Available online at www.sciencedirect.com
SCIENCE DIRECT®

Bulletin of Mathematical Biology 67 (2005) 1031–1047

Bulletin of Mathematical Biology

www.elsevier.com/locate/ybulm

Moments of the Boltzmann distribution for RNA secondary structures

István Miklós[a], Irmtraud M. Meyer[b,*], Borbála Nagy[a]

[a]MTA-ELTE Theoretical Biology and Ecology Modelling Group, H-1117 Budapest, Pázmány Péter sétány 1/c, Hungary
[b]European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

Received 21 July 2004; accepted 17 December 2004

Abstract

We here present a dynamic programming algorithm which is capable of calculating arbitrary moments of the Boltzmann distribution for RNA secondary structures. We have implemented the algorithm in a program called RNA-VARIANCE and investigate the difference between the Boltzmann distribution of biological and random RNA sequences. We find that the minimum free energy structure of biological sequences has a higher probability in the Boltzmann distribution than random sequences. Moreover, we show that the free energies of biological sequences have a smaller variance than random sequences and that the minimum free energy of biological sequences is closer to the expected free energy of the rest of the structures than that of random sequences. These results suggest that biologically functional RNA sequences not only require a thermodynamically stable minimum free energy structure, but also an ensemble of structures whose free energies are close to the minimum free energy.

1. Introduction

In the early seventies, Tinoco and colleagues (Tinoco et al., 1971, 1973) proposed an energy model for RNA folding. In this model, the molar free energy of a particular

* Corresponding author.
E-mail address: irmtraud.meyer@cantab.net (I.M. Meyer).

doi:10.1016/j.bulm.2004.12.003

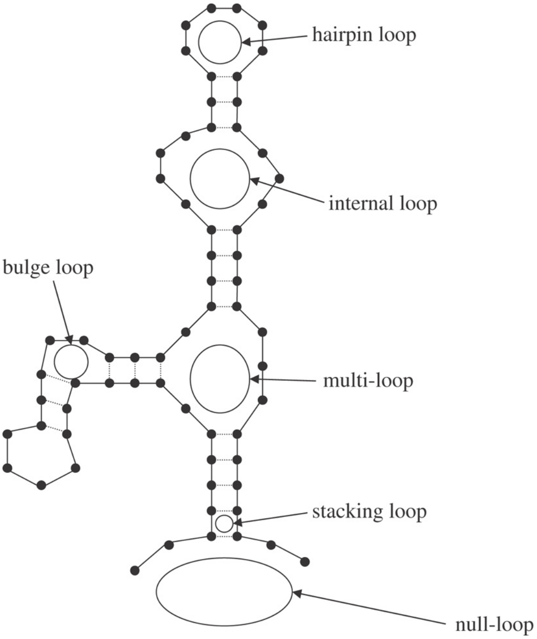

Fig. 1. k-loop decomposition of an RNA secondary structure. Each dot represents an RNA nucleotide, solid lines represent the covalent bonds of the RNA's sugar-phosphate backbone and dotted lines represent the hydrogen bonds between base-pairing nucleotides. The base line that actually does not form a cycle is traditionally called null-loop. Hairpin loops are loops with exactly one hydrogen bound. Multi-loops have more than two hydrogen bonds, while stacking, internal and bulge loops have exactly two hydrogen bonds. A stacking loop consists of only four nucleic acids, internal and bulge loops have more than four nucleic acids. In a bulge loop, the two hydrogen bonds are separated by exactly one covalent bound in one direction. We consider only the following consensus base-pairs `G-C, C-G, A-U, U-A, G-U, U-G`.

secondary structure folding is the sum of independent contributions of base-pair stacking and loop-destabilising terms. Although the number of possible secondary structures grows exponentially with the length of the RNA sequence, dynamic programming algorithms have been developed that find the minimum free energy (mfe) structure for a given sequence in only polynomial time and memory (Nussinov and Jacobson, 1980; Zuker and Stiegler, 1981). The k-loop decomposition (see Fig. 1) was first proposed by Sankoff (Sankoff et al., 1983), and the free energy contributions here follow his definitions. Dynamic programming algorithms that use this parametrisation of free energies are called Zuker–Sankoff algorithms. The state-of-the-art algorithm of this class (Lyngsø et al., 1999) requires $O(l^3)$ time and $O(l^2)$ memory, where l is the length of the sequence.

Although the Zuker–Sankoff algorithm proposes an elegant way for predicting the mfe structure, this prediction may be wrong for several reasons. First, the energy parameters underlying the prediction algorithm are inevitably inaccurate as e.g. slightly different physiological conditions may alter the values of these default energy parameters. The true mfe structure may therefore be only a sub-optimal one with respect to the default parameters. Second, the Zuker–Sankoff algorithm neglects all tertiary interactions as well as pseudo-knots which also decreases the quality of the prediction. A third, more general and important criticism of the Zuker–Sankoff algorithm is that in a biochemical system, not only the mfe structure can be found, but also several sub-optimal ones which may also play a functional role (Ishitani et al., 2003). The free energy of these sub-optimal structures follows a so-called Boltzmann distribution, in which the probability of a certain RNA structure having molar free energy G is proportional to $e^{G/(RT)}$, where R is the universal gas constant and T is the temperature measured in Kelvin degrees. The partition function Z is the normalising constant of the Boltzmann-distribution. As it can be calculated by a dynamic programming algorithm (McCaskill, 1990), we can therefore deduce the exact probability of the RNA structure having any given free energy, in particular the probability of the mfe structure.

Several attempts have already been made to further investigate how well defined the mfe structure is. Zuker (Zuker, 1989) proposed an algorithm which predicts the mfe structure that contains a prescribed, fixed base-pair. Later on, the Vienna group (Wuchty et al., 1999) presented an algorithm which predicts all possible structure whose free energy falls within a prescribed distance from the minimum free energy. In practice, this algorithm can only be used to explore a small free energy interval close to the minimum free energy, since the number of sub-optimal foldings grows exponentially with the distance from the minimum free energy.

