W320–W324 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm258

pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows

Jens Reeder, Peter Steffen and Robert Giegerich*

Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany

Received January 30, 2007; Revised March 28, 2007; Accepted April 8, 2007

ABSTRACT

RNA pseudoknots are an important structural feature of RNAs, but often neglected in computer predictions for reasons of efficiency. Here, we present the *pknotsRG* Web Server for single sequence RNA secondary structure prediction including pseudoknots. *pknotsRG* employs the newest Turner energy rules for finding the structure of minimal free energy. The algorithm has been improved in several ways recently. First, it has been reimplemented in the C programming language, resulting in a 60-fold increase in speed. Second, all suboptimal foldings up to a user-defined threshold can be enumerated. For large scale analysis, a fast sliding window mode is available. Further improvements of the Web Server are a new output visualization using the *PseudoViewer* Web Service or *RNAmovies* for a movie like animation of several suboptimal foldings.

The tool is available as source code, binary executable, online tool or as Web Service. The latter alternative allows for an easy integration into bioinformatics pipelines. *pknotsRG* is available at the Bielefeld Bioinformatics Server (http://bibiserv.techfak.uni-bielefeld.de/pknotsrg).

INTRODUCTION

RNA pseudoknots play an important role in many biological processes. They build the catalytic core of some ribozymes (1,2) and are an important building block of many structural RNAs. Pseudoknots are involved in telomerase activity (3) and they stimulate efficient programmed -1 ribosomal frameshifting (-1 PRF), a mechanism used by a wide range of RNA viruses to encode two proteins within one genomic region. In a recent study (4) over a thousand of potential -1 PRF signals were detected in the yeast genome. The majority of signals however seem to direct the ribosome to premature termination codons. This suggests a mechanism of post-transcriptional gene regulation through the nonsense-mediated mRNA decay pathway. Further genome wide studies have to elucidate this phenomenon. This and many more applications require fast RNA folding algorithms, that are also capable of folding pseudoknots.

Standard RNA folding programs (5,6) neglect pseudoknots for reasons of efficiency. While the standard methods need time proportional to the cube of the input sequence length, pseudoknot prediction is much more demanding. This issue has attracted a large body of bio-informatics work, where all approaches either abandon the model of free energy minimization, or make restrictions on the class of pseudoknots that can be recognized. The well-known algorithm by Rivas and Eddy (7), which is able to predict a restricted class of pseudoknots, needs $O(n^6)$ time and $O(n^4)$ memory space, where n is the sequence length. Even more restrictive, but more efficient by two orders of magnitude is the program *pknotsRG* by Reeder and Giegerich (8), requiring $O(n^4)$ time and $O(n^2)$ space. The new method presented here re-implements and extends this approach in several significant ways.

TOOL

The program *pknotsRG* enriches the usual RNA folding routines by the class of simple recursive pseudoknots. It uses the current thermodynamic energy model by the Turner group (5), extended by some pseudoknot specific values. *pknotsRG* offers three basic methods:

(i) Standard MFE folding with pseudoknots: This method returns the structure of minimum free energy (MFE), containing a pseudoknot or not. In the latter case, this structure is identical to the MFE structure computed by *RNAfold* from the Vienna RNA package.

(ii) Enforced folding: The best structure in the folding space that contains at least one pseudoknot is reported. This is especially useful, when a pseudoknot is suspected, but not computed in the MFE folding.

*To whom correspondance should be addressed. Tel: +49 521 106 2913; Fax: +49 521 - 106 - 6411; Email: robert@techfak.uni-bielefeld.de

(iii) Local folding: The pseudoknot with the best energy to length ratio is predicted. This mode helps to identify promising pseudoknot candidates in otherwise misfolded structures.

These methods were already available in the first version of *pknotsRG*. In this work we present an extension of this algorithm, including a complete re-implementation in the C programming language. This yields a drastically faster program and allows to accommodate the additional features with reasonable resource requirements.

The improved run time is best shown by an example: Folding a sequence of length 200 with the old version takes 80 s, whereas the new version finishes within 1 s. Due to lower memory consumption, we are now also able to fold very long sequences without being limited by the computer memory. The folding of 1000 nucleotides requires only 31 MB memory and 10 min.

