Density of States, Metastable States, and Saddle Points Exploring the Energy Landscape of an RNA Molecule

Jan Cupal[a,*] and **Christoph Flamm**[a]
Alexander Renner[a] and **Peter F. Stadler**[a,b]
[a]Institut f. Theoretische Chemie, Univ. Wien
Währingerstr. 17, A-1090 Wien, Austria
Phone: **43 1 40480 661 Fax: **43 1 40480 660
E-Mail: `jan@tbi.univie.ac.at`
[b]The Santa Fe Institute, Santa Fe, New Mexico, U.S.A.

Abstract

Detailed knowledge of the energy landscape of a biopolymer molecule is a prerequisite for understanding its folding kinetics and its final spatial structure. In the case of RNA we consider the energy landscape defined on the set of all secondary structures that can be formed by a given sequence. We show that the exploration of this energy landscapes is computationally feasible. For this prurpose we present a recursive algorithm for computing the complete density of states and discuss its application to tRNA sequences. For shorter sequences a more detailed analysis of the energy surface is possible using a complete list of all secondary structures. In this case we identify metastable states and the saddle points that connect them.

The Density of States

The density of states (d.o.s.), i.e. the energy distribution of suboptimal secondary structures, is of utmost importance for an understanding of the structural versatility of RNA molecules (Higgs 1995).

For short RNA chains it is possible to exhaustively construct all possible secondary structures and to evaluate their energy. To this end we first generate the list of all possible base pairs (Watson-Crick and **GU**). This list serves as the basis set spanning the space of all structures. A complete list of secondary structures is thus obtained by recursively adding additional base pairs and checking if the resulting structures are valid at each step. The algorithm is structured in such a way that additional constraints, such as minimum stack lengths, can be included very easily.

Unfortunately, the number of secondary structures increases exponentially with the chain length (Hofacker, Schuster, & Stadler 1996), and the explicit construction of all secondary structures becomes unfeasible for chain length larger than some 40 nucleotides (see table 1). This limitation can be overcome by the recursive (dynamic programming) scheme outlined in the box on the following page. The key observation (Cupal, Hofacker, & Stadler 1996) is that the d.o.s. of a subsequence $[i, j]$ can be computed recursively from the d.o.s. of all shorter subsequences contained in $[i, j]$.

The algorithm closely resembles the computation of the partition function of an RNA molecule using McCaskill's method (McCaskill 1990). As an example we have computed the complete density of states for the phenylalanine tRNA from yeast, see figure 1. The algorithm is rather demanding both in terms of memory and CPU time: With a fixed energy resolution the computation of the d.o.s. of a RNA sequence of length n requires $\mathcal{O}(n^5)$ operations and memory of $\mathcal{O}(n^3)$.

The energy parameters are implemented with an accuracy of 0.01 kcal/mol. In most cases, however, an energy resolution of 0.1 kcal/mol is sufficient. On the other hand, a resolution coarser than thermal energy ($RT \approx 0.6$ kcal/mol) will hide the most interesting information. The CPU requirements for computing the complete density of states of a number of RNA sequences are compiled in table 1. The recursive scheme is faster than exhaustive enumeration only for chains longer than some 30 nucleotides. Restriction to a limited energy range above ground state leads to a significant reduction of the CPU requirements.

The overall shape of $N(F)$ is Gaussian (figure 1a).

Table 1: Performance Data for the Density of States. CPU times are measured on an `SGI Power Challange R8000` with 1GB memory. All times are in seconds.

n	Sequence	Num. o. Struct.	Dynamic. Prog. 0.01	0.1	Exaust.
8	$(ACGU)_2$	5	8	< 1	< 1
12	$(ACGU)_3$	35	139	2	< 1
16	$(ACGU)_4$	$2.7{\cdot}10^2$	1254	14	< 1
20	$(ACGU)_5$	$2.2{\cdot}10^3$	5049	51	2
24	$(ACGU)_6$	$2.0{\cdot}10^4$	16926	143	33
28	$(ACGU)_7$	$1.8{\cdot}10^5$	41089	329	345
32	$(ACGU)_8$	$1.7{\cdot}10^6$	*	691	1298
35	random	$2.0{\cdot}10^7$	*	804	10181
40	$(ACGU)_{10}$	$1.6{\cdot}10^8$	*	1791	38387
76	tRNA-phe	$1.5{\cdot}10^{16}$	*	28678	*

