Bioinformatics Advance Access published March 5, 2009

RNATOPS-W: A Web Server for RNA Structure Searches of Genomes

Yingfeng Wang [1], Zhibin Huang [1], Yong Wu [1], Russell L. Malmberg [2,3], and Liming Cai [1,3,*]

[1]Department of Computer Science, University of Georgia, Athens, GA 30602.
[2]Department of Plant Biology, University of Georgia, Athens, GA 30602
[3]Institute of Bioinformatics, University of Georgia, Athens, GA 30602.

Associate Editor: Prof.Ivo Hofacker

ABSTRACT

Summary: RNATOPS-W is a web server to search sequences for RNA secondary structures including pseudoknots. The server accepts an annotated RNA multiple structural alignment as a structural profile and genomic or other sequences to search. It is built upon RNATOPS (Huang *et al.,* 2008), a command line C++ software package for the same purpose, in which filters to speed up search are manually selected. RNATOPS-W improves upon RNATOPS by adding the function of automatic selection of an hidden Markov model (HMM) filter and also a friendly user interface for selection of a substructure filter by the user. In addition, RNATOPS-W complements existing RNA secondary structure search web servers that either use built-in structure profiles or are not able to detect pseudoknots. RNATOPS-W inherits the efficiency of RNATOPS in detecting large, complex RNA structures.

Availability: The web server RNATOPS-W is available at website `www.uga.edu/RNA-Informatics/?f=software&p=RNATOPS-w`. The underlying search program RNATOPS can be downloaded at `www.uga.edu/RNA-Informatics/?f=software&p=RNATOPS`.

Contact: cai@cs.uga.edu

Supplementary Material: The online SUPPLEMENTARY MATERIAL contains additional experimental data.

1 INTRODUCTION

Searching genomes using computational methods has become important for prediction and annotation of non-coding RNAs (Lowe and Eddy, 1997; Rivas and Eddy, 2001; Rivas *et al.,* 2001; Washietl *et al.,* 2005; Hofacker, 2006; Griffiths-Jones, 2007) Profile based RNA structure search is an often used approach for this purpose. However, for large, complex RNA molecules such as those containing pseudoknots, the search task has proven difficult. Typically, some existing web servers for RNA structure search consider pseudoknots whose profiled are predefined and fixed with the search program (Zhang *et al.*, 2005); other available programs allow user-defined profiles but are limited to pseudoknot-free structures (Nawrocki and Eddy, 2007; Klein and Eddy, 2003; Griffiths-Jones, 2003). Web servers with the capability to accept user-defined profiles for arbitrary pseudoknot structure searches are not available. This is due to the lack of appropriate RNA pseudoknot models that can permit efficient algorithms for structure-sequence alignment, a bottleneck task. Search programs can usually be speeded up with filtering methods that can quickly remove genome segments unlikely to contain the desired pattern in the profile (Bafna and Zhang, 2004; Zhang *et al.,* 2005; Lowe and Eddy, 1997; Weinberg and Ruzzo, 2006), but even with a significant speed-up (e.g., with a 99% genome reduction), searching for an complex RNA structure with a pseudoknot may still take hours, if not days, on a typical bacterial or yeast genome.

Our previous work (Song, Y. *et al.*, 2005) introduced a graph-theoretic modeling method for profiling RNA secondary structures including pseudoknots. With this model, we were able to design a very efficient structure-sequence alignment algorithm, ideal for RNA pseudoknot search on genomes, and implemented it in an RNA structure search program called RNATOPS (Huang *et al.*, 2008). One advantage of RNATOPS is its high efficiency searching for large RNA or complex structures including pseudoknots, while maintaining accuracy comparable to other search programs that are only capable of detecting pseudoknot-free structures. To further speed up searches, RNATOPS also executes the whole structure search on filtering results. However, filters (i.e., subsequence or substructure profiles) can only be manually selected. This paper presents a web server version of RNATOPS, called RNATOPS-W with a new built-in function for automatic HMM filter selection. The web server also allows an interactive selection of any substructure as a filter through a user-friendly interface.

2 PROGRAM FEATURES

This section presents the filtering functions of the web server RNATOPS-W and its interface features. We refer the reader to our previous work (Huang *et al.*, 2008; Song *et al.*, 2005) for detailed discussions on the search methods and algorithms used by RNATOPS.

2.1 Filtering Method

RNATOPS-W incorporates a function of automatic HMM filter selection; the selected filter is used to speed up the search program. The filter selection chooses a conserved region as an HMM filter from the given RNA structural profile (a set of structurally aligned RNA sequences). Our filter selection method was built from two previous approaches used to identify conserved amino acids in protein sequences (Capra and Singh, 2007; Song, B. *et al.*, 2005);

*To whom all correspondence should be addressed.

it replaces the overall amino acid distribution in the BLOSUM62 alignments with the nucleotide distribution in the given RNA alignment. In addition, our method ignores columns containing more than 50% gaps instead of the 30% used in the first method (Capra and Singh, 2007). Scores are assigned to columns based on their degree of conservation, with higher scores for more conserved columns. Based on these scores, an automatic peak detection algorithm (Song, B. *et al.,* 2005) is then applied to find a conserved region. In selecting such a region, an 'ignored' column is also re-considered if both its neighboring columns are considered for the conserved region. The selected conserved region is then used to produce a profile HMM filter.

