BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–10

Energy landscape of k-point mutants of an RNA molecule

P. Clote [1,2]*, J. Waldispühl [1,3,4,†], B. Behzadi [3], J.-M. Steyaert [3],

[1] Department of Biology, Higgins 355, Boston College, Chestnut Hill, MA 02467, USA,
[2] Department of Computer Science (courtesy appt.), Boston College, Chestnut Hill, MA 02467, USA, [3] LIX, Ecole Polytechnique, 91128 Palaiseau Cedex, FRANCE, [4] LIAFA, Université Denis Diderot, 2 place Jussieu, 75251 Paris Cedex, FRANCE.

ABSTRACT

Motivation: A k-point mutant of a given RNA sequence $s = s_1, \ldots, s_n$ is an RNA sequence $s' = s'_1, \ldots, s'_n$ obtained by mutating exactly k-positions in s; i.e. Hamming distance between s and s' equals k. To understand the effect of pointwise mutation in RNA (Schuster et al., 1994; Clote et al., 2005b), we consider the distribution of energies of all secondary structures of k-point mutants of a given RNA sequence.

Results: Here we describe a novel algorithm to compute the mean and standard deviation of energies of all secondary structures of k-point mutants of a given RNA sequence. We then focus on the tail of the energy distribution, and compute, using the algorithm AMSAG (Waldispühl et al., 2002), the *k-superoptimal* structure; i.e. the secondary structure of a $\leq k$-point mutant having least free energy over all secondary structures of all k'-point mutants of a given RNA sequence, for $k' \leq k$. Evidence is presented that the k-superoptimal secondary structure is often closer, as measured by base pair distance, and two additional distance measures, to the secondary structure derived by comparative sequence analysis than is the Zuker (Zuker and Stiegler, 1981; Zuker, 2003) minimum free energy structure of the original (wild-type or unmutated) RNA.

Keywords: RNA, secondary structure, energy landscape, pointwise mutations, partition function.

Contacts: clote@bc.edu, Jerome.Waldispuhl@polytechnique.edu

Webserver: http://clavius.bc.edu/~clotelab/RNAmutants/

1 INTRODUCTION

A k-point mutant of a given RNA sequence $s = s_1, \ldots, s_n$ is an RNA sequence $s' = s'_1, \ldots, s'_n$ obtained by mutating exactly k-positions in s; i.e. Hamming distance between s and s' equals k, denoted by $d_H(s, s') = k$. To understand the effect of pointwise mutation on RNA secondary structure, we consider the distribution of energies of all secondary structures of k-point mutants of a given RNA sequence. In Section 1, we describe an algorithm to compute the mean and standard deviation of energies of all secondary structures of all k-point mutants of a given RNA sequence. Since there are exponentially many sequence/structure pairs, the energies cannot in general be enumerated, hence we introduce a novel algorithm to compute the partition function $Z_k(T)$ for all k-point mutants of a given RNA at temperature T in degrees Kelvin. By dynamic programming, our algorithm computes $Z_k(T)$ for fixed k in time $O(n^3)$, and additionally computes $Z_k(T)$ for all k simultaneously in time $O(n^4)$. From statistical mechanics, we see that the average free energy $\langle E_k \rangle$ of all secondary structures of all k-point mutants of a given RNA molecule is equal to RT^2 times the partial derivative of $\ln Z_k(T)$, and hence can be approximated for all k in time $O(n^4)$. Although the current algorithm uses the Nussinov-Jacobson energy model (Nussinov and Jacobson, 1980), our novel partition function and energy computation for k-point mutants is non-trivial and gives biologically interesting results. In future work, we will extend our algorithm to the Turner energy model (Xia et al., 1999). The current paper establishes proof of concept for a novel tool to explore aspects of RNA sequence and structure evolution.

In Section 2.2, we focus on the tail of the energy distribution of all secondary structures of k-point mutants of a given RNA sequence. By extending the S-attribute grammar formalism for the AMSAG software tool of (Waldispühl et al., 2002), we effectively compute the minimum free energy, with respect to (a technical restriction of) the Turner energy model (Matthews et al., 1999), over all secondary structures of all k'-point mutants of a given RNA sequence, where $k' \leq k$. We present evidence that the minimum free energy structure of certain k-point mutants is closer, as measured by base pair distance and two other metrics,[1] to the secondary structure derived by comparative sequence analysis than is the Zuker minimum free energy structure (Zuker and Stiegler, 1981; Zuker, 2003) of the original (unmutated or wild type) RNA.[2] One reason that the minimum free energy structure of a k-point mutant is at times closer to the structure derived by comparative sequence analysis could be that noncanonical base pairing and elements of tertiary structure are captured by the mutation of certain bases.

Before proceeding to the next section, we recall the formal definition of secondary structure.

*Corresponding author: clote@bc.edu

†Corresponding author: waldispu@lix.polytechnique.fr

[1] Analysis using base pair distance is presented in this paper. See web supplement for additional analysis using two alternative metrics (coarse tree edit distance and weighted coarse tree edit distance), both supported by `RNAdist` from the Vienna RNA Package.

[2] In (Ding et al., 2005) it is shown that often the structure derived by comparative sequence analysis is closer to the *centroid* of low-energy RNA structures obtained by *sampling* or *stochastic backtracking*. The approach in this paper is quite different from that of (Ding et al., 2005).

DEFINITION 1. *A secondary structure S on RNA sequence $s_1, \ldots, s_n$ is defined to be a set of ordered pairs corresponding to base pair positions, which satisfies the following requirements.*

1. Watson-Crick or GU wobble pairs: *If (i,j) belongs to S, then pair (s_i, s_j) must be one of the following canonical base pairs: (A,U), (U,A), (G,C), (C,G), (G,U), (U,G).*
2. Threshold requirement: *If (i,j) belongs to S, then $j-i > \theta$.*
3. Nonexistence of pseudoknots: *If (i,j) and (k,ℓ) belong to S, then it is not the case that $i < k < j < \ell$.*
4. No base triples: *If (i,j) and (i,k) belong to S, then $j = k$; if (i,j) and (k,j) belong to S, then $i = k$.*

Generally, the threshold θ, or minimum number of unpaired bases in a hairpin loop, is taken to be 3. For more background on dynamic programming, RNA secondary structure, energy and partition functions, see (Clote and Backofen, 2000). For reasons of space, certain technical details and additional data relevant to this paper can be found at the web supplement `http://clavius.bc.edu/~clotelab/RNAmutants/`.