Our first extension is the capability to compute near-optimal structures. Often the native structure of an RNA is not predicted as the MFE structure. This may stem from uncertainties in the energy parameters, or the molecule may not reach its MFE structure, either due to interactions with another molecule or by ending in a kinetic trap. In such a situation, alternative near-optimal structures may be better predictions. All methods of *pknotsRG* can now be run in a suboptimal mode, where all suboptimal solutions up to a user-defined threshold are computed. This is done in a way similar to *RNAsubopt* (9) from the Vienna RNA package for pseudoknot-free structures.

Another useful extension for large scale analysis, is a sliding window technique with adjustable window size and window position increment. For each position of the window the analysis is performed according to the user settings (e.g. enforced mode with 20% suboptimals). When the window is shifted, only the non-overlapping part of the new window has to be computed, the old part is be re-used. Depending on the window size and increment, this saves a considerable amount of time, much better than a wrapper program that folds each window independently.

Web server integration

The *pknotsRG* Web Server (http://bibiserv.techfak.uni-bielefeld.de/pknotsrg) offers an easy and comfortable way to use the *pknotsRG* algorithm. Input has to be provided in a FASTA format containing one single RNA sequence. All non-nucleotide characters (i.e not A,C,G,T,U,N in upper or lower case) are removed from the sequence. The nucleotide 'T' is treated as 'U' during the folding routine, 'N's are always unpaired. The user selects one of the three basic modes and optionally an energy threshold for suboptimal enumeration. If he wants to use the window mode, at least the window size has to be specified. The default window increment is set to one, unless otherwise specified. Depending on the user selection, there are three different ways how the result is displayed:

(i) MFE structure only: The computed secondary structure and its energy are displayed in Vienna (Dot-Bracket) Notation. Here, a pair of opening and closing parenthesis denotes a base pair, a dot represents a single stranded nucleotide. A 2D visualization is generated using the *PseudoViewer* Web service (10). Also, a link to a Connect formatted file, originally used by Michael Zukers *mfold* (11), is provided. An example output is shown in Figure 1.

(ii) Suboptimal solutions: The list of suboptimal solutions is displayed as an *RNAmovie* (12). *RNAmovies* displays the individual suboptimal structures one after another with an adjustable number of intermediate transition steps. This allows for a smooth morphing from one structure to the next. Additionally, all suboptimal structures are displayed as Vienna strings along with their respective energy value at the bottom of the result page.

(iii) Window mode: For each window, the start and end position is displayed, followed by the result of the window folding. This may be one structure (as Vienna string), or many if the suboptimal mode is also selected. There is currently no vizualization provided, when using the window mode. However, the user can resubmit the sequence of an interesting window on its own. Then, the vizualizations explained in (i) and (ii) are automatically generated.

To ensure proper functionality two restrictions are imposed on the *pknotsRG* Web Server:

