PROBCONS: probabilistic consistency-based multiple alignment of amino acid sequences

Chuong B. Do, Michael Brudno, and Serafim Batzoglou*

Department of Computer Science, Stanford University, California, 94305-9010, USA.

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Obtaining an accurate multiple alignment of protein sequences is often difficult when amino acid percent identity is low. In this paper, we present PROBCONS, a practical tool for protein multiple sequence alignment, based on an algorithm that combines HMM-derived posterior probabilities with consistency-based alignment techniques. On the BAliBASE benchmark alignment database, PROBCONS demonstrates a statistically significant improvement in accuracy compared to several leading alignment programs while maintaining practical running times. Source code of the program is freely available under the GNU Public License at http://probcons.stanford.edu/.

Keywords: multiple alignment, HMMs, consistency

Contact: serafim@cs.stanford.edu

INTRODUCTION

Protein multiple sequence alignments are crucial in many computational biology applications, including structure prediction (Jones 1999), phylogenetic analysis (Phillips et al. 2000), identification of conserved domains (Attwood 2002), and characterization of protein families (Sonnhammer et al. 1998). However, when sequence identity falls below 30%, called the 'twilight zone' of protein alignments, the accuracies of most automatic sequence alignment methods drop considerably (Rost 1999, Thompson et al. 1999b). As a result, alignment quality is often the limiting factor in comparative modeling studies (Jaroszewski et al. 2002).

Progressive alignment approaches, which sequentially merge aligned subsets of sequences into a complete multiple alignment using pairwise steps, are commonly used for reasons of algorithmic efficiency (Feng and Doolittle 1987). Unfortunately, such strategies are highly prone to errors at early stages of the alignment. To combat this, consistency-based alignment techniques use intermediate sequence information to improve the quality of pairwise comparisons (Gotoh 1990, Morgenstern et al. 1998, Notredame et al. 2000); when aligning two sequences, consistency-based approaches use shared homology with outgroup sequences as a guide for distinguishing between coincidental and real sequence conservation.

In this paper, we present PROBCONS, a protein multiple alignment tool that performs consistency-based progressive alignment while accounting for all suboptimal alignments with posterior-probability–based scoring (Durbin et al. 1998) based on hidden Markov models (HMMs) (Krogh et al. 1994, Eddy 1995). Other features of the program include the use of double affine insertion scoring, guide tree calculation via semi-probabilistic hierarchical clustering, optional iterative refinement, and unsupervised Expectation-Maximization (EM) parameter training. On the BAliBASE (Thompson et al. 1999a) reference dataset, PROBCONS gives statistically significant improvements in alignment quality when compared to several leading alignment tools, including CLUSTALW (Thompson et al. 1994), DIALIGN (Morgenstern et al. 1998), and T-Coffee (Notredame et al. 2000), while maintaining comparable running times. Source code for our system is publicly available under the GNU Public License at http://probcons.stanford.edu/.

ALGORITHM

Given a set S of N sequences, PROBCONS applies the following procedure:

PROBCONS algorithm

1. (*Initial alignment*) For every pair of sequences x and y,
   a. Compute a posterior probability table using the HMM specified in the **HMM Topology** section, containing the posterior probabilities $P(x_i \sim y_j \mid x, y)$ for matching each letter x_i of one sequence against each letter y_j of the other.
   b. Compute the expected accuracy of the alignment, $E(x, y)$, defined to be the sum of posterior match probabilities along the highest summing path divided by the length of the shorter sequence.
2. (*Consistency transformation*) Simultaneously update all posterior probability matrices using the transformation
$$P'(x_i \sim y_j \mid x, y) = \frac{1}{N} \sum_{z \in S} \sum_{k} P(x_i \sim z_k \mid x, z) P(z_k \sim y_j \mid z, y) .$$
   Repeat this step for a total of two iterations.
3. (*Guide tree*) Given the expected accuracies for each pairwise alignment, compute a guide tree T using the

* To whom correspondence should be addressed.

following greedy hierarchical clustering procedure:
   a. Initially, place each sequence in its own cluster.
   b. Merge clusters x and y with maximum expected alignment accuracy. When the new cluster xy is formed, define its expected accuracy with any other cluster z to be $E(x, y)(E(x, z) + E(y, z)) / 2$.
   c. Repeat until only one cluster remains.
4. (*Progressive alignment*) Perform progressive multiple alignment according to the guide tree T using a sum-of-pairs objective function consisting of the sum of the re-estimated $P(x_i \sim y_j \mid x, y)$ terms for all aligned residue pairs; as before, no insertion penalties are used in calculating the highest summing path.
5. (*Iterative refinement*) Randomly partition the sequences in the current multiple alignment into two groups and realign. Repeat this step for a total of 100 iterations.