2 METHODS

2.1 Partition functions of k-point mutants

An RNA sequence $s = s_1, \ldots, s_n$ is a word over the alphabet $\{A, C, G, U\}$; here n denotes the length of s. Throughout this section, $s = s_1, \ldots, s_n$ will denote a fixed RNA sequence, which, when clear from context, may be omitted as an explicit parameter in certain functions. For any $1 \leq i \leq j \leq n$, let $s(i,j)$ denote the subsequence of s from positions i to j; i.e. $s(i,j) = s_i, \ldots, s_j$. The Hamming ball $\mathcal{H}_k(s(i,j))$ of radius k is defined as the set of k-point mutants of $s(i,j)$. Note that the number of elements $|\mathcal{H}_k(s(i,j)|$ equals $\binom{j-i+1}{k} \cdot 3^k$, since there are $\binom{j-i+1}{k}$ many manners in which to choose a k-element subset of $\{i, \ldots, j\}$; for each fixed choice of k positions, a k-mutant is produced by replacing each of k chosen nucleotides of $s_i, \ldots, s_j$ by one of 3 different nucleotides. For any RNA sequence s, let $SS(s)$ denote the set of all secondary structures for s, and let $E(s,S)$ denote the energy of secondary structure S for sequence s. In Section 2.1, the energy function E is given for the Nussinov-Jacobson model (Nussinov and Jacobson, 1980; Clote and Backofen, 2000), defined below. With this notation, given RNA sequence $s = s_1, \ldots, s_n$, and any $0 \leq k \leq n$, we define $Z_k^T(i,j)$ to be the partition function at absolute temperature T of the collection of all secondary structures on all subsequences $s'(i,j) = s'_i, \ldots, s'_j$ where the Hamming distance between $s(i,j)$ and $s'(i,j)$ equals k; i.e.

$$Z_k^T(i,j) = \sum_{u \in \mathcal{H}_k(s(i,j))} \sum_{U \in \mathcal{SS}(u)} e^{-E(u,U)/RT}. \qquad (1)$$

Finally, define $Z_k(T) = Z_k^T(1,n)$. Letting $\mathcal{S}_k(s(i,j))$ denote the set of all sequence/structure pairs $S = (u,U)$, where the length of u is $j-i+1$, $d_H(s(i,j),u) = k$ and U is a valid secondary structure for u, we can equivalently write $Z_k^T(i,j) = \sum_{S \in \mathcal{S}_k(s(i,j))} e^{-E(S)/RT}$ in place of equation (1).

We now describe an efficient algorithm to compute the partition function $Z_k(T)$. Define $Z_0^T(i,j) = 1$ for all i,j such that $j - i \leq 3$; otherwise define $Z_0^T(i,j) = \sum_{S \in \mathcal{S}_0} e^{-E(S)/RT}$, where the sum is over all sequence/structure pairs for 0-mutants of $s_i, \ldots, s_j$. For $1 \leq k \leq n$, $1 \leq i \leq j \leq n$, if $j - i + 1 < k$, then define $Z_k^T(i,j) = 0$, since sequence length is too small to allow k distinct mutatations. For $1 \leq k \leq n$, $1 \leq i \leq j \leq n$, if $j - i \leq \theta$, then define $Z_k^T(i,j) = \binom{j-i+1}{k} \cdot 3^k$. In this case, since $i \leq j \leq i + \theta = i + 3$, the empty secondary structure is the only possible structure; when $j - i + 1 \geq k$, we must account for all possible k-point mutants, where each has only the empty secondary structure.

Assuming that $Z_k^T(i',j')$ has been defined for all values $0 \leq k \leq n$ and all $1 \leq i' < j' \leq n$ such that $j' - i' < j + 1 - i$, we define $Z_k^T(i, j+1)$ for $1 \leq k \leq n$ by the recursive definition given in Figure 1. Here the energy contribution $a_{x,y}$ due to base pairing nucleotides x, y is given by[3]

$$a_{x,y} = \begin{cases} -3 & \text{if } \{x,y\} = \{G,C\} \\ -2 & \text{if } \{x,y\} = \{A,U\} \\ -1 & \text{if } \{x,y\} = \{G,U\} \\ +\infty & \text{otherwise} \end{cases} \qquad (7)$$

Equation (2) is the contribution to the partition function when the nucleotide at position $j+1$ does not base pair. In particular, $Z_k^T(i,j)$ is the contribution made by subsequence $s_i, \ldots, s_{j+1}$, where there are k pointwise mutations in the region $s_i, \ldots, s_j$. By mutating s_{j+1} to one of three other nucleotides and considering only $k-1$ pointwise mutations in the region $s_i, \ldots, s_j$, we have the term $3Z_{k-1}^T(i,j)$. Equation (3) is the contribution to the partition function made by base pairing s_{j+1} to some intermediate nucleotide s_r, where neither has been mutated, and recursively considering the contribution due to c mutations in the region $s_i, \ldots, s_{r-1}$ and $k-c$ mutations in the region $s_{r+1}, \ldots, s_j$. Equation (4) is the contribution to the partition function made by mutating the nucleotide s_{j+1} and enforcing a base pairing with an intermediate unmutated nucleotide s_r, where c mutations occur in the region $s_i, \ldots, s_{r-1}$ and $k-1-c$ mutations occur in $s_{r+1}, \ldots, s_j$. Similarly, equation (5) is the contribution to the partition function where the intermediate position s_r has been mutated, and which base pairs with the unmutated nucleotide s_{j+1}. Finally equation (6) is the contribution when there is a base pairing between positions $j+1$ and intermediate position r, yet both nucleotides s_{j+1} and s_r at these positions have been mutated.

It is well-known that rates of mutation differ between *transitions* (purines to purines, or pyrimidines to pyrimidines) and *transversion* (purines to pyrimidines, or pyrimidines to purines). By adding appropriate weights in the summation over $x \in \{A,C,G,U\} - \{s_r\}$, resp. $x \in \{A,C,G,U\} - \{s_{j+1}\}$ in lines (4), (5), (6) of Figure 1, it is possible to incorporate such mutation rates into the partition function calculation.

[3] The values of $a_{x,y}$ could all equal -1, or in fact arbitrary negative values, if x, y can base pair. If x, y cannot base pair, then $a_{x,y} = \infty$, which prevents any contribution for this pair to the partition function in Figure 1.

$$Z_k^T(i,j+1) = Z_k^T(i,j) + 3 \cdot Z_{k-1}^T(i,j) \tag{2}$$

$$+ \sum_{r=i}^{j-\theta} e^{-a_{s_r,s_{j+1}}/RT} \cdot \left(\sum_{c=0}^{k} Z_c^T(i,r-1) \cdot Z_{k-c}^T(r+1,j) \right) \tag{3}$$

$$+ \sum_{x \in \{A,C,G,U\}-\{s_{j+1}\}} \sum_{r=i}^{j-\theta} e^{-a_{s_r,x}/RT} \cdot \left(\sum_{c=0}^{k-1} Z_c^T(i,r-1) \cdot Z_{k-1-c}^T(r+1,j) \right) \tag{4}$$

$$+ \sum_{r=i}^{j-\theta} \sum_{x \in \{A,C,G,U\}-\{s_r\}} e^{-a_{x,s_{j+1}}/RT} \cdot \left(\sum_{c=0}^{k-1} Z_c^T(i,r-1) \cdot Z_{k-1-c}^T(r+1,j) \right) \tag{5}$$

$$+ \sum_{r=i}^{j-\theta} \sum_{x \in \{A,C,G,U\}-\{s_r\}} \sum_{y \in \{A,C,G,U\}-\{s_{j+1}\}} e^{-a_{x,y}/RT} \cdot \left(\sum_{c=0}^{k-2} Z_c^T(i,r-1) \cdot Z_{k-1-c}^T(r+1,j) \right) \tag{6}$$

Fig. 1. Recursive computation of the Boltzmann partition function of k-point mutants, according to the Nussinov-Jacobson model.