None of the existing algorithms, including Zuker's and the Vienna group's, are capable of predicting the entire Boltzmann distribution or global features derived from it. In this paper, we propose an algorithm which can answer the following questions:

- What is the expected free energy of the RNA molecule?
- What is the variance of the free energy for that RNA molecule?

By calculating these two values and by comparing them to the minimum free energy of a given RNA molecule, we then investigate the following, biologically interesting questions:

- Is the expected free energy value close to the minimum free energy value, i.e. how well is the mfe structure defined?
- Can we distinguish random RNA sequences from RNA sequences that are known to have a functional secondary structure?

As the results of this paper will show, biological sequences differ significantly from random sequences. However, none of the statistics proposed in this paper has sufficient separating power to distinguish individual biological RNA sequences from individual random ones.

The rest of this paper is organised in the following way: we first introduce the novel dynamic programming algorithm that calculates the expected value and variance of the Boltzmann distribution of free energies, prove the goodness of the algorithm and analyse its complexity (see section "Algorithm"). We then describe the biological and random data sets and define the four statistics with which we compare the biological and random data sets (see section "Data and methods"). The "Results" section contains the plots for all statistics and investigates the statistical significance of the results. We conclude the paper with the "Discussion" in which we explain our results, discuss their implications and propose directions for further research.

2. Algorithm

We here propose a novel dynamic programming algorithm which is capable of calculating the expected value $E_B[G]$ and the variance $V_B[G]$ of the molar free energy G in the Boltzmann distribution B for a given fixed RNA sequence L. The expected value and variance are by definition

$$E_B[G] = \sum_S \frac{e^{\frac{-G(S)}{RT}} G(S)}{Z}$$

$$V_B[G] = \sum_S \frac{e^{\frac{-G(S)}{RT}} (G(S) - E_B[G])^2}{Z}$$

where Z is the partition function, R is the universal gas constant, T is the temperature in Kelvin degrees, $G(S)$ is the molar free energy of a particular structure S of the fixed RNA sequence L and where we sum over all possible secondary structures S of L. The partition function Z is by definition

$$Z := \sum_S e^{\frac{-G(S)}{RT}}.$$

Our novel dynamic programming algorithm calculates the following two quantities

$$X := \sum_S e^{\frac{-G(S)}{RT}} G(S) \tag{1}$$

$$Y := \sum_S e^{\frac{-G(S)}{RT}} G^2(S) \tag{2}$$

from which we can easily deduce the expected value and variance, since

$$E_B[G] = \frac{X}{Z}$$

$$V_B[G] = E_B[G^2] - E_B^2[G] = \frac{Y}{Z} - \frac{X^2}{Z^2}.$$

We use the Wuchty algorithm (Wuchty et al., 1999) for calculating the minimum free energy secondary structure and its well known variant, the McCaskill algorithm (McCaskill, 1990), for obtaining the partition function Z. The Wuchty algorithm is a variant of the Zuker–Sankoff algorithm (Sankoff et al., 1983; Zuker and Sankoff, 1984) that considers each structure exactly once. Our claim is that we can calculate X and Y by introducing two more sophisticated variants of the Wuchty algorithm.

We proceed by first explaining the Wuchty algorithm as well as the variant that is capable of calculating the partition function Z before introducing the novel modifications that allow us to calculate X and Y and proving that we thereby indeed obtain X and Y.

The calculations of the Wuchty algorithm are performed by three different functions:

- The *basic functions*: These functions take as input a sub-sequence L_i of the RNA sequence L and return as output the free energy G_i for one of the following elementary

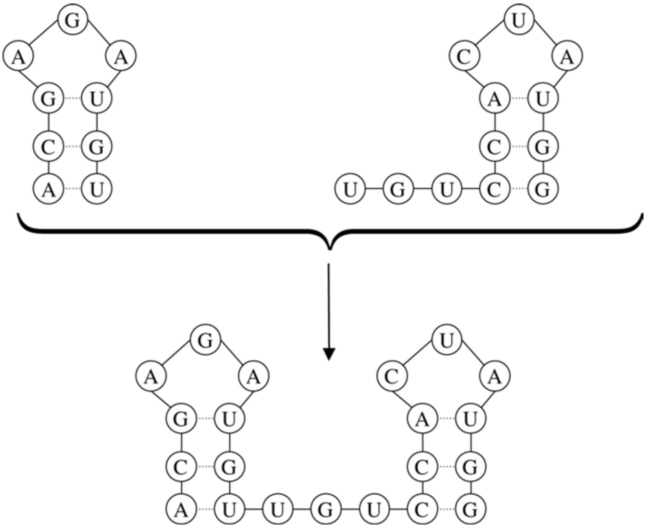

Fig. 2. Merging two sub-strings with secondary structures into a longer sub-string.

secondary structure elements: a hairpin loop, a bulge loop, an internal loop, a multi-loop, a stacking loop or a null-loop being composed of L_i, see Fig. 1.

- The *merge function*: This function takes as input two sub-sequences L_i and L_j and their minimum free energies M_i and M_j and returns as output the free energy M_{ij} of the merged string L_{ij} which is simply the sum of the two individual free energies, i.e. $M_{ij} = M_i + M_j$. This merging either results in the concatenation of two sub-strings, see for example, Fig. 2, or an extension of a helix with a stacking loop, etc. This function may impose constraints on the structures.
- The *choice function*: This function takes as input one sub-string L_i with two competing secondary structures whose free energies are $M_i(a)$ and $M_i(b)$ and returns as output the lower of the two free energies, i.e. $\min\{M_i(a), M_i(b)\}$, see Fig. 3. This function may impose constraints on the structures.

