Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W325–W329***
doi:10.1093/nar/gkm303

taveRNA: a web suite for RNA algorithms and applications

Cagri Aksay[1], Raheleh Salari[1], Emre Karakoc[1], Can Alkan[2] and S. Cenk Sahinalp[1,*]

[1]Lab for Computational Biology, SFU, Canada and [2]Department of Genome Sciences, University of Washington

Received January 31, 2007; Revised April 3, 2007; Accepted April 14, 2007

ABSTRACT

We present taveRNA, a web server package that hosts three RNA web services: alteRNA, inteRNA and pRuNA. alteRNA is a new alternative for RNA secondary structure prediction. It is based on a dynamic programming solution that minimizes the sum of energy density and free energy of an RNA structure. inteRNA is the first RNA–RNA interaction structure prediction web service. It also employs a dynamic programming algorithm to minimize the free energy of the resulting joint structure of the two interacting RNAs. Lastly, pRuNA is an efficient database pruning service; which given a query RNA, eliminates a significant portion of an ncRNA database and returns only a few ncRNAs as potential regulators. taveRNA is available at http://compbio.cs.sfu.ca/taverna.

INTRODUCTION

Until recently RNA was thought to have only two functions: (i) primarily as an information transmitter between DNA and proteins in the form of a messenger RNA (mRNA) and (ii) as a catalyser or an information decoder in protein synthesis in the form of a ribosomal RNA (rRNA) or a transfer RNA (tRNA). The discovery that RNA molecules can regulate gene expression completely altered this simplistic picture (1). This important discovery, in particular the discovery of RNA interference (RNAi), the post-transcriptional silencing of gene expression via interactions between mRNAs and their antisense RNAs (by A. Fire and C. Mello) was awarded with the 2006 Nobel Prize for Physiology or Medicine.

Recent studies have revealed that antisense RNAs are only a very small subset of non-coding RNAs (ncRNAs). A large fraction of the genome sequence (up to 10% in the human genome) appears to give rise to RNA transcripts that do not code for proteins (2). The functionality of many such ncRNAs are only scarcely known.

Regulatory ncRNAs that are generally responsible for regulating gene expression exhibit an exact or partial complementarity to their target mRNAs. Their interaction forms a complex that consists of several non-contiguous helical segments which prevent ribosomal access to the target mRNA. Generally, regulatory ncRNAs contain one or more stem loop structures that are (almost) complementary to specific sequences in the target mRNAs. Interaction with a target RNA is either initiated at such a loop structure of the antisense RNA and a loop structure from the target (forming kissing loop pairs) or between a loop structure and a single-stranded segment of the complementary RNA.

As the number of ncRNAs and in particular regulatory RNAs increase it has become of crucial importance to establish software tools that can help identify their functionality. For this purpose we introduce taveRNA, a web-based computational tool set that can help identify structure and functionality of ncRNA molecules. taveRNA involves tools whose algorithmic foundations were developed by Simon Fraser University's Lab for Computational Biology over the past few years. The tools aim to solve the following key problems:

1. RNA secondary structure prediction problem, which asks to compute all basepairs established by a given (nc)RNA sequence.
2. RNA–RNA interaction prediction problem, which asks to predict the joint secondary structure between two given RNA sequences; a joint secondary structure between two RNAs can be represented via the set of internal and external base pairs established between the interacting RNAs.
3. ncRNA search problem which asks to compute all ncRNAs that can establish stable secondary structures (and in principle play a role in the regulation of) with a given query mRNA.

taveRNA involves web interfaces to the algorithmic tools we have developed for each one of the above problems: (1) *alteRNA*, an alternative thermodynamic model-based RNA secondary structure prediction

*To whom correspondence should be addressed. Tel: +1-604-268-7040; Fax: +1-604-291-3045; Email: cenk@cs.sfu.ca

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

program, (2) *inteRNA*, a program for predicting the joint secondary structure of two interacting RNAs and (3) *pRuNA*, a tool for quickly identifying all potential regulatory ncRNAs for a query mRNA.

