Comparative analysis of RNA genes: the caRNAc software

Helene Touzet

Comparative analysis of RNA genes: the caRNAc software

Hélène Touzet

Abstract

RNA genes are ubiquitous in the cell and are involved in a number of biochemical processes. Because there is a close relationship between function and structure, software tools that predict the secondary structure of noncoding RNAs from the base sequence are very helpful. In this article, we focus our attention on the inference of conserved secondary structure for a group of homologous RNA sequences. We present the caRNAc software which enables the analysis of families of homologous sequences without prior alignment. The method relies both on comparative analysis and thermodynamic information.

Key Words: RNA, *in silico* folding, structure prediction, comparative analysis, thermodynamic model

1. Introduction

It is now well-acknowledged that noncoding RNAs play an essential role in many cellular processes (e.g. protein synthesis, regulation), even if the function of the majority of

RNAs remains to be elucidated (1). Many of noncoding RNAs have characteristic secondary structures that are highly conserved in evolution. Identifying conserved structure is the first step towards the comprehension of the function of the molecule. Computational approaches provide unexpansive and efficient tools for that purpose.

From a historical perspective, there are two main complementary approaches to address RNA folding prediction : thermodynamic models and phylogenetic models.

The secondary structure of an RNA molecule depends onsy the formation of base pairings: Watson-Crick (A-U and G-C), wobble (G-U) or even non canonical pairings. In the thermodynamic approach, the fundamental assumption is that the molecule adopts a globally minimum free energy structure. Negative stabilizing energies are assigned to the stacking of base pairs in helices and destabilizing energies are assigned to unpaired elements, such as bulge or multi-branched loops. In this model, the folding problem amounts to searching for the set of base pairs that minimizes the free energy level (2). This strategy is implemented in the Mfold (3,4) and RNAfold (5) programs. The main limitation of these methods is that the right structure may be overwhelmed by a large number of potential structures having equivalent, or even better, energy level. Furthermore, there is presently no thermodynamic parameters that deal with pseudoknots within this model.

The other line of research for structure inference is phylogenetic analysis. The main idea of this approach is to extract information from the similarities and differences between different, but homologous, RNA sequences. Phylogenetic analysis relies on the assumption that the spatial structure of a molecule is more highly conserved than is its sequence. In other words, the sequence is free to change during evolution. In terms of

secondary structure, this means that mutation of a base involved in a pairing should generally be compensated by a change in its pairing partner. This guarantees that the ability of the both bases to form isosteric base pairs is retained. This phenomenon is called covariation, or compensatory mutation. If sufficient numbers of sequences are available, these covariations can be identified statistically directly from a multiple sequence alignment. The list of structures determined by comparative analysis is long: ribosomal RNAs, transfer RNAs, RNase P RNAs, HACA box RNAs, snoRNAs, etc. (6). The drawback of pure phylogenetic approaches is that they need a large number of related sequences (more than 10) to be theoretically sound. Furthermore, the accuracy of the result strongly depends on the quality of the multiple alignment. Automatically aligning RNA sequences is a difficult issue (7).

The purpose of the caRNAc software is to achieve more flexibility than pure comparative methods by combining both thermodynamic and phylogenetic information. caRNAc does not require any prior alignment between sequences. This implies that it can successfully handle sequences with low level of conservation (from 60%). The full algorithm is described in more detail in (8,9). A comprehensive comparison of main folding programs, including caRNAc, can be found in (10).

2. Materials

caRNAC is available on a web site. All you need is a W3C compliant web browser (Firefox, Internet Explorer, Mozilla,…). Frequent users also may download the platform and install it locally. CaRNAc requires a C compiler. All source codes are available on the web site.

3. Method

3.1 Getting started

The web site is accessible at http://bioinfo.lifl.fr/carnac. Choose the « web server » section in the main menu. The « examples » section provides several data sets and commented results.

The input submission form contains three main fields.

Enter a name for the sequence (optional)

This name serves as a label for the output page.

Enter the RNA sequences

Your data set should include at least two distinct RNA sequences, and sequences should be in FASTA format. A sequence in FASTA format consists of a single-line description, followed by lines of sequence data. The first character of the description line is a greater-than (">") symbol in the first column. Figure 1 gives an example with three tRNA sequences, that will serve as a guideline for the remaining of this presentation. All non-alphabetic characters are removed. IUPAC symbols are not supported. Sequences may be pasted or uploaded from a file.

Enter your E-mail address

This address is used to send the identifier of the job to the user once the job is completed.

