W10–W13 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 15 May 2008*
doi:10.1093/nar/gkn278

R-Coffee: a web server for accurately aligning noncoding RNA sequences

Sébastien Moretti[1], Andreas Wilm[2], Desmond G. Higgins[2], Ioannis Xenarios[1] and Cédric Notredame[3,*]

[1]Swiss Institute of Bioinformatics (SIB), Quartier Sorge - Genopode, UNIL, CH-1015 Lausanne, Switzerland, [2]The Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Ireland and [3]Centre for Genomic Regulation (CRG), UPF, Barcelona, Spain

Received January 30, 2008; Revised April 16, 2008; Accepted April 26, 2008

ABSTRACT

The R-Coffee web server produces highly accurate multiple alignments of noncoding RNA (ncRNA) sequences, taking into account predicted secondary structures. R-Coffee uses a novel algorithm recently incorporated in the T-Coffee package. R-Coffee works along the same lines as T-Coffee: it uses pairwise or multiple sequence alignment (MSA) methods to compute a primary library of input alignments. The program then computes an MSA highly consistent with both the alignments contained in the library and the secondary structures associated with the sequences. The secondary structures are predicted using RNAplfold. The server provides two modes. The slow/accurate mode is restricted to small datasets (less than 5 sequences less than 150 nucleotides) and combines R-Coffee with Consan, a very accurate pairwise RNA alignment method. For larger datasets a fast method can be used (RM-Coffee mode), that uses R-Coffee to combine the output of the three packages which combines the outputs from programs found to perform best on RNA (MUSCLE, MAFFT and ProbConsRNA). Our BRAliBase benchmarks indicate that the R-Coffee/Consan combination is one of the best ncRNA alignment methods for short sequences, while the RM-Coffee gives comparable results on longer sequences. The R-Coffee web server is available at http://www.tcoffee.org.

INTRODUCTION

The increasing interest in small noncoding RNAs (ncRNAs) has led to a renewed interest in the development of computational methods dedicated to the analysis of this important class of molecules. Multiple alignment of ncRNAs is a basic ingredient of many homology-based analyses (1), including gene prediction (2), consensus structure prediction (3) and phylogeny reconstruction (4). While closely related sequences can be treated like regular DNA sequences, more distantly related ncRNA sequences (e.g. <70% sequence identity) require more sophisticated methods to align them correctly (5,6). To maximize accuracy, RNA comparison models must take into account secondary structures, which are mainly formed by Watson–Crick base pairs between residues in a sequence. A number of well established thermodynamics-based methods can produce reasonably accurate *ab-initio in-silico* predictions from single sequences alone [e.g. RNAfold (7) or Mfold (8)]. One remarkable property of these structures is their propensity to evolve through compensated base-pair mutations i.e. by joined mutations, which maintain structure, but reduce sequence identity. This process has two important consequences for the *in-silico* analysis. On the one hand, they hamper the homology-based analysis because they make it possible for sequences to diverge rapidly while maintaining the same fold. On the other hand, they result in a very useful signal in multiple alignments (if they are accurate) where columns exhibiting a high level of covariation can be expected to be base-paired.

Properly incorporating these characteristics would be the best way to design the search strategies needed to discover new ncRNA families, or to extend the scope of existing ones. An ideal search method would determine the structure while carrying out the alignment. Doing so in an efficient manner remains, however, very challenging, and it is often more efficient to determine a putative structure beforehand and use this structure later on to scan existing genomes or targets. Of course, the determination of the structure is itself a problem, often addressed through the use of multiple sequence alignments (MSAs). In practice, any alignment method can be used, but the

*To whom correspondence should be addressed. Tel: +34 93 316 02 71; Fax: +34 93 316 00 99; Email: cedric.notredame@crg.es

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

better the agreement between the alignment and the structure, the better the final model. The problem of assembling RNA MSAs has recently received a lot of attention and several methods have been proposed over the last few years to address this key issue (9–14). Most of these methods are more or less greedy variations of the original Sankoff algorithm (15). Published some 20 years ago, this algorithm defines the Holy Grail of multiple RNA analysis: the simultaneous folding and alignment of a group of homologous sequences, combined in such a way that the alignment and the structure prediction may benefit by informing and guiding one another. Unfortunately, merging the dynamic programming recursions for folding and aligning the sequences results in prohibitive time and space complexity that make the algorithm unusable in practice. Only a few simplified and constrained implementations of this algorithm exist, one being Consan (16), a program able to fold and align pairs of relatively short sequences.