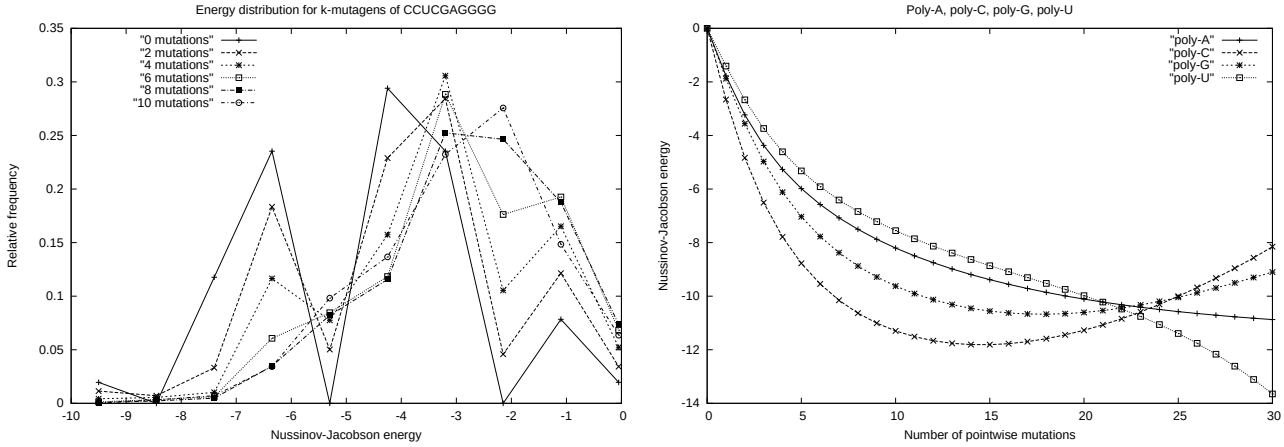

Fig. 2. *(i)* Left panel: Nussinov-Jacobson energy distribution for all k-point mutants, for even $0 \leq k \leq 10$, of the 10-mer prefix `CCUCGAGGGG` from SECIS element `fruA`. Energies range from -9 to 0. Distributions obtained by creating histograms of energies produced by brute force enumeration. *(ii)* Right panel: This figure depicts the relation between number k of mutations (x-axis) and average energy of all k-point mutants of toy 30 nt. sequences consisting only of a single type of nucleotide (i.e. poly-A, poly-C, poly-G and poly-U sequences of length 30). Order of the curves, from top to bottom: poly-U, poly-A, poly-G and poly-C. Corresponding graph for $\leq k$-mutants of the same sequences is similar (not shown, but see web supplement for data) though all four curves converge together as the number of mutations approaches the maximum of 30. (Such convergence for $\leq k$-mutants is of course necessarily always the case.) Average energy values obtained at $T = 310$ Kelvin using method of Figure 1.

Ostensibly it is the case that $Z_k^T(1,n)$ can be computed simultaneously for all values of $k \in \{0,\ldots,n\}$ in $O(n^4)$ time and $O(n^3)$ space by using dynamic programming. Indeed, for each value of k, we fill in a table for all $1 \leq i < j \leq n$ by increasing order of $|j-i|$. For small values of $k \leq k_0$, the algorithm runs in $O(n^3)$ time and $O(n^2)$ space.

Let $Z_k(T)$ denote the partition function at absolute temperature T for the free energy of all secondary structures of all k-point mutants of a given RNA sequence s, where R denotes the universal gas constant 8.3146 $JK^{-1}\text{mol}^{-1}$; i.e. $Z_k(T) = \sum_{S \in \mathcal{S}_k} e^{-E(S)/RT}$, where $\mathcal{S}_k$ denotes the set of all sequence/structure pairs $S = (u,U)$, where u is a k-point mutants of RNA sequence s, and U is a secondary structure for u. Assume that energy $E(S)$ of any given sequence/structure pair S is independent of temperature.[4] Let $\langle E_k \rangle$ denote the expected energy of all

[4] Energy $E(S)$ of a structure is independent of temperature under the Nussinov-Jacobson energy model (Nussinov and Jacobson, 1980; Clote and Backofen, 2000). In the Turner energy model (Matthews et al., 1999; Xia et al., 1999), free energy depends on enthalpy and entropy, where especially the entropic term is temperature-dependent. In our future extension to the Turner model, we distinguish between *table temperature* and *formal temperature*, where table temperature designates the temperature for which base stacking free energies and loop energies are retrieved by

sequence/structure pairs $S \in \mathcal{S}_k$, and let $Pr_{k,T}[S]$ denote the Boltzmann probability of the sequence/structure pair S.

THEOREM 2. *Given any RNA sequence $s = s_1, \ldots, s_n$, for each $0 \leq k \leq n$,*

$$\langle E_k \rangle = RT^2 \cdot \frac{\partial}{\partial T} \ln Z_k(T). \tag{8}$$

PROOF.

$$\begin{aligned} \langle E_k \rangle &= \sum_{S \in \mathcal{S}_k} Pr_{k,T}[S] \cdot E(S) \\ &= \sum_{S \in \mathcal{S}_k} \frac{e^{-E(S)/RT}}{Z_k(T)} \cdot E(S) \end{aligned}$$

Now

$$\begin{aligned} \frac{\partial}{\partial T} \ln Z_k(T) &= \frac{\partial}{\partial T} \ln \sum_{S \in \mathcal{S}_k} e^{-E(S) \cdot T^{-1}/R} \\ &= \frac{1}{Z_k(T)} \cdot \sum_{S \in \mathcal{S}_k} \frac{E(S)}{RT^2} \cdot e^{-E(S)/RT} \\ &= \frac{1}{RT^2} \cdot \frac{\sum_{S \in \mathcal{S}_k} E(S) \cdot e^{-E(S)/RT}}{Z_k(T)} \\ &= \frac{1}{RT^2} \cdot \sum_{S \in \mathcal{S}_k} Pr_{k,T}[S] \cdot E(S) = \frac{\langle E_k \rangle}{RT^2} \end{aligned}$$

The proof of the following theorem is identical to that of Theorem 2.

