

iPARTS2: an improved tool for pairwise alignment of RNA tertiary structures, version 2

Chung-Han Yang^{1,2,†}, Cheng-Ting Shih^{3,†}, Kun-Tze Chen³, Po-Han Lee³, Ping-Han Tsai³, Jian-Cheng Lin³, Ching-Yu Yen³, Tiao-Yin Lin² and Chin Lung Lu^{3,*}

¹Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 30050, Taiwan,

²Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 30050, Taiwan and

³Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan

Received February 20, 2016; Revised April 30, 2016; Accepted May 4, 2016

ABSTRACT

Since its first release in 2010, iPARTS has become a valuable tool for globally or locally aligning two RNA 3D structures. It was implemented by a structural alphabet (SA)-based approach, which uses an SA of 23 letters to reduce RNA 3D structures into 1D sequences of SA letters and applies traditional sequence alignment to these SA-encoded sequences for determining their global or local similarity. In this version, we have re-implemented iPARTS into a new web server iPARTS2 by constructing a totally new SA, which consists of 92 elements with each carrying both information of base and backbone geometry for a representative nucleotide. This SA is significantly different from the one used in iPARTS, because the latter consists of only 23 elements with each carrying only the backbone geometry information of a representative nucleotide. Our experimental results have shown that iPARTS2 outperforms its previous version iPARTS and also achieves better accuracy than other popular tools, such as SARA, SETTER and RASS, in RNA alignment quality and function prediction. iPARTS2 takes as input two RNA 3D structures in the PDB format and outputs their global or local alignments with graphical display. iPARTS2 is now available online at <http://genome.cs.nthu.edu.tw/iPARTS2/>.

INTRODUCTION

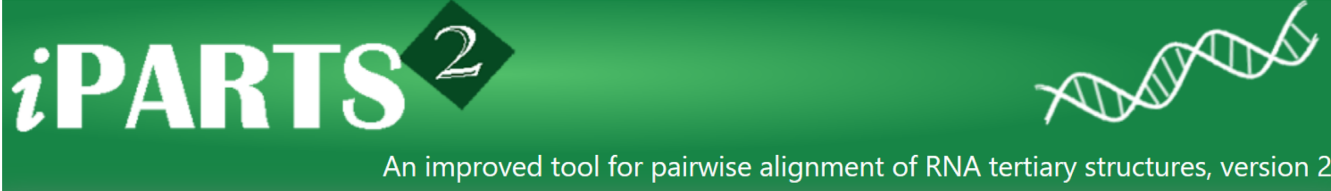
In addition to transmission of genetic information from DNA to proteins, RNA is capable of performing a wide range of biological functions in cells, including catalysis, genetic control and molecular recognition (1). Because the functions of RNAs are largely determined by their diverse three-dimensional (3D) structures, tools capable of

efficiently and accurately comparing two RNA 3D structures are important in computational structural biology. Currently, several popular and useful tools of aligning two RNA 3D structure have been proposed based on heuristic approaches, such as SARA (2,3), iPARTS (4), SETTER (5,6) and RASS (7,8). Both SARA and iPARTS align two RNA 3D structures by using a similar approach, which reduces the 3D structures into one-dimensional (1D) sequences according to some local structure features in the nucleotide backbone conformation (i.e. backbone unit vectors used in SARA and backbone pseudo-torsion angles used in iPARTS) and then applies traditional sequence alignment algorithms to align the resulting 1D sequences (2–4). As to SETTER, it divides the RNA 3D structure into non-overlapping local structural units, called generalized secondary structure units (GSSUs), and then obtains their structural alignment by using a comparison method based on a distance measured by RMSD (Root Mean Square Deviation) transformation between all possible pairs of GSSUs (5,6). RASS develops a method based on elastic shape analysis, which treats the structures of RNAs as 3D curves with their 1D nucleotide sequence encoded on additional three dimensions, so that the structural alignment of two RNAs is performed in a joint sequence-structure space of six dimensions (7,8).

The method we used to implement iPARTS (4) is the so-called structural alphabet (SA)-based approach, which uses an SA of 23 letters to reduce RNA 3D structures into 1D sequences of SA letters and applies traditional sequence alignment to these SA-encoded sequences for determining their global or local similarity. In fact, the accuracy performance of our iPARTS largely depends on the quality of the SA, which was constructed from a list of 117 RNA 3D structures using the pseudo-torsion angles of their nucleotide backbones. It has been shown that for RNAs, two pseudo-torsion angles (η and θ) are sufficient to describe the backbone conformation of each nucleotide (9). Actually, during the last 5 years after the introduction of our iPARTS, several

*To whom correspondence should be addressed. Tel: +886 3 573 1205; Fax: +886 3 573 1201; Email: cllu@cs.nthu.edu.tw

†These authors contributed equally to the paper as first authors.



