BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 4 2006, pages 500–503
doi:10.1093/bioinformatics/btk010

Sequence analysis

# RNAshapes: an integrated RNA analysis package based on abstract shapes

Peter Steffen[1,*], Björn Voß[3], Marc Rehmsmeier[2], Jens Reeder[1] and Robert Giegerich[1]

[1]Faculty of Technology and [2]Center for Biotechnology (Ce BiTec), Bielefeld University, 33594 Bielefeld, Germany and [3]Institute of Biology II, Experimental Bioinformatics, Freiburg University, Schänzlestr. 1, 79104 Freiburg, Germany

Received on October 25, 2005; revised on December 13, 2005; accepted on December 14, 2005
Advance Access publication December 15, 2005
Associate Editor: Golan Yona

## ABSTRACT

**Summary:** We introduce RNAshapes, a new software package that integrates three RNA analysis tools based on the abstract shapes approach: the analysis of shape representatives, the calculation of shape probabilities and the consensus shapes approach. This new package is completely reimplemented in C and outruns the original implementations significantly in runtime and memory requirements. Additionally, we added a number of useful features like suboptimal folding with correct dangling energies, structure graph output, shape matching and a sliding window approach.

**Availability:** RNAshapes is freely available at http://bibiserv.techfak.uni-bielefeld.de/rnashapes/ as C source code, and as compiled binaries for the most common computer architectures. For Microsoft Windows, we also offer a graphical user interface with convenient access to the complete functionality of the package.

**Contact:** psteffen@techfak.uni-bielefeld.de

## 1 INTRODUCTION

Abstraction is our major mental aid to master complexity. When we speak about functional structures of RNA, we speak of long hairpins for miRNA precursors, of clover leaf structures for tRNA, of neighboring hairpins with attenuators, etc. and we often do not care about individual base pairs or helix sizes. For programs comparing RNA structures, it has long been suggested to represent larger structures as trees at different levels of detail (Shapiro, 1988) (subroutine `b2Shapiro` in the Vienna RNA package (Hofacker *et al*., 1994)).

RNA structure prediction algorithms, however, are ignorant of abstraction and either deceive us with a single, minimum free energy prediction, or overwhelm us with a plethora of near-optimal structures, most of which are very similar and thus redundant.

RNA shape abstraction maps structures to a tree-like domain of shapes, retaining adjacency and nesting of structural features, but disregarding helix lengths. Shape abstraction integrates well with dynamic programming algorithms, and hence it can be applied during structure prediction rather than afterwards. This avoids exponential explosion and can still give us a non-heuristic and complete account of properties of the molecule's folding space. Rather magically, some long and hard-studied problems become easy.

So far, we have approached three problems with the use of abstract shapes:

(1) Computation of a small set of representative structures of different shapes, complete in a well-defined sense (Giegerich *et al*., 2004).

(2) Computation of accumulated shape probabilities (B. Voß, R. Giegerich and M. Rehmsmeier, manuscript under review).

(3) Comparative prediction of consensus structures, as an alternative to the over-expensive Sankoff Algorithm: RNAcast (Reeder and Giegerich, 2005).

So far, the first two applications have only been available as implementations in the functional programming language Haskell, with strong limitations on input sequence length. RNAcast was only available as an online tool. Here, we introduce a complete reimplementation of these approaches in the programming language C. This new implementation is around 50–100 times faster than the original implementations, has lower memory requirements and can be used with significantly longer input sequences. It combines all three tools in one single package. We also included a number of additional features.

In the following, we will shortly review the notion of abstract shapes and explain where its power comes from. We will then provide an overview of the problems that can be approached in the new way.