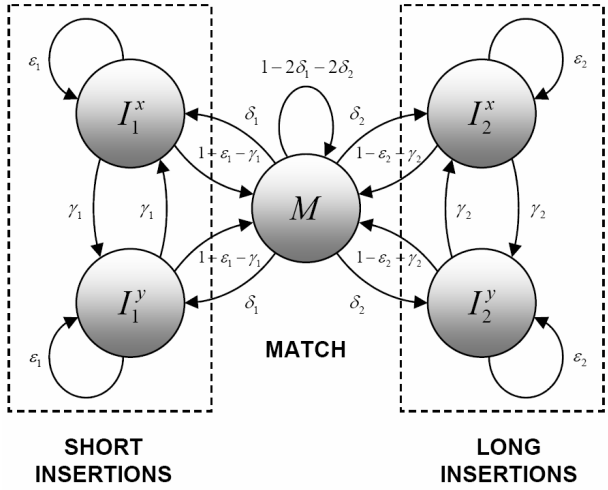

Fig. 1. Topology of pair-HMM.

HMM Topology

Given sequences x and y, the probability distribution over all possible global alignments is specified by a pair-HMM with states for matches (M), short insertions (I_1^x and I_1^y) and long insertions (I_2^x and I_2^y), with transitions given in **Fig. 1**.

Emission probabilities for amino acid pairs in the match states and single amino acid residues in the insertion states are based on statistics for the BLOSUM62 scoring matrix (Henikoff and Henikoff 1992). The forward and backward algorithms compute $P(x_i \sim y_j, x, y)$ and $P(x, y)$ directly, from which $P(x_i \sim y_j \mid x, y)$ may be calculated using Bayes's rule. Unsupervised expectation-maximization (EM) training over unaligned sequences may be used to estimate parameters.

For N sequences of length L, a naïve implementation of the algorithm has time complexity $O(N^3L^3)$ and space complexity $O(N^2L^2)$; by using sparse matrix representations for posterior probability matrices and imposing lower bound cutoffs on the retained values, these may be reduced to $O(N^3L^2)$ and $O(N^2L + L^2)$ in practice, respectively.

RESULTS

To test the PROBCONS aligner, we evaluated its performance on the BAliBASE 2.0 benchmark alignment database of 141 protein structural alignments (Thompson et al. 1999a, 1999b). The database is organized into five reference sets, reflecting different types of alignment problems that arise in practice. The ALN_COMPARE program (Notredame et al. 2000) was used to score alignments according to both the *sum-of-pairs score* (SP), the percentage of aligned core block residues also aligned in the reference, and the *column score* (CS), the percentage of aligned core block columns also aligned in the reference.

Comparison to existing tools

We compared the results of the PROBCONS aligner to those of CLUSTALW 1.83 (Thompson et al. 1994), DIALIGN 2.2.1 (Morgenstern et al. 1998), and T-Coffee 1.37 (Notredame et al. 2000). PROBCONS was run using two pairs of insertion states, two repetitions of the consistency transformation, and 100 iterations of iterative refinement. For testing, two-fold cross-validated EM training was used to estimate two separate parameter sets for testing. Note that in this unsupervised training, the *unaligned* sequences were used to derive all transition and initial state probabilities through 20 iterations of the Baum-Welch procedure. All other programs were run with default parameters.

From the results in **Table 1** and **Table 2**, PROBCONS shows a clear advantage over the other methods. Applying the Wilcoxon matched-pairs signed-ranks test over all 141 alignments indicated that PROBCONS performed significantly better than the other three methods (with $p < 10^{-5}$ for both SP and CS measures). As an aside, we also ran tests (see **Table 3**) which showed that (1) using two pairs of insertion states instead of one, (2) applying the consistency transformation, (3) using iterative refinement, and (4) performing additional rounds of EM training on each set of related proteins before aligning to estimate better sequence-specific parameters all gave measurable improvements in alignment quality.

CONCLUSION

PROBCONS combines several algorithmic techniques that contribute to high-accuracy protein multiple alignment. First, PROBCONS introduces the use of posterior-probability–based scoring for consistency-based progressive alignment. Other features include the use of two insertion state pairs for modeling both short and long insertions, construction of a guide tree based on expected accuracy rather than phylogenetic distance, and optional steps of iterative refinement or unsupervised EM retraining on each multiple alignment. As demonstrated, PROBCONS provides a dramatic improvement in alignment quality over current aligners while maintaining practical running times. Source code for the portable C++ implementation of the algorithm is freely downloadable from the website, http://probcons.stanford.edu/.

ACKNOWLEDGMENTS

We thank Mahathi Mahabhashyam and Sandhya Kunnatur for help in program development. CBD was partly supported by a Siebel Fellowship. MB was partly supported by an NSF Graduate Fellowship. SB and CBD were supported in part by NSF grant EF-0312459.

Table 1: Comparison of BALIBASE performance for DIALIGN, CLUSTALW, T-Coffee, and PROBCONS. The time required to run on the entire BAliBASE data is reported. The best result in each column is shown in bold.