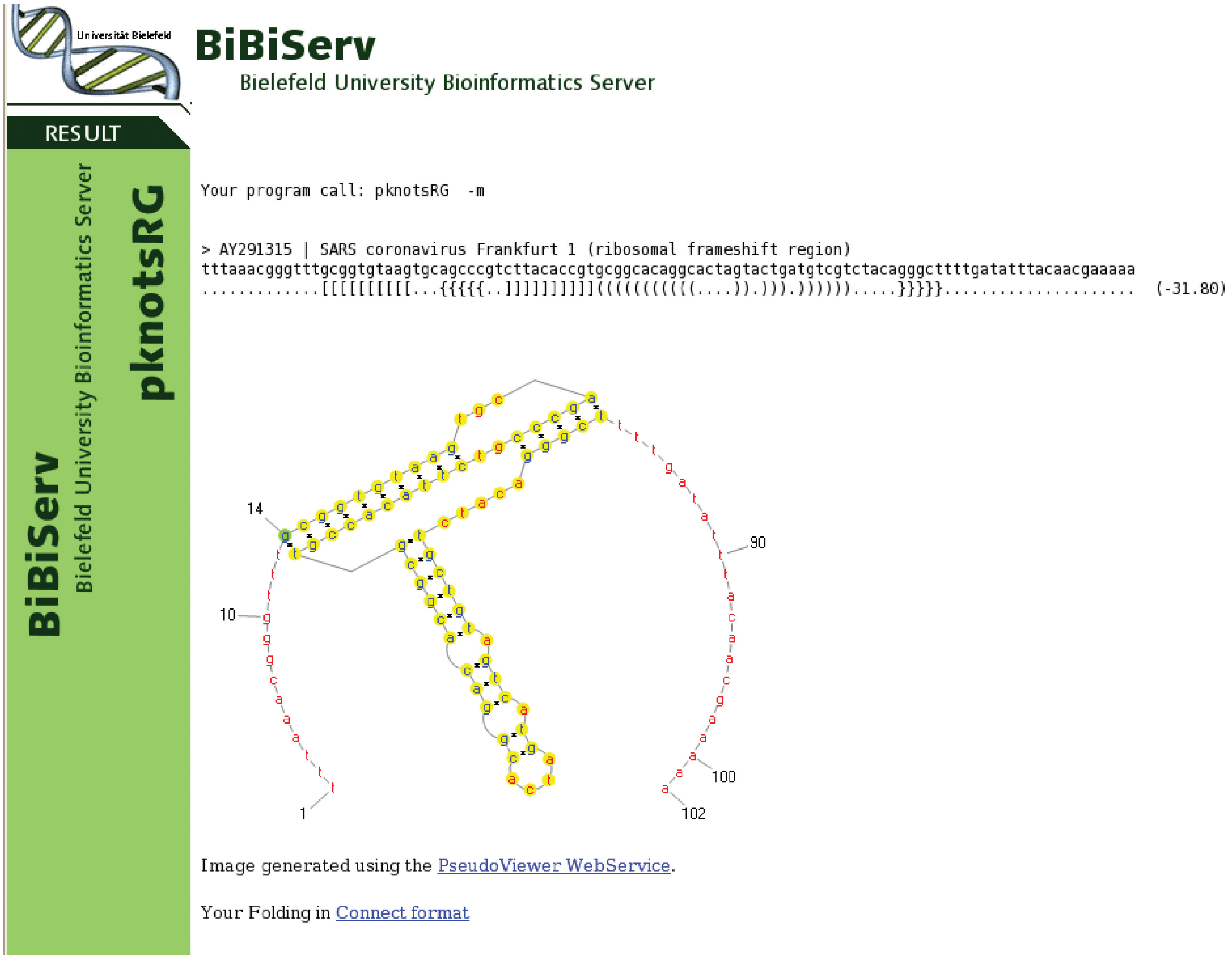
(i) A maximum of 800 nucleotides is allowed for one submission. This guarantees the computation to finish within a few minutes. Note that smaller requests usually are answered instantly.

(ii) The suboptimal folding space of an RNA sequence grows exponentially with the energy threshold and sequence length. We prevent the Web service of returning too much data, by limiting the number of returned structures to the first 100, respectively, 1000 lines for the window mode.

The web interface is a convenient means for manual computation, but not designed for automated folding, i.e. integrated in an Internet-based bio-informatics pipeline. For this purpose, we provide a Web Service interface. This is particularly useful for researchers, who either cannot or do not want to install *pknotsRG* locally on their machines.

The interface is accessible from any computer in any programming language implementing the SOAP protocol. The Web Service Description Language (WSDL) file describes the functionality of the Web Service and can be found at (http://bibiserv.techfak.uni-bielefeld.de/wsdl/pknotsRG.wsdl). Several software development environments can automatically generate program code from the WSDL file for the integration of a Web Service into a local application. We provide an example client written in Java on the project homepage.

The extensions presented here have been integrated into the original *pknotsRG* program in 2005 and 2006. Our server statistics indicate 1068 online uses and 773 downloads for local installation in 2006.


Universität Bielefeld
BiBiServ
Bielefeld University Bioinformatics Server
RESULT
BiBiServ
Bielefeld University Bioinformatics Server
pknotsRG
14
10
1
90
100
102

```
Your program call: pknotsRG  -m

> AY291315 | SARS coronavirus Frankfurt 1 (ribosomal frameshift region)
tttaaacgggtttgcggtgtaagtgcagcccgtcttacaccgtgcggcacaggcactagtactgatgtcgtctacagggcttttgatatttacaacgaaaaa
.............[[[[[[[[[...{{{{{..]]]]]]]]]((((((((((((....)).))).))))))).....}}}}}.....................  (-31.80)
```

Image generated using the PseudoViewer WebService.

Your Folding in Connect format

Figure 1. Web Server output for prediction of the SARS-1 programmed ribosomal frameshift signal (GenBank: *AY29135*).

