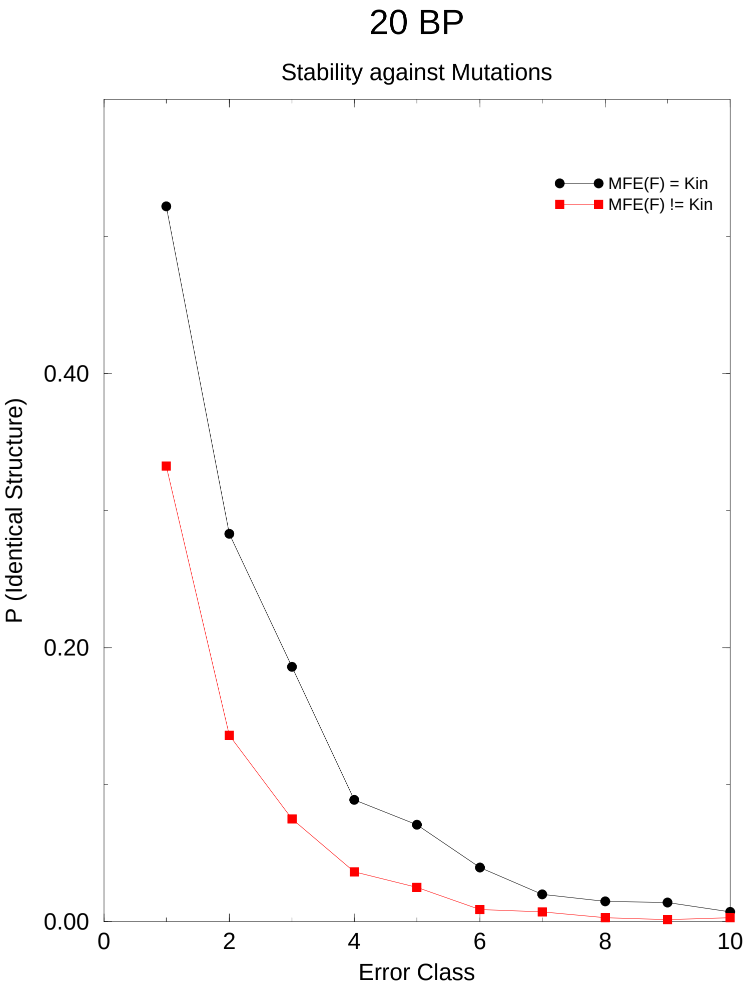


Figure 41: The percentage of neutral neighbours for **GCAU** sequences as a function of the error class, $\nu = 20$

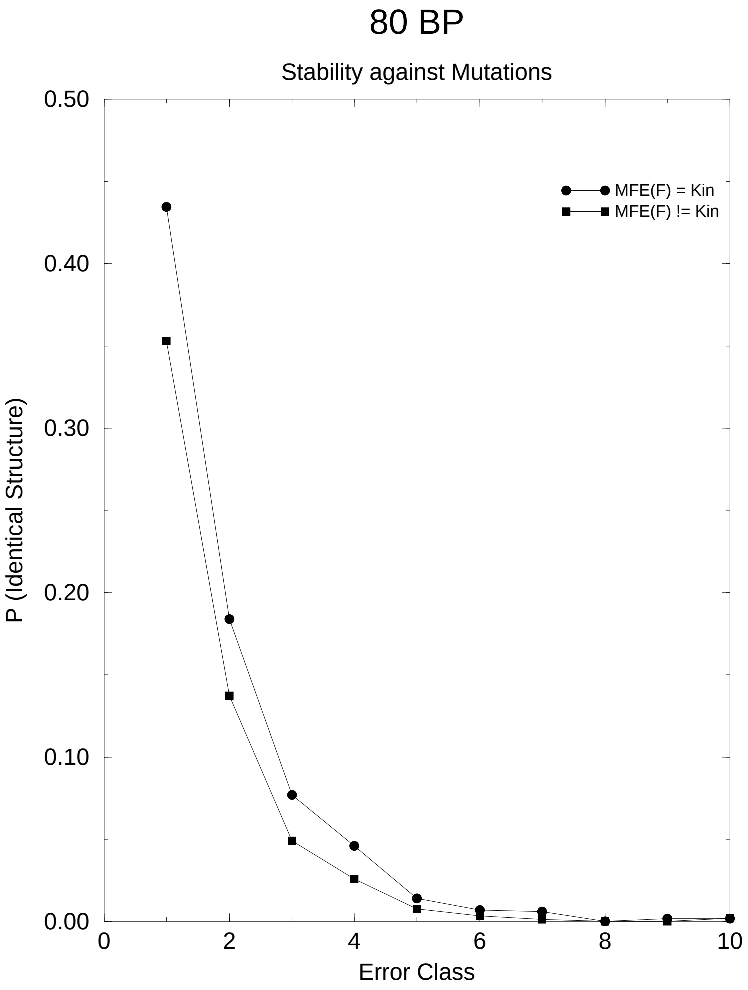


Figure 42: The percentage of neutral neighbours for **GCAU** sequences as a function of the error class, $\nu = 80$

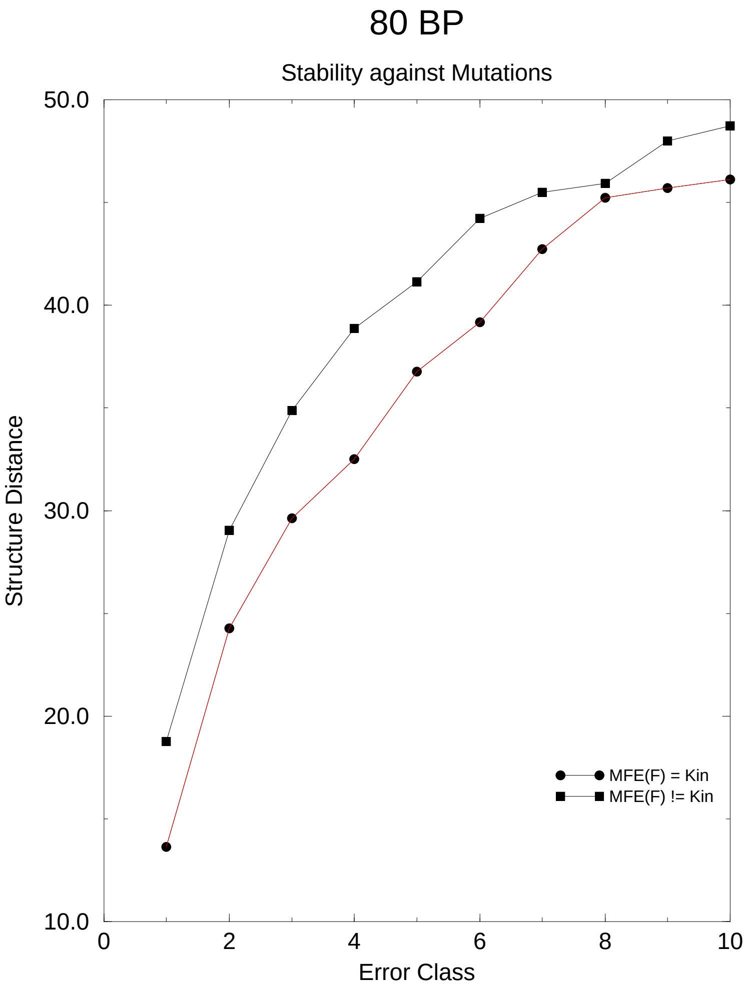


Figure 43: The mean tree distance for **GCAU** sequences as a function of the error class, $\nu = 80$

8.2 Activation Energy of Melting

The thermic stability of the minimal free energy structure is determined by the activation barrier for melting the secondary structure. It is assumed that the main part of the activation energy stems from opening secondary structure base pairs.

Numerical simulation for the determination of the stability of the global optimum were performed in the following way:
An RNA molecule is folded into the thermodynamic structure with the MFE(F) algorithm. The mean escape time from the basin of attraction of this structure is then determined at different temperatures. The mean escape time was measured by performing a simulated annealing scheme. The moves allowed were either to open an existing base pair or to close a base pair chosen randomly. The acceptance rate of each move was proportional $\exp \frac{-(E_1-E_2)}{RT}$ with E_1 denoting the free energy of the start configuration, E_2 the free energy of the target configuration, R the universal gas constant and T the temperature. The number of steps M necessary to move out of the basin of attraction (that is till a configuration was reached where only energy gaining moves were necessary to reach another (metastable) conformation) were monitored.

Fig. 44 shows a plot of M as a function of $\frac{1}{T}$. It is seen that there exist an exponentional relationship between M and $\frac{1}{T}$. That means that in this Arrhenius plot the slope yields the effective activation energy E_{eff} of the melting process. The lowest E_{act} and the mean activation energy value of all escapes was also monitored in each computer simulation.

Numerical simulations were carried out for 13 randomly generated RNAs of chain length 40 and the slopes of each run plotted against the lowest E_{act} found. This is shown in fig. 45. There exist a linear relationship between the slopes and the lowest activation energies found in the simulations. This is consistent with the fact that the slopes represent the activation energies of the melting processes.

No such relationship could be detected in the plot of the energy of the minimum free energy structure vs the slopes (fig. 46).

Fig. 47 shows the distribution of M for an RNA with chain length 77 in the high temperature limit. This is equivalent to diffusion without an external potential. The distribution gives information about the size of the basin of attraction. A still better method to gain insight into the size of the basin of attraction is to plot the distribution of the number of base pairs necessary to open and/or close to leave the optimum. Such a distribution is shown in fig. 48.

The distribution of the activation energies for each reaction path is shown in fig. 49 for an RNA with $\nu = 77$. This plot gives information about the height of the activation energy barrier surrounding the basin of attraction.

These results show that it is possible to characterize the basin of attraction of a secondary structure by the following methods:

- The escape rate as a function of $\frac{1}{T}$ gives the activation energy of melting
- The distribution of the energy values in the high temperature limit gives the heights of the activation barrier
- The structure distances in the high temperature limit give information about the extension of the basin of attraction

Used together these methods characterize in detail the basin of attraction. The application is not restricted to the thermodynamic structure.

The energy of the transition structure with the lowest free energy determines the velocity of leaving the thermodynamic optimum - all the other reaction paths leading over higher energetic transition structures contribute only marginally. For a different point of view on the melting statistics of RNA see [83].

