R. Hofstaedt and T. Lengauer and M. Loeffler and D. Schomburg (eds.)
Computer Science and Biology 96, Univ. Leipzip 1996, pp. 184-186

DYNAMIC PROGRAMMING ALGORITHM FOR THE DENSITY OF STATES OF RNA SECONDARY STRUCTURES

Jan Cupal*[1], Ivo L. Hofacker+ and Peter F. Stadler*†

*Theoretische Biochemie @ Institut für Theoretische Chemie
Universität Wien, Währingerstraße 17, A-1090 Wien, Austria
+Beckman Institute, University of Illinois, Urbana, IL 61801, U.S.A.
†The Santa Fe Institute, 1399 Hyde Park Rd. Santa Fe, NM 87501, U.S.A.

A dynamic programming algorithm for the computation of the complete density of states of RNA secondary structures is presented. CPU and memory requirements scale as n^3m^2 and n^2m, respectively, where n is the chain length and m is the number of energy bins.

The `Vienna RNA Package` [4; 5] is an efficient implementation developed for the computation of the minimum free energy structure [8], sets of suboptimal structures [7], and the complete matrix of base pairing probabilities [6]. All these algorithms are based on a common dynamic programming scheme. The density of states, i.e., the energy distribution of suboptimal secondary structures, is of utmost importance for an understanding of the structural versatility of RNA molecules [3]. Higgs found that d.o.s. of evolved sequences such as tRNAs differ significantly from random RNA sequences. His studies were based, however, on a non-recursive algorithm using a drastically simplified energy model for the RNA secondary structures [2]. We show here that the same dynamic programming scheme that underlies all folding algorithms can be extended to a rigorous computation of the complete density of states. The key observation 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 $[i, j]$, see the box below for details. The algorithm is based on the standard energy model for RNA secondary structures, see e.g. [1]. All code is written in Ansi C.

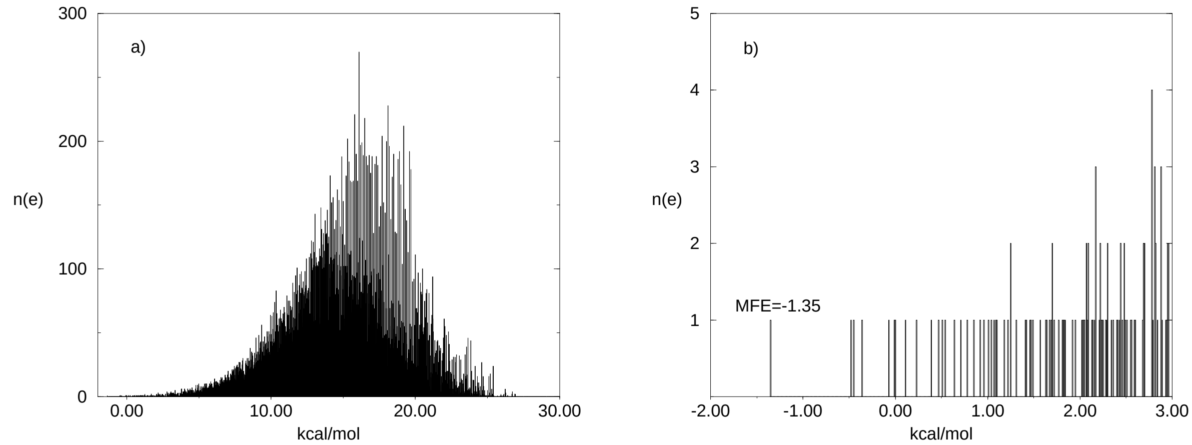

Figure 1: Density of States of the RNA sequence `GUCGUAGUCGAUGCUAGCUGAUCAC` with chain length $n = 25$ (a). The ground state structure `((((((((....))))..))))...` has energy $E = -1.35$kcal/mol. Total number of structures is 70727. The enlargement (b) of the most interesting energy range highlights the large energy gap of 0.87 kcal/mol between the ground state and the most stable suboptimal structure.

[1]Phone: **43 1 40480 661   Fax: **43 1 40480 660   E-Mail: jan@tbi.univie.ac.at

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

Recursion for the calculation of the density of states.