Without any constraints on the energy functions, it would be impossible to calculate the mfe structure in polynomial time. Practical algorithms therefore use linear functions for multi-loop energies to get a polynomial running time (Zuker and Sankoff, 1984). Auxiliary variables are introduced for calculating free energies for multi-loops, see for example (Wuchty et al., 1999), and hence *basic functions* calculate free energies for these auxiliary variables. The same is true for null-loops.

The dynamic programming of the Wuchty algorithm starts with small sub-strings and progresses toward longer sub-strings. Longer sub-strings emerge as shorter sub-strings are merged and constraints are eliminated after obtaining optimal solutions with constraints. Once the minimum free energy of the entire sequence is known, the corresponding mfe structure can be deduced with a trace-back algorithm.

Since the constraints in the above Wuchty algorithm are set up in such a way that each of the possible secondary structures is counted exactly once, the partition function

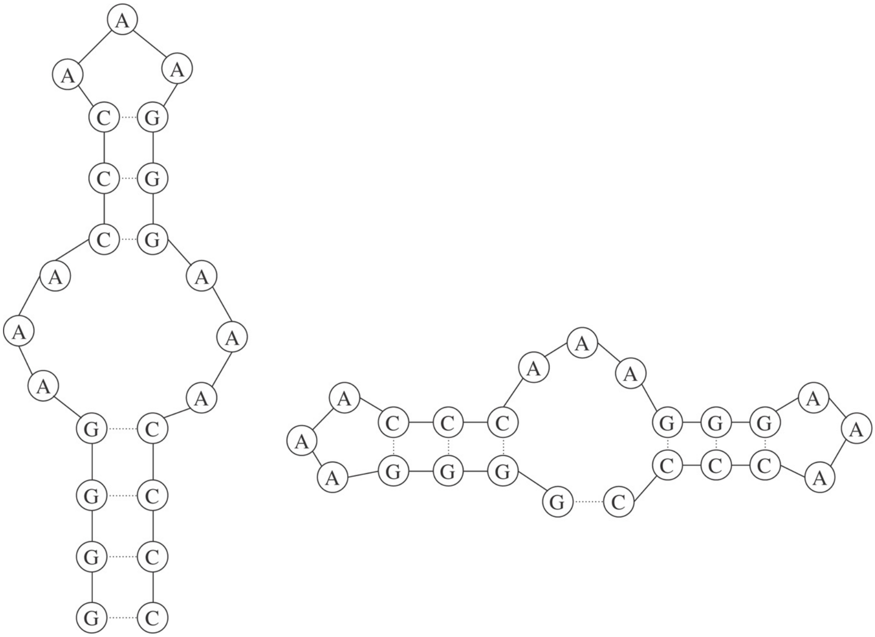

Fig. 3. Two competing structures for the same sub-string. The structure on the left is the optimal structure if the first and last nucleotides have to be paired and close a stacking loop. The structure on the right is the optimal structure if the first and last nucleotides have to close a multi-loop. Both structures are candidates for the optimal structure in which the first and last nucleotides have to be base-paired.

Z can be calculated using a variant of the above Wuchty algorithm (McCaskill, 1990): Instead of operating with the raw free energies G in the above functions, the corresponding exponentiated values $e^{G/(RT)}$ are used. The *choice function* above adds the energies of two competing structures rather than choosing the minimum of the two. The addition of energies is replaced by the multiplication of the corresponding exponentiated values.

The novel modifications to the Wuchty algorithm that we propose in order to calculate X in (1) and Y in (2) are summarised in Table 1. The first column indicates the overall quantity that is calculated with the algorithm (M for the minimum free energy, Z for the partition function and X and Y for Eqs. (1) and (2)). The remaining columns contain the entities that are calculated by the three functions (e.g. the basic function calculates the minimum free energy M_i of sub-string L_i if the algorithm is used to calculate the minimum free energy M of the entire input RNA sequence L, whereas it calculates the partition function Z_i of sub-string L_i if the algorithm is used to calculate the partition function Z).

We now prove the merge function formulae for X and Y in Table 1. Our reasoning is similar to that of the proof for the partition function Z (McCaskill, 1990). When we merge two sub-strings L_i and L_j using the merge function, we have to calculate for X

$$\sum_{S_i} \sum_{S_j} (G(S_i) + G(S_j)) e^{\frac{-(G(S_i)+G(S_j))}{RT}}$$

Table 1
Output quantities of the three functions within the dynamic programming algorithm

Output	Output of the			Algorithm
	Basic function	Merge function	Choice function	
M	$M_i = G_i$	$M_i + M_j$	$\min\{M_i(a), M_i(b)\}$	Wuchty
Z	$Z_i = \mathrm{e}^{\frac{-G_i}{RT}}$	$Z_i Z_j$	$Z_i(a) + Z_i(b)$	McCaskill
X	$X_i = \mathrm{e}^{\frac{-G_i}{RT}} G_i$	$X_i Z_j + X_j Z_i$	$X_i(a) + X_i(b)$	Miklós–Meyer–Nagy
Y	$Y_i = \mathrm{e}^{\frac{-G_i}{RT}} G_i^2$	$Y_i Z_j + 2X_i X_j + Y_j Z_i$	$Y_i(a) + Y_i(b)$	

M indicates a minimum free energy, Z a partition function and X and Y the quantities defined in Eqs. (1) and (2), the indexes i and j refer to sub-strings L_i and L_j of the input RNA sequence L, respectively, and a and b indicate two competing secondary structures. Please refer to the text for more explanation.

