Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W675–W677***
doi:10.1093/nar/gkm267

eProbalign: generation and manipulation of multiple sequence alignments using partition function posterior probabilities

Satish Chikkagoudar[1], Usman Roshan[1,*] and Dennis Livesay[2]

[1]Department of Computer Science, New Jersey Institute of Technology and [2]Department of Computer Science and Bioinformatics Research Center, University of North Carolina at Charlotte

Received February 2, 2007; Revised March 29, 2007; Accepted April 8, 2007

ABSTRACT

Probalign computes maximal expected accuracy multiple sequence alignments from partition function posterior probabilities. To date, Probalign is among the very best scoring methods on the BAliBASE, HOMSTRAD and OXBENCH benchmarks. Here, we introduce eProbalign, which is an online implementation of the approach. Moreover, the eProbalign web server doubles as an online platform for post-alignment analysis. The heart-and-soul of the post-alignment functionality is the Probalign Alignment Viewer applet, which provides users a convenient means to manipulate the alignments by posterior probabilities. The viewer can also be used to produce graphical and text versions of the output. The eProbalign web server and underlying Probalign source code is freely accessible at http://probalign.njit.edu

INTRODUCTION

Multiple sequence alignments are frequently employed for analyzing biomolecular sequences. Their application spans a wide range of problems such as phylogeny reconstruction, protein functional site detection, and protein and RNA structure prediction (1). The research literature is abundant with programs and benchmarks for multiple sequence alignment, particularly for protein data. Traditionally, ClustalW (2) is the most popular program used for multiple sequence alignment; while BAliBASE (3) is a likely the most commonly used benchmark of protein alignments.

MAFFT, Probcons and Probalign are recent alignment strategies that are among recent programs with the highest accuracies on BAliBASE and other common benchmarks (i.e. HOMSTRAD (4) and OXBENCH (5). Both Probcons (6) and Probalign (7) compute maximal expected accuracy alignments using posterior probabilities. In Probcons, posterior probabilities are derived using an HMM whose parameters that have been estimated via supervised learning on BAliBASE unaligned sequences. Probalign, which is largely based on the Probcons scheme, derives the posterior probabilities from the input data by implicitly examining suboptimal (sum-of-pair) alignments using the partition function methodology for alignments (see (7) for a full description of the algorithm). Probalign alignments have been shown to have a statistically significant improvement over Probcons, MAFFT (8) and MUSCLE (9) on all three alignment benchmarks introduced above (7).

We present here eProbalign, a web server that automatically computes Probalign alignments; eProbalign also provides a convenient platform to visualize the alignment, generate images, and manipulate the output by average column posterior probabilities. The average column posterior probability (which is discussed further below) can be considered a measure of column reliability where columns with higher scores are more likely to be correct and perhaps biologically informative.

INPUT PARAMETERS

eProbalign takes as input unaligned protein or nucleic acid sequences in FASTA format. eProbalign checks the dataset to make sure it conforms with IUPAC nucleotide and amino acid one letter abbreviations. White space between residues/nucleotides in the sequences are stripped and the cleaned sequences are passed on to the queuing system. The user can specify gap open, gap extension, and thermodynamic temperature parameters on the eProbalign input page (Figure 1). The input page provides a brief description of the parameters (help link) and links to the standalone Probalign code with publication and datasets.

The three Probalign parameters on the input page are used for computing the partition function dynamic programming matrices from which the posterior probabilities are derived. This is the same as computing a set of

*To whom correspondence should be addressed. Tel: +1-973-596-2872; Fax: +1-973-596-5777; Email: usman@cs.njit.edu

(suboptimal) pairwise alignments (for every pair of sequences in the input) and then estimating pairwise posterior probabilities by simple counting. The thermodynamic temperature controls the extent to which suboptimal alignments are considered. For example, all possible suboptimal alignments would be considered at infinite temperature, whereas only the single best would be used at a temperature of zero. The affine gap parameters are used for the pairwise alignments. Subsequently, Probalign computes the maximal expected accuracy alignment from the posterior probabilities in the same way that Probcons does (6).

OUTPUT AND ALIGNMENT COLUMN RELIABILITY

The eProbalign output provides three options for viewing and analyzing the alignment (Figure 2). The alignment can be viewed in (i) FASTA text format, (ii) pdf graphical format, and (iii) the Probalign Alignment Viewer (PAV) applet (Figure 4). Each column of the alignment in the pdf file and in the applet is colored in a shade of red according to the average column posterior probability. Bright red indicates probability close to one whereas white indicates close to zero (see Figure 4 for an example on a real BAliBASE dataset).

The average column posterior probability is defined as the sum of posterior probabilities of all pairwise residues in the column normalized by the number of comparisons (6). The top row of the alignment in the pdf and applet displays the average column posterior probabilities multiplied by ten and floored to the lower integer (Figure 4). For example, a score of 1 indicates that the probability is between 0.1 and 0.2.