RNA secondary structure prediction is one of the earliest problems considered in computational biology. Despite a 30-year-effort towards its solution, the best algorithms and software for RNA structure prediction are still far from perfect. The most common approach to the RNA secondary structure prediction problem is the standard thermodynamic approach. This general methodology suggests to compute the secondary structure of an RNA molecule by minimizing the total free energy of its substructures such as *stems*, *loops* and *bulges*. *alteRNA* is an *alternativ*e thermodynamic approach to the RNA secondary structure prediction problem, which aims to minimize a *linear combination of total free energy and total energy density* using the dynamic programming formulation proposed in (3), in contrast to the available alternatives such as Mfold (4), RNAscf (5) and alifold (6), which all employ the standard thermodynamic approach.

As mentioned earlier, there are a number of algorithms for predicting the secondary structure of a *single* RNA molecule. However, until recently no such algorithm existed for reliably predicting the *joint* secondary structure of two interacting RNA molecules, or measuring the stability of such a joint structure. At SFU Lab for Computational Biology, we developed *inteRNA* the first program that aims to compute the joint structure prediction of two given RNA sequences through minimizing their total free energy, which is a function of the topology of the joint structure. The algorithmic foundations of inteRNA were introduced in (7). inteRNA aims to minimize the joint free energy under a number of energy models with growing complexity. Our default energy model is based mostly on stacked pair energies given by Mathews *et al.* (8)—the free energy model employed by Mfold. Because there is very little thermodynamic information on kissing loop sequences in the literature, inteRNA employs the approach used by Rivas and Eddy (9) to differentiate the thermodynamic parameters of 'external' bonds from the 'internal' bonds, i.e. we simply multiply the external parameters with a *weight* slightly smaller than 1. Note that the default algorithmic engine used by inteRNA requires substantial resources in terms of running time and memory which could be prohibitive for predicting joint structure of sufficiently long RNAs. Thus we made available an alternative heuristic engine that aims to capture the joint secondary structure of interacting RNAs. This heuristic approach is based on the observation that the substructures that are available independently in each RNA are mostly preserved in their joint secondary structure.

Regulatory ncRNAs play important roles in controlling gene expression via establishing stable joint structures with target mRNAs and prohibiting ribosomal access. As many novel regulatory RNA classes are being discovered, the determination of the exact functionality of all regulatory ncRNAs is one of the most significant scientific challenges to be tackled in the coming years. Our software tool *pRuNA* is developed for helping to address this challenge: it aims to identify all potential regulatory ncRNAs that can establish stable joint structures with a query mRNA (10). An important component of pRuNA is a sequence filter which eliminates a significant fraction of the available ncRNA collection and retains only the most likely ncRNA candidates for forming a stable joint structure with the query mRNA. We note that the majority of regulatory ncRNAs are only partially complementary to their target mRNA sequences. Typically these RNAs contain at least two short (5–7 nt) motifs on their loop structures that are complementary to the specific single-stranded locations in the target RNAs (11). Thus pRuNA first identifies all pairs of 5-mer motifs from the ncRNA loop sequences that are complementary to a pair of 5-mer motifs in the loop sequences of the query mRNA. The ncRNAs that include such 5-mer motif pairs are then tested against the query mRNA via the use of inteRNA and those which can establish stable joint structures are returned by the program.

alteRNA: AN RNA SECONDARY STRUCTURE PREDICTION TOOL BASED ON AN ALTERNATIVE THERMODYNAMIC APPROACH

Input and output

alteRNA requires a sequence in FASTA format, where the sequence is represented as a string of characters from the alphabet $\Sigma = \{A,C,G,U,T\}$ (the characters are case insensitive). The length of the input RNA sequence is limited by 500 nt. There are two user-specified parameters, σ and b, which will be explained in detail later.

alteRNA outputs the results in three different forms:

1. The predicted RNA structure in dot–parenthesis format. The sequence is given from 5′ to 3′ end, and the structure is given with matching parenthesis denoting a base pair and a dot denoting an unbounded base.
2. The predicted RNA structure in Connect(.ct) file format, which is the standard Mfold output format. The sequence length is given in the first line. For each nucleotide i, there is a line which consists of: the line number (i), the letter denotion of the nucleotide, the predecessor base index ($i-1$), the successor base index ($i+1$), the paired base index (0 if unpaired) and the original base index (i).
3. The graphics files for the predicted secondary structure in Postscript(.ps) and GIF format. These graphs are created using NAVIEW (12) and Mfold/plt2.

The secondary structure predicted by alteRNA for 5sRNA is shown in Figure 1a.