which can be transformed into

$$\begin{aligned}
&\sum_{S_i}\sum_{S_j}(G(S_i) + G(S_j))\mathrm{e}^{\frac{-(G(S_i)+G(S_j))}{RT}} \\
&\quad = \sum_{S_i}\sum_{S_j} G(S_i)\mathrm{e}^{\frac{-G(S_i)}{RT}}\mathrm{e}^{\frac{-G(S_j)}{RT}} \\
&\qquad + \sum_{S_i}\sum_{S_j} G(S_j)\mathrm{e}^{\frac{-G(S_j)}{RT}}\mathrm{e}^{\frac{-G(S_i)}{RT}} \\
&\quad = \sum_{S_i} G(S_i)\mathrm{e}^{\frac{-G(S_i)}{RT}} \sum_{S_j}\mathrm{e}^{\frac{-G(S_j)}{RT}} \\
&\qquad + \sum_{S_j} G(S_j)\mathrm{e}^{\frac{-G(S_j)}{RT}} \sum_{S_i}\mathrm{e}^{\frac{-G(S_i)}{RT}} \\
&\quad = X_i Z_j + X_j Z_i
\end{aligned}$$

where

$$X_i = \sum_{S_i} G(S_i)\mathrm{e}^{\frac{-G(S_i)}{RT}} \qquad Z_i = \sum_{S_i}\mathrm{e}^{\frac{-G(S_i)}{RT}}$$

(and similarly with index j instead of i). Similarly, for Y we have to calculate

$$\sum_{S_i}\sum_{S_j}(G(S_i) + G(S_j))^2\mathrm{e}^{\frac{-(G(S_i)+G(S_j))}{RT}}$$

which can be readily transformed to show that it is indeed equal to $Y_i Z_j + 2X_i X_j + Y_j Z_i$. Overall, we can thus conclude that we can use the same dynamic programming algorithm to calculate the minimum free energy M, the partition function Z, X and Y by simply choosing the three underlying functions accordingly.

We implemented the dynamic programming algorithm in C++ in order to take advantage of the underlying structure of the algorithm. Each RNA sequence L corresponds to one instance of an RNA class whose private variables store the respective M_i, Z_i, X_i and Y_i

values for each sub-string L_i. The basic, merge and choice functions are public functions of the RNA class. Each of the three functions takes as input the coordinates of the two sub-strings L_i and L_j and operates on each of the four different types of private variables according to Table 1. In this way, we can calculate all *four* final quantities M, Z, X and Y by traversing the dynamic programming loops only *once*. Our program RNA-VARIANCE uses the free energy parameters of the latest MFOLD version (Mathews et al., 1999; Zuker, 2003) and makes use of the fast internal loop calculation by Lyngsø et al. (Lyngsø et al., 1999). It requires $O(l^3)$ time and $O(l^2)$ memory for analysing an RNA sequence of length l. As our method is based on the Wuchty algorithm, it does not take tertiary structure or pseudo-knots into account. The source code is available on request from the authors. It takes about 10 min and 49 MB memory to analyse an RNA sequence of 120 nucleotides length on a Pentium4 2.0 GHz computer.

3. Data and methods

Due to the theoretical and practical limitations of our algorithm, we selected biological RNA sequences whose known structures do not contain pseudo-knots and which are relatively short (100–200 nucleotides long) in order to be able to analyse a large data set in a reasonable time. We therefore chose the following three data sets: 593 precursor miRNA sequence from the miRNA data base (Griffiths-Jones, 2004), 86 tRNA sequences of the *Escherichia coli* K12 strain (Genbank accession number: 16127994) from the Genbank database (Benson et al., 2004), and 57 5S rRNA sequences from the 5S rRNA database (Szymanski et al., 2002). We did not merge these datasets, because they have very different structures and their nucleotide compositions differ significantly from each other.

It is well known that RNA secondary structures are mainly stabilised by stacking base-pairs. The minimal free energy of a folded RNA sequence therefore highly depends on the distribution of neighbouring nucleotide pairs since the stacking energies contribute more to the free energy of helices than the hydrogen bonds between complementary nucleotides. For each of the three data sets, we therefore generated a large set of random sequences (500 sequences) which has the same dinucleotide statistics and length distribution as the corresponding biological data set (see Tables A.1–A.3). This was done using a first order Markov chain.

We calculated the minimal free energy M, partition function Z and the expected value $E_B[G]$ and variance $V_B[G]$ of the free energies of the Boltzmann distribution of secondary structures for each of the biological and random sequences. We then calculate four different statistics for each biological data set and its corresponding random set:

- the minimum free energy of each sequence, normalised by the length of each sequence such that sequences with average length are normalised by 1 (see distributions titled “minimum energies”);
- the logarithm of the mfe structure’s probability, i.e. $\log_{10}\left(\frac{\exp(-G_{\mathrm{mfe}}/RT)}{Z}\right)$ (see distributions titled “log probabilities”);
- the free energy distance between the minimum free energy and the expected free energy value of the remaining free energy distribution normalised as above by the length of the sequence (see distributions titled “deviations”). The expectation value of the remaining

Table 2
T-tests for the hypothesis that random and biological data have the same expectation value

	Random	Biological	*p* value
		miRNA data set	
Minimum energies	−28.07 ± 0.17	−43.98 ± 0.31	1.31E−238
Log probabilities	−4.67 ± 0.041	−2.37 ± 0.04	6.93E−247
Deviations	2.87 ± 0.03	1.64 ± 0.02	8.02E−203
Variances	4.48 ± 0.05	2.37 ± 0.04	8.12E−183
		E. coli K12 tRNA data set	
Minimum energies	−27.65 ± 0.22	−32.93 ± 0.32	2.85E−15
Log probabilities	−3.89 ± 0.05	−3.07 ± 0.10	8.35E−08
Deviations	2.38 ± 0.03	2.08 ± 0.05	0.00176
Variances	3.54 ± 0.04	2.96 ± 0.08	2.08E−05
		5S rRNA data set	
Minimum energies	−45.26 ± 0.28	−53.23 ± 0.63	4.54E−18
Log probabilities	−5.66 ± 0.06	−4.83 ± 0.11	3.11E−07
Deviations	5.63 ± 0.06	5.29 ± 0.18	0.0316
Variances	3.56 ± 0.04	−3.22 ± 0.10	0.0016

For each of the four statistics, we report the expected value and its standard error as well as the *p*-value which is the probability that the random and biological distributions have the same expected value.

free energy distribution is calculated by

$$\frac{\sum_S G(S)\mathrm{e}^{\frac{-G(S)}{RT}} - \mathrm{e}^{\frac{-G_{\mathrm{mfe}}}{RT}} G_{\mathrm{mfe}}}{Z - \mathrm{e}^{\frac{-G_{\mathrm{mfe}}}{RT}}};$$

- the variance of the free energies of the Boltzmann distribution (see distributions titled “variances”).