$$
\begin{aligned}
N^B_{ij}(\epsilon) &= \delta(\mathcal{H}(i,j),\epsilon) \;+ \sum_{i<j<k<l} N^B_{kl}(\epsilon - \mathcal{I}(i,j,k,l)) \;+ \\
&\quad + \sum_{i<k<j} \left[\sum_{\epsilon'} N^M_{i+1,k-1}(\epsilon')\, N^{M1}_{k,j-1}(\epsilon - \epsilon' - \mathcal{M}_C) \right] \\
N^{M1}_{ij}(\epsilon) &= \sum_{i<l\leq j} N^B_{il}(\epsilon - \mathcal{M}_B(j-l) - \mathcal{M}_I) \\
N^M_{ij}(\epsilon) &= \sum_{i<k\leq j} \left[\sum_{\epsilon'} N^M_{i,k-1}(\epsilon')\, N^{M1}_{kj}(\epsilon - \epsilon') \right] + \\
&\quad + \sum_{i\leq k\leq j} N^{M1}_{kj}(\epsilon - \mathcal{M}_B(k-i)) \\
N^A_{ij}(\epsilon) &= \sum_{i<l\leq j} N^B_{il}(\epsilon) \\
N_{ij}(\epsilon &= \delta(0,\epsilon) + N^A_{ij}(\epsilon) + \\
&\quad + \sum_{i\leq k<j} \left[\sum_{\epsilon'} N_{ik}(\epsilon') N^A_{k+1,j}(\epsilon - \epsilon') \right]
\end{aligned}
$$

Recursion for the calculation of the density of states: Calligraphic symbols denote energy parameters for different loop types: hairpin loops $\mathcal{H}(i,j)$, interior loops, bulges, and stacks $\mathcal{I}(i,j,k,l)$; the multi-loop energy is modeled by the linear ansatz $\mathcal{M} = \mathcal{M}_C + \mathcal{M}_I \cdot \text{degree} + \mathcal{M}_{\mathrm{M}} \cdot \text{unpaired}$, e.g. (Zuker & Sankoff 1984). The number $N^B_{ij}(\epsilon)$ of substructures on the substring $[i,j]$ with energy ϵ subject to the condition that i and j form a base pair is determined recursively from smaller fragments. The contributions depend on the type of the secondary structure element as a consequence of the energy model. The base pair (i,j) can be the closing pair of a hairpin, it may close an interior loop (or extend a stack), or it might close a multi-loop. The auxiliary variables N^M and N^{M1} are necessary for handling the multi-loops (McCaskill 1990), N^A helps reducing the CPU requirements. The unconstrained d.o.s. of the substring $[i,j]$ is stored in $N_{ij}(\epsilon)$. The first term accounts for the unpaired structure. The second term collects all structures that consist of a single component, possibly with an unpaired "tail" at the 3' end. The final term arises from the formal construction of multi-component structures from a 1-component part at the 3' side and an arbitrary structure at the 5' side.

This is not surprising since F is composed of a large number of additive contributions. The overwhelming majority of structures has positive energy, hence only a small subset of all possible structures is physically important. The ground state of all sequences that we have considered so far is unique both at a resolution of 0.1kcal/mol and 0.01kcal/mol. However, in general there is a substantial number of structures with a few RT above the ground state. It is also worth noting that there is a strong correlation between the size of the energy gap between the ground state and the first "excited state" and the fraction p_0 of ground state structure in thermodynamic

