doi:10.1093/nar/gkm227

PROMALS web server for accurate multiple protein sequence alignments

Jimin Pei[1,*], Bong-Hyun Kim[2], Ming Tang[1] and Nick V. Grishin[1,2]

[1]Howard Hughes Medical Institute and [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 6001 Forest Park Road, Dallas, Texas 75390-9050, USA

Received January 31, 2007; Revised March 22, 2007; Accepted March 28, 2007

ABSTRACT

Multiple sequence alignments are essential in homology inference, structure modeling, functional prediction and phylogenetic analysis. We developed a web server that constructs multiple protein sequence alignments using PROMALS, a progressive method that improves alignment quality by using additional homologs from PSI-BLAST searches and secondary structure predictions from PSIPRED. PROMALS shows higher alignment accuracy than other advanced methods, such as MUMMALS, ProbCons, MAFFT and SPEM. The PROMALS web server takes FASTA format protein sequences as input. The output includes a colored alignment augmented with information about sequence grouping, predicted secondary structures and positional conservation. The PROMALS web server is available at: http://prodata.swmed.edu/promals/.

INTRODUCTION

The quality of multiple sequence alignments directly affects their applications in similarity searches, structure modeling, functional prediction and phylogenetic analysis. Preparing accurate multiple alignments for distantly related proteins (e.g. sequence identity below 20%) remains a difficult task. Fast accumulation of database protein sequences also poses a demand to improve alignment speed. Aligning all sequences together by dynamic programming is not feasible for large numbers of sequences (1). Progressive alignment methods reduce the problem of aligning multiple sequences to making a limited number of pairwise alignments. Although progressive methods can be fast, errors made at early stages are not corrected. Classic progressive methods such as ClustalW (2) can give reasonable results for similar sequences, but fail to produce accurate alignments for divergent sequences (3).

In recent years, extensive research has been conducted to improve alignment quality for progressive methods. Refinement after progressive steps is an effective way of correcting alignment errors (4,5). Consistency-based alignment strategy (6) derives a better scoring function before the progressive alignment steps. ProbCons (7) introduced and MUMMALS (8) implemented a probabilistic treatment of consistency derived from pairwise alignment hidden Markov models. Additional information from protein structures and database homologs can lead to further improvement of alignment quality (5,9–11).

We developed PROMALS (12), a progressive method that combines recent advanced techniques to improve multiple alignment quality, especially for distantly related proteins. PROMALS integrates additional information from database searches and secondary structure predictions into a new hidden Markov model that aligns profiles. The alignment scoring function of PROMALS is based on probabilistic consistency among profile–profile comparisons. PROMALS has shown improved results as compared to other leading methods, such as SPEM (13), MUMMALS, ProbCons and MAFFT (12).

Here, we describe the PROMALS web server for multiple protein sequence alignments. In addition to alignment construction, this server outputs useful information about predicted secondary structures, sequence grouping and positional conservation for target sequences.

PROMALS MULTIPLE ALIGNMENT PROCEDURE

Being a progressive method, PROMALS sets the order of pairwise alignments according to a tree built by a *k*-mer counting method (4). To improve alignment speed, PROMALS has two alignment stages for easy and difficult cases, as first implemented in our program PCMA (14). In the first stage, highly similar sequences are progressively aligned in a fast way with a weighted sum-of-pairs measure of BLOSUM62 (15) scores. This procedure results in a set of pre-aligned groups that are relatively divergent from each other. In the second

*To whom correspondence should be addressed. Tel: +214 645 5951; Fax: +214 645 5948; Email: jpei@chop.swmed.edu

Table 1. Evaluation of alignment methods on SABmark and PREFAB benchmarks

Method	SABmark-twi(209/7.7)	SABmark-sup(425/8.3)	PREFAB (1682/45.2)
PROMALS	**0.391**	**0.665**	**0.790**
SPEM	0.326	0.628	0.774
MUMMALS	0.196	0.522	0.731
ProbCons	0.166	0.485	0.716
MAFFT-linsi	0.184	0.510	0.722
MUSCLE	0.136	0.433	0.680
ClustalW	0.127	0.390	0.617