We plotted these four statistics for all three data sets and tested for each statistics the hypothesis that the biological and random sequences have the same expected values.

4. Results

All three types of RNA sequences, i.e. the miRNA sequences, the *E. coli* tRNA sequences and the 5S rRNA sequences, show the same qualitative behaviour for all four statistics (see Figs. 4–6). However, we observe different quantitative results for the three data sets, see Table 2. The biological sequences have lower minimum free energies and their mfe structures have a higher probability than the random sequences. In addition, the Boltzmann distribution of the biological sequences has a smaller variance and the free energy difference between the mfe and expected energy is smaller than for the random sequences. These differences between the biological and the random sequences are significant for all four statistics, see Table 2, especially for the miRNA sequences whose *p*-values are very small. However, none of the statistics has sufficient separating power to distinguish individual biological from individual random sequences.

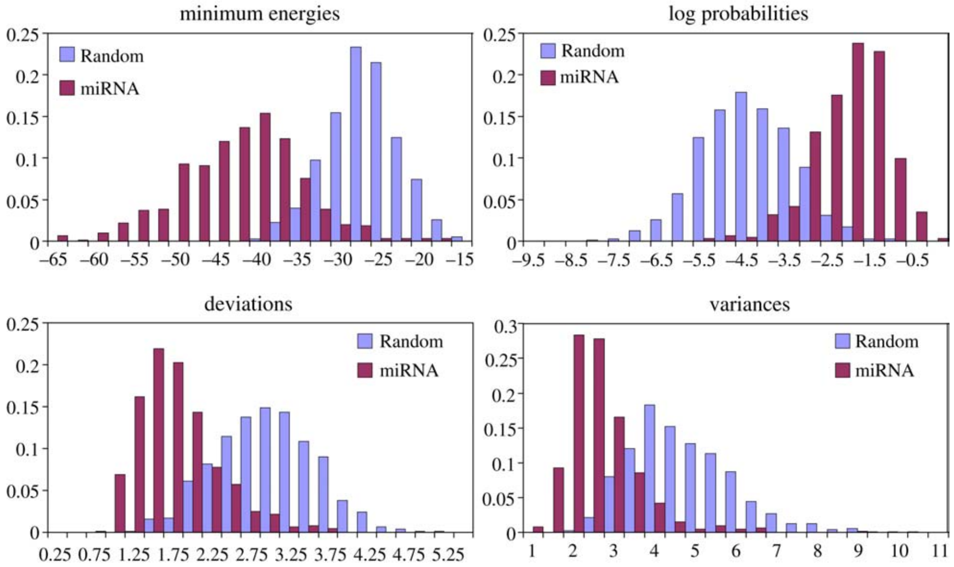

Fig. 4. Comparison of the biological and random miRNA sequences. All free energies are given in kcal/mol, the variance is given in $\text{kcal}^2/\text{mol}^2$. Please see the text for a definition of the four statistics.

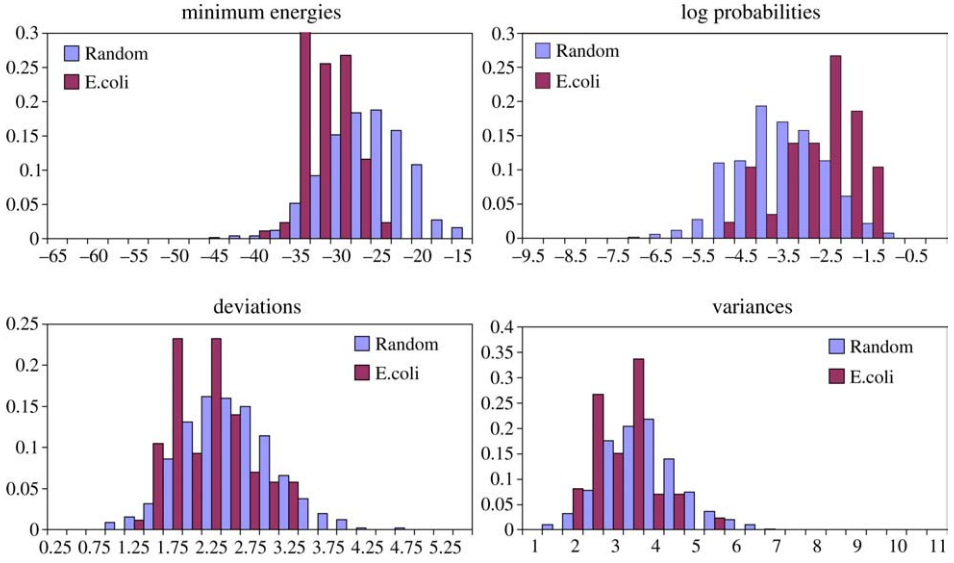

Fig. 5. Comparison of the biological and random *E. coli* K12 tRNA sequences. Please see the text and the caption of Fig. 4 for more information.

