Nucleic Acids Research Advance Access published June 27, 2014

doi: 10.1093/nar/gku459

SARA-Coffee web server, a tool for the computation of RNA sequence and structure multiple alignments

Paolo Di Tommaso[1,2], Giovanni Bussotti[3], Carsten Kemena[4], Emidio Capriotti[5], Maria Chatzou[1,2], Pablo Prieto[1,2] and Cedric Notredame[1,2,*]

[1]Comparative Bioinformatics, Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain, [3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [4]Evolutionary Bioinformatics Group, Institute for Evolution and Biodiversity, University of Münster, Hüfferstraße 1, 48145 Münster, Germany and [5]Division of Informatics, Department of Pathology, University of Alabama at Birmingham, 35249 Birmingham (AL), USA

Received January 31, 2014; Revised May 07, 2014; Accepted May 8, 2014

ABSTRACT

This article introduces the SARA-Coffee web server; a service allowing the online computation of 3D structure based multiple RNA sequence alignments. The server makes it possible to combine sequences with and without known 3D structures. Given a set of sequences SARA-Coffee outputs a multiple sequence alignment along with a reliability index for every sequence, column and aligned residue. SARA-Coffee combines SARA, a pairwise structural RNA aligner with the R-Coffee multiple RNA aligner in a way that has been shown to improve alignment accuracy over most sequence aligners when enough structural data is available. The server can be accessed from http://tcoffee.crg.cat/apps/tcoffee/do:saracoffee.

INTRODUCTION

Phylogeny reconstruction and homology-based annotation are two of the most common modeling procedures in biology. Both of them require the assembly of accurate multiple sequence alignments (MSA). In this work, we introduce a web server dedicated to 3D structure-based multiple RNA sequence alignment using the SARA-Coffee package (1). SARA-Coffee allows the combination of sequences with and without known tertiary structure. It is a suitable companion tool for any modeling technique that can benefit from a structurally accurate ribonucleic acid (RNA) MSA. This includes construction of profile stochastic context free grammar (SCFG) using packages like infernal (2,3).

Accurately aligning RNA sequences is especially important in a context where recent developments in genomics have fueled the detection of thousands of expressed loci and challenged the long held view that non-coding RNA (ncRNA) functions were supported by a much smaller number of genes than their protein-coding counterparts. Between 2008 and 2014, the ENSEMBL non-coding RNA (ncRNA) gene count has grown from 5732 (October 2008) to 22 643 (January 2014), thus overtaking proteins. It remains unclear, however, which proportion of these genes simply results from spurious transcription. One should also bear in mind that the term ncRNA encompasses a very heterogeneous gene population, including about 14 000 Long ncRNAs—of mostly unknown function—and a bit less than 9000 short ncRNAs often better biologically characterized. This last group includes snoRNAs, microRNA precursors, transfer RNA (tRNA) and in general most of the highly structured ncRNAs catalogued in the RNA families database (RFAM) (4). SARA-Coffee has been specifically designed for this category of ncRNA genes.

SARA-Coffee's main characteristic is to combine sequence and structure pairwise alignment methods into a unified multiple sequence alignment, using SARA (5) as a 3D structure pairwise alignment engine. When used in pure 3D structural aligner mode SARA-Coffee is limited by the small number of available RNA PDB structures. These merely constitute ∼3% of all PDB entries (2'941 out of 99'775 in March 2014) and their pace of determination remains significantly slower than that of proteins—with a doubling time of ∼7 years, as compared to 5 years for proteins. Yet, in its sequence/structure hybrid mode, SARA-Coffee can be used to combine available 3D structural information with the very large number of novel uncharacterized ncRNA sequence reported by genome sequencing projects.