We conducted two types of experiments to test the performance of our filter selection method. On synthetic genomes generated by embedding real RNA sequences taken from the profile into randomly generated nucleotides, with the automatically selected HMM filter, RNATOPS-W never missed a real RNA sequence. In the search time test on real genomes, automatically selected HMM filters drastically speeded up the whole structure search (by at least three orders of magnitude) in contrast to randomly generated HMM filters, which found too many false positive filter hits to yield an efficient whole structure search. We have also conducted tests on the HMM filters constructed directly from the full length alignment of structure profiles and compared their performance with our automatically generated filters. The experiments indicate that with (sequentially) conserved RNA profiles, HMM filters generated from the full length alignment have a lower false positive rate than automatically generated HMM filters. On sequentially less conserved RNA profiles, the latter has a higher accuracy. Both filters are sensitive. However, in either case of RNA profiles, searching with the filters selected by our algorithm are about one magnitude faster than searching with a filter from the full length alignment. These test results also indicate that HMM filters automatically generated by RNATOPS-W can maintain both efficiency and the accuracy. Test results and comparisons for automatically generated filters, random filters, full-length alignment filters, and manually selected filters are shown in the SUPPLEMENTARY MATERIAL .

2.2 Interface Features

To use RNATOPS-W for RNA structure search, the user is asked to submit an RNA structure profile (i.e., a set of structurally aligned training RNAs) in pasta format (Huang *et al.,*, 2008) and target genomes in fasta format. These data can be in either a file or an input text box to be uploaded in the start page. By default, RNATOPS-W automatically selects an HMM filter for the given structure profile. The user can also opt to select manually his/her own filter, by specifying the beginning and ending regions of any consecutive substructure from the given structure profile. After the submission of the input and an filter option, the server searches the target genomes with the filter and then searches the filtered hits for whole structure matches. Each search request is given a ticket number with which the user can retrieve later a search result file from a provided link or from the start page.

For each search request, the result file contains information for all search hits that "match" the structure profile. For each hit, the following information is produced: the name of the genome containing the hit, the hit sequence, its position in the genome, the score of the hit sequence, the fold conforming to the structure profile, and the structural alignment between the hit sequence and the structure profile. The output also contains the parameter settings for each whole search request and the total time used in the search.

Additional options are provided for the user to redefine parameters pertinent to the search algorithm to achieve a desired search accuracy. The user, instead of choosing the "All default" option, can select "Adjust parameters". These parameters mostly concern setting priors for stochastic modeling of individual stems and loops in the structure profile and improving the qualities of candidates found for individual stems.

RNATOPS-W provides users a friendly web-interface to perform searches of genomes for RNAs on the basis of their structural profile, including pseudoknots. It adds functionality by automated selection of filters to speed up the search.

ACKNOWLEDGEMENT

This work was supported in part by NIH BISTI grant No: R01GM072080-01A1. A part of the server interface was implemented with help from Mark Wilson.

REFERENCES

Bafna, V. and Zhang, S. (2004) FastR: fast database search tool for non-coding RNA. *Proceedings of the 3rd IEEE Computational Systems Bioinformatics Conference*, 52-61.

Capra, J.A. and Singh, M. (2007) Predicting functionally important residues from sequence conservation, *Bioinformatics*, **15**, 1875-1882.

Griffiths-Jones, S. (2007) Annotating Noncoding RNA Genes, *Annual Review of Genomics and Human Genetics*, **8**, 279-298.

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

Hofacker, I.L. (2006) RNAs everywhere: genom-wide annotation of structured RNAs, *Genome Informatics*, **17(2)**, 281-282.

Huang, Z., Wu, Y., Robertson, J., Liang, F., Malmberg, R.L., and Cai, L. (2008) Fast and accurate search for non-coding RNA pseudoknot structures in genomes, *Bioinforamtics*, **24**, 2281-2287.

Klein, R.J. and Eddy, S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences, *BMC Bioinformatics*, **4:44**.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Research*, **25**, 955-964.

Nawrocki, E.P. and Eddy, S.R. (2007) Query-dependent banding (QDB) for faster RNA similarity searches, *PLoS Computational Biology*, **3:e56**.

Rivas, E. and Eddy, S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis, *BMC Bioinformatics*, **2:8**.

Rivas, E., Klein, R.J., Jones, T.A. and Eddy, S.R. (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics, *Current Biology*, **11**, 1369-1373.

Song, B., Choi J., Chen G., Szymanski, J., Zhang G., Tung A.K.H., Kang, J., Kim, S. and Yang, J. (2005) ARCS: an aggregated related column scoring scheme for aligned sequences, *Bioinformatics*, **19**, 2326-2332.

Song, Y., Liu, C., Pan, F., Malmberg, R.L. and Cai, L. (2005) Tree decomposition based fast searching for RNA structures with and without pseudoknots, *Proceedings of IEEE Computational Systems Bioinformatics Conference*, 223-234.

Washietl, S., Hofacker, I.L., and Stadler, P.F. Fast and reliable prediction of noncoding RNAs, *Proceedings of National Academy of Sciences*, **102**, 2454-2459.

Weinberg, Z. and Ruzzo, W.L. (2006) Sequence-based heuristics for faster annotation of non-coding RNA families, *Bioinformatics*, **22**, 35-39.

Zhang, S., Hass, B., Eskin, E., Bafna, V. (2005) Searching genomes for noncoding RNA using FastR, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **2**, 366-379.