

RNA Secondary Structure Boltzmann Distribution (2005; Miklós, Meyer, Nagy)

Rune B. Lyngsø, University of Oxford, www.stats.ox.ac.uk/~lyngsoe

INDEX TERMS: RNA secondary structure, Gibbs free energy, moments of Boltzmann distribution.

1 SYNONYMS

Also known as *Full Partition Function* of RNA secondary structures.

2 PROBLEM DEFINITION

This problem is concerned with computing features of the Boltzmann distribution over RNA secondary structures in the context of the standard Gibbs free energy model used for RNA Secondary Structure Prediction by Minimum Free Energy (cf. corresponding entry). Thermodynamics state that for a system with configuration space Ω and free energy given by $E : \Omega \mapsto \mathbf{R}$, the probability of the system being in state $\omega \in \Omega$ is proportional to $e^{-E(\omega)/RT}$ where R is the universal gas constant and T the absolute temperature of the system. The normalising factor

$$Z = \sum_{\omega \in \Omega} e^{-E(\omega)/RT} \tag{1}$$

is called the *full partition function* of the system.

Over the past several decades, a model approximating the free energy of a structured RNA molecule by independent contributions of its secondary structure components has been developed and refined. The main purpose of this work has been to assess the stability of individual secondary structures. However, it immediately translates into a distribution over all secondary structures. Early work focused on computing the pairing probability for all pairs of bases, i.e. the sum of the probabilities of all secondary structures containing that base pair. Recent work has extended methods to compute probabilities of base pairing probabilities for RNA heterodimers [2], i.e. interacting RNA molecules, and expectation, variance and higher moments of the Boltzmann distribution.

2.1 Notation

Let $s \in \{A, C, G, U\}^*$ denote the sequence of bases of an RNA molecule. Use $X \cdot Y$ where $X, Y \in \{A, C, G, U\}$ to denote a base pair between bases of type X and Y , and $i \cdot j$ where $1 \leq i < j \leq |s|$ to denote a base pair between bases $s[i]$ and $s[j]$.

Definition 1 (RNA Secondary Structure). *A secondary structure for an RNA sequence s is a set of base pairs $\mathcal{S} = \{i \cdot j \mid 1 \leq i < j \leq |s| \wedge i < j - 3\}$. For $i \cdot j, i' \cdot j' \in \mathcal{S}$ with $i \cdot j \neq i' \cdot j'$*

- $\{i, j\} \cap \{i', j'\} = \emptyset$ (each base pairs with at most one other base)
- $\{s[i], s[j]\} \in \{\{A, U\}, \{C, G\}, \{G, U\}\}$ (only Watson-Crick and G, U wobble base pairs)

- $i < i' < j \Rightarrow j' < j$ (base pairs are either nested or juxtaposed but not overlapping)

The second requirement, that only canonical base pairs are allowed, is standard but not consequential in solutions to the problem. The third requirement states that the structure does not contain pseudoknots. This restriction is crucial for the results listed in this entry.

2.2 Energy Model

The model of Gibbs free energy applied, usually referred to as the nearest-neighbour model, was originally proposed by Tinoco *et al.* [10,11]. It approximates the free energy by postulating that the energy of the full three dimensional structure only depends on the secondary structure, and that this in turn can be broken into a sum of independent contributions from each loop in the secondary structure.

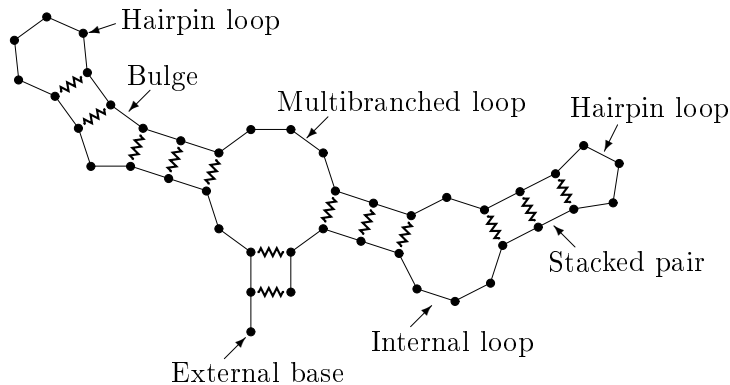


Figure 1: A hypothetical RNA structure illustrating the different loop types. Bases are represented by circles, the RNA backbone by straight lines, and base pairs by zigzagged lines.