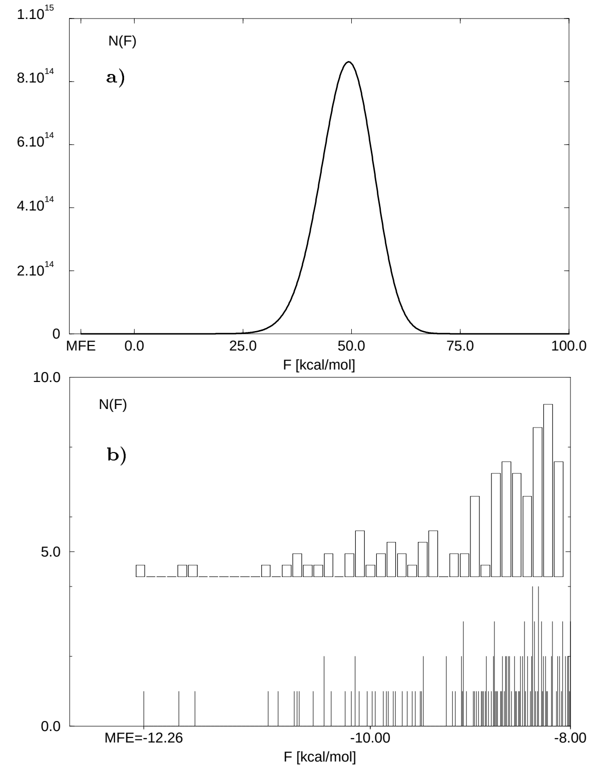

Figure 1: **a)** Full Density of States of tRNA$^{\mathrm{phe}}$ from yeast (n=76) computed with an energy resolution of 0.1 kcal/mol. The total number of structures, $14,995,224,405,213,184$ emphasizes the need for a recursive approach. Less than $1.77 \cdot 10^6$ structures have negative energy, the reference state being the open structure. The minimum energy structure is the familiar *clover leaf* with $E = -12.26$ kcal/mol. **b)** The lower figure shows the region above ground state in resolution of 0.1 and 0.01 kcal/mol. The ground state is unique. There is, however, a moderate number of suboptimal structures within 1 kcal/mol above the ground state.

equilibrium. The latter quantity can be obtained directly from the partition function (McCaskill 1990; Hofacker *et al.* 1994).

Higgs (Higgs 1995) found that the density of states of natural (evolved) sequences such as tRNAs differs significantly from random RNA sequences. His studies were based on a non-recursive algorithm using a drastically simplified energy model (Higgs 1993; 1995). Our own computations support his conclusions: The energy gap between the ground state and the first "excited" structure of E.coli tRNA(phe) is 1.45 kcal/mol, while mutants that have the same ground state structure (namely the familiar clover leaf) have an average energy gap of only 0.15 kcal/mol. A more detailed analysis of tRNA structures will be presented elsewhere.

Local Optima and Saddle Points

A major disadvantage of the recursive algorithm is the fact that we obtain no information about the individual

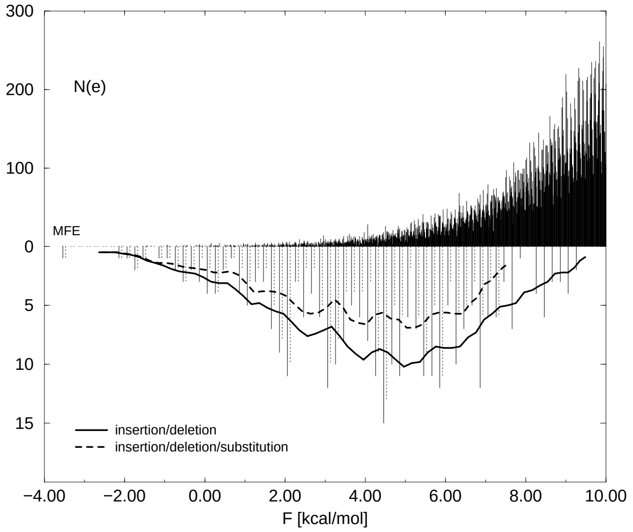

Figure 2: Density of states and density of local minima for sequence ACUAGUCGCGGGGAAUACCUUGGUUCCAAC. The ground state energy is $F = -3.54$ kcal/mol. There are a total of 671276 valid secondary structures. With insertions/deletions there are 320 metastable states, when substitutions are allowed this number reduces to 203.

structures that correspond to the low energy states. A detailed investigation of the energy landscape is therefore limited for the moment to sequences for which we can produce the complete list of all structures.

Table 2 lists the most stable structures of a typical RNA sequence.

A structure is a *local minimum* (or *metastable*) if its energy is lower than the energy of all *neighboring* structures. The ruggedness of the energy landscape thus is critically influenced by the *definition* of neighborhood: what are the elementary operations that interconvert secondary structures. In this contributions we consider two *move sets*:

(A) Opening or closing of a single base pair only.

