View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Qucosa - Publikationsserver der Universität Leipzig

Computational Chemistry with RNA Secondary Structures

Christoph Flamm[†], Ivo L. Hofacker[†], Peter F. Stadler[♣,†]

[†]Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstraße 17, A-1090 Wien, Austria

[♣]Bioinformatik, Institut für Informatik, Universität Leipzig, Kreuzstrasse 7b, D-04103 Leipzig, Germany

Abstract. The secondary structure for nucleic acids provides a level of description that is both abstract enough to allow for efficient algorithms and realistic enough to provide a good approximate to the thermodynamic and kinetics properties of RNA structure formation. The secondary structure model has furthermore been successful in explaining salient features of RNA evolution in nature and in the test tube. In this contribution we review the computational chemistry of RNA secondary structures using a simplified algorithmic approach for explanation.

Keywords: Nucleic Acids, RNA Folding, Structure Prediction, RNA Evolution.

1. Introduction

Computational Chemistry is often used as a synonym for Quantum Chemistry. On the other hand, a relatively small number of measurably physical parameters together with the knowledge of the structure formula often allows quite accurate predictions of other thermodynamic, kinetic, or functional properties of a molecule. As a example consider Hammet's classical theory of substituent effects expressed in terms of σ and ρ parameters, see e.g. [24]. The existence of such semi-empirical laws is also the basis of QSAR methods [11]. In this sense much of the working knowledge of the organic chemist can be regarded as a coarse grained picture of the underlying quantum-theory of molecules and their reactions.

Nucleic acids are unique among molecular systems because they admit a level of description that is coarse grained even further: their *secondary structures* are sufficient to predict sequence specific thermodynamic and kinetic properties without recourse to an atom-by-atom model of the molecule. Here we review the questions and computational techniques that can be employed at this level of description.

This contribution is organized as follows: In section 2 we outline so-called nearest neighbor energy model for RNA and DNA structures. Then we consider the basic dynamic programming algorithms for obtaining various aspects of the thermodynamics of nucleic acid structure formation and stability. Since the algorithms become

rather complicated due to the many case distinctions implicit in the standard energy model we instead use a simplified variant, the so-called maximum circular matching problem, to make the basic ideas transparent. The basic bioinformatics questions that can be posed for nucleic acid structures — structural alignments and pattern search — give rise to algorithmic solutions that are close relatives of the folding algorithms as we shall see in sections 5 and 6. Finally we will see that even the kinetics of the folding process can be described consistently at the level of secondary structures.

2. Secondary Structure Graphs and Their Free Energies

We consider nucleic acid structures at a coarse-grained level, representing each nucleotide by a single point. Instead of spatial coordinates, only covalent and non-covalent contacts (the latter correspond to specific hydrogen bonds) are used. In other words, only the DNA or RNA sequence and the list of *base pairs* enter our considerations.

A *secondary structure* Ψ is a special type of contact structure, represented by a list of base pairs (i, j) with $i < j$ on a sequence x, such that for any two base pairs $[i, j]$ and $[k, l]$ with $i \leq k$ holds:

(i) $i = k$ if and only if $j = l$, and
(ii) $k < j$ implies $i < k < l < j$.

The first condition simply means that each nucleotide can take part in at most one base pair. The second condition forbids knots and pseudo-knots. While pseudo-knots are important in many natural RNAs [44], they can be considered part of the tertiary structure for our purposes. We will therefore neglect them for the purpose of this presentation. The restriction to knot-free structures is necessary for the efficient dynamic programming algorithms discussed below.

The two conditions above imply that secondary structures form a special type of graphs. In particular, a secondary structure graph is *sub-cubic* (i.e., the vertex degree is at most three) and *outer-planar*. The latter property means that the structure can be drawn in the plane in such a way that all vertices (which represent the nucleotides) are arranged on a circle (the molecule's backbone), and all edges (which represent the bases pairs) lie inside the circle and do not intersect, see Fig. 1.

The physico-chemical basis for the course grained computational chemistry of nucleic acids is the possibility to compute the free energy of structure formation given the sequence and the list of base pairs. Note that a secondary structure as defined above corresponds to an *ensemble* of conformations of the molecule at atomic resolution restricted to a certain base pairing (hydrogen bonding) pattern. For example, no information is assumed about the spatial conformation of unpaired regions. The entropic contributions of these restricted conformations have to be taken into account, hence we are dealing with (temperature dependent) free energies here.

The energy of an RNA secondary structure is assumed to be the sum of the energy contributions of all "loops", i.e., the faces of the planar drawing of the structure. This decomposition has a solid graph theoretical foundation [23]: the loops form

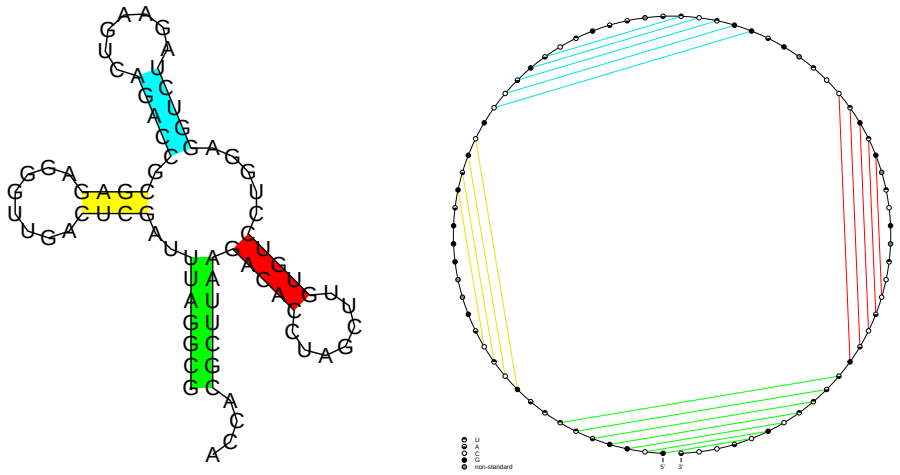

Figure 1. Secondary structure of phenylalanine-tRNA from yeast as conventional drawing and in circular representation. The chords in the circular representation must not cross in secondary structure graphs.