The problem of aligning RNA sequences is rather complex. Even though RNA molecules often contain evolutionary conserved secondary structures, their primary structure signal is rarely as strong as the one resulting from the protein three-letters meta-alphabet. As a consequence, higher

*To whom correspondence should be addressed. Tel: +34 933160271; Fax: +34 933160099; Email: cedric.notredame@crg.eu.

levels of sequence identity are required to infer homology––60% as compared to 25% in proteins (6–9). Integrating additional information, e.g. secondary structure signal, into the alignment process is possible but can be computationally expensive. The first attempt to do so was described by Sankoff and later improved using profile Stochastic Context Free Grammar (SCFG) (2). Over the years, many methods have been reported for this purpose; a non-exhaustive list includes Foldalign (10), Stemloc (11), Consan (11), LocARNA (12), R-Coffee (13,14) and MAFFT (15). All of them constitute an attempt to capture the secondary structure signal contained in RNA molecules in order to build more informative sequence alignment models. These models can then be passed to SCFG profile building tools in order to build co-variation models like the ones used by RFAM.

These multiple alignment heuristics only address the issue of aligning sequences using experimental or predicted secondary structure information. In some recent work, we have shown that one can also combine sequence and experimental 3D structural information (1) into an MSA. This process requires the ability to superpose RNA 3D structures, a problem similar in nature to protein structure comparison and for which several pairwise comparison tools have been developed, including SARA (5), DIAL (16), iPARTS (17), ARTS (18) and R3D Align (19). Our main motivation when developing SARA-Coffee has been to turn these pairwise 3D structure aligners into a method able to deliver 3D-based multiple sequence alignments. We did so by integrating SARA within the T-Coffee consistency framework. It must be stressed that this approach is generic enough to be applied to any of the above-mentioned RNA 3D pairwise aligners.

In a previous study, SARA-Coffee was extensively validated on a purpose built reference dataset named BralDART (1) made of 41 RNA families containing between 4 and 71 members with known 3D structures. This dataset was assembled so as to only include high quality X-ray structures and exclude any discontinuous structure, on which the algorithm is expected to perform poorly. Our benchmarks indicate that SARA-Coffee is significantly more accurate than alternative sequence based methods, even those using predicted secondary structures. Overall, SARA-Coffee was reported to be over 10% points more accurate than primary structure based aligners, judging from its capacity to properly align pairs of interacting residues. On dense secondary structures—in which 70% or more of the nucleotides form Watson and Crick base pairs—SARA-Coffee outperforms all tested methods, including the ones using predicted secondary structure information. It is about 3% points more accurate than MXSCARNA (20), the second best method in this benchmark. Detailed structural analysis using a distance-Root Mean Square Deviation (dRMSD) measurement also indicates that SARA-Coffee produces alignments significantly superior to all methods tested in the study. Overall SARA-Coffee alignments have a dRMSD 10 to 20% lower than alternative models (i.e. 4.53 Å against 5.11 Å for LocARNA, the next best method in terms of 3D modeling).

ALGORITHM

SARA-Coffee (1) produces 3D structure based multiple RNA sequence alignments. Its algorithm can be described as a combination between SARA (5), a pairwise RNA structural aligner and R-Coffee (14), a T-Coffee (22) based multiple RNA sequence aligner. The algorithm is described extensively in (1) and its general flow can be summarized as follows:

By default, SARA-Coffee takes as input a set of RNA sequences with known 3D structures. The first step of the algorithm involves aligning all possible pairs of the provided sequences using SARA. The result is an alignment library populated with SARA' 3D-structure based pairwise sequence alignments. In practice, the alignment library is a list of all the residue pairs found aligned in any of the compiled pairwise alignments. The second step involves extracting contact information from each input structure using the -p mode of x3dna/find_pairs (21) that reports all base-pairs, including the non-canonical and higher-order (3+). This contact information is then used to extend the alignment library so that its alignments become compatible with 2D and 3D contacts. This is achieved by adding pairs of aligned residues that are implied by the contacts but missing from the library. For instance, let us consider a contact between residues XY in sequence A and another contact between residues WZ in sequence B. If X and W are found aligned in the library, then the contact based extension will involve adding the pair YZ to the library, thus ensuring full compatibility between the library and the contact. If in the library, Y (or Z) are already aligned to other residues, this process will introduce incompatibilities that will be resolved through consistency analysis when incorporating the sequences into the final MSA. Once extended using contact information, the library can be fed to the default T-Coffee algorithm. This contact based extension is the main feature of the R-Coffee algorithm.