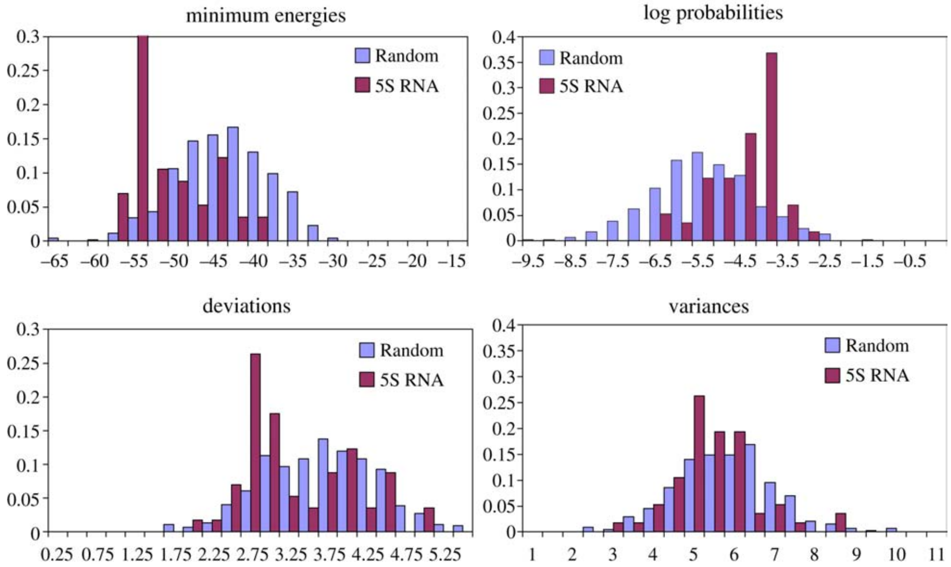

Fig. 6. Comparison of the biological and random 5S rRNA sequences. Please see the text and the caption of Fig. 4 for more information.

5. Discussion

The secondary structure of precursor miRNA sequences is a single hairpin. The miRNA sequences are not post-transcriptionally modified. An enzyme recognises the hairpin structure of the precursor miRNA sequences and excises one strand of the hairpin which corresponds to the mature miRNA. This mature miRNA does not have a distinct secondary structure. In contrast to precursor miRNA sequences, tRNA and 5S rRNA sequences may contain several post-transcriptionally modified nucleotides and, moreover, can change their secondary structure during biochemical reactions (Ishitani et al., 2003). It is therefore not surprising that we find that precursor miRNA sequences differ more significantly from random sequences than tRNA and 5S rRNA sequences.

Even though we can distinguish a *set* of biological sequences from a *set* of random sequences, the overlap between the statistics of the two sets is generally too large to be able to distinguish an *individual* biological sequence from an *individual* random one. This was already known for the statistics of minimum free energies (Workman and Krogh, 1999; Rivas and Eddy, 2000) and our results for the same and novel statistics confirm it. For the log-probabilities of the mfe structures, our results are in conflict with the results reported by the Vienna group. They find that the probability of the mfe structures of *E. coli* K12 tRNA often exceeds 50% (Wuchty et al., 1999), whereas we did not observe a single such case. Their overestimation of this probability can be explained by the fact that they approximated the partition function using a statistics on sub-optimal foldings, whereas we use the exact value of the partition function. They also used different free energy parameters and replaced modified nucleotides with *N* symbols in the sequence analysis, and these nucleotides were not allowed to form base pairs.

Two novel and biologically interesting results of our studies are that biological RNA sequences have a smaller free energy variance than random sequences and that the minimum free energies of biological RNA sequences are closer to the expected free energies than those of random sequences. Both results suggest that nature is not trying to single out the mfe structure by giving it a free energy which is significantly lower than the expected free energy. These results also imply that nature seems to encourage the presence of secondary structures whose energy is close to that of the minimum free energy. These secondary structures may have their own functional roles or may provide an ensemble of secondary structures that biases the future evolution of new structures toward biologically functional structures. Another possible explanation of these findings is that the biological RNA sequences of our data sets may fold co-transcriptionally (Meyer and Miklós, 2004) and that the temporary structures formed during co-transcriptional folding bias the Boltzmann distribution toward the observed small variance values and small deviations between the minimum free energies and the expected free energies (see Fig. 7 for an example). This conjecture is likely to be true for two reasons: first, temporary structures are counted in the Boltzmann ensemble and would result in a bias toward smaller variances and smaller deviations of minimum free energies from the expected free energies. Second, we observe that both variances and deviations for miRNA sequences differ more significantly from random sequences than for tRNA and 5S rRNA sequences and we know that miRNA sequences fold on their own, whereas tRNA and 5S rRNA sequences frequently contain introns and require the presence of several enzymes during folding.

In order to investigate whether or not the observed small variances and deviations are due to close variants of the mfe structure, we did two things. First, we excluded the mfe structure from the calculation of the expectation and variance values in order not to be biased by its typically large probability within the Boltzmann distribution. Second, we repeated the entire analysis by counting only maximally base-pairing structures (i.e. structures whose helices cannot be extended by more base-pairs) in the Boltzmann distribution and got the same qualitative results (data not shown). We can therefore conclude that the observed small variance and deviation values have to be attributed to secondary structures that differ significantly from the mfe structure.

The overall conclusion from our results is that biological RNA sequences are engineered under two, conflicting constraints: the probability of the mfe structure is maximised while the difference between the minimum free energy and the expected energy of the Boltzmann ensemble (excluding the mfe structure) is minimised. The first constraint encourages energetically stable structures, whereas the second one seems to ensure proper functionality. The competition between the two constraints may be the reason why we cannot distinguish individual biological and random sequences based on the statistics we investigated.