We have recently developed a novel algorithm named R-Coffee (17), an extension of the original T-Coffee algorithm (18). In T-Coffee, a collection of alignments is assembled in order to build a library containing one entry (constraint) for each pair of residues observed to be aligned in the input alignments. These constraints do not have to be compatible with one another and the purpose of the algorithm is to build a final MSA as consistent as possible with the library. This library can be filled up using alignments from any source, including global and local pairwise alignments, as used in the original T-Coffee, structure-based sequence alignments [Expresso (19)] or even the output of several MSA packages [M-Coffee (20)]. R-Coffee works under a slightly modified principle and contains two novel components designed for the integration of structural information. First of all, the predicted (or known) secondary structures of the sequences are used to complement the library. For instance, if two nucleotides from one side of a stem are aligned in the primary library, then the corresponding base-paired nucleotides from the other side of the stem are automatically added to the library, whether they were observed to be aligned in the input alignments or not. The second component is the evaluation itself. The cost for matching two residues is now computed by taking into account both the score for matching these two residues and the score for matching their corresponding base-paired residues, if they both are part of a secondary structure. These two components add very little overhead to the computational cost but make it straightforward to incorporate structural information within the process. This association of an RNA sequence with a predicted secondary structure constitutes a novel extension of the template-based alignment principle recently reviewed in (21).

A key property of R-Coffee is its ability to be combined with any existing method. In the original description (17), we showed that R-Coffee was able to improve on all existing MSA methods, as judged by benchmarking on the BRAliBase reference dataset (5). Our results indicate that the combination of pairwise Consan alignments (16) with R-Coffee produces the most accurate RNA alignments of all methods tested on BRAliBase. While this protocol is, due to the computational complexity of Consan, restricted to relatively short sequences, our results also show that results of comparable albeit slightly lower accuracy can be obtained by combining regular sequence aligners (MAFFT, MUSCLE and ProbConsRNA) by means of R-Coffee in a protocol reminiscent of M-Coffee (20).

The purpose of this new server is to make it possible for occasional users to access the full range of possibilities of R-Coffee without having to set up all the required packages, some of which can be challenging to install and run. The server is fully integrated into the T-Coffee Vital-IT server, a powerful framework that has been serving the T-Coffee application for several years.

METHODS

R-coffee algorithm

R-Coffee follows the same principle as T-Coffee to compute its MSAs. The main difference between T-Coffee and R-Coffee is the use of the R-Coffee scoring scheme and the reliance of the process on RNAplfold (22) predicted secondary structures. In R-Coffee, the primary library can be assembled using two approaches. In one approach a combination of standard MSA methods like MAFFT (23), ProbConsRNA (24) and MUSCLE (25) can be used. The second approach combines RNA specific pairwise alignments produced by Consan (16), a highly accurate pairwise RNA alignment method.

The algorithm starts by building all the pairwise or multiple alignments using the selected methods (Figure 1). These alignments are then turned into a so-called primary library of pairwise residue scores just as in default T-Coffee. The library is meant to be used as a position-dependent substitution matrix, so that the score for matching two residues does not depend on the specific residue type, but rather on the relationships as described in the library. Then all possible local base pairing are computed by means of RNAplfold (using the default folding span of 100 nucleotides). While the T-Coffee original library only contains pairs of nucleotides observed in the pairwise alignments, the R-Coffee library also contains extra pairs implied by the secondary structure prediction. For instance if two nucleotides W1 and W2 were found aligned, and if each of these nucleotides is predicted to be part of a Watson and Crick secondary structure (W1 pairing with C1 and W2 with C2), then the pair C1–C2 will be added to the library, or its weight will be increased if it is already part of the library.

The final score for matching two residues W1–W2 is obtained through the extension process. In T-Coffee this score is set equal to the sum of the weight of all the alignments including this pair, either directly or indirectly through a third sequence. In R-Coffee, a similar measure is used but the final score is set to be the maximum of the considered pair against that of the corresponding pair induced by the secondary structure (i.e. the maximum of W1–W2, C1–C2). This heuristic ensures that as much information as possible from the predicted structures finds its way into the final alignment. Given the R-Coffee extended library and the R-scoring scheme, the multiple