An improved tool for pairwise alignment of RNA tertiary structures, version 2

[Help](#)

Input RNA molecules :

- RNA Molecule 1 :

PDB/NDB ID : or upload PDB file : No file selected.

chain ID : , from : to :
- RNA Molecule 2 :

PDB/NDB ID : or upload PDB file : No file selected.

chain ID : , from : to :

Parameters :

- Alignment :
- Gap open penalty :
- Gap extension penalty :
- Number of suboptimal alignment(s) :

Figure 1. The web interface of iPARTS2.

hundreds of new RNA 3D structures have been determined and already deposited in the PDB/NDB databases (10,11). These newly determined RNA 3D structures should benefit us to improve the accuracy of our iPARTS by constructing a new and sufficiently high quality SA. In addition, as was reported in the study of RASS (7,8), both 1D nucleotide sequences and 3D structures of RNAs need to be taken into account when determining their functions, because 1D sequence carries side chain information of nucleotides, 3D structure carries the backbone geometry information of nucleotides and both types of information are different and can play important roles in determining RNA functions.

In this study, we have re-implemented our previous tool iPARTS as a new web server named iPARTS2 (meaning iPARTS version 2) by constructing a totally new SA, which consists of 92 elements with each element carrying both information of base (1D) and backbone geometry (3D) for a representative nucleotide, from a representative and sufficiently non-redundant list of 876 atomic-resolution RNA 3D structures with 65154 nucleotides in total (12). This SA is significantly different from the one used in iPARTS, because the latter, constructed by using 117 crystal RNA

structures with 9527 nucleotides, consists of only 23 elements, each of which carries only the backbone geometry information of a representative nucleotide. Like in iPARTS, we also equip iPARTS2 with two capabilities of aligning two RNA 3D structures: (i) global alignment that can be used to determine their overall structural similarity and (ii) local alignment that can be used to find their locally similar sub-structures. It is worth mentioning here that the function of local alignment in iPARTS2 is unique when compared with other tools SARA, SETTER and RASS, because they all provide the function of global alignment only. For validation, we have used a benchmark dataset FSCOR with 419 RNA 3D structures to test our iPARTS2 and compare the accuracy performance of its global alignment with its previous version iPARTS, as well as other popular tools SARA, SETTER and RASS. Our experimental results have finally shown that our current iPARTS2 indeed outperforms its previous version iPARTS and also achieves better accuracy than SARA, SETTER and RASS in RNA alignment quality and function prediction.

MATERIALS AND METHODS

In this study, we have implemented iPARTS2 by using an improved SA-based algorithm as follows. First, 63402 non-terminal nucleotides from the RNA 3D Hub non-redundant list (version 1.89) of 876 RNA 3D structures (12) were classified into 23 conformation clusters according to their backbone pseudo-torsion angles. Basically, nucleotides in the same cluster are structurally similar in backbone geometry. Next, 23 capital letters were used to represent the center nucleotides of these 23 clusters and for each letter, four different background colors were further used to separately represent four possible base types A, G, C and U of the corresponding center nucleotide. As a result, we constructed an SA of 92 elements with each element (a letter on a colored background) carrying both information of backbone geometry (letter) and base (background color) for a representative nucleotide. Finally, the SA was used to reduce input RNA 3D structures into 1D SA-encoded sequences and a traditional sequence alignment, such as global alignment (without penalty to end gaps) (13) or local alignment (14), was applied to them for determining their global or local similarity. In addition, for the accuracy of aligning two SA-encoded sequences, the statistical method proposed by Henikoff and Henikoff (15) was applied to derive a BLOSUM-like substitution matrix that can reward more similar SA-encoded sequences with high scores. We refer the reader to the Supplementary Data for the details of the above improved SA-based algorithm. It is worth mentioning here that the local alignment algorithm we used to implement iPARTS2 is slightly different from the one used in iPARTS, because we further utilized the technique mentioned in (16) to modify the local alignment algorithm such that the local alignments returned by iPARTS2 are non-intersecting, where two alignments are said to be *non-intersecting* if they do not have a match or mismatch in common. Usually, non-intersecting local alignments of RNA structures are more of practical interest to the user.