The form then proposes several parameters that determine the final predictions The

default values lead to the most reliable results in average.

Eliminate redundant sequences : By default, caRNAc discards sequences that are too close (more than 98% of identity). Uncheck the box to fold all the sequences.

Take GC content into consideration : When this option is selected, caRNAc uses variable energy thresholds for stems according to the average GC percent of the involved sequence.

Allow isolated stems :When this option is selected, caRNAc permits the creation of stem in one sequence alone, without any counterpart in any other sequence. This option may give better results if there is a large evolutionary distance between structures, or when the sequences are of radically different lengths. But it is time and space consuming. So it should be selected with caution.

The folding is launched by pressing the <RUN> button. Each job is assigned a unique identifier (ID). The computation of the putative foldings ranges from a few seconds for short sequences (less than 300 nt) to several minutes for longer sequences (up to 2000 nt). When the job is completed, the results are displayed on a new page, and an alert email is sent to the user. Results are available for 24 h and may be retrieved with the ID using the « retrieve a result with an ID » section in the main menu.

3.2 Output page

For each sequence, the predicted secondary structure is given in five formats, which are summarized in Figure 2. Note that all structures need not to be identical : the program is robust to minor variations in the structure between the sequences.

Connect notation (ct) : It provides a textual description of the base pairings. The syntax is as follows : columns 1, 3, 4, and 6 redundantly give sequence indices, column 2 gives the sequences and column 4 gives « j » in position « i » if « (i,j) » is a base-pair, otherwise this is zero. The heading of the file contains the size of the sequence and its name (found in the FASTA sequence).

Jpeg file : This file is generated from the CT file using the freely distributed drawing tool Naview (11). It contains a graphical two-dimensionnal representation of the secondary structure.

Postscript file : This is a conversion of the Jpeg file to the Postscript format. This format is ready-to-print.

List of constraints : This text file gives an equivalent formulation of the structure. Each line contains the specification of one stem: « F i j k» means that there is a helix of length k formed between the positions [i,i+k-1] and [j-k+1,j]. This format is useful to specify a list of initial constraints for the Mfold and Kinefold programs (see note 1).

Bracket notation : It consists of two lines. The first line contains the sequence. The second line contains the set of associated pairings encoded by brackets and dots. A base pair between base « i » and « j » is represented by a « (» at position « i » and a «) » at position « j ». Unpaired bases are represented by dots. The lack of pseudoknots in the secondary structure ensures that this notation defines a unique folding. This format is widely-used in the Vienna Package (5).

If no structure is detected then the message “No structure found” is displayed. The first explanation is that the sequences actually do not share a common structure. Unfortunately, there are other cases where caRNAc fails to infere correctly the structure. We shall see it in the next section (notes 1, 2 and 4).

RNAfamily (button <Visualize all foldings with RNAfamily>) allows the user to display all foldings at once. RNAfamily is a JAVA applet that is devoted to the visualization of multiple RNA sequences. It creates a plot using linear backbone representation. This is a concise representation that makes it convenient to compare several related structures at a glance. Each class of equivalent helices is assigned a color. RNAfamily includes the following functionnalities : zooming, scrolling, selecting a stem, displaying the nucleotidic content. Figure 3 gives a snapshot of RNAfamily.

It is also possible to download an archive storing all result files.

4. Notes

We give here a list of pitfalls and limitations of the method. When possible, we suggest alternative programs that may prove to be more appropriate within the context. We also give further hints and rules of thumbs to maximise information from caRNAc output.

1. The predicted structure contains large unpaired regions.

The philosophy of caRNAc is to privilege selectivity to sensibility. So it may happen that the prediction misses some stems. But these stems may be recovered afterwards with external programs, such as Mfold. This is the case of the structure inferred for the three tRNA sequences (we choose this example on purpose). On the one hand, the base pairs inferred by caRNAc are globally correct, but there is obviously one stem missing to form the cloverleaf structure (Figure 2). This corresponds to the loop from position 45 to 65. One the other hand, the results obtained with Mfold alone are very poor on that data set (Figure 4- A and B). But combining caRNAc and then Mfold gives a better result (Figure 4-C). For the combination of caRNAc and Mfold, download all results using the contraint format, and paste it in the Mfold web server in the box « constraint information ». This opportunity is also especially attractive with the kynetic-based Kinefold program that allows for pseudoknots (14). Kinefold supports the same format for the list of contraints.