T-Coffee uses a tree based progressive alignment algorithm. It starts by estimating the similarity between every pair of sequences counting words of size four and then uses neighbor joining to turn this distance matrix into a binary guide tree. Sequences are then incorporated into the MSA following the guide tree. The guide tree being binary, its resolution involves aligning at each node either two sequences, a sequence and a profile or two profiles, until reaching the root where the full MSA is resolved. At each node the pairwise alignments are computed using the Gotoh dynamic programming algorithm. The main characteristic of T-Coffee is the ability to use the library described above instead of a standard substitution matrix. When using the library, the cost for aligning two symbols is set to be equal to the number of time these symbols are found aligned in the library, either directly or by combining any two pairs of aligned residues (i.e. the pair XW of aligned residues may be supported by a combination of the two library pairs XK and KW, K being a residue from a third sequence). This consistency analysis helps deciding between conflicting library pairs that may result from the secondary structure based extension. No gap penalty scheme is needed at that stage.

This same algorithm also makes it possible to combine structures and sequences. When doing so, the library is built in a similar way, but in that case SARA is only applied onto sequence pairs having both an experimental 3D structure. In all other cases, a pair-HMM (proba_pair) adapted from Probcons is used to produce the pairwise comparisons and to populate the library with residue pairs having a high alignment posterior probability (22). In the next step, the contact list for sequences with no experimental 3D structures is replaced with a secondary structure prediction, as provided by RNAplfold from the Vienna package (23). The rest of the algorithm (extension and alignment computation) is identical.

SARA-COFFEE WEB SERVER

The SARA-Coffee web server is part of the T-Coffee web platform; its access is free and unrestricted, with no login procedure. The server is accessible from http://tcoffee.crg.cat/apps/tcoffee/do:saracoffee with any standard Internet browser (Mozilla Firefox 5+, Google Chrome, Internet Explorer 8+, Safari 6+ and Opera 11+). Results can be retrieved from the web server or received by email if requested. Anonymous jobs can nonetheless remain available from the submission terminal thanks to a cookies cache procedure. Results are kept on the web server for a month after computation and are assigned a permanent URL during this time.

Input

Sequences must be provided in FASTA format. Sequences with a known PDB structure must be named after their PDB identifier, including the chain index separated by an underscore character (i.e. >*PdbID_chain*), other sequences can be given arbitrary names. White spaces are forbidden, all sequences are required to have a different name and the provided primary sequences must match the PDB SEQRES field. The input interface also gives access to an advanced mode that makes it possible to control several output options, including format, case, residue numbering, output order and interleaved format block length. Once submitted, the server runs the default SARA-Coffee onto the provided dataset. Results are returned via the same interface but can also be accessed via the history.

Computation

The main limiting step of SARA-Coffee is the computation of SARA pairwise 3D-structure based sequence alignment. For two tRNA structures, SARA requires about 5 s. On a dataset of 71 tRNA with known 3D structures, the server takes about 4 h, it requires about 2 min on a dataset containing 5 tRNAs only like the one provided as a test.

Output

The result page displays the following items in order:

(i) MSA: Shows the resulting interleaved MSA. This graphic is the HTML rendering of the file **.score_html*, available for download from the next section. The MSA is colored according to the T-Coffee reliability scheme (24) where red and orange bits correspond to alignment portions for which there is a high consistency within the SARA-Coffee pairwise library, while lighter bits (green, yellow and blue) correspond to the less well supported portions, expected to be less accurately aligned.
(ii) Citation: link to the relevant publications when using this server.
(iii) The result files produced by the submitted SARA-Coffee alignment. All files can be downloaded as a single zip file by clicking the 'download all' link. They can also be automatically imported into the user Dropbox account, if available.
(iv) Info: some information about the executed job.
(v) Replay: this feature allows the users to re-run the job while modifying some input options or data.
(vi) Feedback: this feature encourages users to provide some feedback via social media.

