

ILM: A Web Server for Predicting RNA Secondary Structures with Pseudoknots

Jianhua Ruan¹, Gary D. Stormo^{2,1} and Weixiong Zhang^{1,2,*}

¹Department of Computer Science and Engineering and ²Department of Genetics
Washington University in St. Louis, St. Louis, MO 63130, USA
jruan@cse.wustl.edu, stormo@ural.wustl.edu, zhang@cse.wustl.edu

Abstract

The ILM web server provides a web interface to two algorithms, iterated loop matching, and maximum weighted matching, for the analysis of RNA secondary structures with pseudoknots. The algorithms can utilize either thermodynamic or comparative information or both, thus is able to predict for both aligned and individual sequences. Furthermore, the algorithms allow pseudoknots to be predicted efficiently. Several output formats compatible with various RNA structure visualization tools are supported. The service can be accessed at <http://cheetah0.cs.wustl.edu/RNA/>.

Introduction

RNA molecules play many important regulatory, catalytic and structural roles in the cell. A complete understanding of the functions of RNA molecules requires knowledge of their 3D structures. Since it is often difficult to obtain spectrum data for large RNA molecules to inspect their structures, reliable prediction of RNA structures from their primary sequences is highly desirable.

Many computational methods have been developed for RNA secondary structure predictions. Thermodynamic approaches [25, 11] compute the optimal secondary structure for a sin-

gle RNA sequence with globally minimal free energy, and have been successful for relatively short RNAs. When a number of aligned homologous sequences are available, comparative approaches [5, 8, 7, 1] are more reliable than thermodynamic approaches, and have been used to establish the structures of most known RNA families. Recently several methods [14, 13, 10] have combined the advantages of thermodynamic and comparative methods. By taking both thermodynamic stability and sequence covariance into consideration, these methods are able to produce much better prediction accuracies.

On the other hand, relatively few work has been done on predicting pseudoknotted RNA secondary structures. Pseudoknots are important RNA structures and often have important functional roles [6]. However, optimally predicting pseudoknots in RNA secondary structures is difficult and computationally very expensive [16, 19, 22, 15, 2]. A graph algorithm, maximum weighted matching (MWM), has been used as a practical solution for pseudoknot prediction [4, 21, 17]. Although efficient, the prediction accuracy of MWM is low when the number of homologous sequences is small.

Recently, we have developed an algorithm, iterated loop matching (ILM), for predicting pseudoknotted RNA secondary structures [20]. The method can utilize either thermodynamic or com-

parative information or both, thus can be applied to both aligned and individual sequences. Our experiments have shown that the method is accurate and efficient. The software, however, can only be executed with Unix commands and is complex for biologists to use, let alone to understand and set various parameters.

To address the high demand of a service for predicting RNA secondary structures with pseudoknots, we have developed an easy-to-use web interface that provides most of the features of ILM. We have also provided an option for users to choose the MWM algorithm and compare their results.

Several web servers for RNA secondary structure prediction have been introduced in last year's special issue, including MFold [24], Pfold [3], Vienna RNA package [9] and GPRM [12]. ILM differs from the first three in that it supports pseudoknots. GPRM is designed to find common secondary structure elements in a set of homologous RNA sequences. It cannot be applied to a single RNA sequence or a small dataset, e.g., a family of fewer than 10 sequences.

Overview of the service

The technical details of the ILM and MWM algorithms appear in their original publications [20, 21]. Here we only highlight the basic steps of the web service (Figure 1) and the parameters that need user intervention. The first step after reading in the RNA sequences is to generate a scoring matrix, which describes the likelihood of each base-pair. The method to generate scoring matrices includes mutual information, helix plot, extended helix plot and their combinations [20]. In the second step, the ILM or MWM algorithm is applied to predict a secondary structure with pseudoknots, if exists. Finally, the output of the structure prediction can be viewed with some external visualization tools.

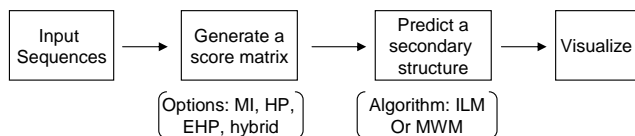


Figure 1. Overview of the service. Abbreviations used: MI, mutual information; HP, helix plot; EHP, extended helix plot; ILM, iterated loop matching; MWM, maximum weighted matching.