6. Conclusions

In this paper, we introduced a novel algorithm which is capable of calculating the expected value and variance of free energies in the Boltzmann distribution of secondary structures for a given RNA sequence. Until now, the Boltzmann distribution could only be approximated. This was done by listing as many as possible sub-optimal foldings

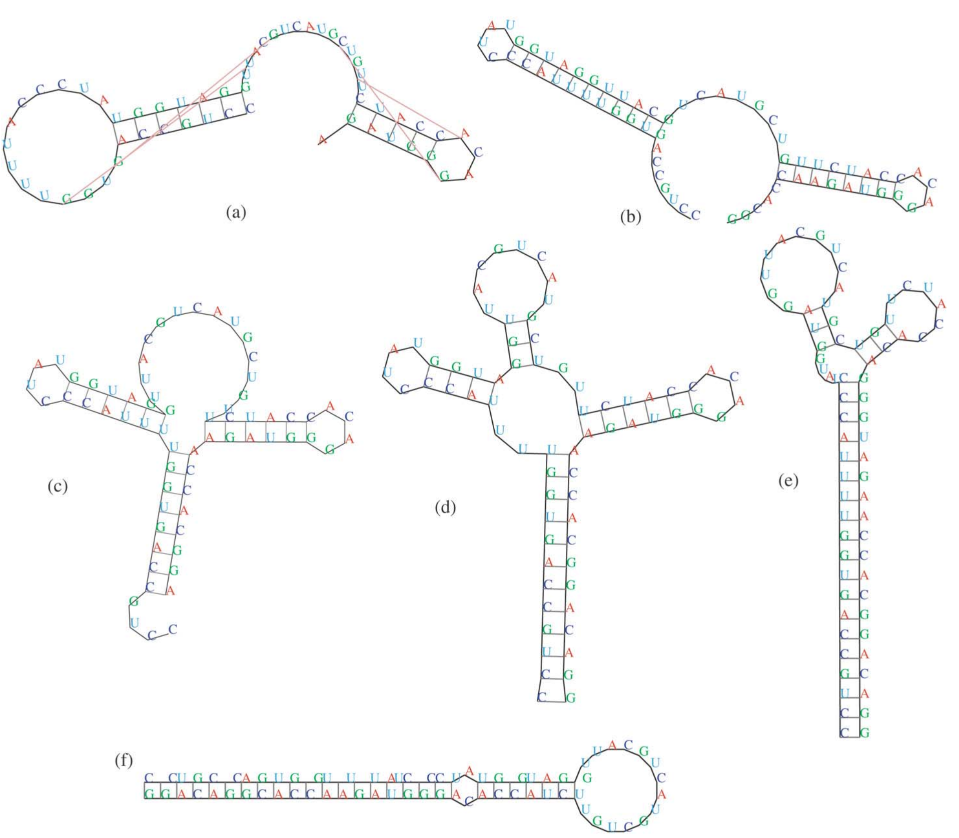

Fig. 7. Co-transcriptional folding of the *Mus musculus* miRNA gene mmu-mir-140, as predicted by the KineFold server (Xayaphoummine et al., 2003): (a) At the beginning of transcription, two helices and pseudoknots (symbolised by connecting lines) are already formed. (b)–(c) Distinct helices form as the transcription continues, keeping the overall temporary structure compact. (d)–(e) At the end of the transcription, the final structure emerges. (f) The final secondary structure is a single hairpin and almost identical to the published structure.

according to their contribution to the Boltzmann distribution. As even short RNA sequences can have hundreds of sub-optimal structures (Wuchty et al., 1999), this calculation is not only an approximation, but also easily becomes prohibitively slow. Our novel algorithm calculates the exact values for the expected value and variance and is only a constant times slower than the algorithm calculating the partition function, and this constant does not depend on how many nearly optimal structures exist. Our algorithm is simple in the sense that it uses the same dynamic programming algorithm as the algorithms that calculate the partition function and the minimum free energy. The algorithm that we proposed can easily be extended to calculate arbitrary higher moments of the Boltzmann distribution without requiring significant changes of the underlying source code. For example, the third moment could be calculated in order to measure the skewness of the Boltzmann distribution. Indeed, if we wish to calculate the k-th moment of the Boltzmann distribution

$$E_B[G^k] \cdot Z = Z_k := \sum_S \mathrm{e}^{\frac{-G(S)}{RT}} G^k(S)$$

we simply have to incorporate the following into the merge function (see Table 1),

$$\sum_{S_i} \sum_{S_j} \mathrm{e}^{\frac{-(G(S_i)+G(S_j))}{RT}} (G(S_i) + G(S_j))^l = \sum_{k=0}^{l} \binom{l}{k} Z_{k,i} Z_{l-k,j}$$

and have to ensure that the $Z_{k,i}$ and $Z_{k,j}$ values have already been calculated for all $k \in \{0, \ldots, l\}$.

Another standard technique to infer RNA secondary structures are stochastic context-free grammars (SCFGs) (Durbin et al., 1998). For SCFGs used in bioinformatics, each derivation tree for an input RNA sequence represents a secondary structure (note that the mapping between structures and derivation trees is only bijective if the grammar underlying the SCFG is unambiguous), and the most likely derivation tree is reported as the predicted secondary structure. Recently, Nebel (2004a,b) investigated which kinds of statistics may be useful for obtaining good RNA secondary structure predictions. He calculated the expected values and variances of several statistics defined by SCFGs. These statistics incorporated both information on the number of different secondary structure elements (e.g. hairpin loops, internal loops, helices etc.) as well as their lengths. Although the dynamic programming algorithms of SCGFs and the free-energy based Zuker–Tinoco model have some similarities, nobody has yet converted the Zuker–Tinoco model into a stochastic context-free grammar as this can only be done in an approximative way. The SCFG analogue of the novel algorithm that we proposed here in order to calculate the free energy moments of the Boltzmann distribution would calculate the moments of log-probabilities of derivation trees for a given RNA sequence. While it is not hard to show that such a dynamic programming algorithm exists, it is an open question whether or not it will be possible to calculate moments of statistics such as e.g. the number of structural elements (helices, loops, etc.) in the Boltzmann distribution.