CONCLUSION

We describe the SARA-Coffee web server, a web based tool able to incorporate 3D structure information within RNA multiple sequence alignments by combining sequence and structural information. SARA-Coffee has been shown to produce accurate alignment as judged from structural analysis. This server makes it possible to combine user's data with publicly available RNA 3D structures, so as to obtain the required models. Future improvements will involve the possibility of uploading user's defined 3D models as well as new output formats providing a local visualization of structural similarity.

ACKNOWLEDGMENT

We thank the reviewers, especially Sean Eddy, for very constructive suggestions, observations and critcisms.

FUNDING

Plan Nacional [BFU2011-28575 to C.N. and P.D.]; Quantomics [KBBE-2A 222664]; Center for Genomics Regulation (CRG); La Caixa Fellowship program (to M.C.); Spanish Ministry of Economy and Competitiveness [BES-2012–051918 to P.B.]. EMBL Interdisciplinary Postdoc (EIPOD) under Marie Curie Actions (COFUND) (to G.B.); Department of Pathology at the University of Alabama at Birmingham (to E.C.). Funding for open access charge: Plan Nacional [BFU2011-28575 to C.N. and P.D.]; Quantomics [KBBE-2A 222664]; Center for Genomics Regulation (CRG); La Caixa Fellowship program (to M.C.); Spanish Ministry of Economy and Competitiveness [BES-2012–051918 to P.B.]. EMBL Interdisciplinary Postdoc (EIPOD) under Marie Curie Actions (COFUND) (to G.B.); Department of Pathology at the University of Alabama at Birmingham (to E.C.); Computational resources are provided by the Center for Genomic Regulation (CRG).
Conflict of interest statement. None declared.

REFERENCES

1. Kemena,C., Bussotti,G., Capriotti,E., Marti-Renom,M.A. and Notredame,C. (2013) Using tertiary structure for the computation of highly accurate multiple RNA alignments with the SARA-Coffee package. *Bioinformatics*, **29**, 1112–1119.
2. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
3. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
4. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, D226–D232.
5. Capriotti,E. and Marti-Renom,M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, i112–i118.
6. Capriotti,E. and Marti-Renom,M.A. (2010) Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*, **11**, 322.
7. Abraham,M., Dror,O., Nussinov,R. and Wolfson,H.J. (2008) Analysis and classification of RNA tertiary structures. *RNA*, **14**, 2274–2289.
8. 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.
9. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
10. Havgaard,J.H., Torarinsson,E. and Gorodkin,J. (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. *PLoS Comput. Biol.*, **3**, 1896–1908.
11. Holmes,I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, **6**, 73.
12. 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.
13. Taly,J.F., Magis,C., Bussotti,G., Chang,J.M., Di Tommaso,P., Erb,I., Espinosa-Carrasco,J., Kemena,C. and Notredame,C. (2011) Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures. *Nat. Protoc.*, **6**, 1669–1682.
14. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**, e52.
15. Katoh,K. and Toh,H. (2008) Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*, **9**, 212.
16. Ferre,F., Ponty,Y., Lorenz,W.A. and Clote,P. (2007) DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Res.*, **35**, W659–W668.
17. Wang,C.W., Chen,K.T. and Lu,C.L. (2010) iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Res.*, **38**, W340–W347.
18. Dror,O., Nussinov,R. and Wolfson,H. (2005) ARTS: alignment of RNA tertiary structures. *Bioinformatics*, **21**(Suppl.2), ii47–ii53.
19. Rahrig,R.R., Leontis,N.B. and Zirbel,C.L. (2010) R3D align: global pairwise alignment of RNA 3D structures using local superpositions. *Bioinformatics*, **26**, 2689–2697.
20. Kiryu,H., Tabei,Y., Kin,T. and Asai,K. (2007) Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, **23**, 1588–1598.
21. Lu,X.J. and Olson,W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protoc.*, **3**, 1213–1227.
22. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
23. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.: AMB*, **6**, 26.
24. Chang,J.M., DiTommaso,P. and Notredame,C. (2014) TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol. Biol. Evol.*, **31**, 1625–1637.