Calligraphic symbols denote energy paramters for different loop types: hairpin loops $\mathcal{H}(i,j)$, interior loops, bulges, and stacks $\mathcal{I}(i,j,k,l)$; the multiloop energy is modeled by the linear ansatz $\mathcal{M} = \mathcal{M}_C + \mathcal{M}_I \cdot \text{degree} + \mathcal{M}_M \cdot \text{unpaired}$, e.g. [8]. 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 multiloop. The auxiliary variables N^M and N^{M1} are necessary for handling the multi-loops [6], 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.

The algorithm is rather demanding both in terms of memory and CPU time: For each of the $\mathcal{O}(n^2)$ subsequences one needs to compute $\mathcal{O}(n)$ convolutions which in turn require $\mathcal{O}(m^2)$ operations, where n is the length of the RNA sequence and m is the number of bins used for storing the density of states for each substructure. Thus a total of $\mathcal{O}(n^3m^2)$ operations is required to compute the $\mathcal{O}(n^2m)$ entries that need to be stored throughout the calculation. With a fixed energy resolution the vector length m becomes proportional to the chain length n resulting in $\mathcal{O}(n^5)$ operations and a memory requirement of $\mathcal{O}(n^3)$.

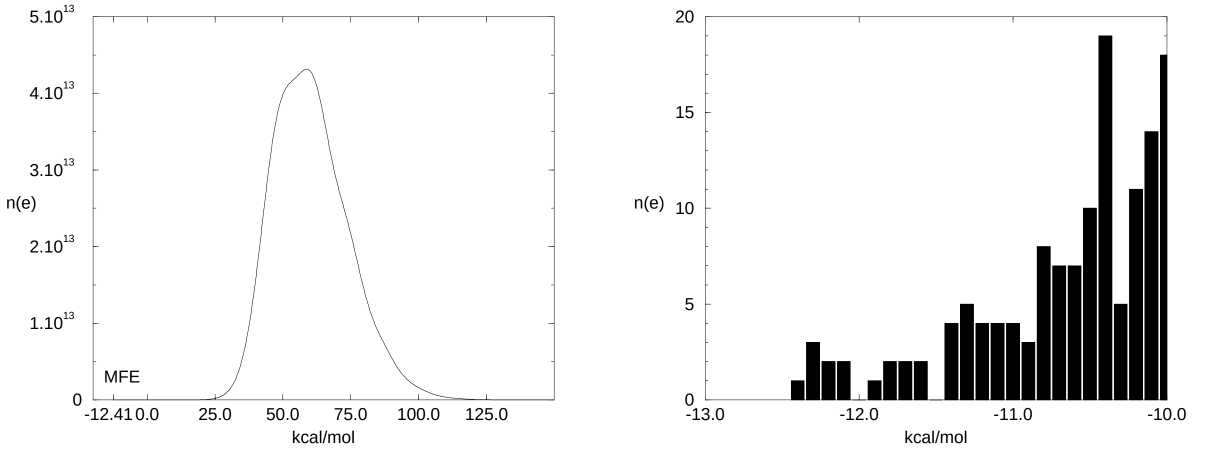

Figure 2: Density of States of $\mathrm{tRNA}^{\mathrm{phe}}$ from yeast ($n = 76$) computed with an energy resolution of 0.1 kcal/mol. The total number of structures, $14,995,224,405,213,184$ emphasizes the need for a recursive approach. The overwhelming majority of structures has very large (even positive) energies implying that they are probably not important biologically; in fact, 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 $E = -12.41$ kcal/mol. The enlargement at the l.h.s. shows that the ground state is again unique. There is, however, a moderate number (18) of suboptimal structures within 1 kcal/mol above the ground state.

The energy parameters are implemented with an accuracy of 0.01 kcal/mol thereby limiting the energy resolution that can be achieved. On the other hand, a resolution coarser than thermal energy ($RT \approx 0.6$ kcal/mol) will hide the most interesting information. Short sequences, $n \leq 35$, can be dealt with at full resolution on a modern workstation (fig.1). A few examples of tRNA sequences have been investigated at a resolution of 0.1 kcal/mol. These calculations took about 9 hours on a single processor of an SGI Power Challenge (fig.2).

Helpful discussions with Walter Fontana and Paul Stolorz are gratefully acknowledged.

References

[1] 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.

[2] P. G. Higgs. RNA secondary structure: a comparison of real and random sequences. *J.Phys.I (France)*, 91(16):2531–2540, 1995.

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

[4] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. `Vienna RNA Package`. `pub/RNA/ViennaRNA-1.03 @ ftp.itc.univie.ac.at`, 1994. (Public Domain Software).

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

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

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

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