USAGE OF IPARTS2

The kernel algorithms of iPARTS2 were written in PHP. Currently, iPARTS2 can be accessed by an easy-to-operate web interface as illustrated in Figure 1. It provides the user two kinds of alignments for comparing two RNA 3D structures: (i) global alignment for determining their whole structural similarity, and (ii) local alignment for finding common similar substructures. Basically, iPARTS2 takes as input two RNA 3D structures, each of which can be either a PDB/NDB ID or a PDB file uploaded by the user, their chain IDs if they have multiple chains, and optionally the starting and ending residue numbers of substructures to be aligned. If required, the user can run iPARTS2 by modifying the default settings of all the parameters, including alignment method (whose default is global alignment), gap open and extension penalties (whose default values are -9 and -1 , respectively), and number of suboptimal alignments (at least one). In the output page, iPARTS2 first shows the details of input RNA molecules and user-specified parameters. Next, iPARTS2 continues to show its running time, as well as its alignment results, including structural alignment

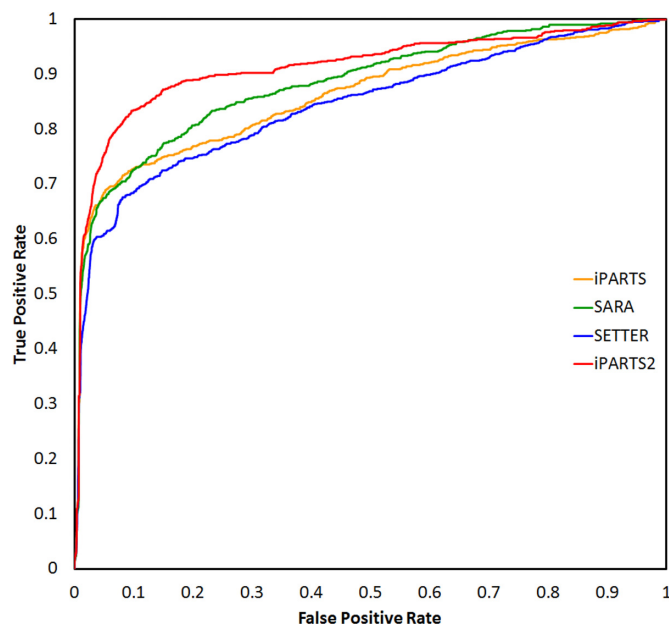


Figure 2. ROC curves for $d = 0$ based on the SAS values of all aligned pairs of RNA 3D structures in the FSCOR dataset, where the AUC values of iPARTS, SARA, SETTER and iPARTS2 are 0.861, 0.883, 0.843 and 0.914, respectively. Note that the AUC value of RASS computed by using the 67006 pairs of RNA 3D structures is 0.892.

score (SAS) (refer to the ‘Experimental Results’ section for its definition) between input RNA 3D structures with corresponding raw score in parentheses, number of aligned nucleotide pairs, RMSD and optimal/suboptimal alignments of their SA-encoded sequences and corresponding RNA sequences. Note that each letter in the aligned SA-encoded sequences is displayed with a colored background, which indicates the base type (A, G, C or U) of the corresponding nucleotide. Finally, iPARTS2 shows a JSmol graphical display (without installing Java plugin) of aligned RNA 3D structures, so that the user can visually view, rotate and enlarge the 3D structures of input RNA molecules and their structural superposition and download their alignment and PDB files. Note that in the JSmol visualization, end-gap residues in global alignment or non-aligned residues in local alignment are displayed in light colors.

EXPERIMENTAL RESULTS

First, we tested iPARTS2 by running its global alignment on a benchmark dataset called FSCOR and evaluated its accuracy in function assignment by comparing its receiver operating characteristic (ROC) curve with those obtained by iPARTS (4) and other existing popular tools, including SARA (2,3) and SETTER (5,6). The FSCOR dataset originally proposed in (3) contains 419 RNA 3D structures that are classified into 168 functional classes. We ran all the tools mentioned above locally by aligning all 87571 pairs of RNA 3D structures in the FSCOR dataset. To take the quality of the structural alignments into account, the ROC curves of all the tools were computed based on a geometric match measure called SAS, which is defined to be $(\text{RMSD} \times 100)/(\text{number of aligned nucleotide pairs})$ (17,18), in-

Table 1. Comparison of average running times for iPARTS, SARA, SETTER, RASS and iPARTS2