THEOREM 3. *Let $U_k(T)$ denote the partition function at absolute temperature T for the square of the free energy of all k-point mutants of RNA sequence s; i.e. $U_k(T) = \sum_{S \in \mathcal{S}_k} e^{-E^2(S)/RT}$, where $\mathcal{S}_k$ denotes the set of sequence/structure pairs (s', S'), where S' ranges over all secondary structures of s', where s' ranges over all k-point mutants of RNA sequence s. Then given any RNA sequence $s = s_1, \ldots, s_n$, for each $0 \leq k \leq n$,*

$$\langle E_k^2 \rangle = RT^2 \cdot \frac{\partial}{\partial T} \ln U_k(T). \tag{9}$$

Note that by this method, we could *in principle* compute all higher order moments of the energy function; our current experiments suggest that numerical instability due to floating point precision renders the computation of $\langle E_k^2 \rangle$ infeasible except for small toy examples.

Given RNA sequence $s = s_1, \ldots, s_n$, our previous discussion entails that we can compute in time $O(n^4)$ for all k both $Z_k(T)$ and $U_k(T)$. Let $\sigma(E_k)$ denote the standard deviation of the distribution of energies $E(S)$ of all minimum free structures S of k-point mutants. By Theorems 2 and 3 we have

$$\begin{aligned} \sigma(E_k) &= \sqrt{\langle E_k^2 \rangle - (\langle E_k \rangle)^2} \\ &= \sqrt{RT^2 (U_k(T) - (Z_k(T))^2} \end{aligned}$$

THEOREM 4. *Given an RNA sequence $s = s_1, \ldots, s_n$, there is an $O(n^4)$ algorithm to approximately compute for all $0 \leq k \leq n$ simultaneously both the mean $\langle E_k \rangle$ and the standard deviation $\sigma(E_k)$ of the energy of all minimum free energy secondary structures of k-point mutants of s at fixed temperature T.*

PROOF. Using the definition of the derivative, we can approximate $\frac{\partial}{\partial T} \ln Z_k(T)$ by

$$\frac{\ln Z_k(T + \Delta T) - \ln Z_k(T)}{\Delta T}$$

for small ΔT, e.g. $\Delta T = 0.0001$. The quality of our approximation depends of course on the function $Z_k(T)$, for which we have no analytical, closed formula. At the current time, we have no knowledge of the *modulus of continuity* of $Z_k(T)$, and hence no current possibility of algorithmically returning an approximation of $\frac{\partial}{\partial T} \ln Z_k(T)$ which is provably within error bound ϵ. By algorithmically sampling values $\ln Z_k(T)$ within a δ-neighborhood of a fixed temperature T, one could provide some sense of the accuracy in the approximate computation of $\langle E_k \rangle$.[5] ■

As pointed out by one of the referees, our work is related to certain investigations in physics concerning the average free energy of disordered systems and spin glasses, where one computes *quenched average* free energy and the *annealed* free energy; cf. (Guerra and Toninelli, 2002) on spin glass thermodynamics, (Orlandini et al., 1999) on biopolymers and (Deutsch and Paladin, 1989) on products of random matrices. Recalling that ensemble free energy $G = -RT \ln Z$, where R is the universal gas constant and T temperature in degrees Kelvin, the quenched average free energy is defined in statistical mechanics by $\langle -RT \ln Z_s \rangle = -RT \langle \ln Z_s \rangle = -RT \sum_s Pr[s] \cdot \ln Z_s$, while the annealed free energy is defined by $-RT \ln \langle Z_s \rangle = -RT \ln \left(\sum_s Pr[s] \cdot Z_s \right)$, where $Pr[s]$ is the probability of state s and the summation is taken over all states s in the microcanonical ensemble. In the context of this paper, a state corresponds to a k-point mutant of a given RNA sequence, while the quenched average corresponds to the average ensemble free energy over all k-point mutants, and annealed energy to the ensemble free energy with respect to the average partition function. Since the log of a sum is unequal to the sum of the logs, the quenched average free energy is in general distinct from the annealed

table look-up from values determined in (Xia et al., 1999), while formal temperature designates the term T appearing in RT and in $Z_k(T)$.

[5] Alternatively, one can compute a least squares quadratic or small-degree polynomial fit for $\ln Z_k(T)$, and subsequently compute the derivative of the fitting polynomial. Current version of our software does not yet do this, but instead computes the previously mentioned finite difference approximation.

k	$\langle E \rangle$ by Boltzmann	$\langle E \rangle$ by enum	Z_k^T	NumSeqSecStr	Num Mutants
0	-4.372569	-4.372549	51.00	51	0
1	-4.042466	-4.042907	1142.06	1142	30
2	-3.724917	-3.724723	11839.53	11839	405
3	-3.431001	-3.428592	75170.10	75167	3240
4	-3.162192	-3.162606	324907.36	324895	17010
5	-2.934739	-2.934487	1002434.38	1002399	61236
6	-2.753071	-2.753521	2243685.30	2243611	153090
7	-2.631960	-2.630359	3610431.22	3610317	262440
8	-2.575835	-2.575795	4012270.30	4012146	295245
9	-2.600944	-2.601228	2792022.35	2791935	196830
10	-2.719101	-2.719846	928128.36	928098	59049

Table 1. Computation of expected energy for RNA sequence `CCUCGAGGGG`, a 10-mer prefix of SECIS element `fruA`. The value k indicates the exact number of pointwise mutations. Expected energy $\langle E \rangle$ is computed by the method of Theorem 4 using Figure 1 (*Boltzmann*) as well as by brute force enumeration of all possible k-point mutants and secondary structures (*enum*). Z_k^T denotes the value of the partition function for `CCUCGAGGGG`, where absolute temperature $T = 10^4$. In computing expected energy by the method of Theorem 2, $T = 10^4$ and $\Delta T = 0.001$. (Small deviations of $\langle E \rangle$ occur when temperature is in the range from 100 to 10^6, and ΔT is in the range from 0.00001 to 0.1.) *NumSeqSecStr* designates the number of sequence/structure pairs actually enumerated in the brute force computation of expected energy. Note that this number is very close to the Boltzmann partition function value. Indeed, as explained below, the Boltzmann probability $Pr_{k,T}[S]$ at absolute temperature $T = 10^4$ is very close to the (uniform) probability $1/Z_k(T)$. Finally, *NumSecStr* designates the the number $\binom{10}{k} \cdot 3^k$ of k-point mutants of `CCUCGAGGGG`. In computing expected energy by brute force enumeration, we computed the energy contribution due to each k-point mutation for each of the 65 possible secondary structures of length 10. At absolute temperature $T = 10^4$, the value $Pr_{k,T}[S] = e^{-E[S)/RT}/Z_k(T) \approx 1$; indeed, $exp(-100/(R * 10000)) = 0.99879801928033662$, where we note that $|E(S)|leq3 \cdot 3 = 9$ for any mutant of `CCUCGAGGGG`. Thus $Pr_{k,T}$ approximately equals the uniform probability distribution; i.e. $Pr_{k,T}[S] \approx 1/Z_k(T)$. There is a gradual discrepancy between $\langle E \rangle$ by Boltzmann and $\langle E \rangle$ by enumeration, for increasingly large values of k. Approximations remain good, even for long RNA sequences, provide values of k are modest.

free energy, except in *self-averaging systems*. In computing energy from the partition function, following Theorem 4, we take $RT^2 \cdot \frac{\partial}{\partial T} Z_k(T)$ instead of $-RT \ln Z_k(T)$; hence from the physics viewpoint of quenched average versus annealed average, we seem to be computing a value analogous to the annealed average energy of all k-point mutants.