Definition 2 (Loops). For $i \cdot j \in \mathcal{S}$, base k is accessible from $i \cdot j$ iff $i < k < j$ and $\neg \exists i' \cdot j' \in \mathcal{S} : i < i' < k < j' < j$. The loop closed by $i \cdot j$, $\ell_{i,j}$, consists of $i \cdot j$ and all the bases accessible from $i \cdot j$. If $i' \cdot j' \in \mathcal{S}$ and i' and j' are accessible from $i \cdot j$, then $i' \cdot j'$ is an interior base pair in the loop closed by $i \cdot j$.

Loops are classified by the number of interior base pairs they contain:

- hairpin loops have no interior base pairs
- stacked pairs, bulges, and internal loops have one interior base pair that is separated from the closing base pair on neither side, on one side, or on both sides, respectively
- multibranched loops have two or more interior base pairs

Bases not accessible from any base pair are called external. This is illustrated in Fig. 1. The free energy of structure \mathcal{S} is

$$\Delta G(\mathcal{S}) = \sum_{i,j \in \mathcal{S}} \Delta G(\ell_{i,j}) \quad (2)$$

where $\Delta G(\ell_{i,j})$ is the free energy contribution from the loop closed by $i \cdot j$. The contribution of \mathcal{S} to the full partition function is

$$e^{-\Delta G(\mathcal{S})/RT} = e^{-\sum_{i,j \in \mathcal{S}} \Delta G(\ell_{i,j})/RT} = \prod_{\ell_{i,j} \in \mathcal{S}} e^{-\Delta G(\ell_{i,j})/RT} . \quad (3)$$

Problem 1 (RNA Secondary Structure Distribution).

INPUT: RNA sequence s , absolute temperature T and specification of ΔG at T for all loops.

OUTPUT: $\sum_{\mathcal{S}} e^{-\Delta G(\mathcal{S})/RT}$, where the sum is over all secondary structures for s .

3 KEY RESULTS

Solutions are based on recursions similar to those for RNA Secondary Structure Prediction by Minimum Free Energy, replacing sum and minimisation with multiplication and sum (or more generally with a *merge function* and a *choice function* [8]). The key difference is that recursions are required to be non-redundant, i.e. any particular secondary structure only contributes through one path through the recursions.

Theorem 1. *Using the standard thermodynamic model for RNA secondary structures, the partition function can be computed in time $O(|s|^3)$ and space $O(|s|^2)$. Moreover, the computation can build data structures that allow $O(1)$ queries of the pairing probability of $i \cdot j$ for any $1 \leq i < j \leq |s|$ [5–7].*

Theorem 2. *Using the standard thermodynamic model for RNA secondary structures, the expectation and variance of free energy over the Boltzmann distribution can be computed in time $O(|s|^3)$ and space $O(|s|^2)$. More generally, the k^{th} moment*

$$E_{\text{Boltzmann}}[\Delta G] = 1/Z \sum_{\mathcal{S}} e^{-\Delta G(\mathcal{S})/RT} \Delta G^k(\mathcal{S}), \quad (4)$$

where $Z = \sum_{\mathcal{S}} e^{-\Delta G(\mathcal{S})/RT}$ is the full partition function and the sums are over all secondary structures for s , can be computed in time $O(k^2|s|^3)$ and space $O(k|s|^2)$ [8].

In Theorem 2 the free energy does not hold a special place. The theorem holds for any function Φ defined by an independent contribution from each loop,

$$\Phi(\mathcal{S}) = \sum_{i \cdot j \in \mathcal{S}} \phi(\ell_{i \cdot j}), \quad (5)$$

provided each loop contribution can be handled with the same efficiency as the free energy contributions. Hence, moments over the Boltzmann distribution of e.g. number of base pairs, unpaired bases, or loops can also be efficiently computed by applying appropriately chosen indicator functions.

4 APPLICATIONS

The original use of partition function computations was for discriminating between well defined and less well defined regions of a secondary structure. Minimum free energy predictions will always return a structure. Base pairing probabilities help identify regions where the prediction is uncertain, either due to the approximations of the model or that the real structure indeed does fluctuate between several low energy alternatives. Moments of Boltzmann distributions are used in identifying how biological RNA molecules deviates from random RNA sequences.

The data structures computed in Theorem 1 can also be used to efficiently sample secondary structures from the Boltzmann distribution. This has been used for probabilistic methods for secondary structure prediction, where the centroid of the most likely cluster of sampled structures is returned rather than the most likely, i.e. minimum free energy, structure [3]. This approach better accounts for the entropic effects of large neighbourhoods of structurally and energetically very similar structures. As a simple illustration of this effect, consider twice flipping a coin with probability $p > 0.5$ for heads. The probability p^2 of heads in both flips is larger than the probability $p(1-p)$ of heads followed by tails or tails followed by heads (which again is larger than the probability

$(1 - p)^2$ of tails in both flips). However, if the order of the flips is ignored the probability of one heads and one tails is $2p(1 - p)$. The probability of two heads remains p^2 which is smaller than $2p(1 - p)$ when $p < \frac{2}{3}$. Similarly a large set of structures with fairly low free energy may be more likely, when viewed as a set, than a small set of structures with very low free energy.