## 2 THE ABSTRACT SHAPES APPROACH

An RNA shape is an abstract representation of an RNA secondary structure. It is inspired by the dot-bracket representation known from the Vienna RNA package (Hofacker *et al*., 1994). Consider

[*]To whom correspondence should be addressed.

the following sequence and two secondary structures from its folding space in dot-bracket representation:

```
AUCGGCGCACAGGACAUCCUAGGUACAAGGCCGCCCGUU
..(((.((..(((....))).(((.....))))))))..
..(((.....(((....))).(((.....)))..)))..
```

The shapes approach offers five abstraction levels—or shape types — ordered in their degree of abstraction. Common to all levels is that they abstract from loop and stack lengths, where unpaired regions are represented by an underscore and stacking regions by a pair of squared brackets. This is the least abstract shape type 1, so the two example secondary structures become

```
_[_[_[_]_[_]]]_
  _[_[_]_[_]_]_
```

The succeeding shape types gradually increase abstraction, ending in type 5, where no unpaired regions are included and nested helices are combined. In this type, our example structures are both represented as

```
[[][]]
```

These abstractions form the basis of all applications of RNA abstract shape analysis. In the following we give an overview of the main applications, all integrated in the new RNAshapes package.

## 2.1 Shape representative analysis

Current RNA folding algorithms either calculate a single, minimum free energy prediction, or a huge number of suboptimal structures, most of which are quite similar and therefore redundant. With shapes, we abstract from the concrete secondary structures and only consider classes of structures that fall into different shapes. The *shape representative* (in short: *shrep*) of a shape is the structure with the minimum free energy inside a shape class.

Figure 1 shows an example program run inside the RNAshapes user interface with the *Natronobacterium pharaonis* tRNA for alanine (gb: AB003409.1/96-167). The predicted *mfe*-structure is one hairpin with internal loops, as depicted in Figure 1 on the left. The biologically active structure is the clover-leaf structure (Figure 1 right). It would appear at position 123 in the energy sorted list of 308 suboptimals, produced by *RNAsubopt* (Wuchty *et al*., 1999) with an energy range of 5 kcal/mol above the *mfe*. Using RNAshapes, we get three shapes in an energy range of 5 kcal/mol, of which the rank 3 *shrep* is the clover-leaf structure.

## 2.2 Shape probabilities

In (Voß *et al*., manuscript under review), we extended the shapes approach to the computation of shape probabilities. The probability of a shape is the sum of the probabilities of all structures that fall into this shape. Several analyses indicate that this approach is quite effective. For example, an analysis of a conformational switch shows the existence of two shapes with probabilities ∼2/3 versus 1/3, whereas the analysis of a micro RNA precursor reveals the hairpin shape with a probability near to 1.0 (Voß *et al*., manuscript under review).

The new implementation contains three approaches for probability analysis, suitable for different input sizes.

*2.2.1 Complete probability analysis* This implies a complete and non-heuristic analysis of the folding space, where the computational effort depends only on the size of the shape space, which is much smaller than the folding space. On a computer with 2 GB main memory, sequences up to a length of ∼300 bases can be processed. The following two approaches relax this restriction.