Average Q-scores of two SABmark data sets ('twi' for 'twilight zone' set, 'sup' for 'superfamily' set) and the PREFAB 4.0 data set are shown. Q-score is the number of correctly aligned residue pairs in the test alignment divided by the total number of aligned residue pairs in the reference alignment. For each data set, the two numbers in the parentheses separated by a slash are the number of alignments tested and the average number of sequences per alignment, respectively. For each data set, PROMALS yields statistically higher accuracy (bold numbers) than any other method (*P*-value < 0.000001) according to Wilcoxon signed rank test. PROMALS and SPEM use secondary structure prediction and database homologs in alignment process, while the other five methods only utilize the input sequences.

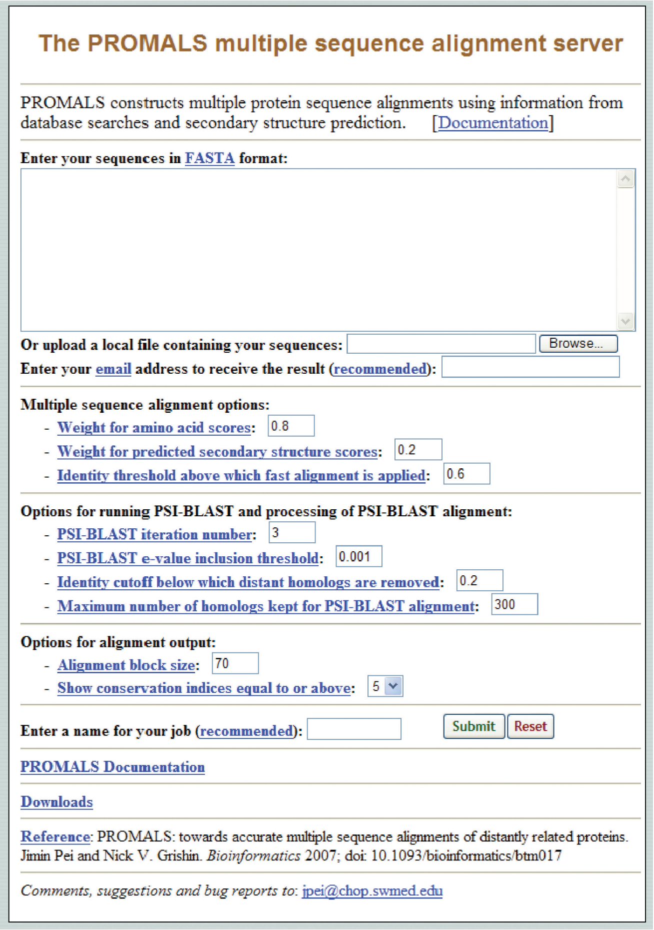
alignment stage, a representative sequence is selected from each pre-aligned group. For each representative sequence, PSI-BLAST (16) is used to identify homologs from the sequence database UNIREF90 (17), and the PSI-BLAST profile (checkpoint file) is used to predict secondary structures by PSIPRED (18). For each pair of representatives, profiles are derived from the PSI-BLAST alignments and PSIPRED secondary structure prediction, and a matrix of posterior probabilities of matches between positions are obtained by a profile–profile hidden Markov model (12). These matrices are used to calculate the probabilistic consistency scoring function, which is used to progressively align the representative sequences. Then the pre-aligned groups obtained in the first stage are merged to the alignment of the representatives. Finally, gap placements in highly gapped regions are refined to make the gap patterns more realistic. The alignment accuracy results of PROMALS and several other methods on SABmark (19) and PREFAB 4.0 (4) benchmarks are shown in Table 1.

PROMALS WEB SERVER

The PROMALS web server is available at: http://prodata.swmed.edu/promals/(Figure 1).