5 OPEN PROBLEMS

As for RNA Secondary Structure Prediction by Minimum Free Energy, improvements in time and space complexity are always relevant. This may be more difficult for computing distributions, as the more efficient dynamic programming techniques of [9] cannot be applied. In the context of genome scans, the fact that the start and end positions of encoded RNA molecule is unknown has recently been considered [1].

Also the problem of including structures with pseudoknots, i.e. structures violating the last requirement in Def. 1, in the configuration space is an active area of research. It can be expected that all the methods of Theorems 3 through 6 in the entry on RNA Secondary Structure Prediction Including Pseudoknots can be modified to computation of distributions without affecting complexities. This may require some further bookkeeping to ensure non-redundancy of recursions, and only in [4] has this actively been considered.

Though the moments of functions that are defined as sums over independent loop contributions can be computed efficiently, it is unknown whether the same holds for functions with more complex definitions. One such function that has traditionally been used for statistics on RNA secondary structure [12] is the *order* of a secondary structure which refers to the nesting depth of multibranching loops.

6 URL to CODE

Software for partition function computation and a range of related problems is available from www.bioinfo.rpi.edu/applications/hybrid/download.php and www.tbi.univie.ac.at/~ivo/RNA/. Software including a restricted class of structures with pseudoknots [4] is available at www.nupack.org.

7 CROSS REFERENCES

RNA Secondary Structure Prediction by Minimum Free Energy, RNA Secondary Structure Prediction Including Pseudoknots, Inside-Outside Algorithms for Stochastic Context Free Grammars.

8 RECOMMENDED READING

- [1] S. BERNHART, I. L. HOFACKER, AND P. STADLER, *Local RNA base pairing probabilities in large sequences*, *Bioinformatics*, 22 (2006), pp. 614–615.
- [2] S. H. BERNHART, H. TAHER, U. MÜCKSTEIN, C. FLAMM, P. F. STADLER, AND I. L. HOFACKER, *Partition function and base pairing probabilities of RNA heterodimers*, *Algorithms for Molecular Biology*, 1 (2006), p. 3.
- [3] Y. DING, C. Y. CHAN, AND C. E. LAWRENCE, *RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble*, *RNA*, 11 (2005), pp. 1157–1166.
- [4] R. M. DIRKS AND N. A. PIERCE, *A partition function algorithm for nucleic acid secondary structure including pseudoknots*, *Journal of Computational Chemistry*, 24 (2003), pp. 1664–1677.

- [5] I. L. HOFACKER AND P. F. STADLER, *Memory efficient folding algorithms for circular RNA secondary structures*, Bioinformatics, 22 (2006), pp. 1172–1176.
- [6] R. B. LYNGSØ, M. ZUKER, AND C. N. S. PEDERSEN, *Fast evaluation of internal loops in RNA secondary structure prediction*, Bioinformatics, 15 (1999), pp. 440–445.
- [7] J. S. MCCASKILL, *The equilibrium partition function and base pair binding probabilities for RNA secondary structure*, Biopolymers, 29 (1990), pp. 1105–1119.
- [8] I. MIKLÓS, I. M. MEYER, AND B. NAGY, *Moments of the boltzmann distribution for RNA secondary structures*, Bulletin of Mathematical Biology, 67 (2005), pp. 1031–1047.
- [9] A. Y. OGURTSOV, S. A. SHABALINA, A. S. KONDRASHOV, AND M. A. ROYTBURG, *Analysis of internal loops within the RNA secondary structure in almost quadratic time*, Bioinformatics, 22 (2006), pp. 1317–1324.
- [10] I. TINOCO, P. N. BORER, B. DENGLER, M. D. LEVINE, O. C. UHLENBECK, D. M. CROTHERS, AND J. GRALLA, *Improved estimation of secondary structure in ribonucleic acids*, Nature New Biology, 246 (1973), pp. 40–41.
- [11] I. TINOCO, O. C. UHLENBECK, AND M. D. LEVINE, *Estimation of secondary structure in ribonucleic acids*, Nature, 230 (1971), pp. 362–367.
- [12] M. S. WATERMAN, *Secondary structure of single-stranded nucleic acids*, Advances in Mathematics, Supplementary Studies, 1 (1978), pp. 167–212.