2. The evolutionary distance is to small (more than 95% identity)**.**

The foundation of comparative analysis is that base pairings should be supported by compensatory mutations. It means that caRNAc is unlikely to find a complete structure if the sequences are very similar, because of the lack of mutations. In this context, it is

wishable to use alternative tools that derive a consensus structure from an alignment. For example, RNAalifold (12) is a good alternative to caRNAc for very similar sequences. The initial multiple alignment can be built with ClustalW (13). Another possibility is to enrich the data set with new sequences at greater evolutionary distance, using similarity searching programs.

3. The evolutionary distance is too high (less than 50% identity).

In this case, caRNAc is not guaranteed to recover a consensus structure because the search space is too wide. The solution here is to select few sequences with a higher conservation rate, if possible. As far as we know, no other program currently deals with such divergent sequences.

4. The structure may contain pseudoknots.

The algorithm of caRNAc is not designed for handling pseudoknots. If sequence are short (less than 70 bases), it might be a major source of error. In this particular context, it is more advisable to use a comparative pseudoknot-friendly program, such as comRNA (15). Note that comRNA is a time and space consuming program compared to caRNAc. It is limited to smaller data sets. For longer sequences, pseudoknots are usually not a problem. Kinefold may be used afterwards to complete the structure and identify potential pseudoknots (we already mentionned this opportunity in note 1).

5. Building an alignment the structures obtained with caRNAc

The multiple alignment tool allows the user to derive a structural alignment, taking into account both primary and secondary structures, from caRNAc output.

6. Discovering if the structure furnished by caRNAc is accurate

Of course, the accuracy rate of caRNAc is not 100%. Benchmark data show that predicted stems are usually correct, as soon as the number of stems is high enough to form a robust structure (like in Figure 2). In this context, some rare missing stems may be recovered afterwards (see note 1). The situation is more complex with sparse structures containing mostly unpaired regions. It is a difficult task to decide if the stems actually exist or if they are false positives occurring by chance. One solution is to compare the energy level of the sparse structure given by caRNAc with randomised equivalent data sets generated with `shuffle-aln.pl` (12). If the free energy is significantly lower with the intial data set, sequences are likely to share a common structure. We plan to integrate this functionality in the web server in the very near future.

The carRNAc website is under constant development. If you have any question, please contact the authors at **carnac@lifl.fr**.

References

1 Eddy S.R.(2001) Non-Coding RNA Genes and the Modern RNA World. *Nature Reviews Genetics*, **2**(12), 919-929

2 Eddy, S.R. (2004) How do RNA folding algorithms work. *Nature Biotechnology* **22**(11),1457-8.

3 Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31** (13), 3406-15

4 Zuker,M., Mathews,D.H. and Turner,D.H. (1999) Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. *RNA Biochemistry and Biotechnology*, J. Barciszewski & B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers

5 Hofacker, IL. (2003) Vienna RNA secondary structure server. *Nucleic Acids Research* **31** (13), 3429-31

6 Brown, J.W. and Ellis, J.C. (2005) Comparative analysis of RNA secondary structure: The 6S RNA. In Handbook of RNA Biochemistry {Wiley-VCH}, A. Bindereif, R. Hartmann, A. Schön, and E. Westhof, eds.

7 Gardner P., Wilm A. and Washietl S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research* **33**(8):2433-2439

8 Perriquet, O., Touzet, H., and Dauchet M. (2003) Finding the common structure shared by two homologous RNAs. Bioinformatics. **19** (1), 108-16

9 Touzet H. and Perriquet O. (2004) CARNAC: folding families of non coding RNAs. *Nucleic Acids Research* **142,** W142-5

10 Gardner, P. and Giegerich R. (2005) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **5** :140, doi:10.1186/1471-2105-5-140

11 Bruccoleri, R. And Heinrich, G. (1988) An improved algorithm for nucleic acid secondary structure display. *Comput. Appl. Biosci.* **4**,167-173

12 Hofacker I.L., Fekete M. and Stadler,P.F. (2002), Secondary Structure Prediction for

Aligned RNA Sequences. *Journal of Molecular Biology* **319**, 1059–1066

13 Higgins D., Thompson J., Gibson T.Thompson J.D., Higgins D.G., Gibson T.J.(1994).CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**,4673-4680

14 Xayaphoummine, A., Bucher, T. and Isambert, H. (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots, *Nucleic Acid Res*.**,** **33**, 605-610