The Probalign Alignment Viewer is a Java applet that provides basic manipulation of the alignment. Basic Java and browser requirements to use the applet are listed on the output page. With the applet the user can opt to view and save the alignment with column posterior probabilities above any specified threshold. This has the benefit of "cleaning up" the alignment by column posterior probabilities, which is unique to eProbalign. The applet also displays posterior probabilities of all columns in a separate window if desired (Figure 3) and provides options to switch between the gapped and ungapped versions of the alignment.

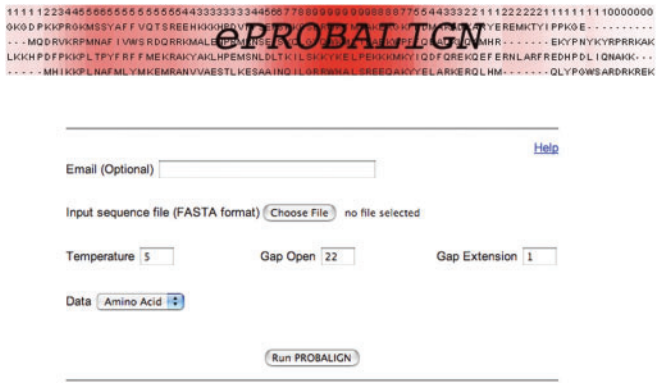

Figure 1. eProbalign input page.

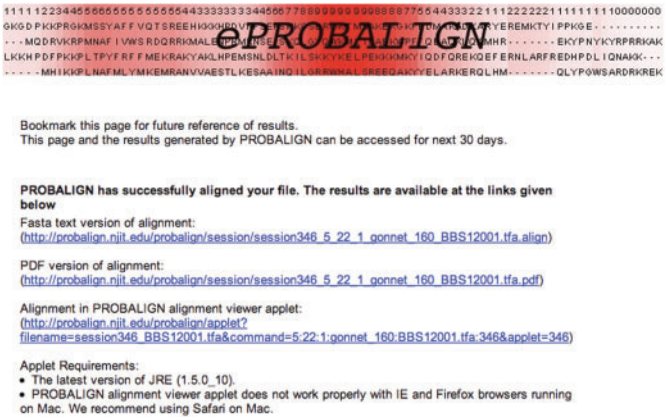

Figure 2. eProbalign output page indicating results are done.

SERVER IMPLEMENTATION

We implement a first-in/first-out queuing system that receives requests for Probalign alignments and processes them accordingly. At most, eProbalign will run two Probalign jobs at once, and it will periodically check the queue for new requests. Alignments that take longer than some defined time limit (10 hours at the time of writing of this paper) are stopped and the user is advised

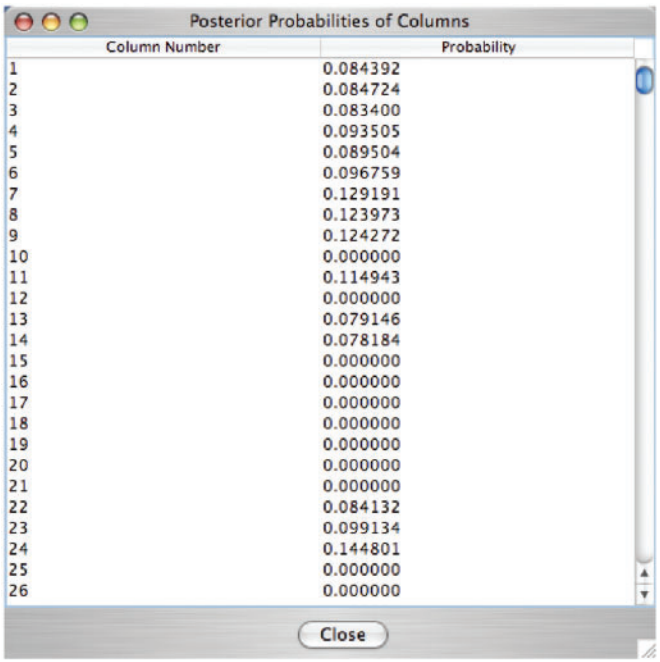

Figure 3. Posterior probability of each column.