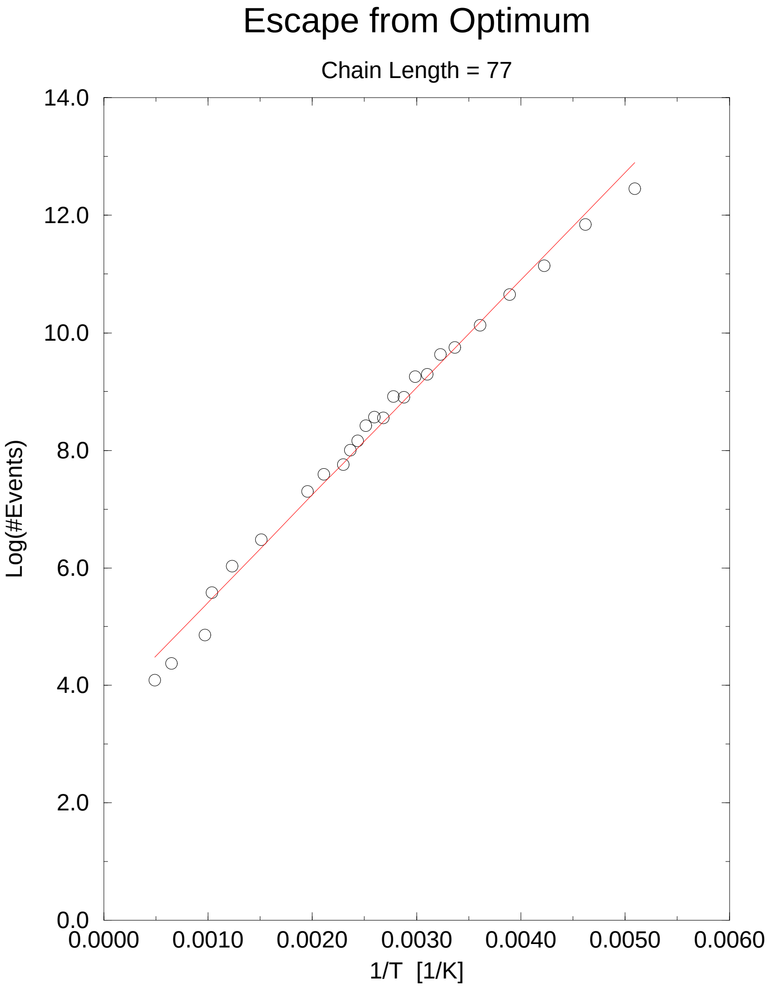


Figure 44: The mean escape time M as a function of $\frac{1}{T}$, **AUGC** alphabet, $\nu = 77$

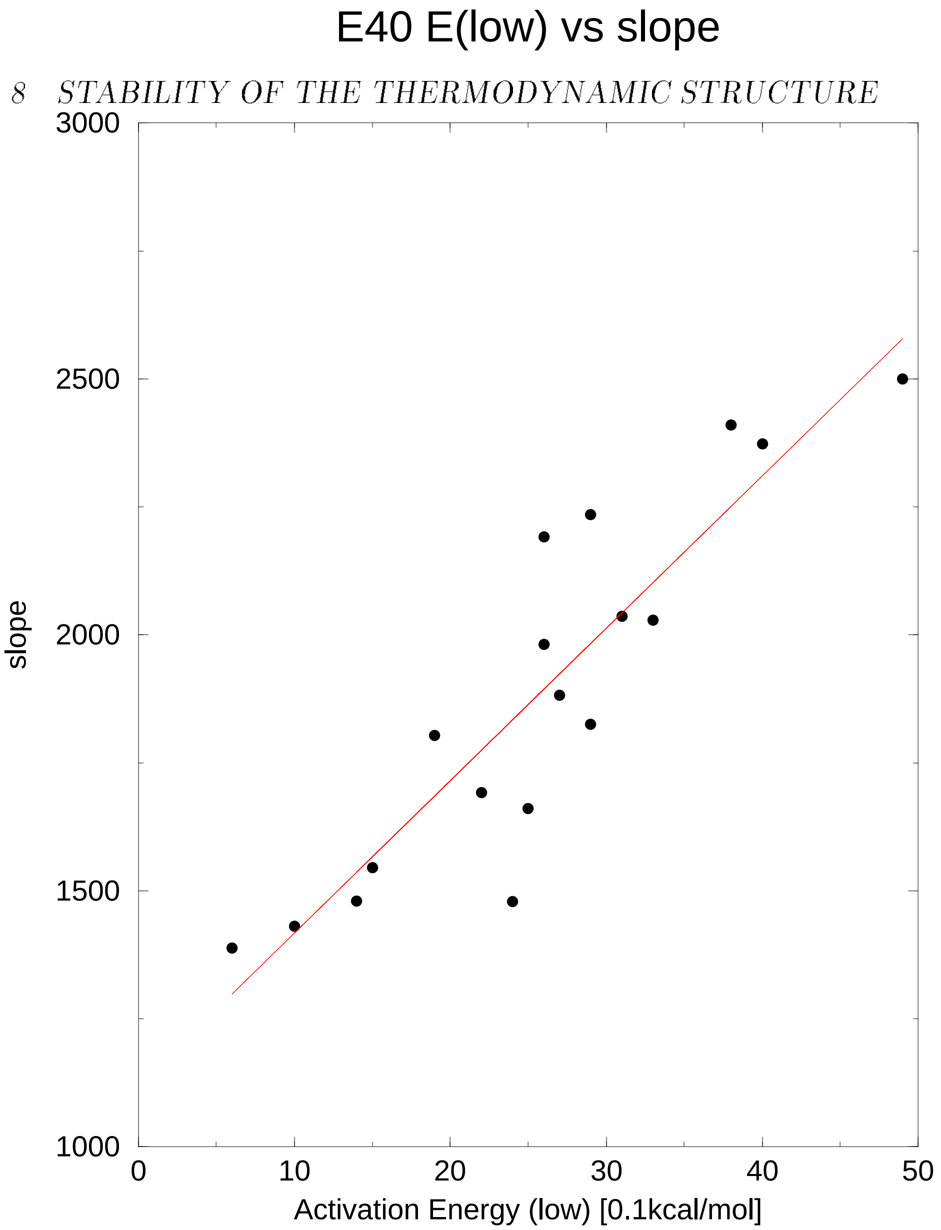


Figure 45: The lowest activation energy found in numerical simulations is plotted vs the slope of the Arrhenius plot, **AUGC** alphabet, $\nu = 40$

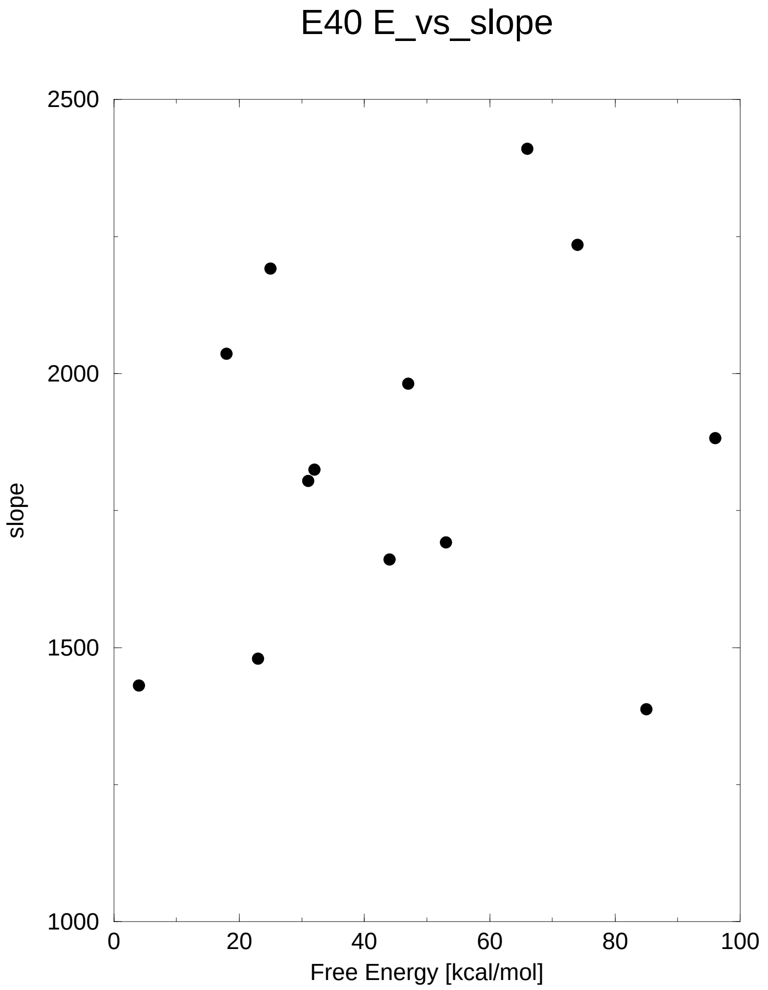


Figure 46: The free energy of the thermodynamic structure is plotted vs the slope of the Arrhenius plot, **AUGC** alphabet, $\nu = 40$

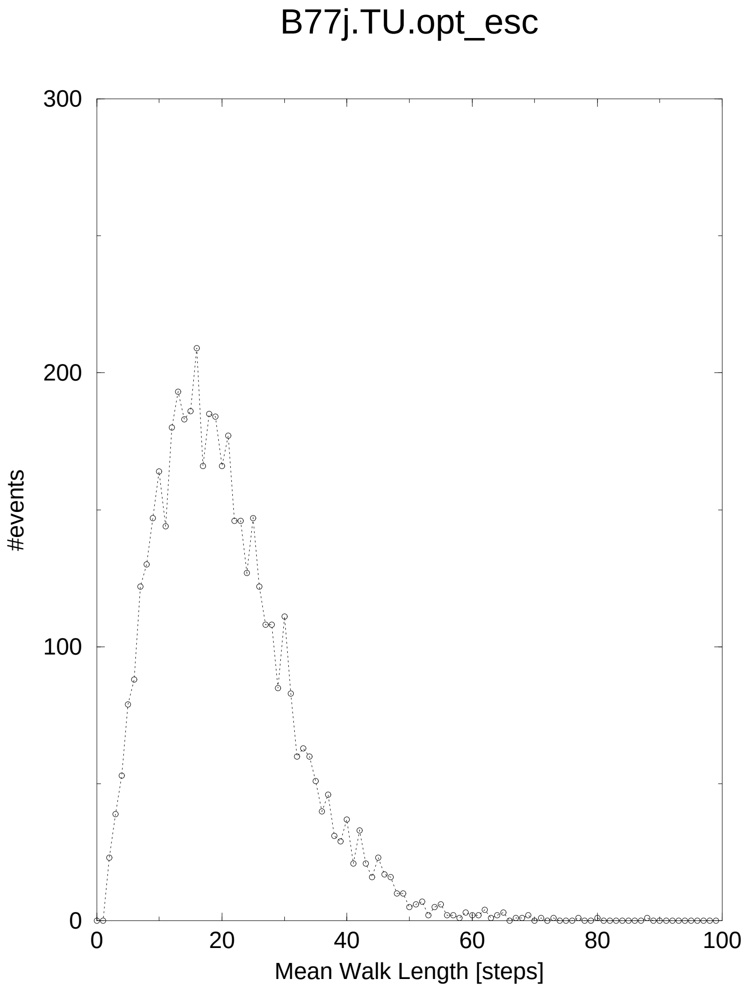


Figure 47: The distribution of steps necessary to escape from the global optimum is shown for an RNA, **GCAU** alphabet, $\nu = 77$ in the high temperature limit