Input

The service can be accessed through a web interface (<http://cheetah0.cs.wustl.edu/RNA/>). The web form takes as input a single RNA sequence or a set of aligned RNA sequences in FASTA format. Users can choose to upload a sequence file or “cut and paste” sequences into the web interface directly. Currently, the maximum length of each individual sequence is 2000 bases and the maximum size of a sequence file is 10 kb.

The web server provides optimized default values for all parameters, which may also be adjusted for different needs. Users may choose to calculate a scoring matrix with only mutual information or (extended) helix plot, or vary the relative weights of the two scoring methods. When the number of sequences is small (< 10), mutual information scores are usually not reliable, thus a lower weight of mutual information should be used. In the case that only a single sequence is provided, mutual information scores are zero for all base-pairs. Therefore only helix plot or extended helix plot can be used under this condition, and the provided weights are ignored. In other cases, a weight ratio of 1:1 is suggested by the server. Users may also select to use either the ILM or MWM algorithm for the prediction and customize parameters for each algorithm, such as minimum helix length and minimum loop length (see text on the website for help on the meaning and suggested value of each parameter).

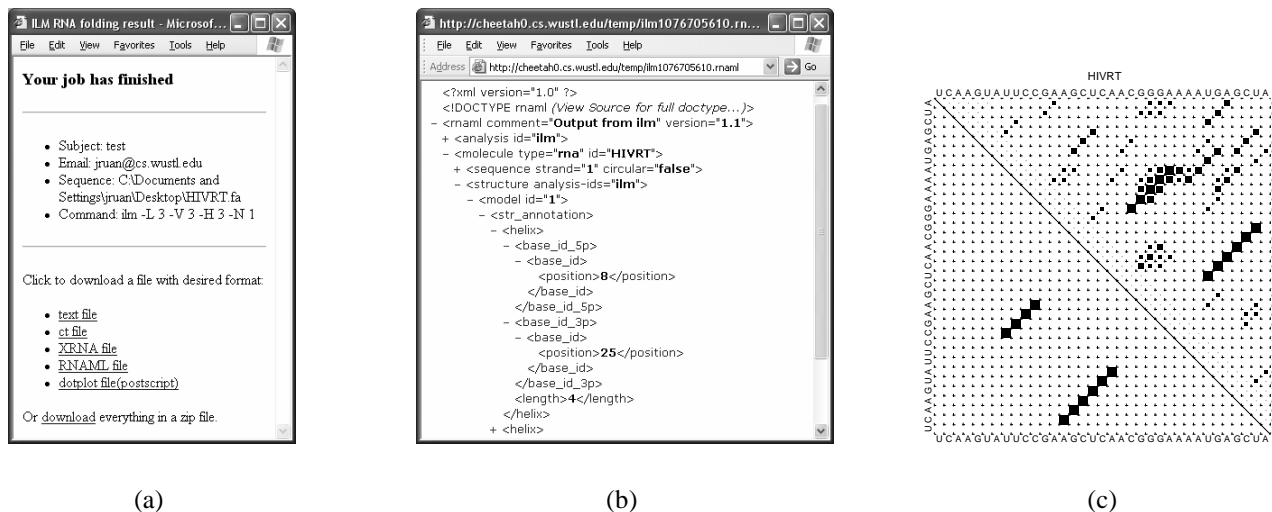


Figure 2. Example output generated by the ILM web server. a), Result page. b), RNAML file. c), Dot plot.

Output

Results for small tasks, i.e., sequence lengths less than 300 bases, will be presented immediately, while a link to the result page for large tasks will be emailed to the user.

The output is self documented (Figure 2). The first part of an output includes user information, dataset name, and parameters specified by the users for reference. The remaining output is the result of a prediction in various formats, including a raw text file depicting the pairing partner for each base. An automatic drawing of RNA secondary structures with pseudoknots is notoriously difficult and many existing visualization tools handle this in a user interactive way. We thus do not attempt to provide a graphic presentation of the secondary structures on our web site. Instead, we generate the output in several formats that can be imported into various RNA secondary structure visualization tools directly. Currently supported formats include .ct file, .xrna file and .rnaml file. The format of .ct files is compatible with several packages, including the RNAviz program [18], which we suggest for drawing pseudoknotted RNA secondary structures. A .xrna file contains a primary sequence and helix

descriptions that can be separately copy-pasted into the XRNA program (<http://rna.ucsc.edu/rnacenter/xrna/xrna.html>). RNAML format has been proposed as a standard for exchanging RNA sequence and structure information between programs [23]. Our RNAML syntax is compatible with the DTD (Document Type Definition) version 1.1 (<http://www-lbit.iro.umontreal.ca/rnaml/current/rnaml.dtd>). Finally, together with the structures, a dot plot is provided in postscript format. This allows users to view a scoring matrix and a predicted structure at the same time. The actual scores are embedded in the postscript file and can be parsed with computer programs.