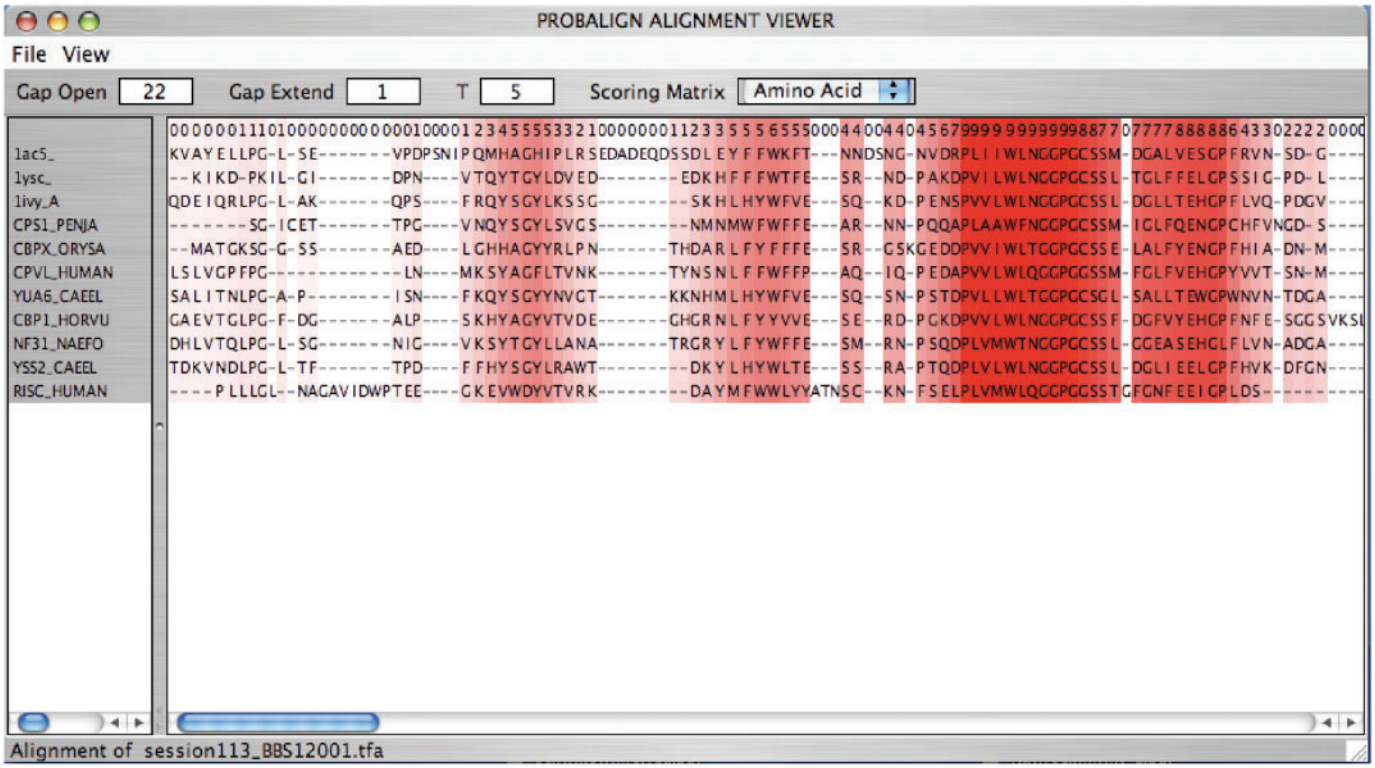

Figure 4. Probalign Alignment Viewer applet.

to download and run the standalone version. This time limit will be increased as the server hardware is upgraded.

SCALABILITY

Currently, eProbalign is installed on a dual processor 2.8GHz Intel Xeon machine with 2GB RAM. With these settings, eProbalign can usually align datasets of up to 20 sequences within one minute. Most BAliBASE 3.0 datasets from RV11 and RV12 also finish within one minute. We have also tested large datasets (in number and length of sequences) from BAliBASE RV30 and RV40 classes on eProbalign. BB30029 and BB30008 from RV30 contain 98 and 36 sequences with lengths from 431 to 852 and 400 to 1155 respectively, and BB40002 from RV40 contains 55 sequences with lengths ranging from 58 to 1502. When the server is idle, eProbalign finished in about 20 minutes on BB30008, 55 minutes on BB30029, and 30 minutes on BB40002. Results may take longer to finish when the server queue is full and multiple jobs are running simultaneously. However, the effect of parallel jobs will diminish as the server moves to a bigger machine in the near future.

ACKNOWLEDGEMENTS

We thank system administrators Gedaliah Wolosh and David Perel who have been helpful with technical issues related to the server. DRL is supported, in part, by NIH R01 GM073082-0181. Funding to pay the open access publication charges for this article was provided by startup funding to DRL from the Bioinformatics Research Center at UNC Charlotte.

Conflict of interest statement. None declared.

REFERENCES

1. Notredame,C. (2002) Recent progresses in multiple sequence alignment: a survey,. *Pharmacogenomics*, **3**, 131–144.
2. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucleic Acids Res.*, **27**, 2682–2690.
3. 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.
4. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Science*, **7**, 2469–2471.
5. Raghava,G.P.S., Searle,S.M.J., Audley,P.C., Barber,J.D. and Barton,G.J. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.
6. Do,C.B., Mahabhashyam,M.S.B., Brudno,M. and Batzoglou,S. PROBCONS: probabilistic consistency based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
7. Roshan,U. and Livesay,D.R. (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, **22**, 2715–2721.
8. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput,. *Nucleic Acids Res.*, **32**, 1792–1797.
9. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.