Methods

It has been shown in (3) that delocalizing the thermodynamic cost of forming an RNA substructure can be achieved by consideration of the *energy density* of the

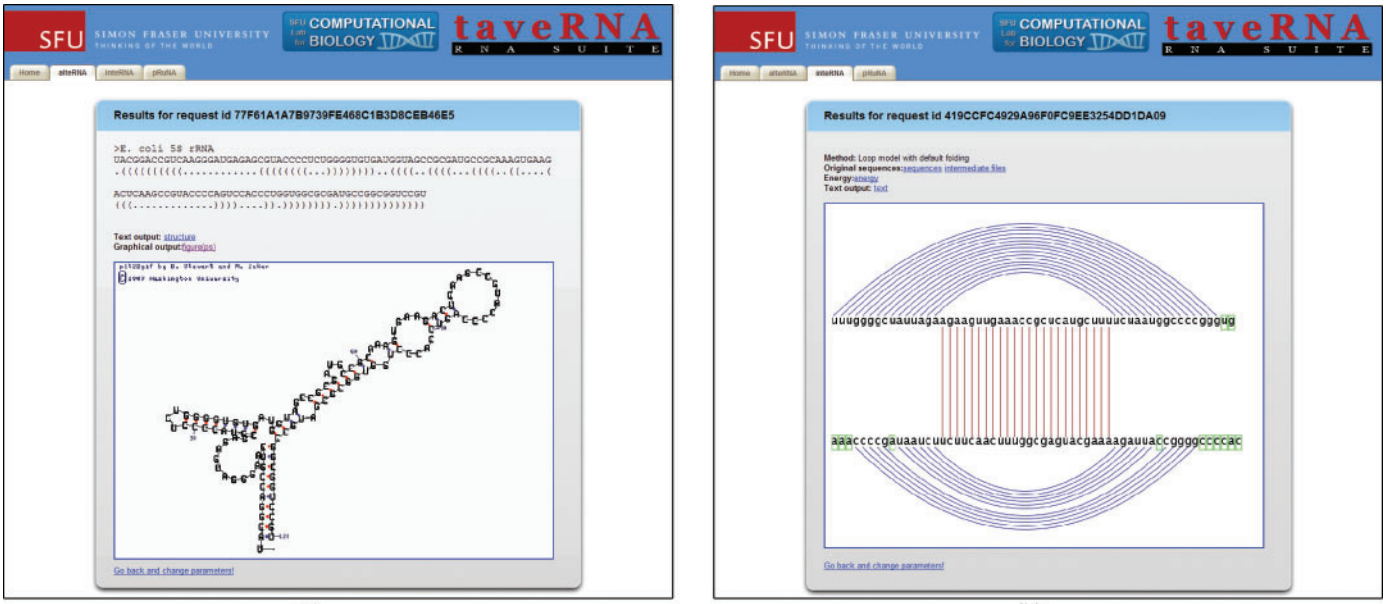

Figure 1. The output pages of taveRNA. (**a**) alteRNA secondary structure prediction for 5sRNA. (**b**) inteRNA joint structure prediction for CopA and CopT.

substructure, which, in return, can improve the secondary structure prediction of several ncRNA families. This alternative thermodynamic approach which we call Densityfold (3) aims to minimize the linear combination of the total free energy and total energy density of all substructures in the RNA molecule. Given an RNA secondary structure, the energy density of a basepair is defined as the free energy of the substructure that starts with the basepair, normalized by the length of the underlying sequence. The energy density of an unpaired base is then defined to be the energy density of the closest basepair that encloses it. Densityfold thus aims to minimize $ED(n) + \sigma \cdot E(n)$, where $ED(n)$ is the *total energy density* of paired and unpaired bases, $E(n)$ is the *total free energy* and n is the length of the RNA sequence. Here σ determines the weight of the contribution of the total free energy in the optimization function. As σ approaches to ∞, the predicted secondary structure gets closer to that implied by the standard thermodynamic approach (employed by Mfold). We have observed that $\sigma = 5$ generally gives the most accurate prediction, in terms of structure edit distance (13), however in some RNA sequences, different σ values can outperform $\sigma = 5$.

