

RNA Secondary Structure Prediction by Minimum Free Energy (2006; Ogurtsov, Shabalina, Kondrashov, Roytberg)

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

INDEX TERMS: RNA secondary structure prediction, Gibbs free energy, base pairs, sparse dynamic programming.

1 SYNONYMS

Also known as *RNA Folding*.

2 PROBLEM DEFINITION

This problem is concerned with predicting the set of base pairs formed in the native structure of an RNA molecule. The main motivation stems from structure being crucial for function and the growing appreciation of the importance of RNA molecules in biological processes. Base pairing is the single most important factor determining structure formation. Knowledge of the secondary structure alone also provides information about stretches of unpaired bases that are likely candidates for active sites. Early work [7] focused on finding structures maximising the number of base pairs. With the work of Zuker and Stiegler [17] focus shifted to energy minimisation in a model approximating the Gibbs free energy of structures.

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.

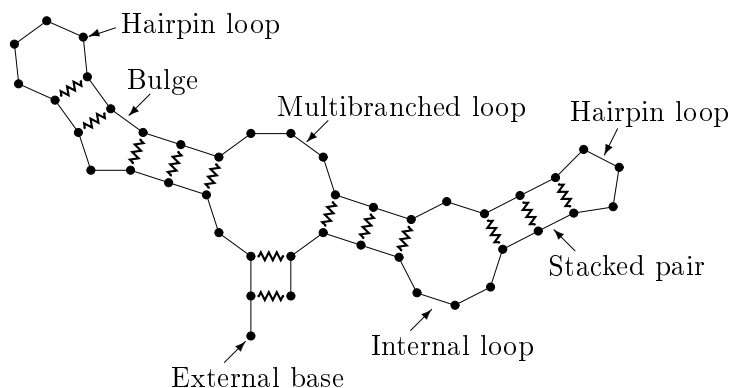


Figure 1: A hypothetical RNA structure illustrating the different loop types. Bases are represented by circles, the RNA backbone by straight lines, and base pairs by zigzagged lines.

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

$$\Delta G(\mathcal{S}) = \sum_{i \cdot j \in \mathcal{S}} \Delta G(\ell_{i,j}) \quad (1)$$

where $\Delta G(\ell_{i,j})$ is the free energy contribution from the loop closed by $i \cdot j$.

Problem 1 (Minimum Free Energy Structure).

INPUT: RNA sequence s and specification of ΔG for all loops.

OUTPUT: $\arg \min_{\mathcal{S}} \{\Delta G(\mathcal{S}) \mid \mathcal{S} \text{ secondary structure for } s\}$

3 KEY RESULTS

Solutions are based on using dynamic programming to solve the general recursion

$$V[i, j] = \min_{k \geq 0; i < i_1 < j_1 < \dots < i_k < j_k < j} \{ \Delta G(\ell_{i,j; i_1 \cdot j_1, \dots, i_k \cdot j_k}) + \sum_{l=1}^k V[i_l, j_l] \}$$

$$W[i] = \min \{ W[i-1], \min_{0 < k < i} \{ W[k-1] + V[k, i] \} \}$$

where $\Delta G(\ell_{i \cdot j; i_1 \cdot j_1, \dots, i_k \cdot j_k})$ is the free energy of the loop closed by $i \cdot j$ and interior base pairs $i_1 \cdot j_1, \dots, i_k \cdot j_k$ and with initial condition $W[0] = 0$. In the following it is assumed that all loop energies can be computed in time $O(1)$.

Theorem 1. *If the free energy of multibranching loops is a sum of*

- *an affine function of the number of interior base pairs and unpaired bases*
- *contributions for each base pair from stacking with either neighbouring unpaired bases in the loop or with a neighbouring base pair in the loop, whichever is more favourable,*

a minimum free energy structure can be computed in time $O(|s|^4)$ and space $O(|s|^2)$ [17].

With these assumptions the time required to handle the multibranching loop parts of the recursion reduces to $O(|s|^3)$. Hence handling the $O(|s|^4)$ possible internal loops becomes the bottleneck.

Theorem 2. *If furthermore the free energy of internal loops is a sum of*

- *a function of the total size of the loop, i.e. the number of unpaired bases in the loop,*
- *a function of the asymmetry of the loop, i.e. the difference in number of unpaired bases on the two sides of the loop,*
- *contributions from the closing and interior base pairs stacking with the neighbouring unpaired bases in the loop,*

a minimum free energy structure can be computed in time $O(|s|^3)$ and space $O(|s|^2)$ [5].

Under these assumptions the time required to handle internal loops reduces to $O(|s|^3)$. With further assumptions on the free energy contributions of internal loops this can be reduced even further, again making the handling of multibranching loops the bottleneck of the computation.

Theorem 3. *If furthermore the size dependency is concave and the asymmetry dependency is constant for all but $O(1)$ values, a multibranching loop free minimum free energy structure can be computed in time $O(|s|^2 \log^2 |s|)$ and space $O(|s|^2)$ [8].*

The above assumptions are all based on the nature of current loop energies [6]. These energies have to a large part been developed without consideration of computational expediency and parameters determined experimentally, although understanding of the precise behaviour of larger loops is limited. For multibranching loops some theoretical considerations [4] would suggest that a logarithmic dependency would be more appropriate.