RESULTS FROM AN APPLICATION EXAMPLE

Recently, an unusual three-stemmed pseudoknot has been identified that promotes programmed -1 ribosomal frameshifting in the SARS Coronavirus (13). The pseudoknot is thought to pause the ribosome during translation, which then shifts back by one nucleotide on the 'slippery site'. This special pseudoknot seems to be conserved in all Coronaviruses, and thus could be a target for anti-viral therapeutics.

The 3-stem topology is predicted by *pknotsRG* as the optimal structure with an energy of −31.8 kcal/mol (Figure 1), using either the MFE or the enforced mode.

CHARACTERIZATION OF THE RECOGNIZED CLASS OF PSEUDOKNOTS

General pseudoknot prediction in energy-based models is a Non-deterministic Polynomial time NP-complete problem (14), and thus requires exponential run time. However, if we impose some restrictions on the helices forming the pseudoknots, we can achieve faster algorithms. The restrictions in *pknotsRG* are motivated by the observation, that most of the currently known pseudoknots are rather simple ones. They consist of only two helices, interacting in a crosswise fashion, as shown in Figure 2. If we allow the unpaired strands (u,v,w in Figure 2) to build secondary structures internally in an arbitrary way, including multiloops and pseudoknots, we call this class *simple recursive pseudoknots*.

Our model further restricts this class by three canonization rules.

(i) **Rule 1:** The 5′ and 3′ part of each pseudoknot helix must have the same length and must not have an interruption. This disallows bulges and internal loops inside pseudoknot stems.

(ii) **Rule 2:** The pseudoknot helices must have maximal length, i.e. if there is a possible base pair at either end, it must be closed.

(iii) **Rule 3:** If due to Rule 2 the helices would overlap, the first helix (a) is prioritized and the second one is shortened.

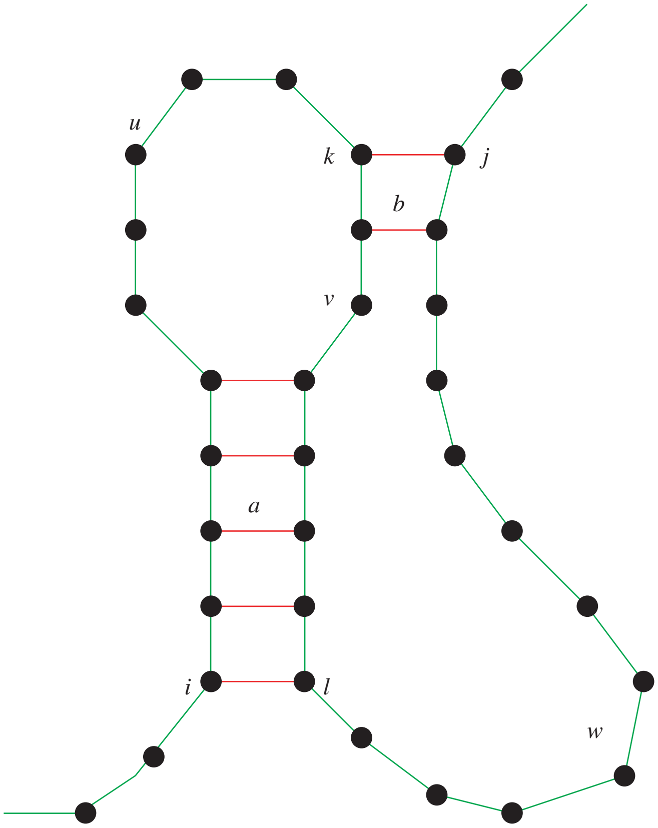


Figure 2. A simple pseudoknot formed by helices *a* and *b*. If the loop regions *u*, *v*, *w* fold further internal structures, including pseudoknots of this type, we have a simple recursive pseudoknot.

We call the resulting class *canonical simple recursive pseudoknots* (csr-pk). Note that all helices not participating in a pseudoknot are not affected by this canonization. A detailed discussion about the effects of canonization can be found in (8).

In an exhaustive analysis of the pseudoknot database PseudoBase (http://biology.leidenuniv.nl/~Batenburg/PKB.html), it was shown, that out of 212 known pseudoknots, 172 are simple recursive (8). Almost 80% of these are even canonical simple recursive pseudoknots. This shows the abundance of the class csr-pk within all validated structures.