We note that the free energy of a multi-branch loop as implied by the standard thermodynamic model is not a linear function and as a result cannot be used in a dynamic programming formulation. Thus we use the same approximate formulation as per Mfold, which, for a given multi-branch loop, L sets:

$$\delta\delta G(L) = a + b \cdot l_s(L) + c \cdot l_d(L) + \delta\delta G_{stack}$$

Here a, b, c are constants in thermodynamic model, l_s is the number of unpaired bases, l_d is the number of base pairs and $\delta\delta G_{stack}$ is the free energy of each branch in the loop. The default setting for b, the coefficient for the number unpaired bases, is 0 in Mfold; this implies that there will be no penalty for the unpaired bases. However, when used with Densityfold, this setting seems to produce large stretches of unpaired bases in multi-branch loops. In order to counter this effect we suggest the user to set b to a small positive value less than 1. alteRNA, our web-based tool for RNA secondary structure prediction via the Densityfold approach allows the user to specify any positive value for σ and b. The default values of σ and b are 5 and 0, respectively and the folding temperature is fixed at 37°C.

inteRNA: A Web-Based Tool For RNA–RNA Interaction Prediction

Input and output

inteRNA requires as an input two RNA sequences in FASTA format. The sequences are represented as strings of characters from the alphabet $\Sigma = \{A,C,G,U,T\}$ (the strings are case insensitive). The length of each input RNA sequence is limited by 150 nt for Stacked Pair Model and by 500 nt for Loop Models. There are three user-specified parameters; *gap penalty*, *maximum subsequence length* and *energy model*; each of these parameters are explained in detail later.

inteRNA reports the results in three different output forms:

1. A text file that contains the base-pair information of the joint secondary structure. Here, each line denotes either a base pair with nucleotide indices or a gap.

A base pair is represented by a quadruple (Sequence, Position, Sequence, Position), where Sequence can be either *S* or *R* (representing the first and the second sequences, respectively) and Position is the base index. Gaps are represented with (*S*,"Gap",*R*,*i*) in sequence *R* and represented with (*S*,*i*,*R*,"Gap") in sequence *S*, where *i* is the index of the free nucleotide.

2. A JPEG image file that can help visualize the joint secondary structure prediction of the input RNAs. The input RNAs are represented by their sequences only. Internal bonds are represented by blue arcs and external bonds are represented by red lines.
3. A text file reporting the calculated free energy of the predicted joint secondary structure.

A sample interaction between CopA and CopT, predicted by inteRNA is shown in Figure 1b.

Method

inteRNA uses a dynamic programming algorithm, which aims to predict the joint secondary structure of two interacting RNA molecules through minimizing their total free energy. In other words, it predicts the secondary structure of both RNA strands and the interaction simultaneously. Since the general joint structure prediction problem is NP-Complete, it is necessary to set some limitations on the structures predicted: inteRNA does not allow any (internal or external) pseudoknots; it does not allow any zig-zag structures either (7).

inteRNA can employ the following two models for approximating the total free energy of the joint structure.

Stacked pair model: This model uses stacking pair energy functions both externally and internally. A *gap penalty* parameter is used to penalize opening a gap. Due to high computational complexity and memory requirements of this model, inteRNA cannot accept inputs with large (>200 nt) sequence lengths with this setting.

Loop model: It has been observed that RNA molecules mostly preserve their independent secondary structure when they interact with other RNA. Interactions thus typically occur between *kissing hairpin loops*. The *loop model* first predicts the secondary structure of each RNA sequence independently using three different approaches: (i) through the standard thermodynamic model (Mfold Loop Model), (ii) through Stacked Pair Model using a single RNA sequence (Default Loop Model) and (iii) through alteRNA (alteRNA Loop Model). It then identifies all independent subsequences of each RNA structure: these are substructures (each implied by a basepair) whose sequence length is less than a specified value. It then computes the joint secondary structure that can be established between each pair of independent subsequences (one subsequence from each RNA) and the free energy of this joint structure. Finally, it finds the set of independent subsequence pairs which can co-exist and minimize the total free energy of the overall secondary structure.

pRuNA: A REGULATORY RNA SEARCH ENGINE

Input and output

pRuNA asks for a query mRNA sequence as well as its corresponding secondary structure in parenthesis format. It returns a list of possible regulating ncRNAs available through the Rfam database (14).