Theorem 4. *If the restriction on the dependency on number of interior base pairs and unpaired bases in Theorem 1 is weakened to any function that depends only on the number of interior base pairs, the number of unpaired bases, or the total number of bases in the loop, a minimum free energy structure can be computed in time $O(n^4)$ and space $O(n^3)$ [13].*

Theorem 5. *All the above theorems can be modified to compute a data structure that for any $1 \leq i < j \leq |s|$ allows us to compute the minimum free energy of any structure containing $i \cdot j$ in time $O(1)$ [15].*

4 APPLICATIONS

Naturally the key application of these algorithms are for predicting the secondary structure of RNA molecules. This holds in particular for sequences with no homologues with common structure, e.g. functional analysis based on mutational effects and to some extent analysis of RNA aptamers. With access to structurally conserved homologues prediction accuracy is significantly improved by incorporating comparative information [2].

Incorporating comparative information seems to be crucial when using secondary structure prediction as the basis of RNA gene finding. As it turns out, the minimum free energy of known RNA genes is not sufficiently different from the minimum free energy of comparable random sequences to reliably separate the two [9, 14]. However, minimum free energy calculations is at the core of one successful comparative RNA gene finder [12].

5 OPEN PROBLEMS

Most current research is focused on refinement of the energy parametrisation. The limiting factor of sequence lengths for which secondary structure prediction by the methods described here is still feasible is adequacy of the nearest neighbour approximation rather than computation time and space. Still improvements on time and space complexities are useful as biosequence analyses are invariably used in genome scans. In particular improvements on Theorem 4, possibly for dependencies restricted to be logarithmic or concave, would allow for more advanced scoring of multibranched loops. A more esoteric open problem is to establish the complexity of computing the minimum free energy under the general formulation of Eq. (1), with no restrictions on loop energies except that they are computable in time polynomial in $|s|$.

6 EXPERIMENTAL RESULTS

With the release of the most recent energy parameters [6] secondary structure prediction by finding a minimum free energy structure was found to recover approximately 73% of the base pairs in a benchmark data set of RNA sequences with known secondary structure. Another independent assessment [1] put the recovery percentage somewhat lower at around 56%. This discrepancy is discussed and explained in [1].

7 DATA SETS

Families of homologous RNA sequences aligned and annotated with secondary structure are available from the Rfam data base at www.sanger.ac.uk/Software/Rfam/. Three dimensional structures are available from the Nucleic Acid Database at ndbserver.rutgers.edu/. An extensive list of this and other data bases is available at www.imb-jena.de/RNA.html.

8 URL to CODE

Software for RNA folding 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 implementing the efficient handling of internal loops of [8] is available from ftp.ncbi.nlm.nih.gov/pub/ogurtsov/Afold.

9 CROSS REFERENCES

RNA Secondary Structure Prediction Including Pseudoknots, RNA Secondary Structure Distributions, (Stochastic) Context Free Grammar Parsing.

10 RECOMMENDED READING

- [1] R. DOWELL AND S. R. EDDY, *Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction*, BMC Bioinformatics, 5 (2004), p. 71.
- [2] P. P. GARDNER AND R. GIEGERICH, *A comprehensive comparison of comparative RNA structure prediction approaches*, BMC Bioinformatics, 30 (2004), p. 140.
- [3] I. L. HOFACKER AND P. F. STADLER, *Memory efficient folding algorithms for circular RNA secondary structures*, Bioinformatics, 22 (2006), pp. 1172–1176.
- [4] H. JACOBSON AND W. H. STOCKMAYER, *Intramolecular reaction in polycondensations. I. the theory of linear systems.*, Journal of Chemical Physics, 18 (1950), pp. 1600–1606.
- [5] 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.
- [6] D. H. MATHEWS, J. SABINA, M. ZUKER, AND D. H. TURNER, *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure*, Journal of Molecular Biology, 288 (1999), pp. 911–940.
- [7] R. NUSSINOV AND A. B. JACOBSON, *Fast algorithm for predicting the secondary structure of single-stranded RNA*, Proceedings of the National Academy of Sciences of the United States of America, 77 (1980), pp. 6309–6313.
- [8] 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.
- [9] E. RIVAS AND S. R. EDDY, *Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs*, Bioinformatics, 16 (2000), pp. 583–605.
- [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] S. WASHIETL, I. L. HOFACKER, AND P. F. STADLER, *Fast and reliable prediction of noncoding RNA*, Proceedings of the National Academy of Sciences of the United States of America, 102 (2005), pp. 2454–59.
- [13] M. S. WATERMAN AND T. F. SMITH, *Rapid dynamic programming methods for RNA secondary structure*, Advances in Applied Mathematics, 7 (1986), pp. 455–464.
- [14] C. WORKMAN AND A. KROGH, *No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution*, Nucleic Acids Research, 27 (1999), pp. 4816–4822.

- [15] M. ZUKER, *On finding all suboptimal foldings of an RNA molecule*, Science, 244 (1989), pp. 48–52.
- [16] M. ZUKER, *Calculating nucleic acid secondary structure*, Current Opinion in Structural Biology, 10 (2000), pp. 303–310.
- [17] M. ZUKER AND P. STIEGLER, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information*, Nucleic Acids Research, 9 (1981), pp. 133–148.