INFO-RNA—a server for fast inverse RNA folding satisfying sequence constraints

Anke Busch* and Rolf Backofen

Albert-Ludwigs-University Freiburg, Institute of Computer Science, Bioinformatics Group, Georges-Koehler-Allee 106, 79110 Freiburg, Germany

ABSTRACT

INFO-RNA is a new web server for designing RNA sequences that fold into a user given secondary structure. Furthermore, constraints on the sequence can be specified, e.g. one can restrict sequence positions to a fixed nucleotide or to a set of nucleotides. Moreover, the user can allow violations of the constraints at some positions, which can be advantageous in complicated cases.

The INFO-RNA web server allows biologists to design RNA sequences in an automatic manner. It is clearly and intuitively arranged and easy to use. The procedure is fast, as most applications are completed within seconds and it proceeds better and faster than other existing tools. The INFO-RNA web server is freely available at http://www.bioinf.uni-freiburg.de/Software/INFO-RNA/

INTRODUCTION

The function of RNA molecules often depends on both the primary sequence and the secondary structure. RNAs are involved in translation (tRNA, rRNA), splicing (snRNA), processing of other RNAs (snoRNA, RNAseP) and regulatory processes (miRNA, siRNA) (1). Furthermore, parts of mRNAs can adopt structures that regulate their own translation (SECIS (2,3), IRE (4)). Since prediction and experimental determination of 3D RNA structures remain difficult, much work focuses on problems associated with its secondary structure, which is the set of base pairs. The problem of predicting the secondary structure of an RNA is called the 'RNA folding problem'. Existing computational approaches are based on a thermodynamic model that gives a free energy value for each secondary structure (5). The structure with the lowest free energy [called the 'minimum free energy (mfe) structure'] is expected to be the most stable one.

Here, we consider the 'inverse RNA folding problem satisfying sequence constraints', which is the design of RNA sequences that fold into a desired structure and fulfill some given constraints on the primary sequence. These constraints can restrict certain positions to fixed nucleotides or to a fixed set of nucleotides. The INFO-RNA web server is applicable to the design of RNA elements that include conserved nucleotides, which are essential for binding of proteins.

METHODS AND USAGE

The INFO-RNA server uses a new algorithm for the INverse FOlding of RNA that involves two steps. The first step contains a new design method for good initial sequences. It is followed by an improved stochastic local search. Both steps are described shortly in the following and more in detail (6).

The initializing step

The input of the algorithm consists of the target structure. During the first step of INFO-RNA, a dynamic programming approach designs an RNA sequence that adopts the lowest energy a sequence can have when folding into the target structure. However, this sequence is not guaranteed to fold into the target structure since this sequence can have another mfe structure. Therefore, the resulting sequence is processed further in a second step.

The local search step

To improve the quality of the sequence generated in the first step, local sequence mutations are made iteratively. In INFO-RNA, this is done by a 'stochastic local search' (SLS) that minimizes the structure distance between the mfe structure of the designed sequence and the target structure. Here, sequence neighbors are tested either in a random order or in an order that depends on the energy difference between the current sequence and the neighbor sequence when folding into the target structure. The higher the difference is, the earlier the mutation is examined. Optionally, the probability of folding into the wanted structure can be optimized as well.

*To whom correspondence should be addressed. Tel: +49 761 2038253; Fax: +49 761 2037462; Email: abusch@informatik.uni-freiburg.de

Novel extensions of the algorithm

In an extension to (6), the INFO-RNA web server can handle a set of user-given constraints on the primary sequence. These constraints have to be fulfilled during both steps of the algorithm. That means, after finishing the initializing step, we get a sequence that adopts the target structure with the lowest energy that is possible if the constraints are fulfilled. During the local search step, only mutations that coincide with the constraints are valid.

If the constraints on the sequence are not strictly fixed, the user can specify some positions where violations of the constraints are allowed. Furthermore, the user can restrict the maximal number of constraints that are violated in the final sequence ($V^{\max}$). This might be useful if one allows violations of two different constraints but wants at most only one of these violations in the designed RNA sequence.

Finally, the INFO-RNA server outputs the best-found RNA sequence satisfying the sequence constraints with at most violations.

