W8–W12 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 22 April 2011*
doi:10.1093/nar/gkr244

PicXAA-Web: a web-based platform for non-progressive maximum expected accuracy alignment of multiple biological sequences

Sayed Mohammad Ebrahim Sahraeian and Byung-Jun Yoon*

Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA

Received February 17, 2011; Revised March 26, 2011; Accepted April 5, 2011

ABSTRACT

In this article, we introduce PicXAA-Web, a web-based platform for accurate probabilistic alignment of multiple biological sequences. The core of PicXAA-Web consists of PicXAA, a multiple protein/DNA sequence alignment algorithm, and PicXAA-R, an extension of PicXAA for structural alignment of RNA sequences. Both PicXAA and PicXAA-R are probabilistic non-progressive alignment algorithms that aim to find the optimal alignment of multiple biological sequences by maximizing the expected accuracy. PicXAA and PicXAA-R greedily build up the alignment from sequence regions with high local similarity, thereby yielding an accurate global alignment that effectively captures local similarities among sequences. PicXAA-Web integrates these two algorithms in a user-friendly web platform for accurate alignment and analysis of multiple protein, DNA and RNA sequences. PicXAA-Web can be freely accessed at http://gsp.tamu.edu/picxaa/.

INTRODUCTION

Multiple sequence alignment (MSA) plays important roles in various problems in molecular biology, such as phylogenetic analysis, predicting the structure of biomolecules, identification of conserved sequence motifs and many others (1,2). Given a set of unaligned sequences, we can find their best alignment by optimizing an objective function that measures the quality of the alignment. For example, we may find the optimal multiple sequence alignment by maximizing the so-called sum-of-pairs (SP) score through dynamic programming (3,4). However, this optimization problem is NP-complete (5) and thus intractable as the number of sequences increases. The same holds true for structural alignment of non-coding RNA (ncRNA) sequences, where we need to consider the structural similarity across sequences, in addition to their sequence similarity, to obtain an accurate sequence alignment that is biologically meaningful (6).

A widely adopted solution for reducing the overall computational complexity is the progressive alignment scheme (7), which tries to construct the multiple sequence alignment in a progressive manner by repetitively performing pairwise alignments according to a guide tree. Many popular MSA algorithms, such as CLUSTALW (8), T-Coffee (9), ProbCons (10), ProbAlign (11), MUSCLE (12), MAFFT (13), MUMMALS (14), and MSAProbs 0.9.4 (15), as well as many RNA structural alignment algorithms, such as Murlet (16), RAF (17), STRAL (18), LocARNA (19), CentroidAlign (20), PMcomp (21), MXSCARNA (22), R-Coffee (23), LARA (24) and MAFFT-xinsi (25) adopt this progressive approach to construct MSAs.

Although the progressive alignment approach is computationally efficient, it tends to propagate early stage alignment errors throughout the alignment process, which may significantly degrade the quality of the final alignment. To address this problem, a number of non-progressive alignment algorithms have been also proposed, where examples include DIALIGN (26), AMAP (27), FSA (28), RNASampler (29), MASTER (30), Stemloc-AMA (31), PicXAA (32) and PicXAA-R (33).

In this article, we introduce PicXAA-Web, a user-friendly web server developed based on PicXAA, a recently proposed MSA algorithm (32) and PicXAA-R (33), an extension of PicXAA for structural alignment of RNA sequences. PicXAA is a probabilistic non-progressive alignment algorithm that finds protein (or DNA) MSAs with maximum expected accuracy. PicXAA greedily builds up the alignment from sequence regions with high local similarity, thereby yielding an

*To whom correspondence should be addressed. Tel: +1 979 845 6942; Fax: +1 979 845 6259; Email: bjyoon@ece.tamu.edu

Table 1. Performance evaluation of PicXAA based on BAliBASE 3.0, IRMBASE 2.0 and SABmark 1.65

Method	BAliBASE 3.0	IRMBASE 2.0	SABmark 1.65	
	SP/CS	SP/CS	Twilight f_D/f_M	Superfamily f_D/f_M
PicXAA-PF	**87.86** / 59.32	89 / 50.08	16.75 / 15.37	49.66 / 41.41
PicXAA-PHMM	86.55 / 56.28	90.76 / **54.48**	17.12 / 14.65	50.37 / 41.13
PicXAA-SPHMM	86.67 / 56.14	72.75 / 33.02	**20.99** / 17.12	**53.53** / **42.77**
ProbAlign	87.61 / 58.82	81.68 / 36.69	15.86 / 13.05	48.66 / 39.82
ProbCons	86.42 / 56.01	85.3 / 42.51	16.64 / 13.55	48.56 / 39.51
MUMMALS	85.53 / 53.85	68.44 / 24.62	19.99 / **18.23**	52.09 / 42.74
MAFFT-linsi	87.22 / 59.28	89.44 / 46.02	17.42 / 13.16	50.47 / 40.01
MAFFT-einsi	87.05 / 58.95	**91.77** / 48.21	17.77 / 13.07	49.94 / 39.23
MSAProbs	87.78 / **60.67**	83.31 / 38.48	17.41 / 13.67	50.51 / 40.75
ClustalW	75.37 / 38.01	26.34 / 2.44	12.87 / 8.72	38.62 / 30.27