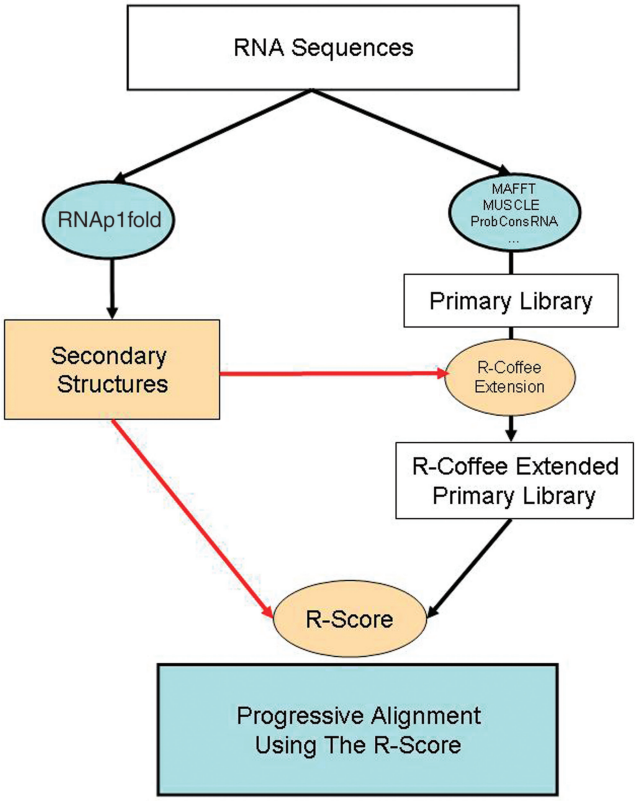

Figure 1. R-Coffee workflow. Secondary structures for each single sequence are predicted by means of RNAplfold. The primary alignment library of pairwise residue scores is computed either using Consan (R-Coffee mode: slow/accurate) or by using a combination of MSA programs (R-Coffee mode: fast/approximate). This library is then extended by the R-Coffee extension, which adds base pairs of aligned residues to the library. The resulting extended library and the so-called R-Score, which again takes the given base pairs into account, are then used to compute a progressive alignment in the same way as default T-Coffee.

alignment is then assembled using the standard progressive alignment strategy of T-Coffee, aligning the sequences two by two with the Needleman and Wunsch algorithm (26) while following the order indicated by a guide tree.

Server and available methods

The R-Coffee server can be accessed on http://www.tcoffee.org. We have provided a regular and an advanced form. The regular form is very straightforward to use. The advanced form gives more control over how to create the primary library.

Regular form

In order to use the regular form users must copy and paste their sequences or upload a sequence file. Once the job has been computed the user will be informed by Email, if an address was entered. This is recommended for long jobs, i.e. when using the slow/accurate option. The Email will contain a link to the results, which will be kept on the server for a week.

Sequences must be uploaded in FASTA format. While the full IUPAC code is, in principle, supported, it can

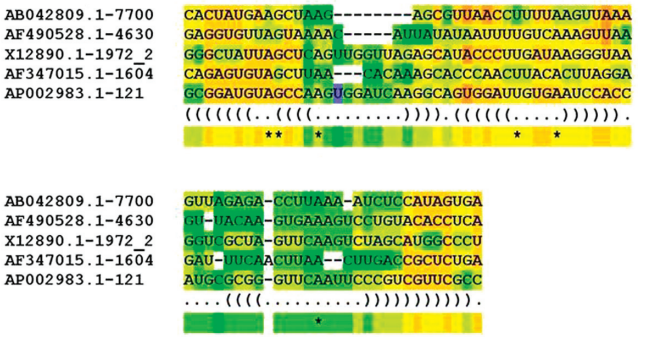

Figure 2. Example alignment output. This colored output was generated by aligning tRNA-aln4 from the BRAliBase benchmark using R-Coffee slow/accurate mode. Colors indicate the consistency of aligned residues with the primary library alignments and the predicted structures: blue to green means low consistency; yellow to red means good consistency. The dot bracket notation below the alignment indicates the consensus structure (predicted with RNAalifold) and was added afterwards.

result in very long run times and users are advised to submit sequences using the canonical bases, as much as possible. RNA can be pasted in its DNA-like form (T instead of U) although the alignments will always be reported back as RNA sequences. The maximum number of sequences and their maximum length depend on the selected options. By default, the fast/approximate mode makes it possible to align up to 50 sequences, 1000 nucleotides long. The slow/accurate mode is limited to 5 sequences of 150 nucleotides.

The slow/accurate mode runs R-Coffee (version 5.68) with Consan (version 1.2) and uses the following t_coffee command line:

t_coffee -seq < RNA seq > -special_mode rcoffee_slow_accurate

The fast/approximate mode uses the same version of R-Coffee to combine multiple alignments produced with MAFFT (version 6.240), MUSCLE (version 3.6) and ProbConsRNA (version 1.10). It runs the following command line:

t_coffee -seq < RNA seq > -special_mode rcoffee_fast_approximate

For both modes, the R-Coffee server generates RNA secondary structure predictions on the fly using RNAplfold (ViennaRNA package, version 1.6.5).

