BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 3 2003, pages 427–428
DOI: 10.1093/bioinformatics/btg008

PCMA: fast and accurate multiple sequence alignment based on profile consistency

*Jimin Pei, Ruslan Sadreyev and Nick V. Grishin**

Howard Hughes Medical Institute, and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Received on August 14, 2002; revised and accepted on October 21, 2002

ABSTRACT

Summary: PCMA (profile consistency multiple sequence alignment) is a progressive multiple sequence alignment program that combines two different alignment strategies. Highly similar sequences are aligned in a fast way as in ClustalW, forming pre-aligned groups. The T-Coffee strategy is applied to align the relatively divergent groups based on profile–profile comparison and consistency. The scoring function for local alignments of pre-aligned groups is based on a novel profile–profile comparison method that is a generalization of the PSI-BLAST approach to profile–sequence comparison. PCMA balances speed and accuracy in a flexible way and is suitable for aligning large numbers of sequences.

Availability: PCMA is freely available for non-commercial use. Pre-compiled versions for several platforms can be downloaded from ftp://iole.swmed.edu/pub/PCMA/.

Contact: jpei@mednet.swmed.edu; grishin@chop.swmed.edu

Multiple sequence alignment is an essential tool in analyzing protein sequences. Accurate and fast construction of multiple alignments is an important task. Progressive multiple alignment methods such as ClustalW (Thompson *et al.*, 1994) implement a greedy strategy that makes pairwise alignment of two sequences or pre-aligned sequence groups at each step. T-Coffee (tree-based consistency objective function for alignment evaluation; Notredame *et al.*, 2000) is a novel progressive method. It attempts to increase alignment accuracy by seeking consistency among a set of global and local pairwise alignments. The scoring function for aligning two sequences or two pre-aligned groups is determined by the whole set of sequences via two processes called library generation and library extension. Although T-Coffee can produce decent alignment accuracy, the consistency measure is time and memory consuming when sequence number is large.

*To whom correspondence should be addressed.

The difficulty of producing accurate alignments is largely determined by sequence diversity. In the work by Thompson *et al.* (1999), most tested programs gave rather accurate alignments if any sequence pair shows above 35% identity. Introducing divergent sequences caused a dramatic accuracy decrease for all tested programs. In practice, the target sequence set usually possesses both closely related and distantly related sequence pairs. We reason that integrating different alignment strategies to align different subsets of sequences can improve alignment efficiency. We have developed a program PCMA (profile consistency multiple sequence alignment) that applies the fast algorithm of ClustalW to align highly similar sequences and the T-Coffee algorithm to align the relatively divergent pre-aligned groups.

PCMA is a tree-based progressive method. The progressive alignment process has two stages. In the first stage, if any two neighboring sequences or pre-aligned groups have average pairwise sequence identity above a certain threshold, e.g. 40%, they are aligned by the ClustalW algorithm to form a new pre-aligned group. At the end of the first stage, similar sequences form pre-aligned groups with relatively low similarity between neighboring groups. In the second stage, consistency measure (library generation and extension) is applied to the pre-aligned groups, in a similar way as in the T-Coffee program. A library consists of global pairwise alignments and local pairwise alignments among the pre-aligned groups. Global pairwise alignments are made using the ClustalW algorithm. Local pairwise alignments are the ten top-scoring non-intersecting local alignments, between each pair of pre-aligned groups, generated using a modified Lalign program from the FASTA package (Huang and Miller, 1991; Pearson, 1998). The scoring function for local alignments is based on a novel profile–profile comparison method (COMPASS). COMPASS (comparison of multiple protein alignments with assessment of statistical significance) constructs optimal local profile–profile alignments and analytically estimates E-values for the detected similarities. The

Table 1. Comparison of alignment accuracy and time

Alignment set (# alignments)	PCMA(40)	T-Coffee	ClustalW1.81
Ref1 > 35% (28)	0.950	0.955	0.948
Ref1 20–40% (31)	0.914	0.894	0.870
Ref1 < 25% (23)	0.385	0.381	0.448
Ref2 (23)	0.545	0.568	0.582
Ref3 (12)	0.574	0.531	0.469
Ref4 (12)	0.697	0.702	0.554
Ref5 (12)	0.877	0.927	0.638
Total (141)	**0.724**	**0.725**	**0.689**
SMART (49)*	**0.852**	**0.841**	**0.780**
Average CPU time for SMART alignments (s)	**805**	**16 284**	**28.2**

All programs were run on the same machine (Dell PowerEdge 8450 server, Pentium III 700 MHz, 4 G RAM). The last row shows the average CPU time for SMART alignments. Others are alignment accuracy, measured as fraction of correctly aligned columns (column score) for the BaliBASE2 alignments (ref1 to ref5) and fraction of correctly aligned residue pairs (sum-of-pairs score) for the SMART alignment (*). For the methods of calculating these scores, refer to Thompson *et al.* (1999). Programs for calculating these scores are available at ftp://iole.swmed.edu/pub/PCMA/evalscore/.

scoring system and E-value calculation are based on a generalization of the PSI-BLAST approach (Altschul *et al.*, 1997) to profile–sequence comparison. Tested along with existing methods, COMPASS shows increased abilities for sensitive and selective detection of remote sequence similarities, as well as improved quality of local alignments. After consistency measure by library extension, the pre-aligned groups are progressively aligned based on the consistency objective function, forming the final alignment.

We tested PCMA on the BaliBASE2 (Bahr *et al.*, 2001) alignments and SMART alignments with the average identity threshold set to 40% (Table 1). Over the 141 BaliBASE2 alignments (ref1 to ref5), PCMA achieves alignment accuracy comparable to T-Coffee (default parameters) and both are significantly better than ClustalW (version 1.81), according to a Wilcoxon signed matched-pair rank test ($P < 0.001$). BaliBASE2 alignments are all of relatively small size (the largest number of sequences is 28). To test the performance of accuracy and speed on larger alignments, we used alignments from the SMART database (Schultz *et al.*, 1998) as benchmarks. SMART alignments, based on profile hidden Markov model searches and manual adjustments, are considered to have good quality. We chose 49 SMART alignments with sequence number between 100 and 200 as of June 2002. Both our program and T-Coffee achieve an average of about 85% correct in aligned pairs according to the sum-of-pairs score measure, which are significantly better than ClustalW (78% correctness). PCMA performs better than T-Coffee for 33 out of 49 alignments. PCMA is about 20 times faster than T-Coffee (Table 1).

To properly balance speed and accuracy, PCMA takes into account both sequence diversity and the user's preferences. The alignment speed is closely related to the average identity threshold, which the user can set. Lowering the threshold can cause more sequences merged in the first stage by ClustalW algorithm and thus increases the alignment speed. At a fixed threshold, the time complexity depends on the diversity of the target set of sequences. Generally, the more diverse the sequence set, the longer it takes to align them.

ACKNOWLEDGEMENTS

We thank Julie Thompson, Toby Gibson and Des Higgins for the ClustalW program, William Pearson and Webb Miller for the Lalign program from the FASTA package. We thank James Wrabl for helpful discussions.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,Z. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bahr,A., Thompson,J.D., Thierry,J.C. and Poch,O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.*, **29**, 323–326.

Huang,X.Q. and Miller,W. (1991) A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.*, **12**, 337–357.

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.

Pearson,W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.

Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.

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. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.