W412–W415 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl312

The ARTS web server for aligning RNA tertiary structures

Oranit Dror*, Ruth Nussinov[1,2] and Haim J. Wolfson

School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel, [1]Sackler Inst. of Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel and [2]Basic Research Program, SAIC-Frederick, Center for Cancer Research, Nanobiology Program, NCI-Frederick, Building 469, Room 151, Frederick, MD 21702, USA

Received February 14, 2006; Revised March 6, 2006; Accepted April 11, 2006

ABSTRACT

RNA molecules with common structural features may share similar functional properties. Structural comparison of RNAs and detection of common substructures is, thus, a highly important task. Nevertheless, the current available tools in the RNA community provide only a partial solution, since they either work at the 2D level or are suitable for detecting predefined or local contiguous tertiary motifs only. Here, we describe a web server built around ARTS, a method for aligning tertiary structures of nucleic acids (both RNA and DNA). ARTS receives a pair of 3D nucleic acid structures and searches for a priori unknown common substructures. The search is truly 3D and irrespective of the order of the nucleotides on the chain. The identified common substructures can be large global folds with hundreds and even thousands of nucleotides as well as small local motifs with at least two successive base pairs. The method is highly efficient and has been used to conduct an all-against-all comparison of all the RNA structures in the Protein Data Bank. The web server together with a software package for download are freely accessible at http://bioinfo3d.cs.tau.ac.il/ARTS.

INTRODUCTION

In recent years there is a fast growing interest in RNA molecules. This stems from the groundbreaking discovery that RNA is not solely a carrier of genetic information, but a key player in a wide range of essential processes within the cell, such as protein synthesis and transport, RNA processing and splicing, gene silencing, and chromosome replication (1–3). RNA is also involved in many pathological processes, like cancerous tumors and retroviral infections as AIDS. Much like proteins, understanding the functions of these active RNA molecules requires methods for analyzing their tertiary structures. However, in contrast to the wide range of 3D structure-based approaches available for proteins (4), a similar field for RNA is only now emerging.

Many methods for structure analysis of RNA have been developed to work at the secondary structure level, that is the level of base pairing (5). In the absence of RNA tertiary structures, such methods provide an excellent starting point for exploring RNA structures. However, their inherent limitation is that they are incapable of predicting and annotating tertiary interactions. These interactions are formed between secondary structure elements and are crucial for establishing the global fold of an RNA (6,7). Fortunately, in the past few years both the number and size of solved RNA tertiary structures has dramatically increased. This has given rise to various computational tools for 3D structural analysis. A variety of methods are available for analyzing and classifying nucleotide conformations and spatial base interactions (8–13). Several other methods have been suggested for measuring the similarity between larger RNA structures, but require the structures to be with the same number of nucleotides and with a predefined correspondence between them (14–16). Fewer methods are available for locating small predefined motifs in larger structures (15,17). These methods are useful for finding new examples of known motifs, but are incapable of discovering novel ones. To date, the problem of identifying a priori unknown common substructures is only partially addressed by a few methods for recognizing recurring 3D contiguous fragments (18,19). Thus, there is a great need for new approaches.

Herein, we present a web server built around the ARTS method [http://bioinfo3d.cs.tau.ac.il/ARTS (20)] for aligning 3D nucleic acid structures. Compared with the current very few comparison tools available for tackling this task, ARTS is suitable for identifying a priori unknown common substructures that may not necessarily be contiguous. The

*To whom correspondence should be addressed. Tel: +972-3-640 5395; Fax: +972-3-640 6476; Email: oranit@post.tau.ac.il common substructures can be either large global folds containing hundreds and even thousands of nucleotides or small local spatial motifs with at least two successive base pairs. ARTS is also highly efficient requiring typically a few seconds for comparing a pair of average-size RNA structures with hundreds of nucleotides. The tool has been used to conduct an all-against-all comparison of all the RNA 3D structures currently available in the Protein Data Bank (PDB) (21). The results can be accessed via the website.

METHOD OUTLINE

The input is a pair of nucleic acid structures represented by the 3D coordinates of their atoms. The phosphate atoms are singled out as critical points and each structure is represented as a set of points in 3D space, where each point is the position of a phosphate atom. Using this representation the problem is a version of the Largest Common Point Set (LCP) problem in Computational Geometry. Namely, the task is to find a rigid transformation (rotation and translation) that superimposes the largest number of phosphate atoms of one structure onto the phosphate atoms of the other one within a predefined bottleneck matching distance (22) error. Although this problem has been studied extensively, the current known exact and approximate algorithms for solving it are impractical, since they require $O(n^{32.5})$ and $O(n^{8.5})$ time, respectively, where n is the number of phosphate atoms (22,23).