Input

The user can paste protein sequences or upload a sequence file. The sequences can be in FASTA format and identical sequence names are not allowed. PROMALS also recognizes CLUSTAL format alignments as input. If such an alignment is provided, it is split into individual sequences and these sequences will be re-aligned by PROMALS. The user can enter a name to identify the submitted job. It is also recommended that the user provide an email address to receive alignment results, as PROMALS can take a considerable amount of time to finish for a large number of divergent sequences, due to the time-consuming steps of running PSI-BLAST searches and profile consistency measure. On a data set of 1785 SCOP (20,21) domain pairs with up to 48 homologs added (the average number of sequences is 41.6 per alignment), the average CPU time of PROMALS is about half an hour under default settings (12). The actual time to finish an alignment job depends on factors such as the number of sequences and their lengths, the diversity among the sequences, the numbers of homologs found in database searches and the server load. It can take several hours for the server to finish aligning a sequence set with a large number of distantly related sequences (>50).



The PROMALS multiple sequence alignment server

PROMALS constructs multiple protein sequence alignments using information from database searches and secondary structure prediction. [Documentation]

Enter your sequences in FASTA format:

Or upload a local file containing your sequences:

Enter your email address to receive the result (recommended):

Multiple sequence alignment options:
- Weight for amino acid scores: 0.8
- Weight for predicted secondary structure scores: 0.2
- Identity threshold above which fast alignment is applied: 0.6

Options for running PSI-BLAST and processing of PSI-BLAST alignment:
- PSI-BLAST iteration number: 3
- PSI-BLAST e-value inclusion threshold: 0.001
- Identity cutoff below which distant homologs are removed: 0.2
- Maximum number of homologs kept for PSI-BLAST alignment: 300

Options for alignment output:
- Alignment block size: 70
- Show conservation indices equal to or above: 5

Enter a name for your job (recommended):

PROMALS Documentation

Downloads

Reference: PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Jimin Pei and Nick V. Grishin. *Bioinformatics* 2007; doi: 10.1093/bioinformatics/btm017

Comments, suggestions and bug reports to: jpei@chop.swmed.edu

Figure 1. Front page of the PROMALS server. The main section allows the user to paste or upload sequences and enter an email address for the results. Options to modify alignment parameters, PSI-BLAST searches and output format are provided. A brief description of each option is available by clicking on the option's name. A document with detailed description of the server is provided. The stand-alone versions of PROMALS can be downloaded from this page.