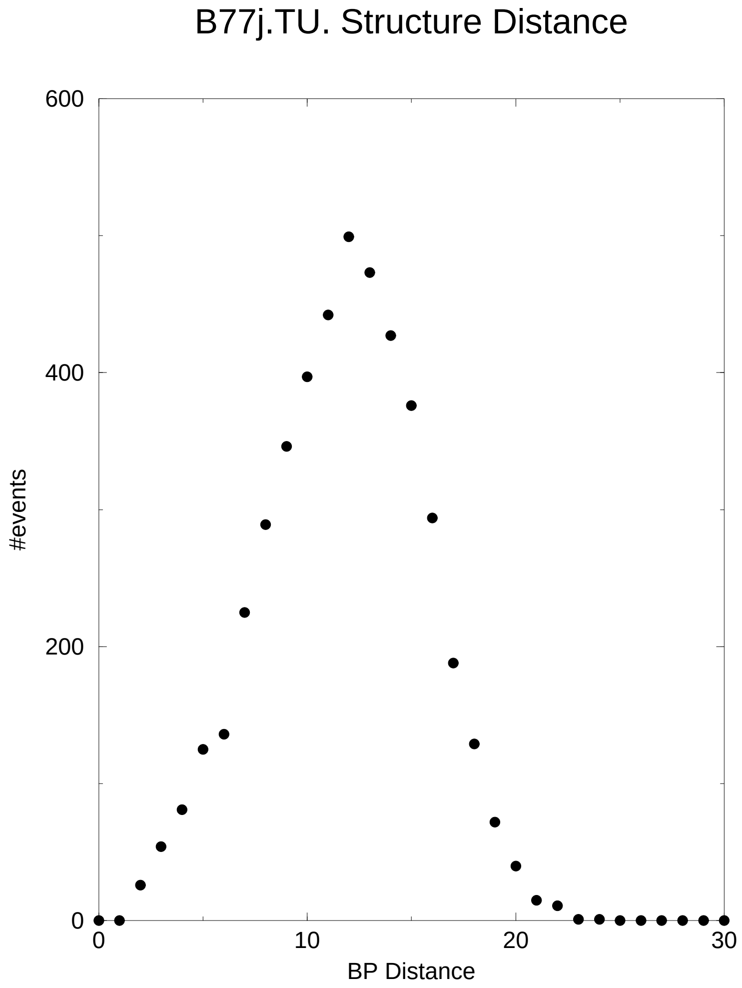


Figure 48: The distribution of structure distances for leaving the global optimum is shown for an RNA, **AUGC** alphabet, $\nu = 77$. It gives detailed information about the size and form of the basin of attraction for the global optimum

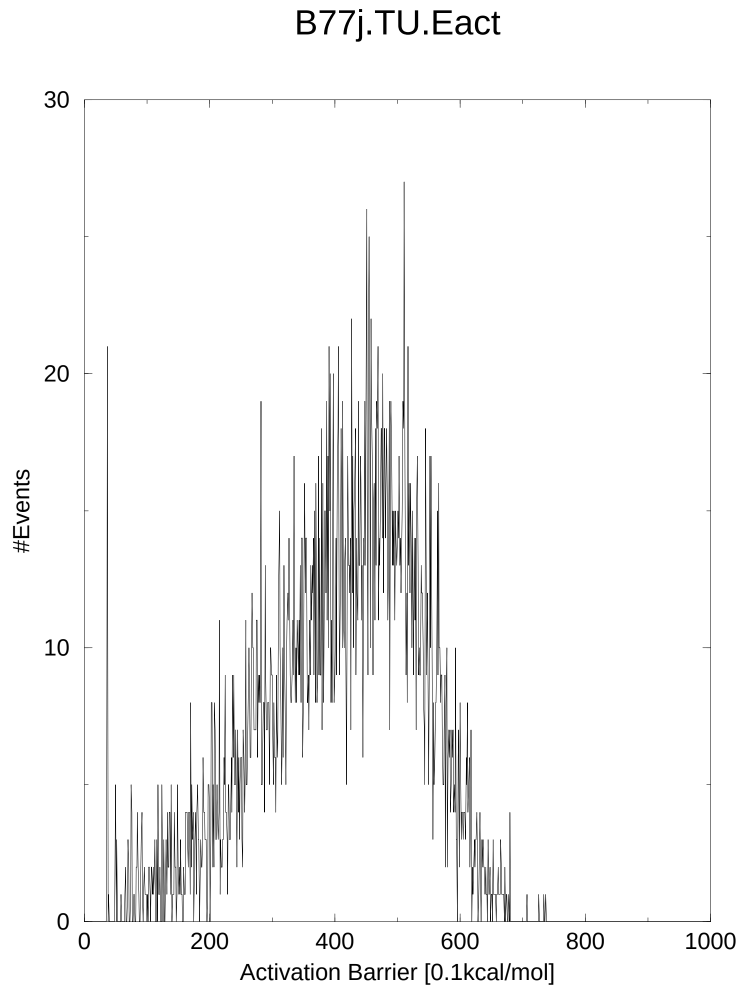


Figure 49: The distribution of activation energies for leaving the global optimum is shown for an RNA, **GCAU** alphabet, $\nu = 77$

9 $5' \rightarrow 3'$ Folding

The folding algorithm used is described in section 4.4.3. The algorithm is designed in such a way that it gives a smooth transition from the Minimal free energy algorithm to a kinetical $5' \rightarrow 3'$ folding. If one chooses the STACKE to be very large or the MINLOOPSZ very small then the algorithm reduces to a minimal free energy folding. On the other hand lowering the STACKE and enlarging the MINLOOPSZ allows stacks to be frozen which are kinetically determined.

By searching systematically the STACKE and the MAXLOOPSZ parameter space we are now in a position to ask the question if RNA folding is kinetically determined or if the RNA molecule finds the thermodynamically most stable structure.

If RNA folding is kinetically determined we should find regions in parameter space where the predictive power of this algorithm is higher than that of the MFE algorithm.

We chose 16S rRNA from different species for this study because these RNAs are long enough to allow for kinetical folding and reliable phylogenetic secondary structures exist for checking the predictive capacity of the algorithm.

9.1 Results

8 different 16S rRNA molecules are folded to investigate the predictive power of the $5' \rightarrow 3'$ folding algorithm and compared with the MFE(F) algorithm. We used phylogenetic structures as reference (kindly provided by Robin Gutel of the University of Colorado, Boulder).

The 16S rRNAs used are:

- arc16smv: Methanococcus vannielli
- arc16sss: Sulfolobus solfataricus
- chl16szm: Zea mays, chloroplast
- eub16stm: Thermatoga maritima
- euk16sgi: Giardia intestinalis
- euk16ssc: Saccheromyces cerevisiae
- mit16sce: Caenorhabditis elegans, mitochondrium
- mit16szm: Zea mays, mitochondrium

Arc16smv and arc16sss are archaebacterial rRNAs, chl16szm is a chloroplast rRNA, eub16stm is from an eubacterium, euk16sgi and euk16ssc are from eukaryots and mit16sce and mit16szm are mitochondrial rRNAs. Thus a diverse group of rRNAs from different organelles and different archae was chosen.

The comparison of the percentage of correctly predicted base pairs of the phylogenetic structures both for the $5' \rightarrow 3'$ folding algorithm and the MFE(F) algorithm is shown in figs. 50, 51, 52, 53, 54, 55 and table 7 for all rRNAs with different values of MAXLOOPSZ and STACKE.

It is easily seen that for small values of STACKE and large values of MAXLOOPSZ for most molecules considered the prediction rate of the $5' \rightarrow 3'$ folding algorithm is worse than with the MFE(F) algorithm. This is due to the fact that stacks with small stacking energies refold very fast and it makes no sense to freeze them. For higher values of STACKE the percentage of correctly predicted base pairs is the same for the $5' \rightarrow 3'$ folding algorithm

	arc16smv	arc16sss	chl16szm	eub16stm	euk16sgi	euk16ssc	mit16szm	mit16sce
12/50	0.2462	0.3421	0.2545	0.5565	0.1432	0.1511	0.1951	0.2011
14/50	0.4457	0.4759	0.2545	0.3646	0.1745	0.1553	0.1704	0.2011
16/50	0.3836	0.4232	0.2950	0.4776	0.1276	0.2422	0.1861	0.2011
18/50	0.5664	0.4167	0.2545	0.3960	0.2161	0.2671	0.1861	0.2011
20/50	0.5664	0.4759	0.2545	0.6119	0.1771	0.2671	0.2040	0.2011
25/50	0.5664	0.4759	0.2545	0.6119	0.1927	0.2857	0.1861	0.2011
12/30	0.2461	0.3421	0.2545	0.5565	0.1432	0.1511	0.1955	0.2011
14/30	0.5654	0.4386	0.2545	0.6119	0.1745	0.1553	0.1701	0.2011
16/30	0.3836	0.4232	0.2545	0.4761	0.1432	0.2422	0.1861	0.2011
18/30	0.5664	0.4161	0.2545	0.5522	0.2161	0.2671	0.1816	0.2011
20/30	0.5664	0.4759	0.2545	0.6119	0.1771	0.2857	0.2040	0.2011
25/30	0.5664	0.4759	0.2545	0.6119	0.1927	0.2857	0.1861	0.2011
12/15	0.2461	0.3421	0.2545	0.6119	0.1432	0.2857	0.1951	0.2011
14/15	0.5664	0.6031	0.2545	0.6119	0.1745	0.1553	0.1861	0.2011
16/15	0.5664	0.4232	0.2545	0.6119	0.1432	0.2857	0.1861	0.2011
18/15	0.5664	0.4167	0.2545	0.6119	0.2109	0.2671	0.1861	0.2011
20/15	0.5664	0.4759	0.2545	0.6119	0.1771	0.2857	0.1861	0.2011
25/15	0.5664	0.4759	0.2545	0.6119	0.1745	0.2857	0.1861	0.2011
MFE(F)	0.5664	0.4759	0.2545	0.6119	0.1432	0.2857	0.1861	0.2011

Table 7: Percentage of correctly predicted base pairs of 16S rRNAs folded with the MFE(F) algorithm and the $5' \rightarrow 3'$ folding algorithm with different choices of STACKE (first number in the 1. column in kcal/mol) and MAXLOOPSZ (second number in the 1. column)

as with MFE(S) - the $5' \rightarrow 3'$ folding algorithm is substantially better only in the case of euk16sgi where MFE(F) predicts only about 14% of all base pairs (see 54).

A summary of the prediction rate is shown in fig. 56. The MFE(F) correctly predicts about 35.8% of all base pairs whereas the $5' \rightarrow 3'$ folding algorithm with the best combination of STACKE (20kcal/mol) and MAXLOOPSZ (30) predicts about 36.4%. This is no substantial improvement.

There is an enormous difference in the free energy value of the phylogenetic structure and the structures generated by MFE(F) or the $5' \rightarrow 3'$ folding algorithm (see table 8). The free energies of the phylogenetic structures are much lower than the free energies of the thermodynamic structure or the structure generated by the $5' \rightarrow 3'$ folding algorithm. Even when the unpaired regions of the phylogenetic structure are filled with a minimum free energy fold (phy(f)) the secondary structure obtained is far less stable than the thermodynamic structure. It is interesting to note that the 16S rRNAs where the difference in the free energies are most pronounced (chl16szm, euk16sgi, euk16ssc, mit16szm and mit16sce) both algorithms, the MFE(F) and the $5' \rightarrow 3'$ algorithm show bad prediction scores.

The reason for the bad prediction and the big differences in free energies can be due to

- Inaccurate energy parameters
- Protein-RNA interaction
- Nonstandard base pairs
- Tertiary interactions (pseudoknots ...)