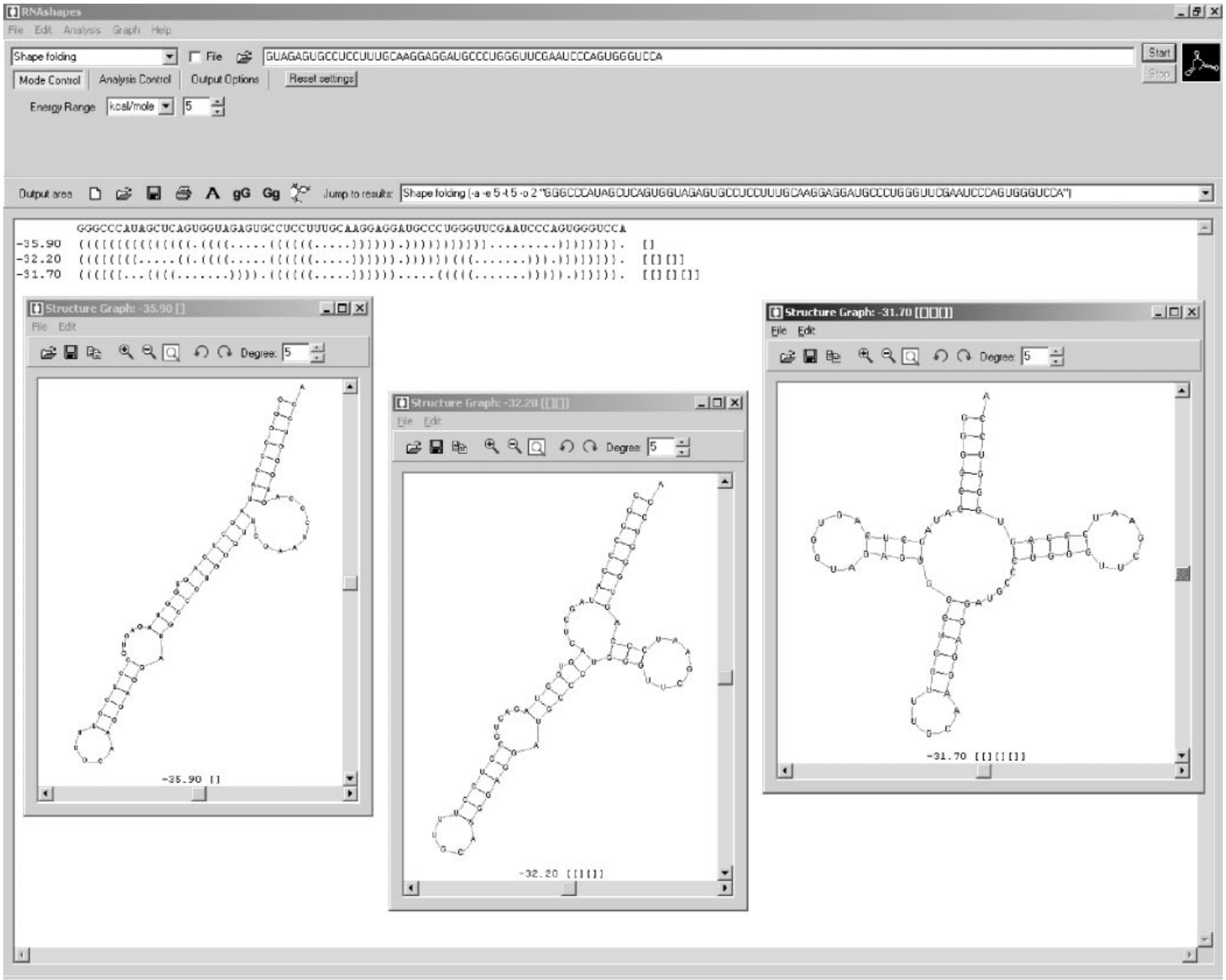
*2.2.2 Sampling shapes probability analysis* The sampling shapes approach works in the same manner as Ding and Lawrence's Sfold program (Ding and Lawrence, 2003). In each step of the recursive backtracing procedure, base pairs and the structural element they belong to are sampled according to their probability, which is obtained from the partition function (McCaskill, 1990). For each sample, we calculate its corresponding shape. The shape probability then results from its frequency in the sample space. A sample size of 1000 (as also used by the Sfold server) is sufficient for high probability shapes. For details see (Voß *et al*., manuscript under review). The sampling approach is computationally feasible with an input length of up to 1500 bases.

*2.2.3 Fast high probability shape analysis* The third option only calculates probabilities for shapes with the lowest free energy shreps. These are often also the shapes of highest probability (but not necessarily so). This mode is implemented as a two-step process. In the first step, the lowest free energy shapes are calculated as in Section 2.1. Then, for each of these shapes, the probability is calculated individually. Since these individual calculations have significant lower resource requirements than the complete probability analysis, it is suitable for input sequences up to ∼500 bases.

## 2.3 Consensus shapes

The well-known Sankoff algorithm (Sankoff, 1985) for simultaneous RNA sequence alignment and folding is currently considered an ideal, but computationally over-expensive method. Available tools implement this algorithm under various pragmatic restrictions (Mathews and Turner, 2002; Havgaard *et al*., 2005). See (Gardner and Giegerich, 2004) for a recent comparative evaluation of these and several further methods.