15 Ji Y., Xu X. and Stormo G.D. (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, **20**(10),1591-1602

16 Höchsmann, M. Voss, B. and Giegerich R.(2004), Pure Multiple RNA Secondary Structure Alignments: A Progressive Profile Approach. *IEEE Transactions on Computational Biology and Bioinformatics* **1**(1)

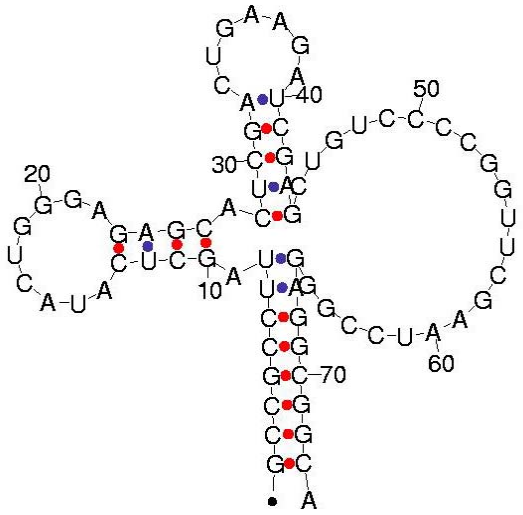
```
>Bacteriophage T4 Thr-tRNA
GCUGAUUUAGCUCAGUAGGUAGAGCACCUCACUUGUAAUGAGGAUGUCGGCGGUUCGAUUCCGUCAAUCAG
CA
>Yeast (S.cerevisiae) mitochondrial Phe-tRNA
GCUUUUAUAGCUUAGUGGUAAAGCGAUAAAUUGAAGAUUUAUUUACAUGUAGUUCGAUUCUCAUUAAGGGC
A
>Halobacterium volcanii Phe-tRNA.
GCCGCCUUAGCUCAUACUGGGAGAGCACUCGACUGAAGAUCGAGCUGUCCCCGGUUCGAAUCCGGGAGGCG
GCA
```

Figure 1 :Three tRNA sequences in FASTA format

A Connect notation (extract)

```
74 Halobacterium volcanii Phe-tRNA.
    1 G      0      2     73      1
    2 C      1      3     72      2
    3 C      2      4     71      3
    4 G      3      5     70      4
    5 C      4      6     69      5
   . . .
   65 G     64     66      0     65
   66 G     65     67      8     66
   67 A     66     68      7     67
   68 G     67     69      6     68
   69 G     68     70      5     69
   70 C     69     71      4     70
   71 G     70     72      3     71
   72 G     71     73      2     72
   73 C     72     74      1     73
   74 A     73      0      0     74
```

B JPEG file(or POSTSCRIPT file)



C List of contraints

```
F 1 73 8
F 10 26 4
F 28 44 5
```

D Bracket notation

```
GCCGCCUUAGCUCAUACUGGGAGAGCACUCGACUGAAGAUCGAGCUGUCCCCGGUUCGAAUCCGGGAGGCGGCA
((((((((.((((.........)))).(((((.......)))))....................)))))))).
```

Figure 2 : Example of output formats for the structure predicted for the third tRNA sequence of Figure 1. This structure is composed of three helices.