Implementation

The web service is implemented with static HTML pages and dynamic CGI scripts implemented in PERL. The programs ILM and MWM are implemented in ANSI C. The server is currently running on a machine with dual AMD Athlon 1.6 GHz CPUs and 2Gb of RAM, running Redhat Linux version 2.4.18 and Apache web server. In the future we plan to use a batch

queuing system to distribute large tasks to other machines.

Acknowledgements

This research was supported in part by NSF grants IIS-0196057 and ITR/EIA-011361 8. GDS was supported by NIH grant HG00249. JR thanks Ivo Hofacker for providing the dotplot routine and Mark Bober for setting up the web server.

References

- [1] V. Akmaev, S. Kelley, and G. Stormo. A phylogenetic approach to RNA structure prediction. In *Proc Int Conf Intell Syst Mol Biol*, pages 10–7. AAAI Press, 1999.
- [2] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1-3):45–62, 2000.
- [3] K. B and H. J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, 31(13):3423–8, 2003.
- [4] R. Cary and G. Stormo. Graph-theoretic approach to RNA modeling using comparative data. *Proc Int Conf Intell Syst Mol Biol*, 3:75–80, 1995.
- [5] D. Chiu and T. Kolodziejczak. Inferring consensus structure from nucleic acid sequences. *Comput Appl Biosci*, 7(3):347–52, Jul 1991.
- [6] E. Dam, K. Pleij, and D. Draper. Structural and functional aspects of RNA pseudoknots. *Biochemistry*, 31(47):11665–1176, Dec 1992.
- [7] B. Gulko and D. Haussler. Using multiple alignments and phylogenetic trees to detect RNA secondary structure. In *Proc Pac Symp Biocomput*, pages 350–67, 1996.
- [8] R. Gutell, A. Power, G. Hertz, E. Putz, and G. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*, 20(21):5785–595, Nov 1992.
- [9] I. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31(13):3429–31, 2003.
- [10] I. Hofacker, M. Fekete, and P. Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–166, Jun 2002.
- [11] I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.
- [12] Y. Hu. GPRM: A genetic programming approach to finding common RNA secondary structure elements. *Nucleic Acids Res.*, 31(13):3446–9, 2003.
- [13] V. Juan and C. Wilson. RNA secondary structure prediction based on free energy and phylogenetic analysis. *J Mol Biol*, 289(4):935–47, Jun 1999.
- [14] R. Luck, S. Graf, and G. Steger. ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res*, 27(21):4208–417, Nov 1999.
- [15] R. Lyngso and C. Pedersen. Pseudoknots in RNA secondary structures. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 201–209. ACM Press, 2000.
- [16] R. Lyngso and C. Pedersen. RNA pseudoknot prediction in energy-based models. *J Comput Biol*, 7(3-4):409–27, 2000.
- [17] R. Page. Comparative analysis of secondary structure of insect mitochondrial small subunit ribosomal RNA using maximum weighted matching. *Nucleic Acids Res.*, 28(20):3839–45, 2000.
- [18] P. D. Rijk, J. Wuyts, and R. D. Wachter. Rnaviz 2: an improved representation of RNA secondary structure. *Bioinformatics*, 19(2):299–300, 2003.
- [19] E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol*, 285(5):2053–2068, Feb 1999.
- [20] J. Ruan, G. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1):58–66, 2004.
- [21] J. Tabaska, R. Cary, H. Gabow, and G. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–69, 1998.
- [22] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori. Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science*, 210(2):277–303, 1999.

- [23] A. Waugh, P. Gendron, R. Altman, J. Brown, D. Case, D. Gautheret, S. Harvey, N. Leontis, J. Westbrook, E. Westhof, M. Zuker, and F. Major. RNAML: a standard syntax for exchanging RNA information. *RNA*, 8(6):707–17, 2002.
- [24] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–15, 2003.
- [25] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–48, Jan 1981.