Dataset	iPARTS	SARA	SETTER	RASS	iPARTS2
tRNA	0.30 s	0.83 s	0.08 s	1.52 s	0.27 s
Ribozyme P4-P6 domain	0.64 s	5.21 s	0.10 s	3.46 s	0.65 s
Domain V of 23S rRNA	3.44 s	1.87 min	0.81 s	9.17 s	3.79 s
16S rRNA	38.16 s	46.53 min	5.30 s	48.09 s	36.69 s
25S rRNA	2.92 min	6.65 h	17.54 s	5.60 min	3.13 min

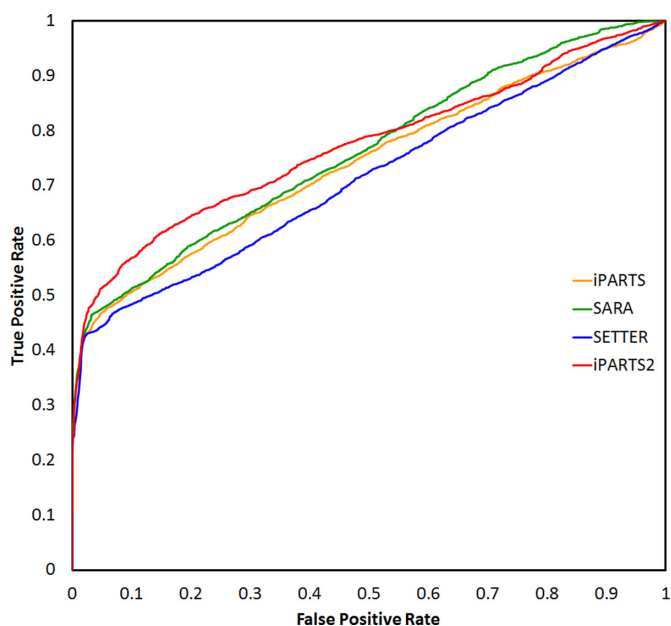


Figure 3. ROC curves for $d \leq 2$ based on the SAS values of all aligned pairs of RNA 3D structures in the FSCOR dataset, where the AUC values of iPARTS, SARA, SETTER and iPARTS2 are 0.740, 0.761, 0.713 and 0.772, respectively. Note that the AUC value of RASS computed by using the 67006 pairs of RNA 3D structures is 0.758.

stead of native alignment score. The reason is that, as suggested in (18), a better structural alignment should match more residues and also have lower RMSD, and the geometric match measure SAS is better than the native alignment score to separate good structural alignments from less good ones. Two RNA structures in the FSCOR dataset are said to be functionally identical if they have the same deepest SCOR classification (i.e. their geodesic distance $d = 0$) or functionally similar if they differ at least in the deepest SCOR classification (i.e. $d \leq 2$). To obtain the ROC curve of each tool, the alignments of all pairs of RNA structures computed by the tool are sorted by their SAS values. A threshold of SAS is then varied between the minimum and maximum of the sorted SAS values for producing the points of the ROC curve. For a fixed threshold, all pairs of aligned RNA structures whose SAS values are above the threshold are assumed positive and all below it negative. Moreover, the pairs assumed positive are counted as true positives (TP) if they are functionally identical ($d = 0$) or similar ($d \leq 2$) and false positives (FP) otherwise; the pairs assumed negative are counted as true negatives (TN) if they are functionally non-identical ($d > 0$) or dissimilar ($d > 2$) and false negatives (FN) otherwise. The point of the ROC curve corresponding to the fixed threshold is then produced by plot-

ting its TP rate $TP/(TP + FN)$ on the y-axis and its FP rate $FP/(FP + TN)$ on the x-axis. As a result, the ROC curves for all the evaluated tools mentioned above are displayed in Figures 2 and 3 for $d = 0$ and $d \leq 2$, respectively. These experimental results have shown that our iPARTS2 outperforms its previous version iPARTS and other tools SARA and SETTER for the function assignment in the FSCOR dataset, because iPARTS2 has the highest AUC values of 0.914 and 0.772 for $d = 0$ and $d \leq 2$, respectively.

Next, we also compared the capabilities of iPARTS, SARA, SETTER and iPARTS2 for the function assignment with RASS (7,8) using the FSCOR dataset. As mentioned before, RASS is a recently developed tool of comparing two RNAs by considering both information of their sequences (bases) and 3D structures (backbone geometry). When running RASS on the FSCOR dataset, however, we noticed that for 20565 pairs among 419 RNA 3D structures, RASS was not able to provide their structural alignments so that their SAS values were not able to be computed. Therefore, for a fair comparison of all the evaluated tools, we calculated their ROC curves only using those 67006 pairs of RNA 3D structures whose structural alignments were able to be provided by RASS. In this situation, iPARTS2 still performs better than all other tools, including RASS, according to the AUC values of their ROC curves (refer to Supplementary Figures S5 and 6). For the results of additional experiments, we refer the reader to the Supplementary Data.