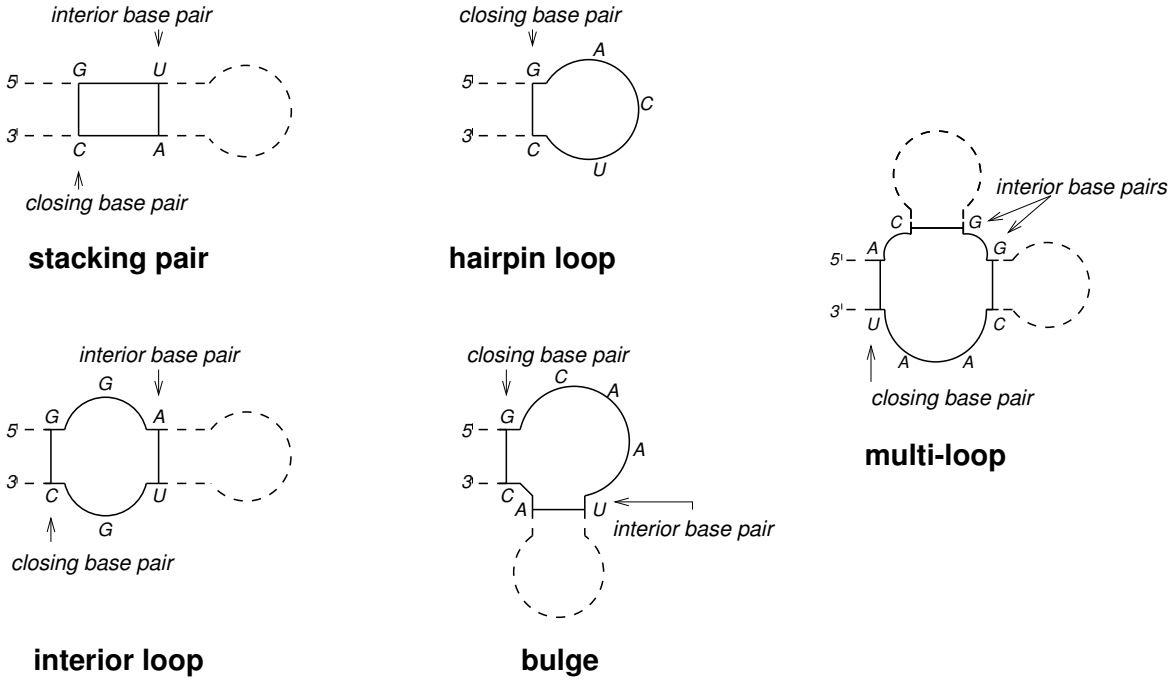

Figure 2. Secondary structure elements that form the basis of the energy model for nucleic acids.

the unique minimal cycle basis of the secondary structure graph. More importantly, however, a large number of careful melting experiments have shown that the energy of structure formation (relative to the random coil state) is indeed additive to a good approximation, see e.g. [8, 21, 41, 25]. Usually, only Watson-Crick (**AU**, **UA**, **CG** and **GC**) and wobble pairs (**GU**, **UG**) are allowed in computational approaches since non-standard base-pairs have in general context-dependent energy contributions that do not fit into the "nearest-neighbor model". Individual non-standard base pairs are therefore treated as special types of interior loops in the most recent parameter sets.

Qualitatively there are two major energy contribution: Stacking of base pairs and loop entropies. Stacking energies can be computed for molecules in the vacuum by means of standard quantum chemistry approaches, see e.g. [32, 22, 14]. The secondary structure model, however, considers only energy differences between folded and unfolded states in an aqueous solution with rather high salt concentrations. As a consequence one has to rely on empirical energy parameters. Loops are destabilizing: the closing base pair restricts the possible conformations of the sequence in the loop relative to the conformations that could be formed by the same sequence segments in a random coil resulting in an entropy loss and thus an increase in free energy.

Here we explain all versions of RNA folding using the simplified *Maximum Circular Matching Problem* paradigm. This will allow us a relatively compact and intelligible representation of the basic idea behind dynamic programming RNA folding algorithms. In section 8 we will briefly return to the realistic energy model.