accurate global alignment that effectively captures the local similarities among sequences. PicXAA-R is an extension of PicXAA, which tries to find an accurate structural alignment of non-coding RNAs (ncRNAs) through a greedy approach. PicXAA-R efficiently constructs an accurate multiple RNA alignment by using both the folding information in each RNA sequence and the local similarities between different RNA sequences. PicXAA-R is one of the fastest algorithms for structural alignment of multiple RNAs, and it consistently yields accurate alignment results. As shown in refs (32,33) through extensive experiments on several widely used benchmark sets, both PicXAA and PicXAA-R outperform many state-of-the art MSA algorithms, in terms of accuracy and efficiency. PicXAA and PicXAA-R are especially effective for aligning sequences that have high local similarities but relatively low overall similarities. Both algorithms (especially, PicXAA-R) scale very well as the number of sequences grows, which makes them suitable for analyzing large datasets.

PicXAA

The main goal of PicXAA is to find the MSA that maximizes the expected number of correctly aligned residue pairs. To this aim, it first computes the posterior pairwise alignment probability $P_a(x_i \sim y_j | \mathbf{x}, \mathbf{y})$ between residues $x_i \in \mathbf{x}$ and $y_j \in \mathbf{y}$ for all sequence pairs $(\mathbf{x}, \mathbf{y})$ in the input sequence set. PicXAA then applies an improved probabilistic consistency transformation that incorporates the information from other homologous sequences in the set to improve the estimated posterior pairwise alignment probability. Next, it sorts the pairwise residue alignments according to their alignment probability into an ordered set $\mathcal{A}$. Using an efficient graph-based technique, PicXAA greedily builds up the alignment by inserting the most probable residue alignment $(x_i, y_j) \in \mathcal{A}$ into the MSA, provided that it satisfies certain consistency conditions. After obtaining the initial alignment, PicXAA goes through a refinement step to improve the alignment quality in sequence regions with low alignment probability.

Table 1 shows the performance of PicXAA on three well-known benchmark datasets: BAliBASE 3.0 (34), IRMBASE 2.0 (26) and SABmark 1.65 (35). In this table, the alignment accuracy is reported based on two different criteria: the SP score, which is the percentage of the correctly aligned residue pairs, and the column score (CS), which is the percentage of the correct columns in the alignment. For comparison, Table 1 also shows the performance of several state-of-the-art MSA algorithms, including ProbAlign 1.1 (10), ProbCons 1.12 (10), MUMMALS 1.01 (14), MAFFT 6.708 (13) with two different options ('-linsi' and '-einsi'), MSAProbs 0.9.4 (15) and CLUSTALW 2.0.10 (8). We report the performance of PicXAA for using three different types of methods for computing the posterior alignment probabilities: (i) partition function (PF), (ii) pair-HMM (PHMM) and (iii) structural pair-HMM (SPHMM). Further details of these methods can be found in ref. (32).

As shown in Table 1, and discussed in ref. (32) in more details, PicXAA consistently yields accurate alignment results on various benchmark datasets with different characteristics. Especially, the advantage of PicXAA stands out more clearly on datasets that consist of sequences with only local similarities, as many progressive methods fail to capture these similarities faithfully.

PicXAA-R

PicXAA-R extends the ideas in PicXAA to the structural alignment of multiple RNA sequences (33). In addition to the pairwise alignment probability $P_a(x_i \sim y_j | \mathbf{x}, \mathbf{y})$ between residues in different sequences, PicXAA-R also incorporates the base-pairing probability $P_b(x_i \sim x_j | \mathbf{x})$ between bases x_i and x_j in the same sequence $\mathbf{x}$.

PicXAA-R employs several probabilistic consistency transformations to obtain improved base pairing and base alignment probabilities by considering both sequence and structural similarities among sequences.