2.2 Superoptimal RNA Structures

In Section 2.1, we considered the mean and standard deviation of the energy distribution of k-point mutants using the Nussinov-Jacobson energy model. The technical novelty of our approach relies on an efficient computation of the partition function $Z_k(T)$ for k-point mutants of a given RNA sequence at a given absolute temperature. By defining $Z^T_{\leq k}(i,j) = \sum_{k'=0}^{k} Z^T_{k'}(i,j)$, the same approach allows a computation of the expected energy of $\leq k$-point mutants, i.e. the average energy of all k'-point mutants, for $k' \leq k$ (see web supplement for sample graphs).

Furthermore, the minimum energy with respect to the Nussinov-Jacobson energy model over all k-point mutants of a given RNA sequence s of length n can be computed in time $O(n^4)$ and space $O(n^3)$ by replacing each occurrence in Figure 1 of a sum (resp. product resp. $e^{-a_{s_x,s_y}/RT}$) by a minimization (resp. sum resp. a_{s_i,s_j}). By backtracking, one could output for each k that secondary structure of a k-point mutant, which has least energy over all secondary structures of all k-point mutants. For economy of space, we suppress formal details (see web supplement).

This discussion leads naturally to the novel concept of *superoptimal* secondary structure, defined as follows. Given RNA sequence s and integer k, a k-superoptimal structure for s is that secondary structure for a $\leq k$-point mutant of s having *least* energy over all secondary structures for all k'-point mutants, for $k' \leq k$. Since the k-superoptimal structure is a secondary structure for a sequence s' which differs from s in at most k positions, we add a post-processing step where mutated bases in s' are replaced by the original bases of s and unauthorized base pairs are removed (e.g. if a base pair (i,j) occurs where mutated nucleotides $s'_i = G$, $s'_j = C$, and if $s_i = C = s_j$, then base pair (i,j) is removed. We define the resulting secondary structure after this post-processing step to be the *k-superoptimal structure*.

In future work, we intend to develop the approach just described to compute k-superoptimal structures with respect to the Turner energy model, for the most recent energy parameters (Xia et al., 1999), including dangles; however, in the present paper, we apply the general software tool, AMSAG, of (Waldispühl et al., 2002) to compute k-superoptimal secondary structures in time $\mathcal{O}(k^2n^3)$ and space $\mathcal{O}(kn^2)$, where n is the length of an input RNA sequence.

The AMSAG framework, pioneered by (Lefebvre, 1995), computes the optimal parse tree for a given sequence and S-attribute grammar. For an application of S-attribute grammars to predict transmembrane protein supersecondary structure, see (Waldispühl and Steyaert, 2005). Here, an S-attribute grammar (Waldispühl et al., 2002) is an extension of a context free grammar, where terminal symbols are assigned an *attribute* $a \in \mathcal{A}$ (attributes are possibly numerical), and for each production or grammar rule $A \rightarrow \alpha_1 \cdots \alpha_r$, an associated rule $f : \mathcal{A}^r \rightarrow \mathcal{A}$ is given which is used to

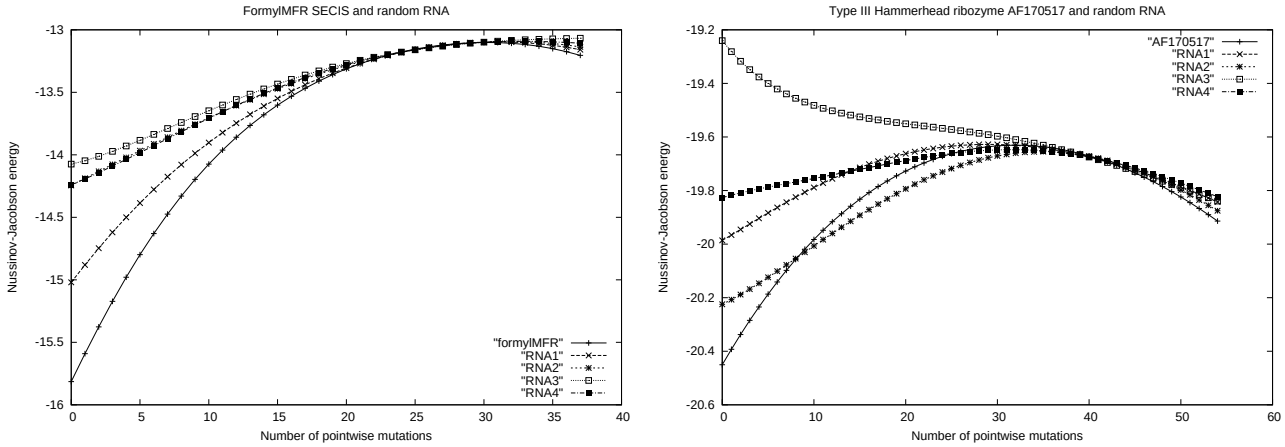

Fig. 3. This figure depicts the relation between number k of mutations (x-axis) and average energy of all k-point mutants for a selenocysteine insertion (SECIS) element and for a hammerhead ribozyme. Average energy values obtained at $T = 310$ Kelvin using our new method to compute the partition function. *(i)* The left panel considers k-point mutants of the 37 nt. SECIS element `formylMFR` with nucleotide sequence `AUGUUGGAGGGGAACCCUGUAAGGGACCCUCCAACAU` four random RNAs of the same dinucleotide frequency, as produced by the implementation in (Clote et al., 2005a) of the algorithm of (Altschul and Erikson, 1985). *(ii)* The right panel considers k-point mutants of the type III hammerhead ribozyme with Rfam accession number AF170517 from Rfam (Griffiths-Jones et al., 2003) and four random RNAs of the same dinucleotide frequency, as produced by the implementation in (Clote et al., 2005a) of the algorithm of (Altschul and Erikson, 1985). Note in both cases that the curve for real RNA generally lies below that of random RNA, and that the average energy of the 0-point mutant (wild-type) is less than that of 1-point mutants, which is less than that of 2-point mutants, etc., suggesting that wild-type sequences have undergone selection against pointwise mutants. Curves for $\leq k$-mutants (where the y-axis displays the expected energy of all k'-point mutants for $k' \leq k$) are similar, except for their convergence toward a single curve when k approaches sequence length. See web supplement for $\leq k$-mutants and additional data.