IMPLEMENTATION

We will briefly sketch the way we implemented these ideas as an extension of the usual dynamic programming (DP) scheme for RNA folding [5,6]. Since only two helices participate in one pseudoknot, we loop over all possible knots in one O(n^4) loop and store the result in a two-dimensional matrix. More detailed, for a pseudoknot between bases *i* and *j*, the algorithm enumerates all *k* and *l*, such that $i<k<l<j$. The first pseudoknot helix (*a* in Figure 2) starts at base pair *i* and *l*, the second (*b*) at position *k* and *j*. Now, we make use of our canonization: The maximal length of a helix can be pre-computed and stored in a two-dimensional array. Therefore, after choosing position *k* and *l*, the algorithm looks up the maximal helix lengths, *h* for (*i,l*) and *h'* for (*k,j*) and checks the applicability of Rule 3. Having the stems fixed, the location of the three enclosed pseudoknot loops (*u,v,w in* Figure 2) follows directly: loop *u* ranges from position $i+h+1$ to $k-1$, loop *v* from $k+h'+1$ to $l-h-1$, and loop *w* from $l+1$ to $j-h'-1$. The best folding for these smaller subsequences has already been computed earlier in the DP scheme. The energy sum of the two pseudoknot helices and the loop folding energies gives us the total pseudoknot energy. The minimal one over all *k* and *l*, is stored is the two-dimensional pseudoknot matrix. This value competes with values of unknotted foldings for the interval (*i,j*).

During the (suboptimal) backtrace procedure the pseudoknot matrix is handled in the same way as the other matrices. Starting with a user-defined energy threshold *thr*, all suboptimal structures having an energy not more than *thr* larger than the optimal structure are backtraced. For each suboptimal structure *s*, the threshold is reduced by the energy difference to the optimal structure and the backtrace procedure is called recursively on the substructures of *s*, yet with a smaller threshold. At some point the threshold falls below 0, which means no further suboptimal structure within the user-defined energy band can arise from this backtrace.

AVAILABILITY

The program is available for download, both as C source code and precompiled binary executable for most common platforms (Solaris, Linux, Windows and Mac OS X). This version has none of the Web Server restrictions. It folds arbitrary long sequences and reports as many suboptimal solutions as the user requests.

The Windows version contains a basic graphical user interface, the other versions provide a powerful interactive command line tool. All versions are available in the download section of the project home page at http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/.

ACKNOWLEDGEMENTS

We thank Jan Krüger for the Web Service implementation of *pknotsRG* and for the initial connection to the *PseudoViewer* Web Service. Funding to pay the open Access publication charges for this article was provided by the Bielefeld University.

Conflict of interest statement. None declared.

REFERENCES

1. Ferré-D'Amaré,A.R., Zhou,K. and Doudna,J.A. (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395**, 567–674.
2. Doudna,J.A. and Cech,T.R. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
3. Theimer,C.A., Blois,C.A. and Feigon,J. (2005) Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. *Mol. Cell*, **17**, 671–682.

4. Jacobs,J.L., Belew,A.T., Rakauskaite,R. and Dinman,J.D. (2007) Identification of functional, endogenous programmed -1 ribosomal frameshift signals in the genome of Saccharomyces cerevisiae. *Nucleic Acids Res.*, **35**(1), 165–174.
5. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
6. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, **125**, 167–188.
7. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
8. Reeder,J. and Giegerich,R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, **5**, 104.
9. Wuchty,S., Fontana,W., Hofacker,I.L. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
10. Byun,Y. and Han,K. (2006) PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res.*, **34**(suppl. 2), W416–422.
11. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary informations. *Nucleic Acids Res.*, **9**(1), 133–148.
12. Evers,D. and Giegerich,R. (1999) RNA movies: Visualizing RNA secondary structure spaces. *Bioinformatics*, **15**(1), 32–37.
13. Plant,E.P., Pérez-Alvarado,G.C., Jacobs,J.L., Mukhopadhyay,B., Hennig,M. and Dinman,J.D. (2005) a three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS Biol.*, **3**(6), e172.
14. Lyngsø,R.B. and Pedersen,C.N.S. (2001) RNA pseudoknot prediction in energy based models. *J. Comp. Biol.*, **7**, 409–428.