ARTS [http://bioinfo3d.cs.tau.ac.il/ARTS (20)] is thus a heuristic method. By exploiting the base pairing and stacking properties of nucleic acids, it is capable of providing biologically relevant solutions in practical running times, even for large compact structures with thousands of nucleotides like the ribosome. Its time complexity is $O(n^3)$. Unlike the LCP problem, the goal is to maximize both the number of superimposed phosphate atoms and the number of superimposed base pairs. The rationale is that more than half of the nucleotides in an average non-coding RNA are involved in base pairing and the stems that they form are evolutionarily more conserved than loops (7). It is thus unlikely that an alignment with a large number of superimposed base pairs will be biologically meaningless, as might happen when solving the pure geometric LCP problem. In the first stage, all the possible local alignments of two successive base pairs between the structures are constructed. Then, a greedy approach is used to extend the local alignments so that a maximal number of phosphate atoms and base pairs will be superimposed. Finally, the global alignments are scored, clustered and ranked, and the highest scoring ones are reported.

We estimate the significance of the obtained alignments by computing the P-value of their score with respect to a random dataset Γ of pairwise alignments. In the current version of the application Γ contains ~245 000 alignments that have been randomly chosen from an all-against-all comparison of all the RNA structures in the PDB. The P-value of an alignment between a pair of RNA structures with n and m nucleotides is computed with respect to all alignments in Γ for which the number of nucleotides in the smallest structure is ±20% $\min(n,m)$. The resulting P-value of the alignment represents the probability that a pairwise alignment for which the size of the smallest structure is similar would receive a higher or equal score by chance.

WEB SERVER

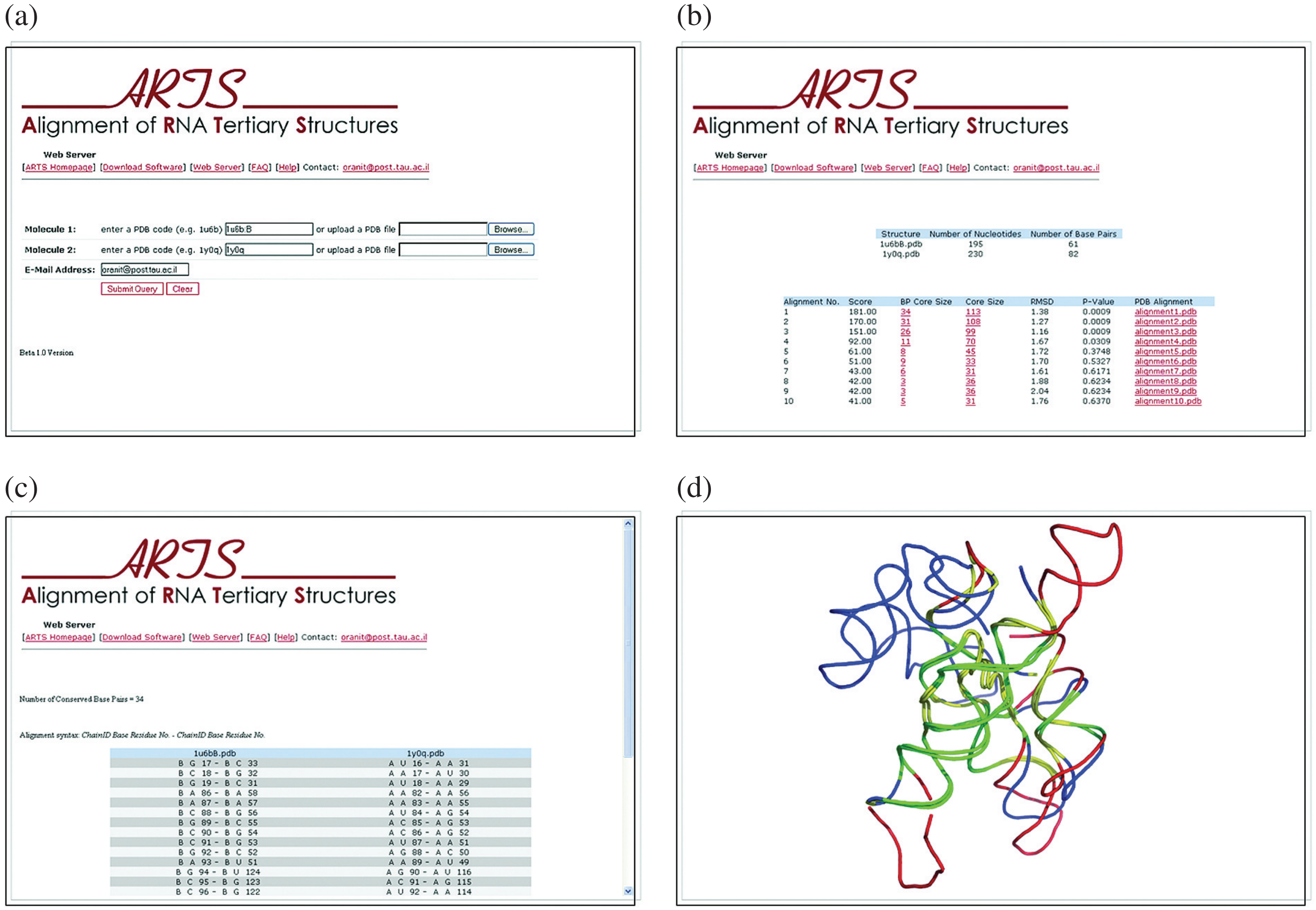
The ARTS web server as well as an accompanied software package are freely available at http://bioinfo3d.cs.tau.ac.il/ARTS.