3. Forward Recursions

We begin our exposition by counting the secondary structures that can be formed by a given sequence $x = (x_1, x_2, \ldots, x_n)$ of length n. We will simply write "(i, j) pairs" to mean that the nucleotides x_i and x_j *can* form a Watson Crick or a wobble pair, i.e., $x_i x_j$ is one of GC, CG, AU, UA, GU, or UG. The basic idea behind all dynamic programming algorithms for RNA folding is the observation that a structure on n nucleotides can be formed in only two distinct ways from shorter structures: Either a structure on $n-1$ nucleotides is extended by an unpaired base, or the n^{th} nucleotide is paired. In the latter case it has a pairing partner, say j such that the (j, n) pair encloses a secondary structure on the sub-sequence from $j+1$ to $n-1$ since base pairs must not cross by definition. The remainder, the interval from 1 to $j-1$ is of course also a secondary structure:

`xxxxxxxxxxxxxx` = `.xxxxxxxxxxxxx` or `(yyyyyyy)zzzzzz`

It is now easy to compute the number N_{ij} of secondary structure on the sub-sequence $x[i..j]$ from positions i to j [42, 43]:

$$N_{ij} = N_{i+1,j} + \sum_{k,\,(i,k)\ \text{pairs}} N_{i+1,k-1} N_{k+1,j} \tag{1}$$

The first term accounts for the case in which position i is unpaired, the terms in the sum consider the base pairs from i to some position k. Because of the "no-pseudoknots" condition both the part of the sequence that is enclosed by the pair (i, k) and the part beyond the base pair form secondary structures that are completely independent of each other: thus we may simply multiply their numbers. This simple combinatorial structure of secondary structures was realized by M. Waterman in the late 1970s [42, 43]. It can be exploited to derive typical structural features of RNA molecules such as expected helix length or distribution of loop types [15].

Restricting ourselves to the number $N_{ij}(\epsilon)$ of structures with a fixed energy ϵ we can immediately generalize eq.(1) to a recursion for the density of states of an RNA

molecule [3, 2].

$$N_{ij}(\epsilon) = N_{i+1,j}(\epsilon) + \sum_{k,\,(i,k)\ \text{pairs}} \sum_{\epsilon'} N_{i+1,k-1}(\epsilon') N_{k+1,j}(\epsilon - \epsilon' - \beta_{ik}) \tag{2}$$

Energy minimization, as the first step towards computing the minimum free energy structure of an RNA molecule was historically the first variant of equ.(1), see [31, 48, 47]. The recursion for the minimal energy E_{ij} of any structure on the subsequence $x[i..j]$ is simply

$$E_{ij} = \max \left\{ E_{i+1,j} + \max_{k,\,(i,k)\ \text{pairs}} \{E_{i+1,k-1} + E_{k+1,j} + \beta_{ik}\} \right\} \tag{3}$$

The free energy parameters are here simplified to contributions β_{ik} for individual base pairs. Variants of this algorithm for the realistic, loop-based, energy model is implemented in Michael Zuker's `mfold` package[47, 46] and in the `Vienna RNA Package` [18, 16] by the present authors. Note that the free energy parameters $\beta_{ik} = \beta_{ik}(T)$ explicitly depend on the temperature as they contain both entropic and enthalpic contributions.

John McCaskill [27] observed that essentially the same recursion can be used to obtain the partition function $Z = \sum_\Psi \exp(-E(\Psi)/RT)$ over all possible secondary structures Ψ. For the partition function Z_{ij} over all structures on sub-sequence $x[i..j]$ one obtains

$$Z_{ij} = Z_{i+1,j} + \sum_{k,\,(i,k)\ \text{pairs}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\beta_{ik}/RT) \tag{4}$$

The partition function is starting point for exploring the thermodynamics of RNA secondary structure formation. The free energy of structure formation, for example is, $\Delta G = -RT \ln Z$. From this we may compute other thermodynamic parameters, e.g. melting curves.

4. Backtracking

Backtracking is the procedure that generates one or more RNA structures in a stepwise fashion based on the information collected in the forward recursions. The basic object is a *partial structure* π consisting of a collection Ω_π of base pairs and a collection Υ_π of sequence intervals in which the structure is not (yet) known. Positions that are known to be unpaired can easily be inferred from this information. The completely unknown structure on the sequence interval $[1, n]$ is therefore $\varnothing = (\emptyset, \{[1, n]\})$ while a structure is *complete* if it is of the form $\pi = (\Omega, \emptyset)$.

Suppose $I = [i, j] \in \Upsilon$ are positions for which the partial structure $\pi = (\Omega, \Upsilon)$ is still unknown. If we know that i is unpaired then $\pi' = (\Omega', \Upsilon')$ with $\Omega' = \Omega$ $\Upsilon' = \Upsilon \setminus \{I\} \cup \{[i + 1, j]\}$. If $(i, k), i < k \leq j$ is a base pair then $\Omega' = \Omega \cup \{(i, k)\}$ and $\Upsilon' = \Upsilon \setminus \{I\} \cup \{[i + 1, k - 1], [k + 1, j]\}$. Here we use the convention that empty intervals are ignored. Furthermore, base pairs can only be inserted within a single interval of the list Υ. We write $\pi' = \pi \blacktriangleleft (i)$ and $\pi' = \pi \blacktriangleleft (i, k)$ for these two cases.

Table 1. Comparison of backtracking recursions for different algorithms.