Usage

The INFO-RNA web server is clearly and intuitively arranged. In order to obtain an RNA sequence folding into a target structure and satisfying some sequence constraints, both (structure and sequence constraints) have to be given. The structure has to be given in bracket notation. Here, a base pair between bases i and j is represented by a '(' at the i-th position and a ')' at position j. Unpaired bases are represented by dots. The sequence constraints have to be entered in IUPAC symbols, where e.g. restricting a position to Y means that a C or a U is allowed there. In addition, the user can choose some positions where the constraints are allowed to be violated during the local search. Besides, the maximal number of positions where the constraints are allowed to be violated in the final sequence can be specified. Furthermore, the user can fix some parameters used during the stochastic local search, e.g. the search strategy of either only minimizing the structure distance or additionally maximizing the folding probability as well as the search order of the sequence positions. Finally, the user can choose whether the results are shown on the web page or send via email. For all options, a comprehensive help and detailed examples are given. Figure 1 shows the output of a typical computation. First, the input data are summarized. Below, the designed sequence is shown including information about its mfe structure, its free energy and its folding probability. Additionally, the user can download the results in FASTA, CT and RNAML format.

RESULTS AND APPLICATION

The INFO-RNA web server allows biologists to design RNA sequences, which fold into a given structure, in an automatic manner. The procedure is fast, as most applications are completed within seconds. As shown in (6), INFO-RNA (not considering sequence constraints) proceeds better and faster than other existing tools. Artificial as well as biological test sets were analyzed. The biological test sets divide into computationally predicted structures for known RNA sequences and structures from the biological literature. INFO-RNA turned out to be the algorithm having the highest succession rates as well as the lowest computation times for all test sets. Additional stability tests showed that the designed sequences are more stable than the biological ones.

The novel extension of INFO-RNA including sequence constraints allows the design of *cis*-acting mRNA elements such as the 'iron responsive element' (*IRE*) and the 'polyadenylation inhibition element' (*PIE*). Both elements have conserved sequence positions in loops. The IRE is essential for the expression of proteins that are involved in the iron metabolism (7). It consists of a stem-loop structure, and the first five nucleotides in the hairpin loop as well as the bulged nucleotides were found to be essential for binding of iron-regulatory proteins. The PIE contains two binding sites for U1A proteins (8). It consists of a stem structure with two asymmetric internal loops that serve as U1A-binding sites (Figure 2). Using the INFO-RNA web server, we designed artificial IREs and PIEs having a much higher folding probability compared to natural elements. While designed sequences for the IRE having a single C bulge fold into the target structure with an average probability of 88%, natural sequences do so only with an average probability of 15%. Regarding IREs having an interior loop with left size 3 and right size 1, the results are similar. Furthermore, the average probability of the designed PIE sequences folding into the target structure is more than 20 times higher than the probability of the natural PIE sequences (Supplementary Figure 1). Besides, all IREs designed by the INFO-RNA web server adopt the wanted structure as its mfe structure whereas only a small fraction of the natural ones does (Supplementary Figure 2).

Furthermore, we demonstrated the usability of the INFO-RNA web server by designing artificial microRNA (miRNA) precursors that are as stable as possible. To this end, artificial miRNA sequences published in (9) were used. Applying the INFO-RNA web server, we designed precursors of these artificial miRNAs as well as of the natural miRNA. All of the designed sequences have a free energy that is at least twice as low as the free energy of the natural precursor sequences. On average, their probability of folding into the target miRNA precursor structure is five times as high as the folding probability of the natural precursor sequences. For more details see Supplementary Table 1.

Other potential application areas are the design of ribozymes and riboswitches (10), which may be used in research and medicine, and the design of non-coding RNAs, which are involved in a large variety of processes, e.g. gene regulation, chromosome replication and RNA modification (11).