Input

The user interface of the web server is straightforward (Figure 1a). It requires the user to enter an Email address and a pair of nucleic acid structures in PDB format (21). The structures can be either uploaded to the server or retrieved from the PDB. In the second case the user has to enter a four-character PDB code, optionally followed by a colon and a list of chain IDs, for instance '1u6b', '1u6b:B' and '1u6b:BC'. In both cases, the structures must contain all atoms and not only the ones on the backbone. The reason is that otherwise hydrogen bonds cannot be computed and these are necessary for finding base pairs. Another requirement is that each structure has at least two successive base pairs.

Output

A typical run of ARTS for comparing a pair of average-size nucleic acid structures with hundreds of nucleotides takes a few seconds. After the run completes, a web page with a summary of the obtained alignments is displayed. In addition, an Email with a link to this web page is sent to the user. Figure 1b displays a summary page obtained for two self-splicing group I introns, the *Azoarcus* pre-tRNA^{Ile} intron with both exons [PDB—1u6b:B (24)] and the Twort ribozyme intron [PDB—1y0q (25)]. The page contains two tables. The upper table shows the name of the compared structures and the number of nucleotides and base pairs in each structure. The bottom table shows the 10 top-ranking alignments sorted in descending order by their score. Besides the score, the following data are presented for each alignment: (i) the number of matched base pairs (BP Core Size); (ii) the total number of matched nucleotides including unpaired ones (Core Size); (iii) the root mean square deviation (RMSD) between the phosphate atoms of the matched nucleotides in the core; (iv) the P-value; and (v) a PDB file with the aligned structures. Clicking on the 'BP Core Size' field of one of the alignments displays a new page with a table of the matched base pairs. Figure 1c shows the page obtained after clicking on the 'BP Core Size' field of the top-ranking alignment in the summary page presented in Figure 1b. The table consists of two columns, one for each structure. Each line corresponds to a match between 2 base-pairs, and each entry provides the chain identifier, base type and residue number of the 2 nucleotides in the corresponding base pair. Clicking on the 'Core Size' field of one of the alignments in the summary page (Figure 1b) displays a similar page with a table of all matched nucleotides (paired and unpaired). A PDB file with the input structures superimposed one onto another will be downloaded or presented by a viewer (if configured) when clicking on the 'PDB Alignment' field of one of the alignments in the summary page. Figure 1d



(a)

Alignment of RNA Tertiary Structures

Web Server

[ARTS Homepage] [Download Software] [Web Server] [FAQ] [Help] Contact: oranit@post.tau.ac.il

Molecule 1: enter a PDB code (e.g. 1u6b) 1u6b:B or upload a PDB file

Molecule 2: enter a PDB code (e.g. 1y0q) 1y0q or upload a PDB file

E-Mail Address: oranit@post.tau.ac.il

Beta 1.0 Version

(b)