$\varnothing \rightarrow \mathfrak{S}$.
while $\mathfrak{S} \neq \emptyset$ **do**
  $\pi \leftarrow \mathfrak{S}$;
  if π is complete **then** output π
  $[i,j] = I \in \Upsilon_\pi$.
  $\pi' = \pi \blacktriangleleft (i+1)$
  if $E(\pi') = E_{\text{opt}}$ **then** $\pi' \rightarrow \mathfrak{S}$; **next**;
  for all $k \in [i,j]$ **do**
  $\pi' = \pi \blacktriangleleft (i,k)$
  if $E(\pi') \leq E_{\text{opt}}$ **then** $\pi' \rightarrow \mathfrak{S}$; **next**;
	$\varnothing \rightarrow \mathfrak{S}$.
	while $\mathfrak{S} \neq \emptyset$ **do**
	  $\pi \leftarrow \mathfrak{S}$;
	  if π is complete **then** output π;
	  for all $[i,j] = I \in \Upsilon_\pi$ **do**
	  $\pi' = \pi \blacktriangleleft (i+1)$
	  if $E(\pi') \leq E_{\text{opt}} + \Delta E$ **then** $\pi' \rightarrow \mathfrak{S}$;
	  for all $k \in [i,j]$ **do**
	  $\pi' = \pi \blacktriangleleft (i,k)$
	  if $E(\pi') \leq E_{\text{opt}} + \Delta E$ **then** $\pi' \rightarrow \mathfrak{S}$;
		$\varnothing \rightarrow \mathfrak{S}$.
		while $\mathfrak{S} \neq \emptyset$ **do**
		  $\pi \leftarrow \mathfrak{S}$;
		  if π is complete **then** output π;
		  for all $[i,j] = I \in \Upsilon_\pi$ **do**
		  $\pi' = \pi \blacktriangleleft (i+1)$
		  $\pi' \rightarrow \mathfrak{S}$ with probability $Z(\pi')/Z(\pi)$
		  for all $k \in [i,j]$ **do**
		  $\pi' = \pi \blacktriangleleft (i,k)$
		  $\pi' \rightarrow \mathfrak{S}$
		  with probability $Z(\pi')/Z(\pi)$

Algorithm B1. Backtracking a single structure [31, 48]	**Algorithm B2.** Backtracking of multiple structures [45]	**Algorithm B3.** Stochastic backtracking [19] `Vienna RNA Package` since version 1.5β.

The energy of a partial structure π is defined as

$$E(\pi) = \sum_{(k,l)\in\Omega} \beta_{kl} + \sum_{I\in\Upsilon} E_{\text{opt}}(I) \tag{5}$$

where $E_{\text{opt}}(I) = E_{ij}$ is the optimal energy for the substructure on the interval $I = [i, j]$

The standard backtracking for the minimal energy folding starts with the unknown structure. Instead of a recursive version we describe here a variant where incomplete structures are kept on a stack $\mathfrak{S}$. We write $\pi \leftarrow \mathfrak{S}$ to mean that π is popped from the stack and $\pi \rightarrow \mathfrak{S}$ to mean that π is pushed onto the stack.

If we want all optimal energy structures instead of a single representative we simply test all alternatives, i.e., we omit the **next** in the algorithm above. It is now almost trivial to modify the backtracking to produce all structures within an energy band $E_{\text{opt}} \le E \le E_{\text{max}}$ above the ground state.

Stochastic backtracking procedure for dynamic programming algorithm such as pairwise sequence alignment are well known [29]. Replacing Z_{ij} by N_{ij} in Algorithm B3 we recover recursions for producing a uniform ensemble of structures similar to the procedure for producing random structures without sequence constraint used in [39].

Note that the probabilities of $\pi \blacktriangleleft (i+1)$ and $\pi \blacktriangleleft (i,k)$ for all k add to 1 so that in each iteration we take exactly one step. Hence we simply fill one structure which we output as soon as it is complete.

5. The Sankoff Algorithm

Many functional classes of RNA molecules, including tRNA, rRNA, RNAse P RNA, SRP RNA, exhibit a highly conserved secondary structure but little sequence homology. In order to compare these molecules between different species it is therefore necessary to take structural information into account.

David Sankoff described an algorithm that simultaneously allows the solution of the structure prediction and the sequence alignment problem [34]. The basic idea is to search for a maximal secondary structure that is common to both molecules. Given a score $\sigma_{ij,kl}$ for the alignment of the base pairs (i, j) and (k, l) from the two sequences (as well as gap penalties γ and scores α_{ik} for matches of unpaired positions) we compute the optimal alignment recursively from alignments of the subsequences $x[i..j]$ and $y[k..l]$. Let $S_{ij,kl}$ be the score of the optimal alignment of these fragments. We have

$$\begin{aligned} S_{ij;kl} = \max\big\{ & S_{i+1,j;kl} + \gamma, S_{ij;k+1,l} + \gamma, S_{i+1,j;k+1,l} + \alpha_{ik}, \\ & \max_{(p,q)\text{ paired}} \{S_{i+1,p-1;k+1,q-1} + \sigma_{ij,pq} + S_{p+1,j;q+1,l}\} \big\} \end{aligned} \tag{6}$$

Backtracking is just as easy as in the RNA folding case. Only now π is a partial alignment of two structures and we insert aligned positions instead of positions in individual structures. More precisely we use

$$\pi \blacktriangleleft \begin{pmatrix} i. \\ \text{-} \end{pmatrix} \quad \pi \blacktriangleleft \begin{pmatrix} \text{-} \\ j. \end{pmatrix} \quad \pi \blacktriangleleft \begin{pmatrix} i. \\ j. \end{pmatrix} \qquad \pi \blacktriangleleft \begin{pmatrix} i(& & j) \\ p(& , & q) \end{pmatrix} \tag{7}$$

The algorithm is unfortunately very expensive, requiring $\mathcal{O}(n^4)$ memory and $\mathcal{O}(n^6)$ CPU time. Currently available software packages such as `foldalign` [10] and `dynalign` [26] therefore implement only restricted versions. The simple, maximum matching style version is used in [17] as an approach to comparing base pairing probability matrices.