The most recently compiled dataset by Freier [26] is not complete. Apart from the approximation of nearest neighbour interaction there are no experimental data available for multiloops which are therefore approximated by a linear function (see equation 1). It is not clear, if real multiloop destabilizing energies are well described by this approximation. In section 5 the dependence of the prediction accuracy on the parameter set was discussed and the importance of it stressed. The inaccuracy and incompleteness of the experimentally determined energy parameter set is the most likely reason for the difference in free energies.

Another point which has still to be determined is the influence on protein-RNA interaction. Ribosomal RNAs are *in vivo* tightly coupled to proteins. Thus it is easy to imagine that the protein binding forces the RNA to adopt another secondary structure.

About 9.6% of the base pairs occuring in the phylogenetic structures involve nonstandard base pairs. For these base pairs no energy parameters are available. In our implementation of the folding algorithm these base pair are approximated by standard base pairs.

Last but not least is is possible that the introduction of tertiary interactions can change the secondary structure.

On the whole the newly introduced $5' \rightarrow 3'$ folding algorithm shows no substantial improvement to the MFE(F) folding. Both algorithms predict for some 16S rRNAs just about 20% of all base pair correctly which occur in the phylogenetic structures. Furthermore the free energies of the predicted secondary structures deviate enormously. These effects suggest that either the parameter set is incomplete and the approximations for multiloops or nonstandard base pairs wrong or that the protein-RNA interaction changes the secondary structure heavily.

From these data it dan not be concluded that RNA folding is kinetically determined. It seems rather that RNA folds into the thermodynamical structure but conclusive evidence can not be given because the predicted structures and their energy values differ too much from the phylogenetic structures.

	arc16smv	arc16sss	chl16szm	eub16stm	euk16sgi	euk16ssc	mit16szm	mit16sce
12/50	-361.2	-488.4	-384.8	-566.2	-459.0	-329.3	-413.2	-65.1
14/50	-421.2	-582.2	-384.8	-477.2	-559.1	-354.4	-435.8	-65.1
16/50	-440.0	-529.7	-377.0	-533.8	-497.5	-357.6	-468.8	-65.1
18/50	-461.6	-571.6	-384.8	-530.2	-525.5	-372.3	-467.4	-65.1
20/50	-461.6	-582.2	-384.8	-567.8	-549.9	-375.2	-468.2	-65.1
25/50	-461.6	-582.2	-384.8	-567.8	-561.4	-375.2	-468.8	-65.1
12/30	-361.1	-488.4	-384.4	-566.2	-459.0	-329.3	-413.2	-65.1
14/30	-461.6	-507.2	-384.8	-567.8	-559.1	-354.4	-435.8	-65.1
16/30	-440.0	-529.7	-384.8	-533.8	-565.6	-357.6	-468.8	-65.1
18/30	-461.6	-571.6	-384.8	-563.7	-525.5	-372.3	-467.4	-65.1
20/30	-461.6	-582.2	-384.8	-567.8	-549.9	-375.2	-468.2	-65.1
25/30	-461.6	-582.2	-384.8	-567.8	-561.4	-375.2	-468.8	-65.1
12/15	-361.6	-488.4	-384.8	-567.8	-459.0	-375.2	-413.2	-65.1
14/15	-461.6	-567.3	-384.8	-567.8	-559.1	-354.4	-468.8	-65.1
16/15	-461.6	-529.7	-384.8	-567.8	-565.6	-375.2	-468.8	-65.1
18/15	-461.6	-571.6	-384.8	-567.8	-556.5	-372.3	-468.8	-65.1
20/15	-461.6	-582.2	-384.8	-567.8	-549.9	-375.2	-468.8	-65.1
25/15	-461.6	-582.2	-384.8	-567.8	-563.9	-375.2	-468.8	-65.1
MFE(F)	-461.6	-582.2	-384.8	-567.8	-565.6	-375.2	-468.8	-65.1
phy	-300.3	-388.5	-198.6	-382.2	-170.2	-18.3	-108.9	+66.2
phy(f)	-365.0	-477.5	-257.2	-471.2	-283.5	-96.8	-305.5	+42.0

Table 8: Free energies of 16S rRNA secondary structures (in 0.1 kcal/mol) obtained by folding with the $5' \rightarrow 3'$ folding algorithm with different values of STACKE (first number in the 1. column in [kcal/mol]) and MAXLOOPSZ (second number in the 1. column), with MFE(F), the phylogenetic structure (phy) and the filled phylogenetic structure (phy(f)) which was generated by filling the unpaired regions of the phylogenetic structure with MFE(F)

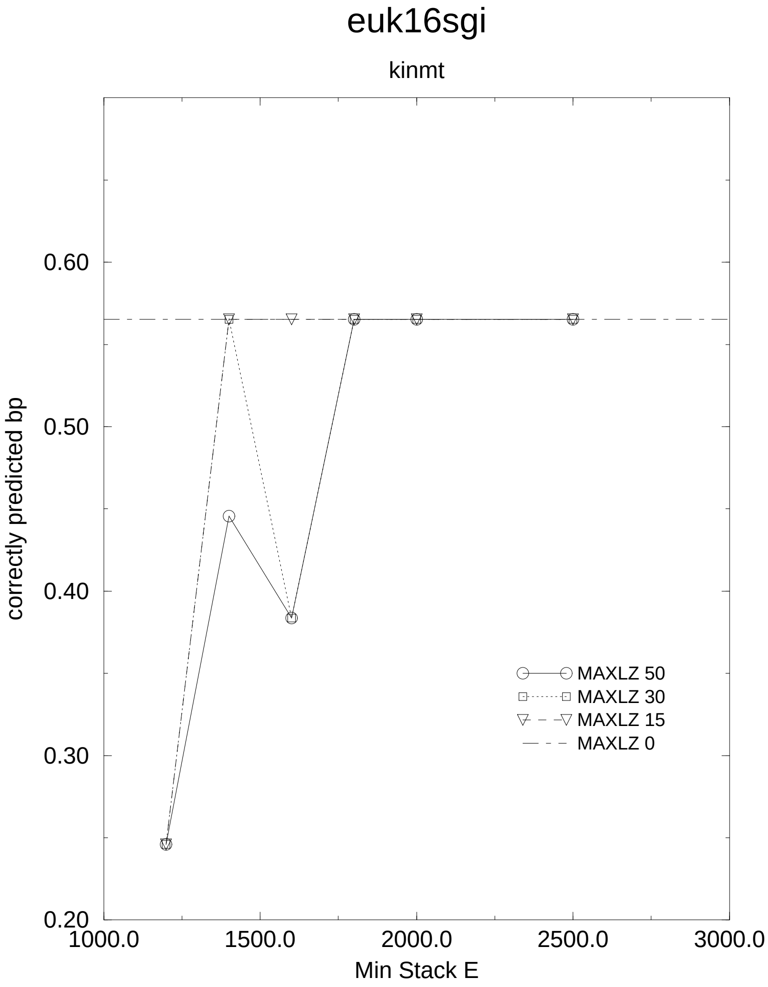


Figure 50: The number of correctly predicted base pairs of 16S rRNA from Methanococcus vannielli folded with the $5' \rightarrow 3'$ folding algorithm with different values of MAXLOOPSZ and STACKE in comparison with MFE(F)

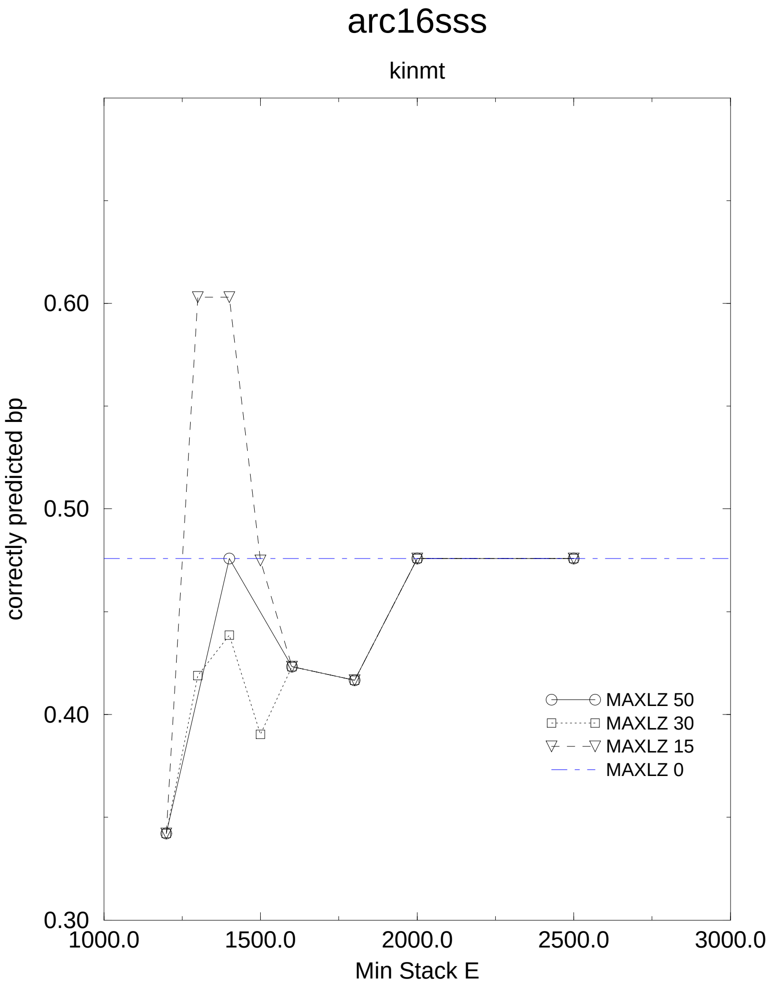


Figure 51: The number of correctly predicted base pairs of 16S rRNA from Sulfolobus solfaticus folded with the $5' \rightarrow 3'$ folding algorithm with different values of MAXLOOPSZ and STACKE in comparison with MFE(F)