Method

At the heart of pRuNA we have a lookup table of size 1024×1024, which represents every possible pair of 5-mer motifs from the 4-letter RNA alphabet. Each entry in the lookup table maintains a pointer to every occurrence of the respective 5-mer motif pair in the loop sequences of the ncRNA structures in the Rfam database. For querying the index, pRuNA first identifies the loop sequences from the mRNA, based on the given secondary structure information. It then extracts each pair of non-overlapping 5-mer motifs from the mRNA loop sequences and checks the lookup table for its reverse complement. The union of the set of ncRNAs that include each motif pair are retained by pRuNA, whereas the remainder of the ncRNAs in the database are eliminated. We have observed that many of the motif pairs in the lookup table occur very frequently in ncRNA sequences and thus have limited information content. We omitted the 5-mer pairs which occur in more than the number (14) of ncRNAs that include a 'typical' motif pair. This increased our pruning efficiency substantially yet had no effect on the quality of the results: we have tested pRuNA mRNAs that are known to be regulated by an ncRNA available through the NPInter database (15). For all 24 of these mRNAs, pRuNA was able to return the known regulatory ncRNA among the set of potential regulators, while achieving a pruning efficiency of 86% on the average. If the search is restricted to *E. coli* ncRNAs, the pruning efficiency for the 22 mRNAs regulated by these ncRNAs go up to 91% on the average. Further details on pRuNA (including the method and experimental results) can be found in (10).

DISCUSSION

taveRNA is a web application implemented in html/php; the underlying algorithms are all implemented in C/C++. Our server is hosted on a Linux machine with 16 GB RAM and 4 AMD Opteron 2200 processors. On an RNA sequence with length n, the running time of alteRNA is $O(n^4)$ and the memory requirement is $O(n^4)$. On two input RNA sequences S and R, where $|S| = n$ and $|R| = m$, the running time of inteRNA is $O(n^3m^3)$ and the memory requirement is $O(n^2m^2)$. It is possible to improve the running time using the loop model to where k is the maximum independent structure size. The running time of pRuNA is linear with the input sequence length and the result set. Because of the large running-time requirements of inteRNA and alteRNA, taveRNA has an e-mail notification option.

taveRNA has been developed by the Lab for Computational Biology at Simon Fraser University and

it is accessible from http://compbio.cs.sfu.ca/taverna. We acknowledge our funding sources, in particular NSERC, Canada Research Chairs Program, Canada Foundation for Innovation and Michael Smith Foundation for Health Research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by NSERC, CIRC Program, CFI, MITACS and MSFHR.

Conflict of interest statment. None declared.

REFERENCES

1. Novina,C.D. and Sharp,P.A. (2004) The RNAi revolution. *Nature*, **430**, 161–164.
2. Claverie,J. (2005) Fewer genes, more noncoding RNA. *Science*, **309**, 1529–1530.
3. Alkan,C., Karakoc,E., Sahinalp,C., Unrau,P.J., Ebhardt,H.A., Zhang,K. and Buhler,J. (2006) RNA secondary structure prediction via energy density minimization. *Proc. RECOMB*, **LNBI 3909**, 130–142.
4. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
5. Bafna,V., Tang,H. and Zhang,S. (2005) Consensus folding of unaligned RNA sequences revisited. *Proc. RECOMB*, **LNBI 3500**, 172–187.
6. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
7. Alkan,C., Karakoc,E., Nadeau,J.H., Sahinalp,C. and Zhang,K. (2005) RNA-RNA interaction prediction and antisense RNA target search. *Proc. RECOMB*, **LNBI 3500**, 152–171.
8. Mathews,D., Sabina,J., Zuker,M. and Turner,D. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
9. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
10. Aksay,C., Karakoc,E., Ho,C.K., Unrau,P.J., and Sahinalp,C. (2007) ncRNA discovery and functional identification via sequence motifs. Technical Report TR 2007-06, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.
11. Hjalt,T. and Wagner,E.G.H. (1992) The effect of loop size in antisense and target RNAs on the efficiency of antisense RNA control. *Nucleic Acids Res.*, **20**, 6723–6732.
12. Bruccoleri,R. and Heinrich,G. (1988) An improved algorithm for nucleic acid secondary structure display. *Comput. Appl. Biosci.*, **4**, 167–173.
13. Lin,G., Ma,B. and Zhang,K. (2001) Edit distance between two RNA structures. *Proc. RECOMB*, 211–220.
14. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
15. Wu,T., Wang,J., Liu,C., Zhang,Y., Shi,B., Zhu,X., Zhang,Z., Skogerbo,G., Chen,L. *et al.* (2006) NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res.*, **34**, D150–D152.