It is straight forward to build a density of states and a counting version from this recursion. Its partition function variant is of particular interest since it could be used to assess the reliability of the structure based alignments. The basic recursion reads

$$Q_{ij;kl} = Q_{i+1,j;kl}e^{\gamma} + Q_{ij;k+1,l}e^{\gamma}Q_{i+1,j;k+1,l}e^{\alpha_{ik}} + \sum_{(p,q)\text{ paired}} Q_{i+1,p-1;k+1,q-1}Q_{p+1,j;q+1,l}e^{\sigma_{ij,pq}} \tag{8}$$

where we assume that the similarity scores γ, α, and σ are already properly scaled with the fictitious temperature RT.

6. Structural Patterns

The partition function formalism can be used to compute the probability that a sequence will form a particular structural pattern. For any pattern let Ω be the set of secondary structures that contain the pattern. We may then compute the partition function over all structures containing that pattern, $Z(\Omega) = \sum_{\Psi\in\Omega} \exp(-E(\Psi)/RT)$, and thus its probability $p(\Omega) = Z(\Omega)/Z$. For simple patterns it is often possible to compute $Z(\Omega)$ efficiently by dynamic programming without much extra effort.

The most common case is the computation of pair probabilities, i.e. $\Omega_{i,j}$ is the set of secondary structures that contain the pair (i, j). To compute these we introduce the partition function $\widetilde{Z}^{ij}$ of structures *outside* the sequence interval $[i, j]$. Since the pair (i, j) divides the structure in two independent halves, we have

$$p_{ij} = \widetilde{Z}^{ij} Z_{i+1,j-1} \exp(-\beta_{ij}/RT)/Z \tag{9}$$

The exterior partition functions $\widetilde{Z}^{ij}$ satisfy the recursion

$$\widetilde{Z}^{ij} = Z_{1,i-1}Z_{j+1,n} + \sum_{k<i;j<l} \widetilde{Z}^{kl} Z_{k+1,i-1}Z_{j+1,l-1} \exp(-\beta_{kl}/RT) \tag{10}$$

Pair probabilities provide a convenient and easy to visualize representation of structure ensembles, Fig. 3.

A similar technique works if the pattern is an arbitrary substructure σ of length d. The probability of finding σ at position i of the molecule is given by $p_i(\sigma) = Z_i^{\sigma}/Z$ with

$$Z_i^{\sigma} = \left(Z_{1,i-1}Z_{i+d,n} + \sum_{k<i;j<l} \widetilde{Z}^{kl} Z_{k+1,i-1}Z_{i+d,l-1} \right) \exp(-E(\sigma)/RT) \tag{11}$$

The case where σ is simply a stretch of d unpaired bases (and thus $E(\sigma) = 0$) is of particular interest when selecting target regions for gene silencing via designed siRNAs. In [4, 5] the same problem is approached by sampling structures using stochastic backtracking. The advantage of the exact approach is that relative rare structural elements, $p < 1/\text{sample-size}$, can be delt with.

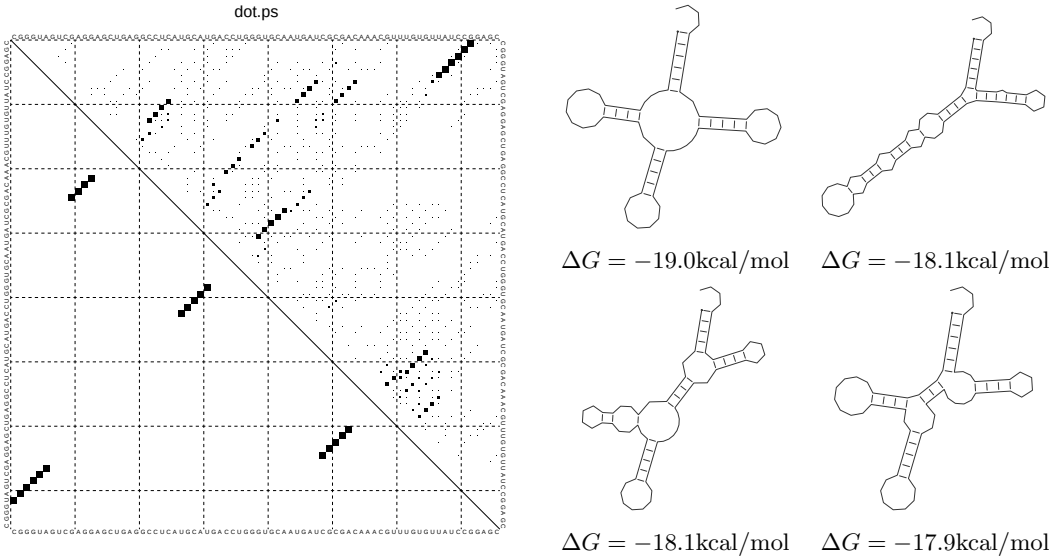

Figure 3. Base pairing probability matrix of an RNA with many nearly degenerate structures. The contact of the ground state structure are shown in the lower-left part of the matrix. The dots in the upper-right part have an area proportional to the pairing probability p_{ij}. Four examples of suboptimal structures within about $2kT$ from the ground state are shown on the left.