Finally, for the running time comparison of all the tools mentioned before, we used five datasets containing two or more RNA 3D structures of various lengths as follows: (i) five tRNA structures (1EHZ:A, 1H3E:B, 1I9V:A, 2TRA:A and 1YFG:A) with an average length of 76 nucleotides, (ii) three ribozyme P4-P6 domains (1GID:A, 1HR2:A and 1L8V:A) with an average length of 157 nucleotides, (iii) two domains V of 23S rRNA (1FFZ:A and 1FG0:A) with an average length of 496 nucleotides, (iv) two 16S rRNA (1J5E:A and 4V4Q:AA) with an average length of 1522 nucleotides and (v) two 25S rRNA (4V7R:B1 and 4V7R:D1) with an average length of 3396 nucleotides. The average running times of all the tools were obtained by running them with their default parameters on local machine with Intel CPUs with 3.4 GHz and 32 GB of RAM under Linux system. As shown in Table 1, SETTER is the fastest tool among all the five tools. However, our iPARTS2, as well as iPARTS, outperforms both SARA and RASS, and it can finish its alignment job in several seconds up to a couple of minutes.

SUMMARY

In this study, we have re-implemented our previous tool iPARTS into a new web server iPARTS2 by constructing a totally new SA of 92 elements, with each element carry-

ing both information of base (1D) and backbone geometry (3D) for a representative nucleotide. According to our experimental results on a benchmark dataset, iPARTS2 indeed outperforms iPARTS and also achieves better accuracy than other popular tools, such as SARA, SETTER and RASS, in RNA alignment quality and function prediction. Therefore, iPARTS2 can serve as a useful tool for aligning two RNA 3D structures, which can further provide insight into structural and functional properties of RNAs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Ministry of Science and Technology of Taiwan [MOST104-2221-E-007-027-MY2, in part]. Funding for open access charge: Ministry of Science and Technology of Taiwan [MOST104-2221-E-007-027-MY2].

Conflict of interest statement. None declared.

REFERENCES

- Gesteland, R.F., Cech, T. and Atkins, J.F. (2006) *The RNA world: the Nature of Modern RNA Suggests a Prebiotic RNA World*, 3rd edn, Cold Spring Harbor Laboratory Press, NY.
- Capriotti, E. and Marti-Renom, M.A. (2008) RNA structure alignment by a unit-vector approach. *Bioinformatics*, **24**, 112–118.
- Capriotti, E. and Marti-Renom, M.A. (2009) SARA: a server for function annotation of RNA structures. *Nucleic Acids Res.*, **37**, W260–W265.
- 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.
- Hoksza, D. and Svozil, D. (2012) Efficient RNA pairwise structure comparison by SETTER method. *Bioinformatics*, **28**, 1858–1864.
- Cech, P., Svozil, D. and Hoksza, D. (2012) SETTER: web server for RNA structure comparison. *Nucleic Acids Res.*, **40**, W42–W48.
- Laborde, J., Robinson, D., Srivastava, A., Klassen, E. and Zhang, J. (2013) RNA global alignment in the joint sequence-structure space using elastic shape analysis. *Nucleic Acids Res.*, **41**, e114.
- He, G., Steppi, A., Laborde, J., Srivastava, A., Zhao, P. and Zhang, J. (2014) RASS: a web server for RNA alignment in the joint sequence-structure space. *Nucleic Acids Res.*, **42**, W377–W381.
- Wadley, L.M., Keating, K.S., Duarte, C.M. and Pyle, A.M. (2007) Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. *J. Mol. Biol.*, **372**, 942–957.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
- Coimbatore Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A.I., Sweeney, B., Zirbel, C.L., Leontis, N.B. and Berman, H.M. (2014) The nucleic acid database: new features and capabilities. *Nucleic Acids Res.*, **42**, D114–D122.
- Leontis, N.B. and Zirbel, C.L. (2012) Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: Leontis, N.B. and Westhof, E. (eds) *RNA 3D Structure Analysis and Prediction*. Springer, NY, pp. 281–298.
- 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.
- Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
- Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 10915–10919.
- Waterman, M.S. and Eggert, M. (1987) A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.*, **197**, 723–728.
- Subbiah, S., Laurents, D.V. and Levitt, M. (1993) Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.*, **3**, 141–148.
- Kolodny, R., Koehl, P. and Levitt, M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.