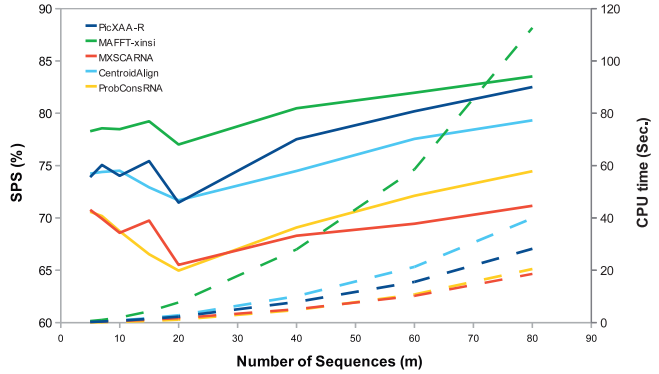

Figure 1. Performance evaluation of PicXAA-R and several other algorithms based on BraliSub and LocExtR data sets. The SPS score is shown in solid lines and the CPU time is shown in dashed lines.

These enhanced probabilities are used in a two-step greedy alignment process. In the first step, PicXAA-R constructs the structural skeleton of the alignment by greedily adding the most probable alignment between base pairs with high base-pairing probabilities. In the next step, PicXAA-R updates the obtained skeleton by successively adding the most probable base alignments. As in PicXAA, the initial MSA obtained from the greedy construction process is further refined to improve the alignment quality in low similarity regions.

Figure 1 compares the performance of PicXAA-R with several well-known RNA sequence alignment algorithms, such as MAFFT-xinsi 6.717 (25), MXSCARNA 2.1 (22), CentroidAlign (20) and ProbConsRNA 1.10 (10). Figure 1 shows the alignment accuracy of the compared algorithms in terms of the SPS score (solid lines) as well as the computational complexity in terms of the CPU time (dashed lines) based on BraliSub and LocExtR datasets (36). The accuracy (and the complexity) of each algorithm is shown as a function of the number of sequences in the alignment, in order to show its scalability. As shown in this figure, and also discussed in more details in ref. (33), PicXAA-R is one of the fastest RNA structural alignment algorithms, which consistently yields highly accurate structural alignment of multiple RNAs, especially for datasets that consist of RNA sequences with high local similarity but low percentage identity.

PicXAA WEB SERVER

PicXAA-Web provides an interactive web-based platform for aligning multiple biological sequences using PicXAA or PicXAA-R. PicXAA can be used for aligning a set of protein (or DNA) sequences and PicXAA-R can be used for structural alignment of RNAs. The user can specify which algorithm to use. The unaligned input sequences (in FASTA format) can be entered either directly through the input window or by uploading a sequence file. Two example inputs are provided by the server, which can be easily loaded into the input window to quickly try out the alignment algorithms.

Figure 2. Sample output pages for (**A**) PicXAA and (**B**) PicXAA-R.

Options and parameter setting

PicXAA-Web allows the user to choose different options and parameters for PicXAA and PicXAA-R. One can also simply use the default parameters.

Options for PicXAA. PicXAA has three options for computing the posterior pairwise residue alignment probabilities $P_a(x_i, y_j|\mathbf{x}, \mathbf{y})$ using one of the following schemes: (i) partition function (PF) (11), (ii) pair-HMM (PHMM) (10) and (iii) structural pair-HMM (SPHMM) (14). The default option is PF. SPHMM can only be used for protein sequences, while PF and PHMM can be used for both protein and nucleotide sequences. The respective advantages and disadvantages of these methods have been discussed in ref. (32). SPHMM is nearly three times more complex than the two other methods, but it usually yields more accurate alignment results when aligning structurally similar proteins. Typically, PF outperforms PHMM for most datasets, while PHMM yields better alignment results for locally similar sequences.

Options for PicXAA-R. For structural alignment of ncRNA, PicXAA-R should be used. The user can

adjust three parameters employed in the PicXAA-R algorithm (33):

(1) Parameter for intra-sequence consistency transformation: the parameter α takes a value between 0 and 1 (default value is $\alpha = 0.4$). $\alpha = 0$ makes the algorithm to ignore the originally estimated base-pairing probabilities and simply use the transformed probabilities, while $\alpha = 1$ makes the algorithm to use the original probabilities and not employ the transformed probabilities. Otherwise, the original and transformed probabilities will be linearly combined.
(2) Parameter for four-way consistency transformation: the parameter β also takes a value between 0 and 1 (default value is $\beta = 0.1$). $\beta = 0$ results in simply using the transformed probabilities, while $\beta = 1$ results in using the original probabilities. Otherwise, the original and transformed probabilities will be linearly combined.
(3) Threshold for identifying reliable basepairs: the parameter T_b specifies the minimum base-pairing probability of the base pairs that should be considered during the structural skeleton construction step [see (33) for details]. T_b should be between 0 and 1, and its default value is set to $T_b = 0.5$. In general, smaller T_b will result in a larger structural skeleton.