7. Kinetic Folding

The folding landscape (or Potential Energy Surface, PES [28, 12]) of a RNA molecule is a complex surface of the (free) energy versus the conformational degrees of freedom. In our case, the allowed conformations are of course the secondary structures which can be formed by a particular RNA sequence; the degrees of freedom are the allowed transformations provided by a "move set", in our case the insertion/deletion (closing/opening) of single base pairs. Two conformations x and y are said to be neighbors if they can be inter-converted by applying a single move from the move set. Instead of a smooth surface defined on a space of real-valued coordinate vectors we are therefore dealing with a landscape on a complex graph [33].

A conformation x is a global minimum if $E(x) \leq E(y)$ for all $y \in X$ and a local minimum if $E(x) \leq E(y)$ for all neighbors y of x. The energy $\hat{E}$ of the lowest saddle point separating two local minima x and y is

$$\hat{E}[x, y] = \min_{\mathbf{p} \in \mathbb{P}_{xy}} \max_{z \in \mathbf{p}} E(z) \tag{12}$$

where $\mathbb{P}_{xy}$ is the set of all paths $\mathbf{p}$ connecting x and y by a series of consecutive transformations taken from the move set. If the energy function is non-degenerate then there is a unique saddle point $s = s(x, y)$ connecting x and y characterized by $E(s) = \hat{E}[x, y]$. To each saddle point s there is a unique collection of conformations $B(s)$ that can be reached from s by a path along which the energy never exceeds $E(s)$. In other words, the conformations in $B(s)$ are mutually connected by paths that never go higher than $E(s)$. This property warrants to call $B(s)$ the *basin of attraction* below the saddle s.

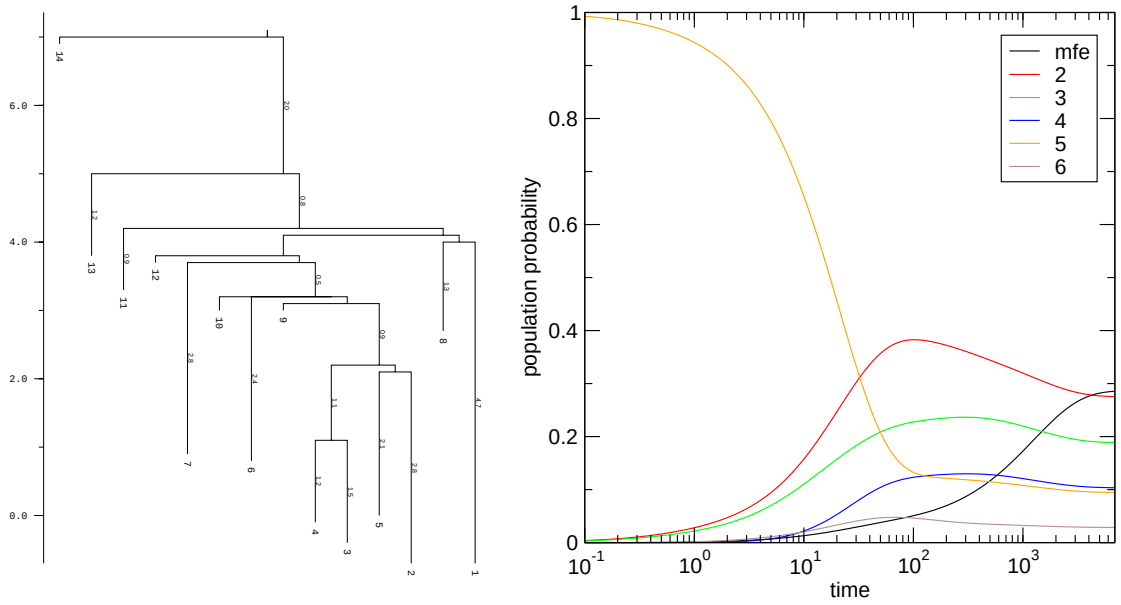

Figure 4. L.h.s.: Barrier tree of short artificial sequence lilly `UAUGCUGCGGCCUAGGC`. The leaves of the tree are the local minima of the energy landscape. R.h.s.: folding kinetics of lilly from the open structure. Population densities p_α of the basins of attraction of the local minima α is shown as a function of time.

Two situations can arise for any two saddle points s and s' with energies $E(s) < E(s')$. Either the basin of s is a "sub-basin" of $B(s')$ or the two basins are disjoint. This property arranges the local minima and the saddle points in a unique hierarchical structure which is conveniently represented as a tree, termed *barrier tree*.

An efficient flooding algorithm [7] starting from an energy sorted list of all low energy conformations [45] is used to identify local minima and saddle-points. In this way it is possible to construct the barrier tree (see Figure 4) of the low-energy portion of the folding landscape. The barrier tree depicts, in a compact form, the likelihood of a conversion between local minima. The leaves of the barrier tree correspond to local minima, while the internal nodes are the energetically highest (saddle) points on a path between any two local minima.