In (Reeder and Giegerich, 2005), we proposed to redefine the consensus structure prediction problem in a way that does not imply a multiple sequence alignment step. For a family of RNA sequences, our method RNAcast explicitly and independently enumerates the near-optimal abstract shape space, and predicts as the consensus an abstract shape common to all sequences. For each sequence, it delivers the thermodynamically best structure that has this common shape. Since the shape space is much smaller than the structure space, and identification of common shapes can be done in linear time (in the number $k$ of shapes considered), the method is linear in the number $s$ of sequences, yielding $O(n^3 \cdot k \cdot s)$ overall. Our evaluation shows that the new method compares favorably with the available alternatives (Reeder and Giegerich, 2005). It is particularly useful on sequences with low conservation, where methods based on sequence alignment cannot be employed. We have now integrated RNAcast into the RNAshapes package.



Shape folding ▾ □ File GUAGAGUGCCUCCUUUGCAAGGAGGAUGCCCUGGGUUCGAAUCCCAGUGGGUCCA

Energy Range kcal/mole ▾ 5

Jump to results: Shape folding (-a -e 5 -t 5 -o 2 "GGGCCCAUAGCUCAGUGGUAGAGUGCCUCCUUUGCAAGGAGGAUGCCCUGGGUUCGAAUCCCAGUGGGUCCA")

```
         GGGCCCAUAGCUCAGUGGUAGAGUGCCUCCUUUGCAAGGAGGAUGCCCUGGGUUCGAAUCCCAGUGGGUCCA
-35.90   ((((((((((((((.((((.....((((((.....)))))).))))))))))).........))))))).   []
-32.20   (((((((.....((.((((.....((((((.....)))))).))))))((.......)).))))))).   [][]
-31.70   ((((((...((((.......)))).((((((.....)))))).....(((((.......))))).)))))).   [][][]
```

Degree: 5

**Fig. 1.** Predicted *shreps* for *N. pharaonis* tRNA-ala in an energy range of 5 kcal/mol above the *mfe*. This energy range holds 308 structures. The figure shows the RNAshapes user interface for Microsoft Windows. The results of the current RNAshapes analysis are shown in the program's output area. To create a structure graph drawing, the user can simply click onto the dot-bracket string of the desired result.

## 3 THE RNAshapes PACKAGE

In addition to the main modes of operation described earlier, the package offers a number of convenient functions:

- Input sequences can either be single sequences, sequence files or multi-sequence files in fasta format. Additionally, interactive user input of sequences is supported.
- Graphical output of secondary structures in postscript format [implemented with program code from the Vienna RNA package (Hofacker *et al*., 1994)].
- Complete suboptimal folding that handles dangling bases correctly.
- A sliding window function for processing whole genomes. Apart from RNAcast, this option can be used with all analysis modes.
- Detailed options to modify the program output.
- Complete control of program functionality by command line options, useful for automatic script processing.
- A graphical user interface for Microsoft Windows with convenient access to the complete functionality of the package. It also offers an interactive visualization of structures from the program output (Fig. 1). The appearance of these structure drawings can be controlled in a very flexible manner (e.g. colors, sizes, fonts, orientation).

## 4 CONCLUSION

RNAshapes offers several powerful RNA analysis tools in one single software package. Owing to the new implementation in C, it is considerably more efficient than the previous separate implementations. With the integrated graphical user interface the package offers enhanced usability, especially for researchers not used to the command line, and hopefully reaches a new community of users.

## ACKNOWLEDGEMENT

Funding to pay the Open Access publication charges for this article was provided by Bielefeld University.

*Conflict of Interest:* none declared.

## REFERENCES

Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.

Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA structure prediction. *BMC Bioinformatics*, **5**, 140.

Giegerich,R. *et al.* (2004) Abstract shapes of RNA. *Nucleic Acids Res.*, **32**, 4843–4851.

Havgaard,J.H. *et al.* (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, **125**, 167–188.

Mathews,D.H. and Turner,D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, **317**, 191–203.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Reeder,J. and Giegerich,R. (2005) Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, **21**, 3516–3523.

Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45** (5), 810–825.

Shapiro,B.A. (1988) An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.*, **4**, 387–393.

Wuchty,S. *et al.* (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.