Output

The output can be generated either in CLUSTALW or in MFA (multiple FASTA) format. Upon submitting a sequence set, a Job ID will be assigned and a progress window will be launched with a link to the result page and additional information about the job being performed. The page will reload every 5s until the output is ready. Once the MSA is ready, the alignment result will automatically appear on the web browser. An email notification containing a link to the final result will be sent, if the user has provided an email address. A link to a downloadable output file is also provided in the result page. Sample output pages are shown in Figure 2. The web server includes a help page that contains comprehensive and easy-to-follow guidelines for using PicXAA-Web. The links to the freely downloadable C++ source code for PicXAA and PicXAA-R are also provided.

CONCLUSIONS

In this article, we introduced PicXAA-Web, a web-based tool for aligning multiple biological sequences, developed based on two recently proposed algorithms: PicXAA and PicXAA-R. PicXAA provides a user-friendly web-platform for accurate alignment and analysis of protein, DNA and RNA sequences, and it can serve as a useful resource for the molecular biology research community.

ACKNOWLEDGEMENTS

The authors would like to thank S. Vahid Faghihi and Jason Knight for their helpful advices on developing the web server.

FUNDING

Funding for open access charge: Texas A&M faculty start-up fund.

Conflict of interest statement. None declared.

REFERENCES

1. Wong,K.M., Suchard,M.A. and Huelsenbeck,J.P. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.
2. Kemena,C. and Notredame,C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, **25**, 2455–2465.
3. 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.
4. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
5. Wang,L. and Jiang,T. (1994) On the complexity of multiple sequence alignment. *J. Comput. Biol.*, **1**, 337–348.
6. Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, **45**, 810–825.
7. Feng,D.F. and Doolittle,R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351–360.
8. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
9. 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.
10. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
11. Roshan,U. and Livesay,D.R. (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, **22**, 2715–2721.
12. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
13. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics*, **9**, 286–298.
14. Pei,J. and Grishin,N.V. (2006) MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. *Nucleic Acids Res.*, **34**, 4364–4374.
15. Liu,Y., Schmidt,B. and Maskell,D.L. (2010) MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics*, **26**, 1958–1964.
16. Kiryu,H., Tabei,Y., Kin,T. and Asai,K. (2007) Murlet: a practical multiple alignment tool for structural RNA sequences. *Bioinformatics*, **23**, 1588–1598.
17. Do,C.B., Foo,C.S. and Batzoglou,S. (2008) A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, **24**, 68–76.
18. 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.

19. 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.
20. Hamada,M., Sato,K., Kiryu,H., Mituyama,T. and Asai,K. (2009) CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score. *Bioinformatics*, **25**, 3236–3243.
21. Hofacker,I.L., Bernhart,S.H. and Stadler,P.F. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.
22. Tabei,Y., Kiryu,H., Kin,T. and Asai,K. (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.
23. 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.
24. Bauer,M., Klau,G.W. and Reinert,K. (2007) Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization. *BMC Bioinformatics*, **8**, 271.
25. 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.
26. Subramanian,A.R., Kaufmann,M. and Morgenstern,B. (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.*, **3**, 6.
27. Schwartz,A.S. and Pachter,L. (2007) Multiple alignment by sequence annealing. *Bioinformatics*, **23**, e24–29.
28. Bradley,R.K., Roberts,A., Smoot,M., Juvekar,S., Do,J., Dewey,C., Holmes,I. and Pachter,L. (2009) Fast statistical alignment. *PLoS Comput. Biol.*, **5**, e1000392.
29. 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.
30. Lindgreen,S., Gardner,P.P. and Krogh,A. (2007) MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics*, **23**, 3304–3311.
31. Bradley,R.K., Pachter,L. and Holmes,I. (2008) Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics*, **24**, 2677–2683.
32. Sahraeian,S.M. and Yoon,B.J. (2010) PicXAA: greedy probabilistic construction of maximum expected accuracy alignment of multiple sequences. *Nucleic Acids Res.*, **38**, 4917–4928.
33. Sahraeian,S.M. and Yoon,B.J. (2010) PicXAA-R: efficient structural alignment of multiple RNA sequences using a greedy approach. *BMC Bioinformatics*, **11(Suppl. 1)**, S38.
34. Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.
35. Van Walle,I., Lasters,I. and Wyns,L. (2005) SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.
36. Wang,S., Gutell,R.R. and Miranker,D.P. (2007) Biclustering as a method for RNA local multiple sequence alignment. *Bioinformatics*, **23**, 3289–3296.