(B) Opening or closing of a single base pair or subsitution of a base pair (i, j) by a base pair (i, k).

Since move set (A) is subset of move set (B) all local optima of move set (B) are also local optima under (A), but not vice versa. Note that the number local optima depends strongly on the choice of the move set (figure 2).

Discussion

It is interesting to compare the metastable states with respect to moveset (A) or (B) to the set of structures that are produced by suboptimal folding algorithms (Zuker 1989). We shall say that a secondary structure s is *Z-suboptimal* if there is no other secondary structure s' with lower energy containing all base pairs that are present in s. Obviously, the ground state is a local minimum with respect to any move set and it is also Z-suboptimal. It is surprising to see, however, that a substantial fraction of the low energy structures

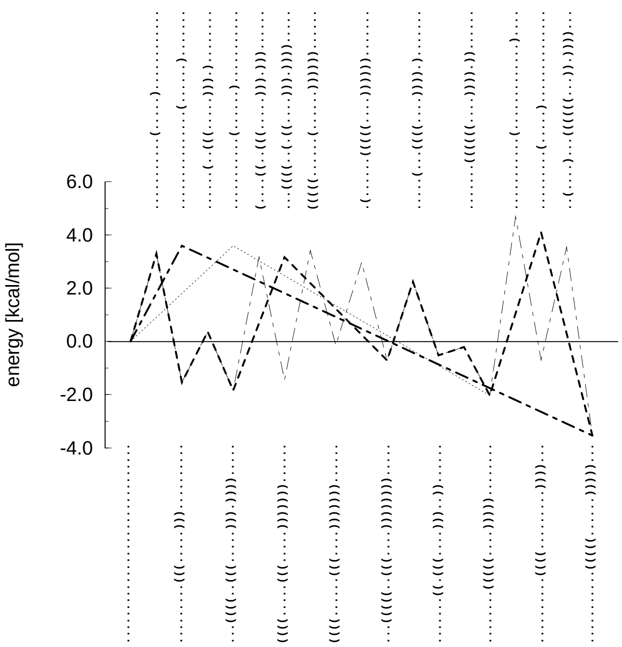

Figure 3: A variety of folding pathways starting with the open (denatured) structure lead to the ground state, among them a pathway with a single saddle point. Alternative foldings, however, have a lower energy barrier. Note that almost all saddle point contain an isolated base pair, i.e., they correspond to the nucleation of a novel stem.

are not Z-suboptimal. In fact, there are metastable states w.r.t. to the move sets (A) and (B) that are not Z-suboptimal, and conversely, some Z-suboptimal structures are not metastable, see table 2.

All configurations that are not local minima or maxima of the energy surface are sometimes called saddle points. For our purposes it is more convenient to use a more restrictive definition of a saddle point: A secondary structure s is *saddle point* if there are at least two local minima that can be reached by downhill walks starting at s. Of course the saddle point with lowest energy that separates the basins of two local minima s and s' is of particular importance.

The data compiled in table 2 can be used to extract folding pathways. In figure 3 we display the three most favorable pathways leading from the open (denatured) structure to the ground state. The first saddle point is determined by the nucleation of the first base pair. Adding base pairs to an established stack leads to lower energies. If the correct base pair is formed in the first step, the ground state is found without further obstacles. However, the energy barrier to the correct folding pathway is not the lowest in our example. Most saddle points encountered along the folding pathways contain an isolated base pair, i.e., they correspond to the nucleation of an novel stem. This is consistent with experimental findings on RNA folding: while the nucleation

Table 2: Energetically Favorable Structures of Sequence `ACUAGUCGCGGGGAAUACCUUGGUUCCAAC`, $n = 30$.