The process of kinetic folding can be modeled as *homogeneous Markov chain*. The probability $P(i,t)$ that a given RNA molecule will have the secondary structure i at time t is given by the master equation

$$\frac{\mathrm{d}P(i,t)}{\mathrm{d}t} = \sum_{j \neq i} [P(j,t)k_{ji} - P(i,t)k_{ij}] \tag{13}$$

where k_{ij} and k_{ji} are the rate constants for the transitions between the two secondary structures i and j in the deterministic description [9]. For short sequences or very restricted subsets of conformations equation (13) can be solved exactly or integrated numerically [38]. Solving the master equation for larger conformation spaces is out of the question. In such cases the dynamics can be obtained by simulating the Markov chain directly by a rejection-less Monte Carlo algorithm [6] and sampling a large number of trajectories.

Alternatively, the barrier tree can be used as a starting point for the definition of a coarse grained dynamics. In the simplest case transition rates between local minima can be modeled by their respective barrier heights in the tree. This approximation completely neglects entropic terms arising from possible multiple paths between the local minima.

8. Realistic Models for RNA Secondary Structure Prediction

All the quantities introduced above for the maximum circular matching problem can be computed similarly for the full energy model. The recursions do however get more complicated, and often require several auxiliary arrays. For illustration we show below the recursions for the minimum energy problem equivalent to Eq. 3.

Let F_{ij} be the optimal free energy on the sequence interval $[i, j]$, and let C_{ij} be the optimal free energy under the condition that i, j form a pair, F^M_{ij} holds the optimum given that $[i, j]$ lies within a multi loop and with at least two helices, while for F^{M1} only one helix is required.

$$\begin{aligned}
F_{ij} &= \min\left\{F_{i+1,j},\ \min_{i<k\le j} C_{ik} + F_{k+1,j}\right\} \\
C_{ij} &= \min\left\{\mathcal{H}(i,j),\ \min_{i<k<l<j} C_{kl} + \mathcal{I}(i,j;k,l),\ F^M_{ij}\right\} \\
F^M_{ij} &= \min\left\{F^M_{i+1,j} + c,\ \min_{i<k\le j} C_{ik} + b + F^{M1}_{k+1,j}\right\} \\
F^{M1}_{ij} &= \min\left\{F^{M1}_{i+1,j},\ \min_{i<k<j} C_{ik} + b + F^{M1}_{k+1,j},\ \min_{i<k\le j} C_{ik} + a + (j-k)c\right\}
\end{aligned} \tag{14}$$

Note that the corresponding recursions for the partition function can be obtained simply by replacing minimum operations with sums and additions with multiplication.

9. Concluding Remarks

RNA secondary structures are routinely used to display, organize, and interpret experimental findings. The dominant role of the secondary structure is also well documented in nature by the conservation of secondary structure elements in evolution [1]. *In vitro* selection experiments with RNA more often than not yield families of selected sequences that share distinctive secondary structure features. Furthermore, secondary structures are folding intermediates in the sense that all helices typically form before tertiary contacts complete the formation of the three-dimensional structure [40]. In addition, dynamical aspects of RNA secondary structure formation, including transitions at the level of RNA secondary structure, can play a crucial role for the understanding of the biological function of RNA [30].

Extensive computational studies of RNA evolution are feasible only because the RNA folding problem is solved efficiently by the relatively simple algorithms described in the previous sections. They have revealed the far-reaching consequences of the principles

of RNA structure formation for evolutionary phenomena, see [36] for a recent review. Four properties of the "genotype-phenotype map" relating RNA sequences and their minimum free energy structures have been predicted [37]:

(i) **More sequences than structures**. There are orders of magnitude more sequences than structures and hence, the map is many-to-one.
(ii) **Few common and many rare structures**. Relatively few common structures are opposed by a relatively large number of rare structures.
(iii) **Shape space covering**. The distribution of sequences that fold into the same structure is approximately random in sequence space. As a result it is possible to define a spherical ball, with a diameter d_{cov} being much smaller than the diameter of sequence space (n), which contains on the average for every common structure at least on sequence that folds into it.
(iv) **Existence and connectivity of neutral networks**. Neutral networks (the sets $G(S)$ of sequences that fold into the same structure S) of common structures are connected in most cases.

Experimental evidence for the existence and the properties of neutral networks in the RNA sequence-structure-function relationship is provided e.g. in [35, 13, 20].

Acknowledgments. This work is supported by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893, and the DFG Bioinformatics Initiative (P.F.S). Thanks to Michael Wolfinger for the lilly example.