Although we introduced and investigated our novel algorithm for the Wuchty algorithm which excludes pseudo-knots, the algorithm itself is not limited to investigating secondary structures without pseudo-knots. The general pseudo-knot prediction problem has been shown to be NP-hard (Lyngsø and Pedersen, 2000). However, several polynomial time algorithms exist which are capable of predicting RNA structures that contain special classes of pseudoknots. For some classes of pseudoknots, variants of these algorithms exist that count each possible structure exactly once, the two most comprehensive ones have been published by Reeder and Giegerich (2004) and Dirks and Pierce (2003). They can be easily extended to calculate arbitrary moments of the Boltzmann distribution including those special types of pseudoknots by incorporating our algorithm into the dynamic programming steps. These algorithms are currently too time-consuming (they run in $O(l^4)$ to $O(l^5)$ time, where l is the length of the sequence), but in a few years time, computers will be fast enough to evaluate them, especially with a careful implementation that uses corner-cutting methods such as those presented by Eppstein et al. (1988) and Hofacker et al. (1994).

Acknowledgements

This research was funded by ESRA grant HAMJW and MRC grant HAMKA. I.M. is further supported by a Békésy György postdoctoral fellowship.

Appendix A. Dinucleotide composition for each of the three data sets

Table A.1
Starting probabilities (top) and transition matrix (bottom) of the Markov model that was used to generate the miRNA set of random sequences

	A	C	G	U
	0.114504	0.259542	0.427481	0.198473
	A	C	G	U
A	0.224466	0.232233	0.302136	0.241165
C	0.300514	0.255327	0.120867	0.323292
G	0.215173	0.228908	0.284500	0.271419
U	0.170892	0.230047	0.330516	0.268545

Table A.2
Starting probabilities (top) and transition matrix (bottom) of the Markov model that was used to generate the *E. coli* tRNA set of random sequences

	A	C	G	U
	0.034884	0.058140	0.441860	0.465116
	A	C	G	U
A	0.189105	0.282490	0.350195	0.178210
C	0.206790	0.319959	0.265432	0.207819
G	0.214393	0.242379	0.328336	0.214893
U	0.183989	0.348315	0.240169	0.227528

Table A.3
Starting probabilities (top) and transition matrix (bottom) of the Markov model that was used to generate the 5S rRNA set of random sequences

	A	C	G	U
	0.222074	0.074468	0.421543	0.281915
	A	C	G	U
A	0.227118	0.271833	0.299011	0.202038
C	0.210826	0.341015	0.272457	0.175702
G	0.232497	0.228276	0.295501	0.243726
U	0.231999	0.260688	0.327385	0.179929

References

Benson, D.A., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Wheeler, D., 2004. GenBank: update. Nucleic Acids Res. 32, D23–D26.

Dirks, R., Pierce, N., 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknot. J. Comput. Chem. 24, 1664–1677.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.

Eppstein, D., Galil, Z., Giancarlo, R., 1988. Speeding up dynamic programming. In: Proc. 29th Symp. Foundations of Computer Science. Assoc. Comput. Mach., pp. 488–496.

Griffiths-Jones, S., 2004. The microRNA registry. Nucleic Acids Res. 32, D109–D111.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, M., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie 125, 167–188.

Ishitani, R., Nureki, O., Nameki, N., Okada, N., Nishimura, S., Yokoyama, S., 2003. Alternative tertiary structure of tRNA for recognition of a post-transcriptional modification enzyme. Cell 113, 383–394.

Lyngsø, R., Pedersen, C., 2000. Pseudoknots in RNA secondary structures. In: Procceedings of RECOMB. Tokyo, Japan, pp. 201–209.

Lyngsø, R., Zuker, M., Pedersen, C., 1999. Fast evaluation of internal loops in RNA secondary structure prediction. Bioinformatics 15 (6), 440–445.

Mathews, D., Sabina, J., Zuker, M., Turner, D., 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288, 911–940.

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

Meyer, I.M., Miklós, I., 2004. Co-transcriptional folding is encoded within RNA genes. BMC Mol. Biol. 10, 5.

Nebel, M., 2004a. Identifying good predictions of RNA secondary structure. In: Proceedings of the Pacific Symposium on Biocomputing. vol. 9. pp. 423–434.

Nebel, M., 2004b. Investigation of the Bernoulli-model for RNA secondary structures. Bull. Math. Biol. 66 (6), 925–964.

Nussinov, R., Jacobson, A., 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc. Natl. Acad. Sci. USA 77, 6309–6313.

Reeder, J., Giegerich, R., 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics 5, 104.

Rivas, E., Eddy, S., 2000. Secondary structure alone is generally not statistically significant for the detection of non-coding RNAs. Bioinformatics 16 (7), 583–605.

Sankoff, D., Kruskal, J., Mainville, S., Cedergren, R., 1983. Fast algorithms to determine RNA secondary structures containing multiple loops. In: Sankoff, D., Kruskal, J. (Eds.), Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley, Reading, MA, pp. 93–120.

Szymanski, M., Barciszewska, M.Z., Erdmann, V.A., Barciszewski, J., 2002. 5S ribosomal RNA database. Nucleic Acids Res. 30, 176–178.

Tinoco, I.J., Borer, P., Dengler, B., Levine, M., Uhlenbeck, O., 1973. Improved estimation of secondary structure in ribonucleic acids. Nat. New Biol. 246, 40–41.

Tinoco, I., Uhlenbeck, O.C., Levine, M.D., 1971. Estimation of secondary structure in ribonucleic acids. Nature 230, 362–367.

Workman, C., Krogh, A., 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res. 27 (24), 4816–4822.

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

Xayaphoummine, A., Bucher, T., Thalmann, F., Isambert, H., 2003. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. Proc. Natl. Acad. Sci. USA 100, 15310–15315.

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

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

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

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