compute the attribute of A from the attributes of $\alpha_1 \cdots \alpha_r$. (See Definition 2 of (Waldispühl et al., 2002) for a formal definition.) In this manner, the root attribute can be computed inductively from the leaf attributes in a parse tree, and hence a numerical value (the root attribute) can be associated with any valid parse tree. In addition to showing that AMSAG efficiently computes the optimal S-attribute parse tree for any given input sequence and S-attribute grammar, (Waldispühl et al., 2002) give an application of AMSAG for RNA secondary structure prediction using energy minimization. While the energy model for RNA in (Waldispühl et al., 2002) is a refinement of the Nussinov-Jacobson energy model (energies for base pairs, rather than *stacked* base pairs, affine costs given to loops), for this paper, we have developed an S-attribute grammar for a technical restriction of the Turner energy model (Matthews et al., 1999) – specifically, the energy rules are those of `mfold 3.0` (Zuker and Stiegler, 1981; Zuker, 2003), whose values are taken from (Matthews et al., 1999), including tetraloop bonus. However, AMSAG, in contrast to `mfold 3.0`, does not include energies for dangles; moreover, for computational efficiency, AMSAG currently considers pointwise mutations at every nucleotide position, but only retains those mutations which form base pairs. In particular, the algorithm does not retain mutated positions in tetraloops which improve the tetraloop bonus (such positions are in the loop region, hence do not base pair).

Although this paper focuses on Hamming distance between RNA sequences, it should be noted that AMSAG can handle more general sequence distance measures, such as *edit distance*, which then allow mutant sequences to be obtained by insertion and deletion of nucleotides. As well, though tables and figures in this paper concern base pair distance between RNA secondary structures, we have performed additional analysis using two forms of *tree edit distance* between secondary structures (see web supplement for data).

comparative seq. anal. mfe structure 12-superoptimal

Fig. 4. *E. coli* 5S rRNA with accession number V00336 from the *Comparative RNA Web Site* (i.e. Gutell's database) (Gutell et al., 2002). From left to right: *(i)* secondary structure obtained by comparative sequence analysis (Gutell et al., 2002), *(ii)* 0-superoptimal structure from AMSAG, i.e. minimum free energy (mfe) structure using energies from Zuker's `mfold 3.0` without dangles (see text), *(iii)* 12-superoptimal structure from AMSAG.

3 DISCUSSION

3.1 Expected energy using the partition function

As shown in Table 1, the method of Theorem 4 efficiently computes expected energy values $\langle E_k \rangle$ for all k-point

mutants, which are quite close to those computed by brute force enumeration. Indeed, following a suggestion of one of the referees, we have computed Z_k^T for the special case where base pairing energy $a_{x,y}$ from equation (7) equals 0 if nucleotides x, y can base pair. In this case, partition function values Z_k^T are *exactly* equal to the values of NumSeqSecStr (data not shown). While our new method requires $O(n^3)$ time to determine $\langle E_k \rangle$ for all $k \leq k_0$, brute force enumeration requires exponential time $O(\sum_{k=1}^{k_0} \binom{n}{k} 3^k \cdot 1.8^n)$. (The binomial coefficient and factor 3^k correspond to the number of k-point mutants, while the exhaustive determination of the free energy of every secondary structure requires approximately 1.8^n steps.) Table 1 shows that for moderate to large absolute temperature T, $Pr_{k,T}[S] \approx 1/Z_k(T)$, where $Z_k(T) \approx |\mathcal{S}_k|$, so for large temperature, the Boltzmann weighted expected energy is very close to the arithmetic average energy.

The left panel of Figure 2 depicts superposed histograms of Nussinov-Jacobson energies of all secondary structures of k-point mutants, for even $0 \leq k \leq 10$, given the 10-mer initial portion `CCUCGAGGGG` of the SECIS element `fruA`. The right panel graphs expected energy, using our Boltzmann approach, of all k-point mutants of poly-A, poly-C, poly-G and poly-U sequences of length 30, where $0 \leq k \leq 30$. Note the order of the curves, where the poly-C curve is on the bottom and poly-U curve on the top. In data on the web supplement, we have computed the same curves, except that for a given value of k on the x-axis, the value on the y-axis is the expected energy for all $\leq k$-point mutants. The general form of the curves and their order remains the same (data not shown). Figure 3 displays curves of expected energy as a function of number k of mutations, for the selenocysteine insertion sequence (SECIS) element `formylMFR` as well as for type III hammerhead ribozyme with Rfam accession number AF170517 from Rfam (Griffiths-Jones et al., 2003). The random sequences were produced using the first author's implementation in (Clote et al., 2005a) of the algorithm of (Altschul and Erikson, 1985) (see (Clote et al., 2005a) for a variety of new results on how structural RNA differs from random RNA). Note the general tendency for random RNA to have higher average energy for its k-mutants, perhaps a general feature for which RNA sequence is under selective pressure (Clote et al., 2005a).

3.2 Experiments using AMSAG

Here, we present results from computer experiments using AMSAG with RNA sequences from the *RNaseP* database (Brown, 1999) – namely *Beta Purple Bacteria* and *Green Non Sulfur Bacteria*. (These classes were selected because the minimum free energy structure obtained by `mfold` predicts a rather different structure than that of the secondary structure obtained by *comparative sequence analysis*; see remarks below. Results for additional classes are available on the web supplement). For each RNA sequence s and each $k \leq 10$, we compute the k-superoptimal free energy (before the post-processing replacement of mutated by wild-type nucleotides and the removal of subsequent unauthorized base pairs) and associated k-superoptimal secondary structure (after the post-processing step). Here, note that the 0-superoptimal structure is just the mfe structure, as determined by Zuker's method. For our experiments, 6 Gb of memory and 2 hours of cpu time on a 666Mhz DEC-alpha were required to compute the k-superoptimal secondary structures and their energies, for $k \leq 10$, for a given RNA sequence of 300 nucleotides.

In data on the web supplement, we show that there is an approximately linear relation between k, the number of mutations, and k-superoptimal free energy δ_k, provided that $k \ll n$, where n denotes RNA sequence length. The slope for graphs of *Beta Purple Bacteria* and *Green Non Sulfur Bacteria* for values $0 \leq k \leq 10$ is similar and approximately equal to -4.93 ± 0.65. This means that on average -4.93 kcal/mol are gained per mutation. Contrasting this amount with the energy gain of -3.3 kcal/mol per additional GC stacked base pair, it follows that multiloops and interior loops must be (radically) modified in the mfe structure for the RNA sequence having an additional mutation – this presents additional evidence of the restructuring of secondary structure for k-mutants. As indicated by similar experiments with a collection of selenocysteine insertion sequence (SECIS) elements from Rfam (Griffiths-Jones et al., 2003) of length 50, the graph of δ_k as a function of k approaches a horizontal asymptote when $k \approx n$.

While experimentally determined RNA secondary structures form the absolute standard, in the absence of a crystal structure, *comparative sequence analysis* is thought to provide a very close approximation to the real structure. Indeed, (Gutell et al., 2002) show that 97% of the base pairs predicted on the basis of comparative sequence analysis are found in the crystal structures of ribosomal RNA. It is thus standard practice (Mathews, 2004), when measuring the accuracy of energy minimization secondary structure algorithms, that a comparison of predicted structure is made with that obtained by comparative sequence analysis.