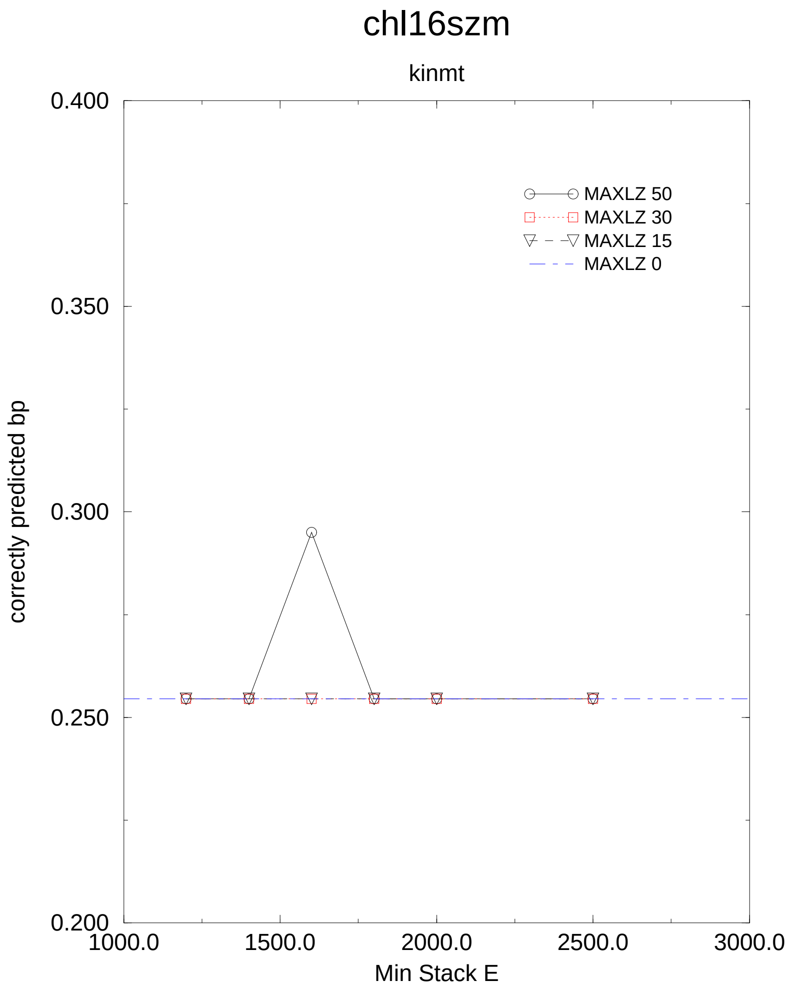


Figure 52: The number of correctly predicted base pairs of 16S rRNA from the chloroplast of Zea mays folded with the 5′ → 3′ folding algorithm with different values of MAXLOOPSZ and STACKE in comparison with MFE(F)

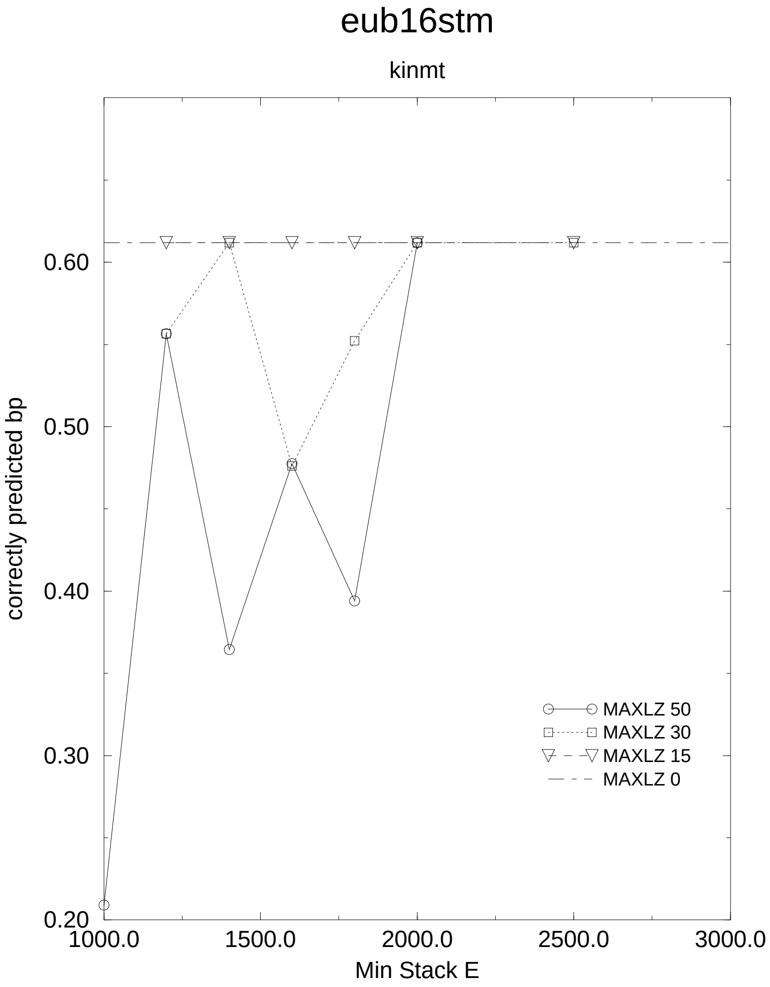


Figure 53: The number of correctly predicted base pairs of 16S rRNA from Thermatoga maritima folded with the $5' \to 3'$ folding algorithm with different values of MAXLOOPSZ and STACKE in comparison with MFE(F)

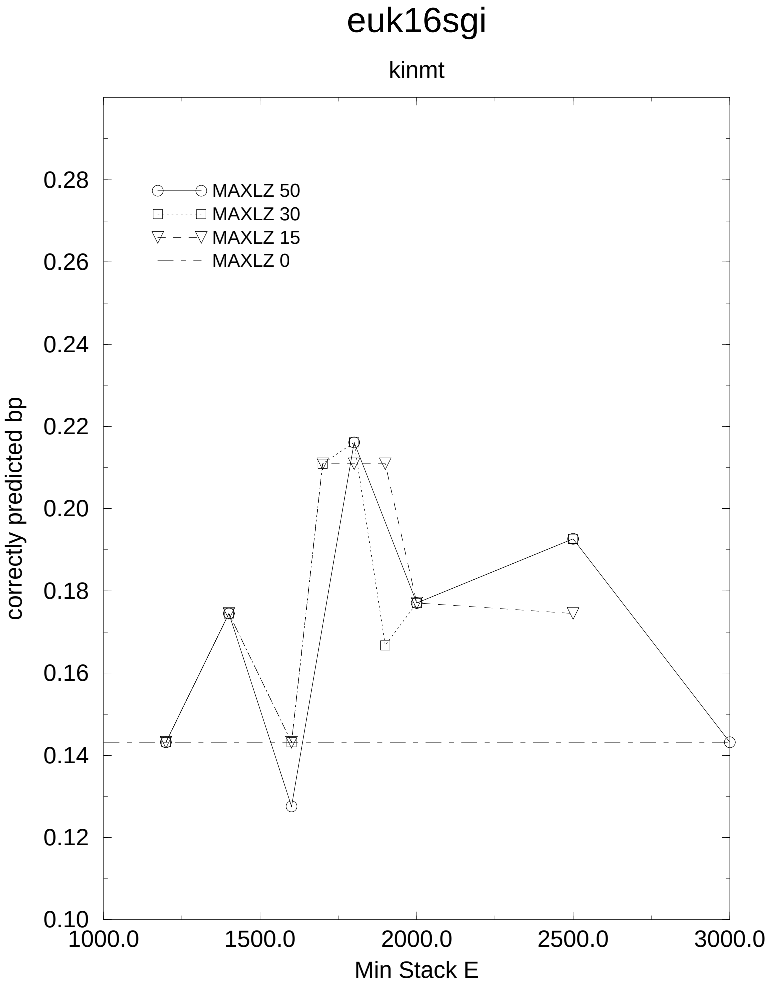


Figure 54: The number of correctly predicted base pairs of 16S rRNA from Giardia intestinalis folded with the $5' \rightarrow 3'$ folding algorithm with different values of MAXLOOPSZ and STACKE in comparison with MFE(F)

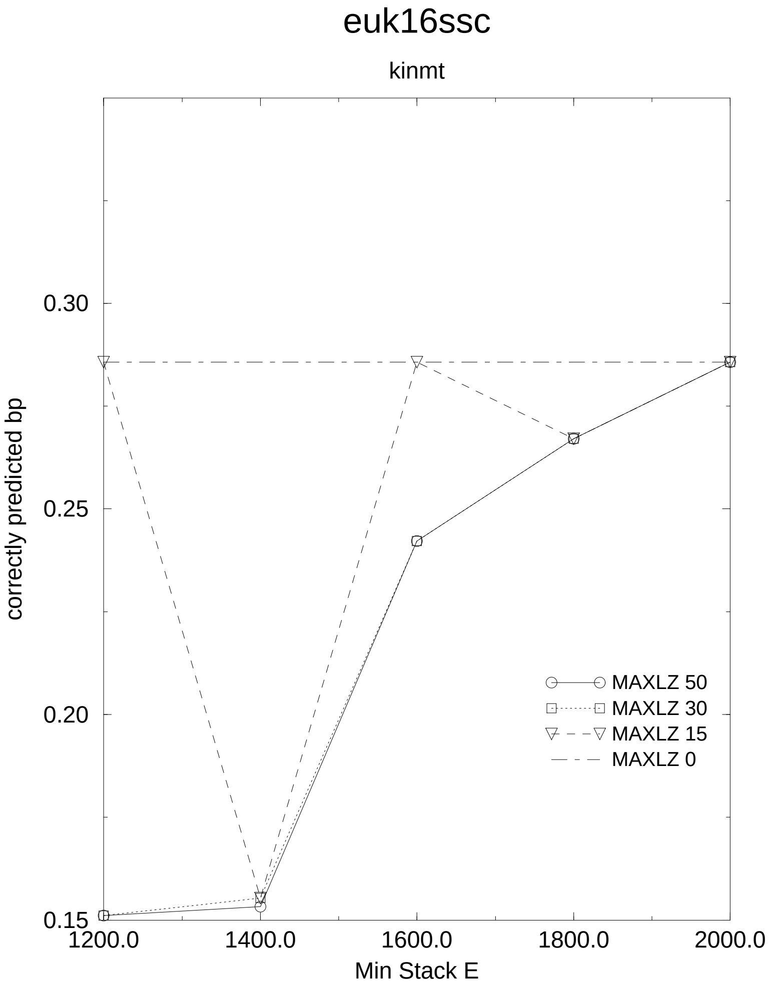


Figure 55: The number of correctly predicted base pairs of 16S rRNA from Saccheromyces cerevisiae folded with the $5' \rightarrow 3'$ folding algorithm with different values of MAXLOOPSZ and STACKE in comparison with MFE(F)

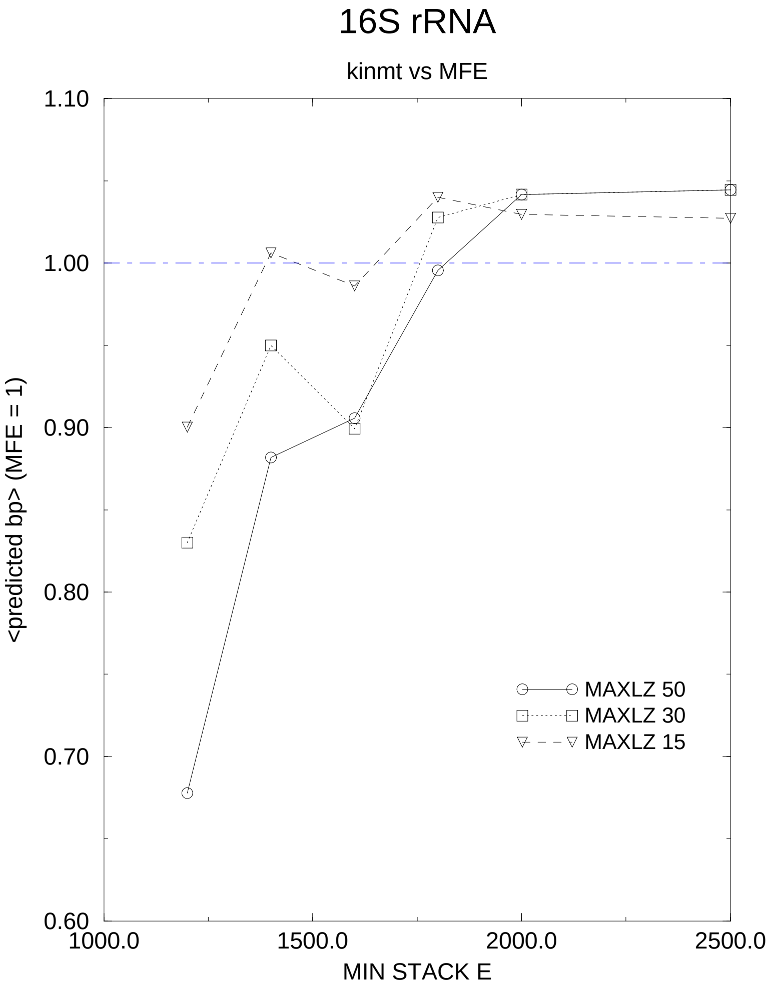


Figure 56: A summary of the percentage of correctly predicted base pairs by the $5' \rightarrow 3'$ folding algorithm for all 16S rRNAs. Prediction rate of MFE(F) = 1