Algorithm	*Ref* 1 (82)		*Ref* 2 (23)		*Ref* 3 (12)		*Ref* 4 (12)		*Ref* 5 (12)		*Overall* (141)		Time
	SP	CS	SP	CS	SP	CS	SP	CS	SP	CS	SP	CS	(mm:ss)
CLUSTALW	86.4	78.3	88.9	40.6	75.5	46.8	81.1	50.4	86.1	63.9	85.4	65.9	**1:05**
DIALIGN	81.3	71.4	85.0	27.9	68.6	34.8	91.2	81.9	94.1	84.5	82.8	63.2	3:04
T-Coffee	86.8	77.9	88.6	38.9	78.8	49.5	91.9	74.9	96.0	90.5	87.6	69.9	24:02
PROBCONS	**90.3**	**83.2**	**91.5**	**48.9**	**85.1**	**63.1**	**95.2**	**85.7**	**98.2**	**92.4**	**91.1**	**76.9**	8:26

Table 2: Percentage of alignments in which each method produced the (1) unique best alignment or (2) the best alignment (two or more methods achieved the same highest accuracy). The best results in each row are shown in bold.

Algorithm	CLUSTALW		DIALIGN		T-Coffee		PROBCONS	
	SP	CS	SP	CS	SP	CS	SP	CS
% unique best alignment	12.1%	9.9%	3.5%	5.7%	14.2%	14.2%	**51.8%**	**46.1%**
% best alignment	19.9%	21.3%	13.5%	19.1%	30.5%	34.0%	**68.1%**	**66.7%**

Table 3: Comparison of BALIBASE performance for PROBCONS variants. The four parameters varied over these runs include: *s*, the number of insertion state pairs in the HMM topology; *c*, the number of consistency transformation applied; *ir*, the number of rounds of iterative refinement via randomized partitioning; and *em*, the number of unsupervised EM iterations used to train sequence-specific parameters for each set before aligning.

s	*c*	*ir*	*em*	*Ref* 1 (82)		*Ref* 2 (23)		*Ref* 3 (12)		*Ref* 4 (12)		*Ref* 5 (12)		*Overall* (141)		Time
				SP	CS	SP	CS	SP	CS	SP	CS	SP	CS	SP	CS	(mm:ss)
1	0	0	0	87.7	79.3	89.1	36.2	83.4	52.3	86.5	63.0	96.2	85.9	88.2	69.1	2:45
2	0	0	0	87.8	79.3	89.8	41.5	83.3	52.6	86.6	63.9	95.2	83.4	88.2	69.9	5:08
2	1	0	0	89.1	81.4	91.2	47.3	85.5	62.3	90.5	73.2	97.5	90.5	90.0	74.3	5:29
2	2	0	0	89.4	81.8	91.5	48.9	85.1	63.1	90.5	73.2	98.2	92.4	90.2	75.0	5:54
2	2	100	0	90.3	83.2	91.5	48.9	85.1	63.1	95.2	85.7	98.2	92.4	91.1	76.9	8:26
2	2	100	1	90.6	83.5	91.7	49.5	85.3	63.8	95.2	85.7	98.2	92.4	91.4	77.2	14:14

REFERENCES

Attwood, T.K. (2002) The PRINTS database: a resource for identification of protein families. *Brief Bioinform*, **3**(3), 252-263.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) *Biological Sequence Analysis*. Cambridge UP, Cambridge, pp. 80-99.

Eddy, S. (1995) Multiple alignment using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol*, **3**, 114-20.

Feng, D.F., and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, **25**, 351-360.

Gotoh, O. (1990) Consistency of optimal sequence alignments. *Bull Math Biol*, **264**, 823-838.

Gotoh, O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol*, **264**, 823-838.

Henikoff, S., and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Nat Acad Sci USA*, **89**, 10915-10919.

Jaroszewski, L., Li, W., Godzik, A. (2002) In search for more accurate alignments in the twilight zone. *Prot Sci*, **11**(7), 1702-1713.

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**(2), 195-202.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994) Hidden Markov models in computational biology: applications to protein modeling. *J Mol Biol*, **235**, 1501-1531.

Mizuguchi, K., Deane, C.M., Blundell, T.L., and Overington, J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Prot Sci*, **7**, 2469-2471.

Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics*, **14**, 290-294.

Notredame, C., Higgins, D.G., and Heringa, J. (2000) T-Coffee: a novel method for multiple sequence alignments. *J Mol Biol*, **302**, 205-217.

Phillips, A., Janies, D., and Wheeler, W. (2000) Multiple sequence alignment in phylogenetic analysis. *Mol Phylogenet Evol*, **16**(3), 317-330.

Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*, **12**(2), 85-94.

Sonnhammer, E.L.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, **26**(1), 320-322.

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.

Thompson, J.D., Plewniak, F., and Poch, O. (1999a) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**(1), 87-88.

Thompson, J.D., Plewniak, F., and Poch, O. (1999b) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, **27**(13), 2682-2690.