F	Structure*	AB	Z
-3.54	...........(((((......)))))...	●●	●
-2.14	.........(.(((((......)))))..)		●
-2.03	........(((((....)))))........	●●	●
-1.84	...((((..(((.....))).))))....	●●	●
-1.70	.((......))(((((......)))))...	●●	●
-1.54	.........(((.....))).........	●●	●
-1.45	...........((((........))))...		
-1.44	((((.....(((.....)))))))......	●●	●
-1.43	...(((((.(((.....)))))))).....		●
-1.35	....(((..(((.....))).)))......		
-1.11	......(.(((((....)))))).......		●
-1.03	.....((..(((.....))).)).......		
-0.95	...(((((.((......)).))))).....	●●	●
-0.94	.......(...(((((......)))))..)		●
-0.94	....((((.(((.....)))))))......		
-0.91	....((.....(((((......))))).))	●●	●
-0.71	...(((((..(((....)))))))).....	●●	
-0.70	..........((((.........))))...	●●	●
-0.86	.........((((....)))).........		
-0.62	.....(((.(((.....)))))).......		
-0.56	............((((......))))....		
-0.52	.......((.(((....)))..))......	●●	
-0.52	.......(((((.....)))..))......	●●	●
-0.47	(((((....((......)))))))......	●●	●
-0.46	....((((.((......)).))))......		

*For the "dot-bracket" notation see (Hogeweg & Hesper 1984; Hofacker *et al.* 1994)

of a helix is a slow, closing additional base pairs is fast cooperative process (Pörschke 1974).

We have shown that a computational exploration of the energy landscape is feasible for RNA molecules of moderate chain length. For short sequences, below some 40 nucleotides, it is possible to generate all structures, while we have to restrict ourselves to computing the density of states for larger molecules. Our approach goes beyond previous investigations in several respects: We use the best available energy model for secondary structures (Freier *et al.* 1986; Jaeger, Turner, & Zuker 1989), and we consider the complete ensemble of all valid secondary structures that can be formed by a given sequence. This is important when one is interested in thermodynamics; Only about two-thirds of the low energy states are Z-suboptimal, see table 2. Suboptiomal folding algorithm are not suitable therefore for computing reliable thermodynamic data.

Metastable states, and hence also saddle points, depend crucially on the choice of the move set. Once local minima and saddle points of an energy landscape are known it is straight forward to compute folding pathways and to simulate the dynamics of folding. Input from experimental work will be necessary to decide whether insertion/deletion/substitution is a realistic choice or if more sophisticated movesets are necessary to correctly describe the topology of the RNA energy surfaces. This is of practical importance for RNA structure prediction: kinetic folding algorithms are of course only as good as their intrinsic model of the energy landscape.

Acknowledgements

This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj. Nos. P10578-MAT and P11065-CHE. Useful discussion with Walter Fontana, Ivo Hofacker, and Peter Schuster are gratefully acknowledged.

References

Cupal, J.; Hofacker, I. L.; and Stadler, P. F. 1996. Dynamic programming algorithm for the density of states of RNA secondary structures. In Hofstädt, R.; Lengauer, T.; Löffler, M.; and Schomburg, D., eds., *Computer Science and Biology 96 (Proceedings of the German Conference on Bioinformatics)*, 184–186. Leipzig (Germany): Univeristät Leipzig.

Freier, S. M.; Kierzek, R.; Jaeger, J. A.; Sugimoto, N.; Caruthers, M. H.; Neilson, T.; and Turner, D. H. 1986. Improved free-energy parameters for prediction of RNA duplex stability. *Proc.Natl.Acad.Sci.USA* 83:9373–9377.

Higgs, P. G. 1993. RNA secondary structure: a comparison of real and random sequences. *J.Phys.I (France)* 3:43.

Higgs, P. G. 1995. Thermodynamic properties of transfer RNA: A computational study. *J.Chem.Soc.Faraday Trans.* 91(16):2531–2540.

Hofacker, I. L.; Schuster, P.; and Stadler, P. F. 1996. Combinatorics of RNA secondary structures. *Discr. Appl. Math.* submitted, SFI preprint 94-04-026.

Hofacker, I. L.; Fontana, W.; Stadler, P. F.; Bonhoeffer, S.; Tacker, M.; and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125(2):167–188.

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

Jaeger, J. A.; Turner, D. H.; and Zuker, M. 1989. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA* 86:7706–7710.

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

Pörschke, D. 1974. Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix. *Biopys. Chem.* 1:381–386.

Zuker, M., and Sankoff, D. 1984. RNA secondary structures and their prediction. *Bull.Math.Biol.* 46(4):591–621.

Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* 244:48–52.