References

[1] T. R. Cech. Conserved sequences and structures of group I introns: building an active site for RNA catalysis — a review. *Gene*, 73:259–271, 1988.
[2] J. Cupal, C. Flamm, A. Renner, and P. F. Stadler. Density of states, metastable states, and saddle points. Exploring the energy landscape of an RNA molecule. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 88–91, Menlo Park, CA, 1997. AAAI Press.
[3] J. Cupal, I. L. Hofacker, and P. F. Stadler. Dynamic programming algorithm for the density of states of RNA secondary structures. In R. Hofstädt, T. Lengauer, M. Löffler, and D. Schomburg, editors, *Computer Science and Biology 96 (Proceedings of the German Conference on Bioinformatics)*, pages 184–186, Leipzig (Germany), 1996. Univeristät Leipzig.
[4] Y. Ding and C. E. Lawrence. Statistical prediction of singletranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucl. Acids Res.*, 29:1034–1046, 2001.
[5] Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucl. Acids Res.*, 31:7180–7301, 2003.
[6] C. Flamm, W. Fontana, I. Hofacker, and P. Schuster. RNA folding kinetics at elementary step resolution. *RNA*, 6:325–338, 2000.
[7] C. Flamm, I. L. Hofacker, P. F. Stadler, and M. T. Wolfinger. Barrier trees of degenerate landscapes. *Z. Phys. Chem.*, 216:1–19, 2002.
[8] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for prediction of RNA duplex stability. *Proc. Natl. Acad. Sci. USA*, 83:9373–9377, 1986.
[9] D. T. Gillespie. A general method for nummerically sinulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[10] J. Gorodkin, L. J. Heyer, and G. D. Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucl. Acids Res.*, 25:3724–3732, 1997.
[11] C. Hansch, A. Leo, and D. Hoeckman. *Exploring QSAR*. American Chemical Society, Washington, 1995.
[12] D. Heidrich, W. Kliesch, and W. Quapp. *Properties of Chemically Interesting Potential Energy Surfaces*, volume 56 of *Lecture Notes in Chemistry*. Springer-Verlag, Berlin, 1991.
[13] D. M. Held, S. T. Greathouse, A. Agrawal, and D. H. Burke. Evolutionary landscapes for the acquisition of new ligand recognition by RNA aptamerts. *J. Mol. Evol.*, 57:299–308, 2003.
[14] P. Hobza and J. Šponer. Towards true DNA base-stacking energies: MP2, CCSD(T), and complete basis set calculations. *J. Amer. Chem. Soc.*, 124:11802–11808, 2002.
[15] I. Hofacker, P. Schuster, and P. Stadler. Combinatorics of secondary structures. *Discr. Appl. Math.*, 89:177–207, 1999.
[16] I. L. Hofacker. Vienna RNA secondary structure server. *Nucl. Acids Res.*, 31:3429–3431, 2003.
[17] I. L. Hofacker, S. Berhart, and P. F. Stadler. Alignment of rna base pairing probability matrices. *Bioinformatics*, 2003. submitted.
[18] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, 125(2):167–188, 1994.
[19] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Mh. Chem.*, 125:167–188, 1994.
[20] Z. Huang and J. W. Szostak. Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer. *RNA*, 9:1456–1463, 2003.
[21] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, 86:7706–7710, 1989.
[22] E. T. Kool. Preorganization of DNA: Design principles for improving nucleic acid recognition by synthetic oligonucleotides. *Chem. Rev.*, 97:1473–1487, 1997.
[23] J. Leydold and P. F. Stadler. Minimal cycle basis, outerplanar graphs. *Elec. J. Comb.*, 5:R16, 1998. See `http://www.combinatorics.org`.
[24] T. H. Lowry and K. S. Richardson. *Mechanism and Theory in Organic Chemistry*. Harper and Row, New York, 1976.
[25] D. Mathews, J. Sabina, M. Zuker, and H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.
[26] D. H. Mathews and D. H. Turner. Dynalign: An algorithm for finding secondary structures common to two rna sequences. *J. Mol. Biol.*, 317:191–203, 2002.
[27] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
[28] P. G. Mezey. *Potential Energy Hypersurfaces*. Elsevier, Amsterdam, 1987.
[29] U. Mückstein, I. L. Hofacker, and P. F. Stadler. Stochastic pairwise alignments. *Bioinformatics*, 2002.
[30] J. H. A. Nagel and C. W. A. Pleij. Self-induced structural switches in RNA. *Biochimie*, 84:913–923, 2002.
[31] R. Nussinov, G. Piecznik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.
[32] R. L. Ornstein, R. Rein, D. L. Breen, and R. D. MacElroy. An optimized potential function for the calculation of nucleic acid interaction energies. I. base stacking. *Biopolymers*, 17:2341–2360, 1978.
[33] C. M. Reidys and P. F. Stadler. Combinatorial landscapes. *SIAM Review*, 44:3–54, 2002.
[34] D. Sankoff. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.*, 45:810–825, 1985.
[35] E. A. Schultes and D. P. Bartel. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science*, 289:448–452, 2000.
[36] P. Schuster. Molecular insight into the evolution of phenotypes. In J. P. Crutchfield and P. Schuster, editors, *Evolutionary Dynamics – Exploring the Interplay of Accident, Selection, Neutrality, and Function*. Oxford University Press, New York, 2003.

[37] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. Lond. B*, 255:279–284, 1994.

[38] M. Tacker, W. Fontana, P. F. Stadler, and P. Schuster. Statistics of RNA melting kinetics. *Eur. Biophys. J.*, 23:29–38, 1994.

[39] M. Tacker, P. F. Stadler, E. G. Bornberg-Bauer, I. L. Hofacker, and P. Schuster. Algorithm independent properties of RNA structure prediction. *Eur. Biophy. J.*, 25:115–130, 1996.

[40] O. C. Uhlenbeck. A coat for all sequences. *Nature Struct. Biol.*, 5:174–176, 1998.

[41] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of rna folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.

[42] M. S. Waterman. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.*, 1:167 – 212, 1978.

[43] M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Math. Biosc.*, 42:257–266, 1978.

[44] E. Westhof and L. Jaeger. RNA pseudoknots. *Current Opinion Struct. Biol.*, 2:327–333, 1992.

[45] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.

[46] M. Zuker. The use of dynamic programming algorithms in RNA secondary structure prediction. In M. S. Waterman, editor, *Mathematical Methods for DNA Sequences*, pages 159–184. CRC Press, 1989.

[47] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.

[48] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.