Alignment of RNA Tertiary Structures

Web Server

[ARTS Homepage] [Download Software] [Web Server] [FAQ] [Help] Contact: oranit@post.tau.ac.il

Structure	Number of Nucleotides	Number of Base Pairs
1u6bB.pdb	195	61
1y0q.pdb	230	82

Alignment No.	Score	BP Core Size	Core Size	RMSD	P-Value	PDB Alignment
1	181.00	34	113	1.38	0.0009	alignment1.pdb
2	170.00	31	108	1.27	0.0009	alignment2.pdb
3	151.00	26	99	1.16	0.0009	alignment3.pdb
4	92.00	11	70	1.67	0.0309	alignment4.pdb
5	61.00	8	45	1.72	0.3748	alignment5.pdb
6	51.00	9	33	1.70	0.5327	alignment6.pdb
7	43.00	6	31	1.61	0.6171	alignment7.pdb
8	42.00	3	36	1.88	0.6234	alignment8.pdb
9	42.00	3	36	2.04	0.6234	alignment9.pdb
10	41.00	5	31	1.76	0.6370	alignment10.pdb

(c)

Alignment of RNA Tertiary Structures

Web Server

[ARTS Homepage] [Download Software] [Web Server] [FAQ] [Help] Contact: oranit@post.tau.ac.il

Number of Conserved Base Pairs = 34

Alignment syntax: *ChainID Base Residue No. - ChainID Base Residue No.*

1u6bB.pdb	1y0q.pdb
B G 17 - B C 33	A U 16 - A A 31
B C 18 - B G 32	A A 17 - A U 30
B G 19 - B C 31	A U 18 - A A 29
B A 86 - B A 58	A A 82 - A A 56
B A 87 - B A 57	A A 83 - A A 55
B C 88 - B G 56	A U 84 - A G 54
B G 89 - B C 55	A C 85 - A G 53
B C 90 - B G 54	A C 86 - A G 52
B C 91 - B G 53	A U 87 - A A 51
B G 92 - B C 52	A G 88 - A C 50
B A 93 - B U 51	A A 89 - A U 49
B G 94 - B U 124	A G 90 - A U 116
B C 95 - B G 123	A C 91 - A G 115
B C 96 - B G 122	A U 92 - A A 114

Figure 1. The ARTS web server. (**a**) The entrance page of the web server. The user is required to enter an Email address and a pair of nucleic acid structures in PDB format (21). The structures can be either uploaded to the server or retrieved from the PDB. In the second case the user has to enter a four-character PDB code, optionally followed by a colon and a list of chain IDs. (**b**) A web page with a summary of the 10 top-ranking alignments obtained for '1u6b:B' and '1y0q' PDB codes. (**c**) The page obtained after clicking on the 'BP Core Size' field of the top-ranking alignment in the summary page presented in (b). (**d**) The superimposition of the input structures displayed by PyMOL (26) after clicking on the 'PDB Alignment' field of the top-ranking alignment in the summary page presented in (b). The backbone of the two structures, PDB:1u6bB and PDB:1y0q, is depicted in red and blue, respectively. The matched base pairs are in green and the matched unpaired nucleotides are in yellow.

shows the superimposition of the input structures displayed by the PyMOL viewer [http://www.pymol.org (26)] after clicking on the 'PDB alignment' field of the top-ranking alignment in the summary page presented in Figure 1b. Scripts for easy use with viewers are provided with the software package. Among them are PyMOL [http://www.pymol.org (26)] and RasMol (27) scripts for displaying the alignments and selecting the matched base pairs (bpcore) and all matched nucleotides including unpaired ones (core).

CONCLUSIONS

We have presented a freely available web server accompanied by a software package for 3D structural alignment of nucleic acids. The web server receives as input a pair of tertiary structures of nucleic acids in PDB format, and searches for a priori unknown common substructures that are not necessarily contiguous. The output consists of the top-ranking superpositions between the two input structures in PDB format and corresponding lists of matched nucleotides in the common substructures. To the best of our knowledge, this is the first web server that performs RNA structural comparisons that are truly 3D and irrespective of the order of the nucleotides on the chain. The only requirement is that there are at least two consecutive base pairs in the match. The algorithm behind the web server is highly efficient, where a typical comparison of two nucleic acids takes a few seconds on a standard PC. An all-against-all comparison of all the RNA structures currently available in the PDB has been carried out and the results can be accessed via the web server. In future work we intend to allow online searches of uploaded structures against the entire PDB.