Figure 4 displays *(i)* the structure obtained by comparative sequence analysis (Gutell et al., 2002), *(ii)* the 0-superoptimal structure produced by AMSAG, i.e. minimum free energy (mfe) structure using energies from Zuker's `mfold 3.0` but without dangles (see Section 2.2 for precise description of energy model), *(iii)* the 12-superoptimal structure produced by AMSAG. Although it is possible that Zuker's `mfold 3.0` or Vienna RNA Package `RNAfold` could produce a better structure than the 0-superoptimal structure obtained by AMSAG (since `mfold` and `RNAfold` support dangles), this is not the case for the sequence from Figure 4. For this 5S rRNA from *E. coli* with accession V00336 from (Gutell et al., 2002), the structure produced by `RNAfold 1.4`, although different than *(ii)* of Figure 4, still has an overall topology quite different than that of *(i)* and *(iii)* (see web supplement).

Motivated by the striking overall improvement of topology produced by the superoptimal structure, in Table 2 (see web supplement for similar tables for other classes), we analyze *sensitivity* and *specificity* when comparing k-superoptimal structures with structures derived by comparative sequence analysis, for value of $0 \leq k \leq 10$. Here, *sensitivity* is defined to be the number of correctly predicted base pairs in the k-superoptimal structure divided by the number of base pairs in the structure derived from comparative sequence analysis. As well, *specificity* is defined to be the number of

RNA id.	score	0	1	2	3	4	5	6	7	8	9	10	best
C-aurantiacus	sens.	48.31	49.44	49.44	56.18	56.18	56.18	56.18	56.18	56.18	56.18	44.94	56.18
	spec.	40.95	44.90	43.14	48.08	47.62	45.87	45.45	44.64	43.86	44.25	35.09	48.08
H-aurantiacus	sens.	34.83	39.33	39.33	39.33	39.33	39.33	38.20	40.45	40.45	40.45	40.45	40.45
	spec.	30.69	34.65	33.98	33.65	35.71	35.00	31.48	32.73	32.43	32.14	31.58	32.73
SMB-B3	sens.	18.31	18.31	18.31	47.89	47.89	47.89	47.89	30.99	29.58	29.58	29.58	47.89
	spec.	15.29	15.12	14.94	36.56	36.17	35.42	35.42	22.45	21.00	20.79	21.43	36.56
T-album	sens.	40.66	40.66	40.66	58.24	58.24	29.67	29.67	29.67	29.67	29.67	29.67	58.24
	spec.	34.91	34.58	33.94	47.75	47.32	23.68	23.48	23.28	23.08	23.08	22.88	47.75
T-roseum	sens.	38.95	49.47	49.47	45.26	45.26	56.84	44.21	55.79	55.79	38.95	38.95	56.84
	spec.	34.91	44.76	44.34	37.39	37.07	46.96	35.59	45.30	44.92	30.58	30.58	46.96
Total	sens.	37.01	40.46	40.46	49.43	49.43	45.98	42.99	43.22	42.99	39.31	37.01	49.43
	spec.	32.01	35.41	34.71	40.80	40.95	37.45	34.19	34.00	33.39	30.32	28.50	40.95

Table 2. This table evaluates the performance of AMSAG for RNA from Green Non Sulfur Bacteria. For each value of $0 \leq k \leq 10$, we report the *sensitivity*, i.e. the number of correctly predicted base pairs in the k-superoptimal structure divided by the number of base pairs in the structure derived by comparative sequence analysis. Additionally, we report the *specificity*; i.e. the number of correctly predicted base pairs in the k-superoptimal structure divided by the number of base pairs in the k-superoptimal structure. The last column gives the score obtained at the level where highest sensitivity is observed.

correctly predicted base pairs in the k-superoptimal structure divided by the number of base pairs in the k-superoptimal structure. In some cases (see for example *B-bronchiseptica* RNA in the Beta Purple Bacteria class, or *SMB-B3* RNA in the Green Non Sulfur Bacteria class), the k-superoptimal structure is quite close to the secondary structure derived by comparative sequence analysis. Moreover, the upper-bound of $k \leq 10$ is often not required to obtain the best results, although in Figure 4, we had to take $k = 12$ to find an approximate resemblance for 5S rRNA with accession number V00336. Our data suggests that competing secondary structures often appear when one admits only a small number of mutations for structural RNA. Figure 5 displays hierarchical clustering obtained by Ward's algorithm for the secondary structure obtained by comparative sequence analysis with k-superoptimal structures, for $0 \leq k \leq 10$. Note that specificity values may be poorer than in reality, because of the fact that structures derived by comparative sequence analysis are determined by covariation of base pairs in high quality multiple sequence alignments, and hence include only a proper subset of the set of experimentally determined base pairs. In particular, the latter structures are generally *not* saturated in the sense of (Zuker and Sankoff, 1984) or locally optimal in the sense of (Clote, 2005). While (Ding et al., 2005) provides a method for improving accuracy of RNA secondary structure prediction, we make no such claim in this paper. In contrast, our interest in this paper is to explore the secondary structure landscape of RNA k-point mutants.

4 CONCLUSION

In this paper, we have described new tools to explore the energy landscape of k-point mutants of a given RNA molecule. Although much is known about the ensemble of low energy structures at thermodynamic equilibrium (McCaskill, 1990; Ding and Lawrence, 2003; Ding et al., 2005), little is known about the energy distribution for secondary structures of k-point mutants. In Section 2.1, we describe a novel algorithm to compute the Boltzmann partition function over all secondary structures of k-point mutants, according to the Nussinov-Jacobson energy model. Using this, we present an efficient method to compute the mean and the standard deviation of the secondary structure energy for k-point mutants. The algorithms to compute the partition function, expected energy, and to compare with brute force enumeration were originally implemented in Python by P. Clote and subsequently translated to C by A. Schreiner. In the future, we intend to extend our partition function and mean energy computation to the newest parameters for the Turner energy model (Xia et al., 1999). One application of this will be to compute the base pairing probability at *positions* i, j in all k-point mutants. While McCaskill's algorithm (McCaskill, 1990) allows one to compute the probability that positions i, j base pair in a given RNA sequence, our approach would allow one to explore critically important base pairing positions and structurally important folds critical for the function of certain *classes* of RNAs. More generally, since structural RNA produces different curves than random RNA when graphing the energy of k-point mutants, cf. Figure 3, a direction of potential application of our work lies in the design of (artificial) RNA sequences, guaranteed not only to have a certain secondary structure (inverse folding), but for which the secondary structure is largely maintained for most k-point mutants. Since the current partition function is non-trivial, we focus on basic concepts in this paper and plan to consider applications and the extension to the Turner energy model in a sequel paper, which will perhaps provide us a better understanding of RNA sequence evolution.