10 Conclusion

It has been shown in this work that the prediction accuracy of folding algorithms depend heavily on the experimentally determined energy set - the energy sets compiled by Salser and the more recent update of Freier *et. al.* The secondary structures predicted with MFE, a dynamic programming algorithm which finds the thermodynamically most stable structure, with different parameter sets differ enormously,for sequences with chain length $\nu = 100$ the number of identically predicted structures is negligibly small. The MFE and a kinetical folding algorithm (KIN) with the same parameter set show better correspondence but even in this case the secondary structures are very different. A third algorithm considered, the maximum matching algorithm, which maximizes the number of base pairs irrespectively of the energy set, predicts structures which have nothing in common with the structures generated by the other algorithms.

In order to obtain generic results for the complex combinatory map of secondary structures it consequently is essential to concentrate on robust properties of this mapping, that is properties which occur with every algorithm and parameter set used.

One such robust result is the distribution of secondary structures which follows Zipf's first law indicating that there exist a few common structures and very many rare ones. This result holds for the two letter **GC** alphabet and the natural **GCAU** alphabet with the frequence of occurence for the most common structures in the **GC** alphabet much higher than in the four letter alphabet.

The statistic of secondary structure elements shows that general features are the same for all algorithms, parameter sets and alphabets:

- The number of base pairs, loops, stacks as well as the mean free energies of secondary structures scales linearly with ν

- The mean stack size and the mean loop size get constant for large ν (the rate of convergence for the KIN algorithm being very slow though)

Although the qualitative features of the structure statistics are robust there exist significant differences between the algorithms, for example the structures generated by the KIN algorithm thend to have less base pairs, larger mean loop and stack size. This structures are less stable than the MFE structures.

The autocorrelation function for all algorithms, alphabets and parameter sets decay exponentially, the approximation as an AR(1) landscape with an autocorrelation length l [85] is thus meaningful in each case. Both l_E and l_S scale linearly with ν. The correlation lengths l_E of the free energy landscapes are always larger than the corresponding l_Ss of the CM of secondary structures. The l_Ss are very small indicating a very rugged landscapes. This would suggest very difficult conditions for adaptation because it is generally assumed that autocorrelation lengths are a good measure for the performance of optimization runs on landscapes. The number of local optima on rugged landscapes are very large and it is easy to get stucked.

This assumption is disproved by the fact that there exist extended neutral nets for each alphabet, algorithm and parameter set. In the case of the biophysical **GCAU** alphabet most networks percolate through the whole sequence space, even with very large networks for the structures which occur at low frequency. The size of neutral networks depends on the alphabet, it is smaller for the two letter alphabets due to combinatorial reasons and on the ration of sequences to structures.

The importance of neutral nets for molecular adaptation can not be overestimated because it suggests a plausible mechanism for optimization: Once a sequence can be found which folds into a secondary structure catalytically active (the probability of randomly finding a suitable sequence is given by the probability of occurence of in the structure distribution) the fine tuning of the molecule is performed by walking via one and two point mutations on this network. The conservation of the secondary structure ensures a high probability of catalytic activity at each step.

This mechanism is important as a theoretical foundation for molecular evolution (and for selex experiments) but it also shows that efficient optimization can take place on a higly rugged landscapes with small autocorrelation lengths. What is more important is the fine structure of the landscape, a measure of which is the extension of neutral nets.

Folding of an RNA molecule *in vivo* proceeds from the $5'$ to the $3'$ direction. This folding process was simulated with a novel folding algorithm (KIN53) to investigate if long RNA sequences are folded into their thermodynamically most stable structure of if they are kinetically determined. The result of this investigation showed that there was no better prediction rate of 16S rRNA secondary structures when using KIN53 compared to the MFE(F) algorithm. This would suggest that kinetical trapping does not play a role in RNA folding, but this result can not be relied on because both algorithms predict only about phylogenetical structure correctly and the free energy values found for these structures and the phylogenetic structures differ enormously. This suggests that the parameter set is either inaccurate and incomplete (especially there are no experimentally determined data available for multiloops) or that the binding of proteins to RNA changes the secondary structure drastically.

Last but not least a methodology was developed to characterize the stability of secondary structures. The escape rate from the secondary structure gives the activation energy for melting, the distribution of the activation energies and the structure distances for all paths characterize in detail the basin of attraction.

A Uniform Sampling on Landscapes

Random walks possess very bad convergence properties of statistical quantities. This is due to the small number of independent samples, which is approximately $\frac{n}{2\lambda+1}$ [18]. It is therefore desirable to find a better sampling procedure. This has been found in the uniform sampling technique.
We start with a random sequence in configuration space and generate a sample of mutants which are hamming distance $h = 1$ to $h = (n-1)$ apart. Then the melting kinetic constants are computed and the difference D_k relative to the chosen reference are stored.

$$D_k = |X_{ref} - X_i(D)|$$

We get the distances D_k with the condition that the sequences are Hamming distance h apart. The conditional probability of finding a distance D given that two sequences have Hamming distance h is

$$p(D|h) \approx \frac{n(h,D)}{\sum_{D=0}^{D_{max}} n(h,D)}. \tag{3}$$

$n(h,D)$ denotes the number of occurences of distance D given h. The autocorrelation function can now be expressed in terms of conditional probabilities

$$\rho(h) = 1 - \frac{\sum_{D=0}^{D_{max}} D^2 n(h,D)}{\sum_{h=0}^{n} \sum_{D=0}^{D_{max}} D^2 n(h,D) p(h)}. \tag{4}$$

$p(h)$ denotes the probability that two randomly chosen sequences have Hamming distance h

$$p(h) = (\kappa - 1)^h \binom{\nu}{h} \kappa^{-\nu} \tag{5}$$

with κ being the number of letters in the alphabet and ν the chainlength. Convergence is much faster achieved with the uniform sampling technique than with random walks and therefore it was used in this work.

B Autocorrelation

Autocorrelation function:

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

X_i and X_{i+k} are random variables belonging to the same stochastic proces s.X_{i+k} is the value of the random variable exactly k times later than X_i. $cov(XY)$ is the covariance of X_i and X_{i+k}. It is defined as

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

An entirely uncorrelated stochastic process has an autocorrelation $\rho = 0$ for $k = 1, 2, \ldots$, a fully correlated process an autocorrelation $\rho(k) = +1$ for $k = 1, 2, \ldots$.

$$-1 \leq \rho(k) \leq +1$$

Due to the stationarity assumption of the stochastic time series, which can be expressed in the form

$$\langle X_i^r X_j^s\rangle = \langle X_{i+k}^r X_{j+k}^s\rangle \qquad \text{for all} \quad i, j, k, r, s \tag{6}$$

the dominator of formula can be simplified:

$$\rho(k) = \frac{\langle (X_i - \langle X_i\rangle)((X_{i+k} - \langle X_{i+k}\rangle)\rangle}{\langle (X_i - \langle X_i\rangle)^2\rangle} \tag{7}$$

For the numerical calculations in this work a variant of equation (7) was used:

$$\rho(k) = 1 - \frac{\langle X_i^2\rangle - \langle X_i X_{i+k}\rangle}{\langle X_i^2\rangle - \langle X_i\rangle^2} \tag{8}$$

In equation (8) it can easily be seen that $\lim_{k\to 0}\rho(k) = 1$ and $\lim_{k\to\infty}\langle X_i X_{i+k}\rangle = \langle X_i\rangle\langle X_{i+k}\rangle = \langle X_i\rangle^2 = 0$, because the two variables X_i and X_{i+k} are independent for sufficiently large k.
In our case X_i represents a rate constant of the configuration $\vec{x}_i$. If the autocorrelation function decays slowly then neighbour-configurations have similar rate constants and the landscape ist smooth. The faster the autocorrelation function decays, the more rugged is the landscape .
If the autocorrelation function can be approximated by an exponential curve $\rho(k) = e^{-\lambda k}$, the correlation length $l = \lambda^{-1}$, that is the k-value where $\rho(k) = e^{-1}$, suffices to fully characterize the autocorrelation function. A large correlation length corresponds to a slowly decaying autocorrelation function.

C Representation of Secondary Structures

C.1 Trees

It is very useful to represent and investigate RNA secondary structures with graphs theoretical means.
A **graph** $G = (X, U)$ consists of a finite set X whose elements p are called vertices (nodes) together with a set U of two element subsets of X the elements q of which are called edges. G then is of the order q and of size p. For easier visualization graphs are very often represented by a planar diagram showing all vertices and edges.
Trees are acyclic graphs where all nodes are connected. It has no crossing edges and can therefore always be drawn in a plane.
Any RNA secondary structure S_i is a two dimensional graph of order $p = \#bp$ and of size $q = \nu$. The concept of trees has already been applied to RNAs by Michael Zuker and David Sankoff in their classical paper on the folding algorithm [93]. It was used for comparisons on a large scale by Bruce Shapiro [77, 78]. Paulien Hogeweg and Danielle Konings exploited another graphical method for the comparison of secondary structures which they called mountain representation [45].
The secondary structure S_i is converted into the corresponding tree T_i by the following rules:

1. Assign to each unpaired base a leaf node, and to each base pair one internal node

2. The parent of each tree-node is the internal node corresponding to that accessible base pair $k \cdot l \quad (k < l)$ which contains the smallest k. The parents of nodes corresponding to external bses and external base pairs is a virtual node representing the root of the tree.

3. Connect each node to its parent

A modification of the tree representation is the homeomorphically irreducible tree **HIT**. A HIT is generted from the original rooted tree by merging all nodes of the same class to one node and weighing this node with the number of merged nodes. The HIT contains all the information of the full tree.
Secondary structure, full tree and HIT are equivalent.