R-Coffee produces the same type of output as T-Coffee, including Phylip, FASTA and MSF formatted alignments. The server also produces a color-coded output (Figure 2) where each nucleotide receives a color indicating the consistency of its alignment with the primary library. Previous analyses made on protein alignments suggest a good correlation between this consistency coding and alignment accuracy.

Advanced form

The advanced form provides the user with a larger number of alternatives e.g. the possibility of selecting the

alignment methods to be combined. Two types of methods are provided: the pair_wise methods that incorporate pairwise alignments within the library and the multiple methods that fill-up the library with multiple alignments of the provided sequences. The advanced mode also provides more output options. It is worth stressing that the incorporation of the consan_pair method to the computation automatically restricts the maximum dataset size to a maximum of 5 sequences, 150 nucleotides long.

Computational platform

R-Coffee is integrated within the Vital-IT web server framework. Every R-Coffee component is compiled for all the Vital-IT cluster architectures and thus can run on more than 200 CPUs. Nevertheless, for very long jobs, or large-scale usage, users are encouraged to install R-Coffee locally. R-Coffee is part of the latest T-Coffee distribution, which is available from http://www.tcoffee.org. T-Coffee is open-source freeware and distributed under the GNU public license.

CONCLUSION AND FUTURE WORK

The R-Coffee web server allows biologists to create high-quality multiple RNA alignments that automatically take predicted secondary structure into account. Depending on the length and number of sequences which have to be aligned the user can choose between the slow and accurate mode, which has been shown to be the top performing program on BRAliBase, and the fast and approximate mode which is fast enough to align larger datasets, while still creating high-quality alignments comparable to other multiple RNA structure alignment programs. Along with the Pmmatch/Pmmulti and the Murlet web server, R-Coffee is one of the few online services for the alignment of multiple RNA sequences taking structure into account. In contrast to other services it is able to align much longer and more sequences (up to 50 sequences with 1000 nucleotides). Future work will include support for user, provided structure/base-pair libraries and automatic consensus structure prediction using external tools, such as RNAalifold. R-Coffee will also be progressively mirrored on other T-Coffee web servers.

ACKNOWLEDGEMENTS

This work was partly supported by funding from the Science Foundation, Ireland. The development of the server was supported by the Vital-IT framework and by the European Union (ICGR-SIB FP6-026204). C.N. thanks the CRG for active support. Funding to pay the Open Access publication charges for this article was provided by Science Foundation Ireland.

Conflict of interest statement. None declared.

REFERENCES

1. Freyhult,E.K., Bollback,J.P. and Gardner,P.P. (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, **17**, 117–125.
2. Meyer,I.M. (2007) A practical guide to the art of RNA gene prediction. *Brief. Bioinform.*, **8**, 396–414.
3. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
4. Hudelot,C., Gowri-Shankar,V., Jow,H., Rattray,M. and Higgs,P. (2003) RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol. Phyl. Evol.*, **28**, 241–252.
5. Gardner,P.P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
6. Wilm,A., Mainz,I. and Steger,G. (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 19.
7. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
8. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
9. Xu,X., Ji,Y. and Stormo,G.D. (2007) RNA sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics*, **23**, 1883–1891.
10. Kiryu,H., Tabei,Y., Kin,T. and Asai,K. (2007) Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, **23**, 1588–1598.
11. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.
12. Dalli,D., Wilm,A., Mainz,I. and Steger,G. (2006) STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics*, **22**, 1593–1599.
13. Hofacker,I.L., Bernhart,S.H.F. and Stadler,P.F. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.
14. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
15. Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
16. Dowell,R. and Eddy,S. (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, **7**, 400.
17. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-Coffee: a method for multiple alignment of non-coding RNAs. *Nucleic Acids Res.*, [Epub ahead of print; April 17].
18. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
19. Armougom,F., Moretti,S., Poirot,O., Audic,S., Dumas,P., Schaeli,B., Keduas,V. and Notredame,C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.*, **34**, W604–W608.
20. Wallace,I.M., O'Sullivan,O., Higgins,D.G. and Notredame,C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.*, **34**, 1692–1699.
21. Notredame,C. (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput. Biol.*, **3**, e123.
22. Bernhart,S.H., Hofacker,I.L. and Stadler,P.F. (2006) Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.
23. Katoh,K., Kuma,K.-I., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
24. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
25. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
26. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.