ACKNOWLEDGEMENTS

We thank Maxim Shatsky for stimulating discussions. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12 400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This Research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The research of O. Dror has been supported by the Eshkol Fellowship funded by the Israeli Ministry of Science. The research of H.J.W. has been supported in part by the Israel Science Foundation (grant No. 281/05) and Hermann Minkowski Minerva Center for Geometry at TAU. Funding to pay the Open Access publication charges for this article was provided by N01-CO-12400 from the NCI.

Conflict of interest statement. None declared.

REFERENCES

1. Storz,G. (2002) An expanding universe of noncoding RNAs. *Science*, **269**, 1260–1263.
2. Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.*, **2**, 919–929.
3. Doudna,J.A. and Cech,T.R. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
4. Wolfson,H.J., Shatsky,M., Schneidman-Duhovny,D., Dror,O., Shulman-Peleg,A., Ma,B. and Nussinov,R. (2005) From structure to function: methods and applications. *Curr. Protein Pept. Sci.*, **6**, 171–183.
5. Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**, 303–310.
6. Batey,R., Rambo,R. and Doudna,J. (1999) Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed. Engl.*, **38**, 2326–2343.
7. Moore,P.B. (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **68**, 287–300.
8. Gendron,P., Lemieuxs,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
9. Lemieux,S. and Major,F. (2002) RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res.*, **30**, 4250–4263.
10. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
11. Lu,X.-J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
12. Lu,X.-J. and Olson,W.K. (1999) Resolving the discrepancies among nucleic acid conformational analyses. *J. Mol. Biol.*, **285**, 1563–1575.
13. Lu,X.-J., Babcock,M. and Olson,W. (1999) Overview of nucleic acid analysis programs. *J. Biomol. Struct. Dyn.*, **16**, 833–843.
14. Reijmers,T.H., Wehrens,R. and Buydens,L.M.C. (2001) The influence of different structure representations on the clustering of an RNA nucleotides data set. *J. Chem. Inf. Comput. Sci.*, **14**, 1388–1394.
15. Duarte,C.M., Wadley,L.M. and Pyle,A.M. (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res.*, **31**, 4755–4761.
16. Huang,H., Nagaswamy,U. and Fox,G.E. (2005) The application of cluster analysis in the intercomparison of loop structures in RNA. *RNA*, **11**, 412–423.
17. Harrison,A.-M., South,D.R., Willett,P. and Artymiuk,P.J. (2003) Representation, searching and discovery of patterns of bases in complex RNA structures. *J. Comput. Aided Mol. Des.*, **17**, 537–549.
18. Hershkovitz,E., Tannenbaum,E., Howerton,S.B., Sheth,A., Tannenbaum,A. and Williams,L.D. (2003) Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. *Nucleic Acids Res.*, **31**, 6249–6257.
19. Wadley,L.M. and Pyle,A.M. (2004) The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res.*, **32**, 6650–6659.
20. Dror,O., Nussinov,R. and Wolfson,H.J. (2005) ARTS: Alignment of RNA tertiary structures. *Bioinformatics*, **21**, ii1–ii7.
21. Berman,H., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T., Weissig,H., Shindyalov,I. and Bourne,P. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
22. Ambühl,C., Chakraborty,S. and Gärtner,B. (2000) Computing largest common point sets under approximate congruence. In *Proceedings of the Eighth Annual European Symposium on Algorithms (ESA)*, Mike Paterson (Ed.), Saarbrücken, Germany, September 5–8, 2000, Lecture Notes in Computer Science 1879, Springer-Verlag, pp. 52-63.
23. Akutsu,T. (1996) Protein structure alignment using dynamic programming and iterative improvement. *IEICE Trans. Inform. Syst.*, **E79D**, 1629–1636.
24. Adams,P.L., Stahley,M.R., Kosek,A.B., Wang,J. and Strobel,S.A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45–50.
25. Golden,B.L., Kim,H. and Chase,E. (2005) Crystal structure of a phage Twort group I ribozyme–product complex. *Nature Struct. Mol. Biol.*, **12**, 82–89.
26. DeLano,W. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.
27. Sayle,R. and Milner-White,E. (1995) RasMol: biomolecular graphics for all. *Trends. Biochem. Sci.*, **20**, 374–376.