C.1.1 Different levels of resolution

For many statistical applications the full secondary structure is overloaded with details, e. g. to distinguish a $tRNA$ from all other secondary structure it is only the clover leaf structure (the over all shape) which is important and not the fine detail of the base pairing or the size of the variable loop. The same holds true for structure distribution of RNAs as will be seen below (Zipf's law). In order to generate different levels of resolution we introduce the concept of coarse graining. Coarse graining can be done either by omitting all nodes with weights smaller than a certain threshold [24] or by ignoring the particular arrangement of unpaired sequences in loops and labelling each stack with only the type of the loop enclosed by it [77, 78]. An even coarser level of description is obtained by ignoring all interior loops and bulges and focussing on the branching structure of the molecule only. For details see fig. 57

Mountain Representation:
Mountain representation was proposed by Hogeweg and Konings as an alternative encoding of secondary structures [37, 45]. It is equivalent to the full tree structure (fine graining) of secondary structures. Starting from the $5'$end of the string every unpaired base is represented by a horizontal line, every upstream bounded base by a ascending line and every downstream bounded

base by a descending line. Matching bases appear at the same height, hairpin loops as flat tops, interior loops and bulges as intermediate plateaus, helices as slopes and branching regions as valleys.

C.2 Distances between Secondary Structures

Tree Edit Distance:
A tree can be transformed into another tree by a series of tree edit operations with predefined costs. These tree edit operations consist of insertion, deletion , substitution and relabelling of bases or basepairs. For the costs of these edit operations see Tab. 9.
The distance $d(T_i, T_j)$ between two trees T_i and T_j is given by the sum of the costs along an editing path which is minimal with respect to this sum.

$$d(T_i, T_j) = d_t(i, j) \tag{9}$$

It can be shown, that this tree distance forms a metric in the space of trees [74]. The tree edit operations preserve the relative traversal order of the tree nodes. Tree editing can therefore be seen as a generalization of sequence alignment. For the limiting case of a tree which consists solely of leaves, tree editing reduces to the standard sequence alignment.

Base Pair Distance bpd:
Another way of determining a distance between secondary structures is to count the base pairs that are different in the two structures. This is a Hamming distance and thus form a metric in the space of secondary structures. As an instance of this distance consider the two secondarystructures
A: (((...(((...))).....))) and
B: (((...((())))))..... .
To transform A into B it is necessary to open 3 base pairs and to close 3 base pairs. The base pair distance is thus six.

Edit Operation	Edit Cost
Single base insertion	1
Base pair insertion	2
Single base deletion	1
Base pair deletion	2
Relabel	0
Substitution	1

Table 9: Costs of tree edit operations

A: (((...(((...))).....))) → ......(((...)))........ bpd = 3

......(((...)))........ → B: (((...(((...))))))..... bpd = 3.

The base pair distance is very useful when comparing secondary structures of the same RNA sequence.

Alignment Distance:

This is the application of standard sequence alignment [74] methods to fine grained structures.

We used this method in order to verify that the specific structure distance chosen does not change the qualitative results presented in this work.

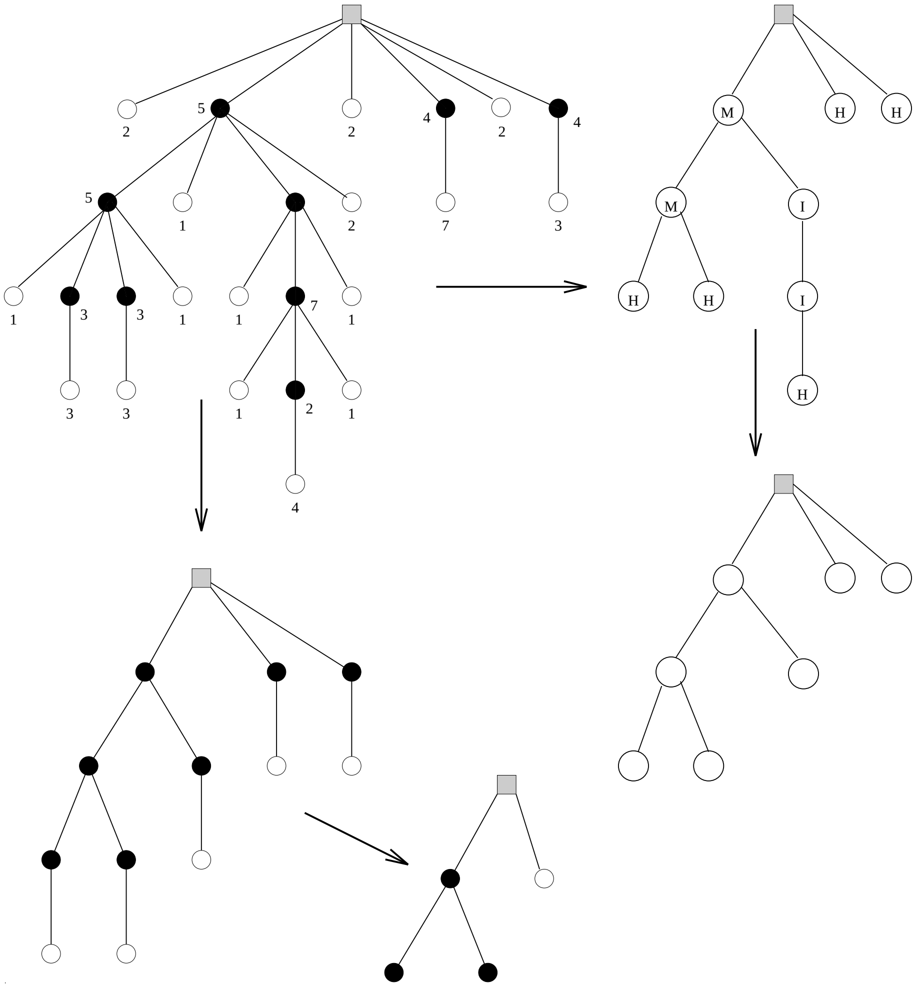


Figure 57: The transformation of the secondary structure to a rooted tree and the transformation to a HIT and coarse graining

D Freier Parameter Set

he parameter set is taken from [39] and [84].

Free energies for stacked pairs(0.1 kcal)						
$5'/3'$	AU	CG	GC	GU	UA	UG
AU	-9	-21	-17	-5	-9	-7
CG	-18	-29	-20	-15	-17	-15
GC	-23	-34	-29	-13	-21	-19
GU	-5	-19	-15	-5	-7	-5
UA	-11	-23	-18	-7	-9	-5
UG	-7	-13	-15	-6	-5	-5

Free energies for loops[1](0.1 kcal)[2]:

---

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

[2] Large destabilizing energies (9999) are assigned to hairpins with less than three unpaired bases

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

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

E Salser Parameter Set

The parameter set is taken from [73].

Free energies for stacked pairs(0.1 kcal)						
$5'/3'$	AU	CG	GC	GU	UA	UG
AU	-12	-21	-21	-3	-18	-3
CG	-21	-48	-30	-13	-21	-13
GC	-21	-43	-48	-13	-21	-13
GU	-3	-13	-13	-3	-3	-3
UA	-18	-21	-21	-3	-12	-3
UG	-3	-13	-13	-3	-3	-3

Free energies for loops[3](0.1 kcal)[4]:

hairpin AU : 0, 999, 999, 80, 75, 69, 64, 66, 68, 69, 70, 70, 71, 72, 73, 73, 74, 74, 75, 75, 76, 76, 76, 76, 76, 77, 77, 77, 77, 77, 79, 79

hairpin GC : 0, 999, 999, 84, 59, 41, 43, 45, 46, 48, 49, 49, 50, 51, 52, 52, 53, 54, 54, 55, 55, 55, 55, 56, 56, 57, 57, 57, 58, 58, 59, 59

bulge : 0, 28, 39, 45, 50, 52, 53, 55, 56, 57, 58, 58, 59, 60, 61, 61, 62, 62, 63, 63, 64, 64, 64, 65, 65, 65, 65, 65, 66, 66, 67, 67

internal loop AU&AU : 0, 10, 18, 26, 33, 38, 42, 43, 44, 45, 46, 46, 48, 48, 49, 49, 50, 50, 51, 51, 52, 52, 52, 52, 52, 54, 54, 54, 54, 54, 56, 56

internal loop GC&GC : 0, 999, 1, 9, 16, 21, 25, 26, 27, 28, 29, 30, 31, 31, 32, 32, 33, 33, 34, 34, 35, 35, 36, 36, 36, 37, 37, 37, 38, 38, 39, 39

internal loop AU&GC : 0. 999, 9, 17, 24, 29, 33, 34, 35, 36, 37, 38, 39, 39, 40, 40, 41, 41, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 46, 46, 47, 47

---

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

[4] Large destabilizing energies (999) are assigned to hairpins with less than three unpaired bases

F List of Abbreviations

CM:	Combinatory map
$C(S)$:	Sequence space
h:	Hamming distance
E_{act}:	Activation energy
G:	Free energy
k_d:	Dissociation rate
KIN:	Kinetic algorithm with the Freier dataset
KIN53:	$5' \rightarrow 3'$ folding algorithm with the Freier dataset
l_E:	Free Energy Autocorrelation length
l_S:	Structure Autocorrelation length
M:	Escape rate
MFE(F):	Minimal free energy folding with the Freier dataset
MFE(S):	Minimal free energy folding with the Salser dataset
MM:	Maximum matching algorithm
ν:	Chain length of the RNA sequence
P:	Stickiness
PCR:	Polymerase chain reaction
R:	Universal gas constant
$\rho(d)$:	Autocorrelation function
T:	Temperature
τ:	Mean dissociation time

References

[1] J.P. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Research*, 18(10):3035 – 3044, 1990.

[2] D.P. Bartel and J.W. Szostak. Isolation of a new ribozyme from a large pool of random sequences. *Science*, 261:1411, 1993.

[3] A.A. Beaudry and G.F. Joyce. Directed evolution of an RNA enzyme. *Science*, 257:635 – 641, 1992.

[4] J.A. Bluestone, S.O. Sharrow, S.L. Epstein, K. Ozato, and D.H. Sachs. Induction of anti-h2 antibodies in the absence of alloantigen exposure by in vivo administration of anti-idiotype. *Nature*, 219:233, 1981.

[5] 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.

[6] F. Buckley and F. Harrary. *Distances in Graphs.* Addison-Wesley, Reading, Ma., 1990.

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

[8] T.R. Cech. Ribozyme engineering. *Current Opinion in Structural Biology*, 2:605, 1992.

[9] M. Chamberlin and T. Ryan. volume 15 of *The Enzymes*, pages 87–108. Academic Press, New York, 1982.

[10] M. Chastain and I. Tinoco Jr. Structural elements in RNA. *Proc. Nucl. Acids Res*, 41:131–177, 1991.

[11] Y.S. Chen. Zipf's law in text modeling. *Int. J. General Systems*, 15:233–252, 1989.

[12] C.T. Chu, D.S. Parris, R.A.F. Dixon, F.E. Farber, and P.A. Schaffer. *Virology*, 98:168, 1979.

[13] Charles Darwin. *The Origin of Species.* reprinted in Penguin Classics, 1859.

[14] G. Dueck and P. Winkler. New concepts and algorithms for portfolio choice. Technical Report 90.06.006, IBM, IBM Germany, Heidelberg Scientific Center, June 1990.

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

[16] A.D. Ellington and J.W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346:818 – 824, 1990.

[17] A. Fernandez. *Eur. J. Biochem.*, 182:161, 1989.

[18] A.M. Ferrenberg, D.P. Landau, and K. Binder. Statistical and systematic errors in monte carlo sampling. *J. Stat. Phys.*, 63:867, 1991.

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

[20] W. Fontana, T. Griessmacher, W. Schnabl, P.F. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Monatshefte der Chemie*, 122:795, 1991.