Input:

General Parameters:

RNA secondary structure: 5'−(((((...(((((......)))))).)))))−3'

Sequence constraints: 5'−NNNNNUGCNNNNNCAGUGHNNNNNCNNNNN−3'

Allowed constraint violations: 5'−000001100000000000100000000000−3'

Maximal number of violations: 2

Parameters of the Stochastic Local Search:

Objective function: *mfe*

Probability of accepting bad mutations: *0.1*

Presort candidates for mutation: *yes*

Results:

1. Designed Sequence

Designed sequence: 5'−GGGCCUUCGCCCCCAGUGAGGGGCCGGCCC−3'

Target structure: 5'−(((((...(((((......)))))).)))))−3'

Constraint violations: *1*

Free energy (target structure): *−19.50 kcal/mol*

Folding probability (target structure): *0.850225*

Base pair distance of the mfe structure to the target structure: *0*

mfe structure: 5'−(((((...(((((......)))))).)))))−3'

Direct output for copy and paste:

Designed sequence → GGGCCUUCGCCCCCAGUGAGGGGCCGGCCC
Target structure → (((((...(((((......)))))).)))))
mfe structure → (((((...(((((......)))))).)))))

Downloadable File Formats (designed sequence + its mfe structure): **.fasta file   .ct file   RNAML**

Figure 1. INFO-RNA web server output. The figure shows the output of a typical computation (design of an IRE with fixed bases in the interior and hairpin loop and a maximum of two constraint violations at three possible sequence positions).

DISCUSSION

We have shown that the INFO-RNA web server is a very fast and successful tool to design RNA sequences, which fold into a given structure and fulfill some sequence constraints. The core of the algorithm was introduced in (6). There, we already showed that INFO-RNA (not considering sequence constraints) proceeds better and faster than other existing tools. Here, we have demonstrated that the INFO-RNA web server, which can handle additional constraints on the primary sequence, also performs well and fast.

Most of the sequences designed by the INFO-RNA web server are highly stable and have very low free energy. This might result from the high GC content that most of the sequences show since G–C base pairs are energetically most favorable. It is not clear whether such highly stable

Figure 2. Structures and conserved sequence positions of a PIE. The figure shows the consensus structure and conserved sequence positions of a PIE that contains two asymmetrical internal loops as binding sites for U1A proteins (U1A-PIE). Conserved sequence positions are highlighted in gray.

structures are always of advantage or how the high GC content may influence the kinetics of the folding process. To reduce the GC content, the user can constrain some positions to A and/or U. In the future, it is desirable to extend the algorithm to allow the user to specify the GC content.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Huttenhofer,A., Brosius,J. and Bachellerie,J.P. (2002) RNomics: identification and function of small, non-messenger RNAs. *Curr. Opin. Chem. Biol.*, **6**, 835–843.
2. Huttenhofer,A., Westhof,E. and Bock,A. (1996) Solution structure of mRNA hairpins promoting selenocysteine incorporation in *Escherichia coli* and their base-specific interaction with special elongation factor SELB. *RNA*, **2**, 354–366.
3. Liu,Z., Reches,M., Groisman,I. and Engelberg-Kulka,H. (1998) The nature of the minimal 'selenocysteine insertion sequence' (SECIS) in *Escherichia coli*. *Nucleic Acids Res.*, **26**, 896–902.
4. Addess,K.J., Basilion,J.P., Klausner,R.D., Rouault,T.A. and Pardi,A. (1997) Structure and dynamics of the iron responsive element RNA: implications for binding of the RNA by iron regulatory binding proteins. *J. Mol. Biol.*, **274**, 72–83.
5. Zuker,M. (1994) Prediction of RNA secondary structure by energy minimization. *Methods Mol. Biol.*, **25**, 267–294.
6. Busch,A. and Backofen,R. (2006) INFO-RNA – a fast approach to inverse RNA folding. *Bioinformatics*, **22**, 1823–1831.
7. Hentze,M.W. and Kuhn,L.C. (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl Acad. Sci. USA*, **93**, 8175–8182.
8. Varani,L., Gunderson,S.I., Mattaj,I.W., Kay,L.E., Neuhaus,D. and Varani,G. (2000) The NMR structure of the 38 kDa U1A protein – PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein. *Nat. Struct. Biol.*, **7**, 329–335.
9. Schwab,R., Ossowski,S., Riester,M., Warthmann,N. and Weigel,D. (2006) Highly specific gene silencing by artificial microRNAs in Arabidopsis. *Plant Cell*, **18**, 1121–1133.
10. Knight,J. (2003) Gene regulation: switched on to RNA. *Nature*, **425**, 232–233.
11. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.