In Section 2.2, we used AMSAG (Waldispühl et al., 2002), which supports the general framework of (multitape) S-attribute grammars, to compute k-superoptimal secondary structures for a technical restriction of the Turner energy model (Matthews et al., 1999). Building on an early program of F. Lefebvre (Lefebvre, 1995), the AMSAG program was designed and implemented in C by J. Waldispühl, B. Behzadi and J.-M. Steyaert. Clustering using Ward's method was carried out in `R` by M. Schuler. Our web server provides a link to the web supplement, and computes the expected

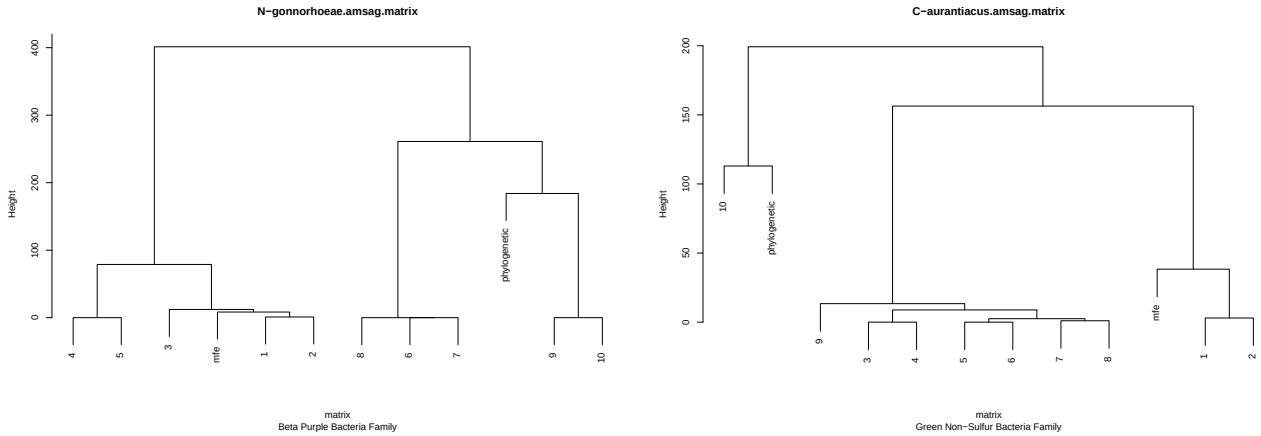

Fig. 5. Clustering of the mfe structure (i.e. 0-superoptimal), the structure derived by comparative sequence analysis, and k-superoptimal structures for $1 \leq k \leq 10$ for one RNA from each of the families *Beta Purple Bacteria* and *Green Non Sulfur Bacteria*, using Ward's method of hierarchical clustering, as implemented in `R`. Additional examples for both Ward's method and for average linkage analysis are available on the web server. Base pair distance between each two pairs of structures is first computed, with subsequent application of Ward's method.

energy for all k-point mutants, as well as the k-superoptimal secondary structures for a given RNA.

Observing that the k-superoptimal structure is often closer to the phylogenetic structure than is the (Zuker) mfe structure, we applied machine learning measures of sensitivity and specificity to quantify this phenomenon. Use of superoptimal structures is *not* meant to be an accurate algorithm for improving RNA secondary structure prediction, but rather a means to provide alternative potentially important candidate structures. In future work, we plan to compute superoptimal secondary structures with respect to the Turner energy model (Xia et al., 1999) (with dangles and without technical restrictions) using the approach described at the beginning of Section 2.2. We intend to provide a better understanding of superoptimal structures, to determine the location of pointwise mutations in the sequence associated with the superoptimal structure, as well as to characterize the structural changes which occur in k-superoptimal structures as they near the phylogenetic structure.

Finally, we would like to thank anonymous referees, for very valuable comments and suggestions.

REFERENCES

Altschul, S. and Erikson, B. (1985). Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol*, 2(6):526–538.

Brown, J. W. (1999). The ribonuclease p database. *Nucleic Acids Res.*, 27(314).

Clote, P. (2005). An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. *J. Computational Biology*, 12(1):83–101.

Clote, P. and Backofen, R. (2000). *Computational Molecular Biology: An Introduction*. John Wiley & Sons. 279 pages.

Clote, P., Ferrè, F., Kranakis, E., and Krizanc, D. (2005a). Structural RNA has lower folding energy than random rna of the same dinucleotide frequency. *RNA*, 11(5):578–591.

Clote, P., Gasieniec, L., Kolpakov, R., Kranakis, E., and Krizanc, D. (2005b). On realizing shapes in the theory of RNA neutral networks. *J. Theoretical Biology*, 236(2):216–227.

Deutsch, J. and Paladin, G. (1989). Product of random matrices in a microcanonical process. *Physical Review Letters*, 62(7).

Ding, Y., Chan, C., and Lawrence, C. (2005). RNA secondary structure by centroids in a Boltzmann weighted ensemble. *RNA*. in press.

Ding, Y. and Lawrence, C. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31(24):7280–7301.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. (2003). Rfam: an RNA family database. *Nucleic Acids Res.*, 31(1):439–441.

Guerra, F. and Toninelli, F. (2002). The thermodynamic limit in mean field spin glass models. *Commun. Math. Phys.*, 230:71–79.

Gutell, R., Lee, J., and Cannone, J. (2002). The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol.*, 12:301–310.

Lefebvre, F. (1995). An optimized parsing algorithm well-suited to rna folding. In press, A., editor, *Proceedings of the Third International Conference on Intelligent Systems for Molecular B iology*, pages 222–230.

Mathews, D. (2004). Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10:1178–1190.

Matthews, D., Sabina, J., Zuker, M., and Turner, D. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940.

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

Nussinov, R. and Jacobson, A. (1980). Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences, USA*, 77(11):6309–6313.

Orlandini, E., Tesi, M., and Whittington, S. (1999). A self-avoiding walk model of random copolymer adsorption. *J. Phys. A: Math. Gen.*, 32:469–477.

Schuster, P., Fontana, W., Stadler, P., and Hofacker, I. (1994). From sequences to shapes and back: A case study in RNA secondary structures. *Proc. R. Soc. Lond. B*, 255:279–284.

Waldispühl, J., Behzadi, B., and Steyaert, J.-M. (2002). An approximate matching algorithm for finding (sub-)optimal sequences in

s-attributed grammars. In *Proceedings of the first European Conference on Computational Biology, ECCB 2002*, volume 18 of *Bioinformatics*, pages 250–259. OXFORD University Press.

Waldispühl, J. and Steyaert, J.-M. (2005). Modeling and predicting all α-transmembrane proteins including helix-helix pairing. *Theoretical Computer Science*, 335(1):67–92.

Xia, T., J. SantaLucia, J., Burkard, M., Kierzek, R., Schroeder, S., Jiao, X., Cox, C., and Turner, D. (1999). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–35.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415.

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

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148.