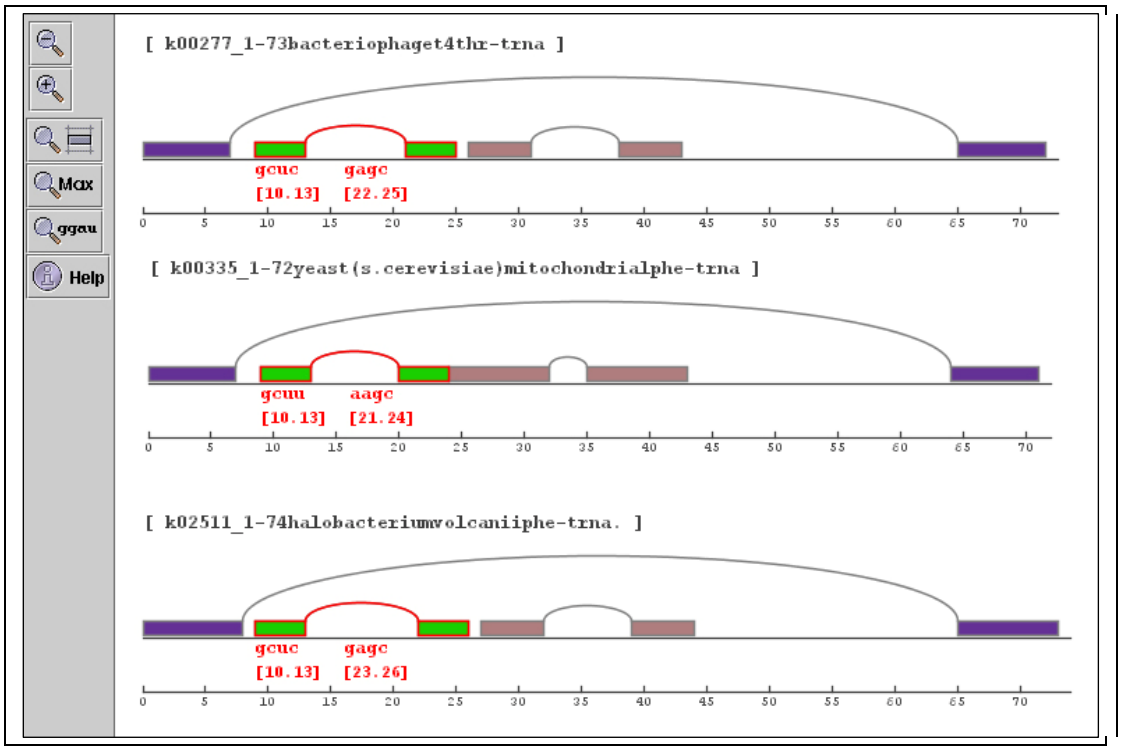


Figure 3 : Snapshot of RNAfamily

It shows the common structure for the three tRNA sequences of Figure 1. Clicking on a stem displays the nucleotidic content of the stem (here the green stems). Clicking on <ggau> in the left menu displays the nucleotidic content of all sequences.

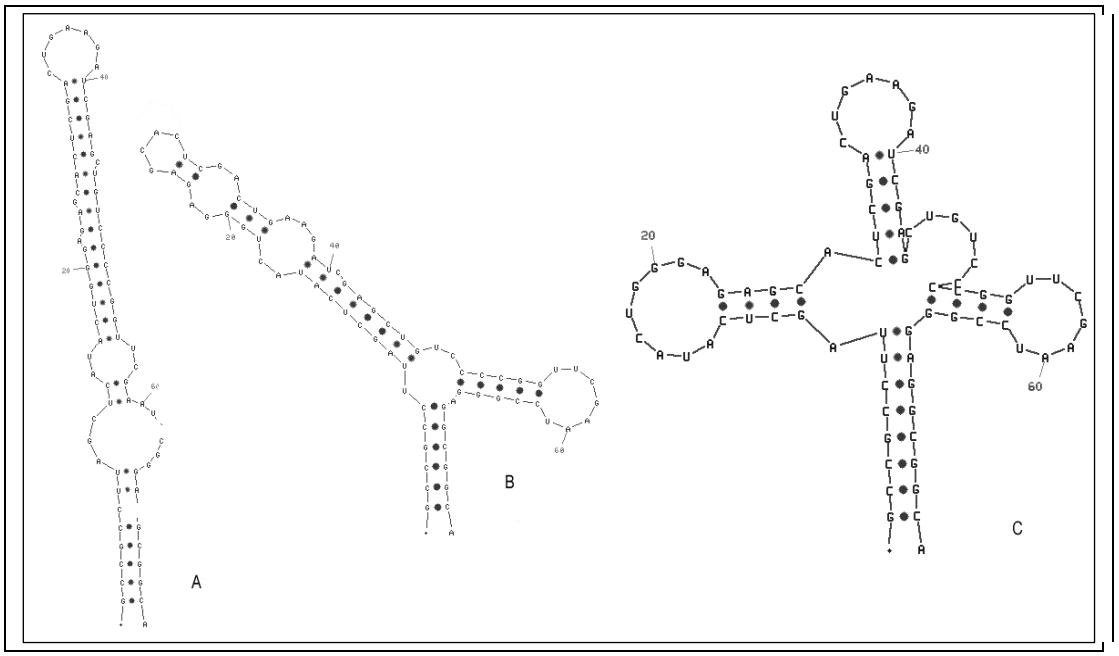


Figure 4 :Example of combination of caRNAc and Mfold

The first two structures (A and B) are the best two results given by Mfold alone for the third tRNA sequence of Figure 1. The last structure (C) is obtained with Mfold using constraint information produced by caRNAc (file C in Figure 2). In this case, Mfold correctly completes the structure and identifies the fourth stem that is missing in caRNAc output. This leads to the typical clover leaf structure (the acceptor stem is on the top).