[21] W. Fontana, D.A.M. Konings, P.F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *submitted to Biopolymers*, 1993.

[22] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaptation. *Phys.Rev.A*, 40(6):3301–3321, 1989.

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

[24] 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. *Phys. Rev. A*, 47:2083–2099, 1993.

[25] H. Frauenfelder, S.G. Sligar, and P. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598, 1991.

[26] 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.

[27] D.E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, Reading (Mass.), 1989.

[28] M. Gouy, P. Martiere, C. Papanicolau, and J. Ninio. Prediction des structures secondaires dans les acides n ucleiques: aspects algorithmiques et physiques. *Biochimie*, 67:67–79, 1985.

[29] A. P. Gultyaev. The computer simulation of RNA folding involving pseudoknot formation. *Nucleic Acids Research*, 19(9):2489 – 2494, 1991.

[30] R.R. Gutel. *Evolutionary Characteristics of 16S and 23S rRNA sequences*. World Scientific, Singapore, 1992.

[31] R.R. Gutel. Comparative studies of RNA: Inferring higher order structure from patterns of sequence variation. *Current Opinion in Structural Biology*, 3:313–322, 1993.

[32] R.R. Gutel and C.R. Woese. Higher order structural elements in ribosomal rRNAs: Pseudoknots and the use of non-canonical pairs. *PNAS*, 87:663–667, 1990.

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

[34] R.W. Hamming. *Coding and Information Theory*. Prentice Hall, Englewood Cliffs, 1986.

[35] L. He, R. Kierzek, J. SantaLucia, A.E. Walter, and D.H. Turner. Nearest-neighbour parameters for g-u mistatches. *Biochemistry*, 30:11124–11132, 1991.

[36] I. Hofacker, P. Schuster, and P.F. Stadler. Combinatorics of secondary structures. *submitted to SIAM J. Disc. Math.*, 1993.

[37] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nuc. Ac. Res*, 12:67–74, 1984.

[38] J.J. Hopfield and D.W. Tank. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79:2554, 1982.

[39] John A. Jaeger, Douglas H. Turner, and Michael Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, 86:7706, 1989.

[40] S. A. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.*, 128:11, 1987.

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

[42] S.A. Kauffman. *The Origin of Order.* Oxford University Press, New York, Oxford, 1993.

[43] S.H. Kim, F.L. Suddath, G.J. Quigley, A. McPherson, J.L. Sussman, A.H.J. Wang, N.C. Seeman, and A. Rich. Three dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science*, 185:435, 1974.

[44] S. Kirkpatrick, D.C. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671, 1983.

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

[46] A.I. Lamond and T.J. Gibson. Catalytic RNA and the origin of genetic systems. *TIG*, 6:145–149, 1990.

[47] E. Lawler, A. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys. *The Traveling Salesman Problem.* Wiley, New York, 1985.

[48] S.-Y. Le and M. Zuker. Common structures of the 5' non-coding RNA in enterovirus and rhinovirus. *J.Mol.Biol.*, 216:729, 1990.

[49] C.A. Macken and A.S.Perelson. *Affinity Maturation on Rugged Landscapes*, volume IX of *Santa Fe Institute Studies in the Science of Complexity*, page 93. Addison-Wesley, Redwood City, Ca., 1989.

[50] A. Malhotra, H.A. Gabb, and S.C. Harvey. Modeling large nucleic acids. *Current Opinion in Structural Biology*, 3:241–246, 1993.

[51] B. Mandelbrot. An information theory of the statistical structure of language. In *Proceedings of the Symposium on Applications of Communications Theory*, London, 1953. Butterworths.

[52] Hugo M. Martinez. An RNA folding rule. *Nucleic Acids Research*, 12(1):323 – 334, 1984.

[53] M.D. Matteucci and H.L. Heyneker. *Nucleic Acids Research*, 11:3113, 1983.

[54] A.M. Maxam and W. Gilbert. *PNAS*, 74:560–564, 1977.

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

[56] G. Parisi Mezard M. and M. A. Virasoro. *Spin Glass Theory and Beyond*, volume 9 of *World Scientific Lecture Notes in Physics*. World Scientific, Signapore, 1987.

[57] S.L. Miller and L.E. Orgel. *The Origin of Life on the Earth*. Prentice Hall, 1974.

[58] D.R. Mills, R.L. Peterson, and S. Spiegelman. An extracellular darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Nat. Acad. Sci. USA*, 58:217, 1967.

[59] A:A. Mironov, L:P. Dyakonova, and A.E. Kister. A kinetic approach to the prediction of RNA secondary structures. *Journal of Biomolecular Structure and Dynamics*, 2(5):953 – 962, 1985.

[60] A.A. Mironov and A.E. Kister. RNA secondary structure formation during transcription. *Journal of Biomolecular Structure and Dynamics*, 4(1):1 – 9, 1986.

[61] J. Ninio. Prediction of pairing schemes in RNA molecules - loop contributions and energy of wobble and non wobble pairs. *Biochimie*, 61:1133–1169, 1979.

[62] R. Nussinov, G. Pieczenik, J.R. Griggs, and D.J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35:68, 1978.

[63] A.J. Oparin. *The Origin of Life on Earth.* Academic Press, New York, 1957.

[64] N.R. Pace, G.J. Olson, D.K. Smith, and B.D. James. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA - a review. *Gene*, 82:65–75, 1989.

[65] A.S. Perelson and G.F. Oster. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J. theoret. Biol.*, 81:645 – 670, 1979.

[66] 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.

[67] J.M. Pipas and J.E. McMahon. Method for predicting RNA secondary structures. *PNAS*, 72:2017–2021, 1975.

[68] C.W.A. Pleij and L. Bosch. *RNA Pseudoknots: Structure, Detection and Prediction*, page 289. Academic Press, methods in enzymology 180 edition, 1989.

[69] D. Poerschke. Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix. *Biophys. Chem.*, 1:381, 1974.

[70] D. Poerschke. Elementary steps of base recognition and helix-coil transitions in nucleic acids. In I. Pecht and R. Riegler, editors, *Molecular Biology, Biochemistry and Biophysics Vol. 24*, pages 191–218. Springer, 1977.

[71] A.M. Pyle, F.L. Murphy, and T.R. Cech. *Nature*, 358:123, 1992.

[72] W. Saenger. *Principles of Nucleic Acid Structure.* Springer, London, 1984.

[73] W. Salser. Globin messenger RNA sequences - analysis of base-pairing and evolutionary implications. *Cold Spring Harbour Symp. Quant. Biol.*, 42:985–1002, 1977.

[74] D. Sankoff and J.B. Kruskal. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison Wesley, London, 1983.

[75] J. SantaLucia, R. Kierzek, and D.H. Turner. Context dependence of hydrogen bond free energy revealed by substitutions in an rna hairpin. *Science*, 256:217–219, 1992.

[76] P. Schuster. Complex optimization in an artificial RNA world. In D. Farmer, C. Langton, S. Rasmussen, and C. Taylor, editors, *Artificial Life II*, volume XII of *SFI Studies in the Science of Complexity.* Addison-Wesley, 1991.

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

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

[79] D. Shortle and D. Botstein. volume 100 of *Methods Enzymol.*, page 457. 1983.

[80] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 17:213–253, 1971.

[81] Peter F. Stadler and Wolfgang Schnabl. The landscape of the traveling salesman problem. *Phys. Letter A*, 161:337, 1992.

[82] L. Stryer. *Biochemistry.* Freeman, 1990.

[83] M. Tacker, W. Fontana, P.F. Stadler, and P. Schuster. Statistics of RNA melting kinetics. *submitted to Eur. J. Biophys.*, 1993.

[84] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Ann. Rev. Biophys. Biophys. Chem.*, 17:167–192, 1988.

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

[86] J.A. Wells, M. Vasser, and D.B. Powers. *Gene*, 34:315, 1985.

[87] C.R. Woese, R. Gutel, R. Gupta, and H.F. Noller. *Microbiol. Rev.*, 47:621–669, 1983.

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

[89] R.A. Zakour and L.A. Loeb. *Nature*, 295:708, 1982.

[90] Y. Zhou, X. Zhang, and R.H. Ebright. *Nucleic Acids Research*, 19:6052, 1991.

[91] G.K. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, Reading(Mass.), 1949.

[92] M. Zuker, J.A. Jaeger, and D.H. Turner. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nuc. Ac. Res.*, 19:2707–2714, 1991.

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

Curriculum Vitae

Name : Tacker Manfred
Geburtsdatum : 21. September 1962

Schulbildung:

1968-1976 : Volks- und Hauptschule
1976-1980 : Fachschule für Technische Chemie, Wien 17, Rosensteingasse; Abschluß mit ausgezeichnetem Erfolg
Sept.1983-Juni1987 : Abendschule am BRG für Berufstätige, Wien 15; Matura mit ausgezeichnetem Erfolg
Okt.1987-April1992 : Diplomstudium der Biochemie an der Universität Wien
Juli1991-April1992 : Diplomarbeit am Institut für Theoretische Chemie der Universität Wien Thema: Schmelzkinetik von RNA; Abschluß des Studiums mit ausgezeichnetem Erfolg
ab April 1992 : Dissertation am Institut für Theoretische Chemie der Universität Wien

Berufliche Laufbahn:

Juni1980-Feb.1981 : Präsenzdienst beim LWStR 25
März1981-Mai1981 : Laborant bei Woidich&Co, Wien 19;
Juni1981-Nov.1985 : Chemotechniker bei Bender&Co, Wien 12;
Dez.1985-Mai1986 : Diagnostikaberater bei Pharmacia&Co, Wien 23;
Juni1986-Sept.1987 : Chemotechniker im Sandoz Forschungsinstitut, Wien 23;
Aug.1988-Jan.1989 : Werksstudent am Sandoz Forschungsinstitut;
Juli1989-Sept.1989 : FEBS-Fellow an der St. George's Hospital Medical School, London, UK;

Publikationen:

[1] Fowler, P.D., M. Tacker, G.St.J. Whitley, A. Meager, S.S. Nussey and A.P. Johnstone (1990) *Expression of Intercellular Adhesion Molecule-1 (ICAM-1) on Human Thyroid Cellines Correlated with their Binding of Lymphoblasts.* Mol. Cell. Endocr. 71:55-61.

[2] Fontana, W., P.F. Stadler, E. Bauer, T. Griesmacher, I.L. Hofacker, M. Tacker, P. Tarazona, E.D. Weinberger and P. Schuster (1993) *Statistical Properties of RNA Free Energy Landscapes.* Phys. Rev. E 47:2083-2099.

[3] Tacker, M., W. Fontana, P.F. Stadler and P. Schuster (1993) *Statistics of RNA Melting Kinetics* submitted to Eur. J. Bioph.

[4] Hofacker, I.L., W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker and P. Schuster (1993) *Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package).* SFI-preprint 93-07-44. Mh. Chem. in press.

[5] Tacker, M., P.F. Stadler and P. Schuster (1993) *Robust Properties of RNA Landscapes.* in preparation.

[6] Tacker, M. and P.F. Stadler (1993) *RNA: Genotype and Pheotype.* in Proceedings of the Studies in Complexity, Santa Fe Institute. Addison-Wesley, The Advanced Book Program, Redwood City.