Alignment options

A number of alignment options are provided in the web page. One important parameter is the identity threshold that determines the partition of fast alignment stage and slow alignment stage, and thus balances alignment quality and speed. Lowering this threshold can cause more

```
Conservation:                       9   9 965 67             7                   66  99  6 6   9
Q394B3_BURS3_207_388        VIADRFADVSVLFADIVDFTGFSAGMRPEQLVEMLNEIFTGFDTIADHCGLEKIKTIGDAYMAAAGLPVP
Q7U096_MYCBO_208_376        KPLPGARQVTVAFADLVGFTQLGEVVSAEELGHLAGRLAGLARDLTA-PPVWFIKTIGDAVMLVCPDP--
Q1TAV7_9MYCO_205_370        LPLPGAREVTVMFADLVGFTRLGEAVPPEKLEQLARRLGDLARELAV-APVRFVKTIGDAVMLVSTDP--
Q5UFR5_MYCAV_46_217         ARVTPDGRVVILFTDIEESTALNERIGDRAWVKLISSHDKLVSDLVRRQSGHVVKSQGDGFMVAFARP--
Q1TAD0_9MYCO_95_261         ARLAPHGRVAILFSDIEDSTALNNRIGDRAWARLIGRHARSVQRHVREHDGHVVKSQGDGFMVAFASP--
Q60Z17_CAEBR_310_503        FTMNLMTNVSILFADIAGFTKMSSNKSADELVNLLNDLFGRFDTLCRLRGLEKISTLGDCYYCVAGCPEP
ADCY9_CHICK_374_562         FKMQQIEQVSILFADIVGFTKMSANKSAHALVGLLNDLFGRFDRLCEDTKCEKISTLGDCYYCVAGCPEP
Q1URK5_9MYCO_143_320        SGKPANPEVTLVFSDLVGFSSWALTAGDDATLRLLRRVAQAYEPPLLEAGGRIVKRMGDGSMAVFTDP--
Consensus_ss:                      eeeeeeee   hhhhhhh  hhhhhhhhhhhhhhhhhhhhh   eeeeee  eeeeee

Conservation:                           6    5               9 9   9              9  6  95 9 56
Q394B3_BURS3_207_388        AADHATRAAHMALDMIDALARFNAAR-HCNLKLRIGINSGEVVAGVIGKRKFIYDLWGAAVNLASRMESQ
Q7U096_MYCBO_208_376        -----APLLDTVLKLVEVVDTD-----NNFPRLRAGVASGMAVSR-------AGDWFGSPVNVASRVTGV
Q1TAV7_9MYCO_205_370        -----AALLEAALALLDAATSD-----AEFPRLRVGLAVGQAVSR-------AGDWFGSPVNLASRVTGA
Q5UFR5_MYCAV_46_217         -----EQAVRCGIELQRALRRNANRKRHEEIRVRIGIHMGRSVRR-------GDDLFGRNVAMAARVAAQ
Q1TAD0_9MYCO_95_261         -----ENAVRCAIALQHSLRRRPN-----GIRVRIGIHTGKSVRR-------GEDLFGRNVALAARVAAE
Q60Z17_CAEBR_310_503        CDDHACRTVEMGLDMIVAIRQFDIDR-GQEVNMRVGIHTGKVMCGMVGTKRFKFDVFSNDVTLANEMESS
ADCY9_CHICK_374_562         RADHAYCCIEMGLGMIKAIEQFCQEK-KEMVNMRVGVHTGTVLCGILGMRRFKFDVWSNDVNLANLMEQL
Q1URK5_9MYCO_143_320        -----GTAVRAVLNAMAAVRSVEID--GYSPRMRVGVHTGRPQRI-------GSDWLGVDVNTAARVMER
Consensus_ss:                    hhhhhhhhhhhhhhhhhhh       eeeeeeeeeeeeee         eee  hhhhhhhhhh

Conservation:               6  9   75 56                             7              5
Q394B3_BURS3_207_388        GVAGRVQVTDATRVMLGEA----------FVFEERGLIAAKGMG------EFRTWFVVG
Q7U096_MYCBO_208_376        ARPGAVLVADSVREALGDAPE-----ADGFQWSFAGPRRLRGIRG-----DVRLFRVRR
Q1TAV7_9MYCO_205_370        ARPGTVLVSESVREAVGDD--------ERFSWSYAGARHLKGIRG-----EVKLFRARR
Q5UFR5_MYCAV_46_217         AAGGEILVSQPVRDALSRSD--------GIRFDDGREVELKGFSG-----TYRLFAVLA
Q1TAD0_9MYCO_95_261         ADGGEILVSEAVRDAVAGAD--------GVSIGDGREVSLKGFSG-----KHHLYVVSA
Q60Z17_CAEBR_310_503        GVAGRVHVSEATAKLLKGLYEI----EEGPDYDGPLRMQVQGTERRVKPESMKTFFIKG
ADCY9_CHICK_374_562         GVAGKVHISEATAKYLDDRYE---------MEDGKVTERVGQSAVADQLKGLKTYLISG
Q1URK5_9MYCO_143_320        ATRGGLIVSQATLDRIPAEELAALNVTVKRQRRQVFSLKPDGVPP-----ELGMYRVRR
Consensus_ss:                   eeee hhhhhh              eeeee  eeee           eeeeeee
```

Figure 2. An example of colored alignment produced by the PROMALS server. These sequences are adenylate/guanylate cyclase catalytic domains selected from the PFAM database (Accession number: PF00211) (23). The first line in each alignment block begins with 'Conservation:' and shows conservation index numbers for conserved positions. The last line in each block begins with 'Consensus_ss:' and shows the consensus secondary structure predictions ('h': α-helix; 'e': β-strand). Each representative sequence has a magenta name and is colored according to PSIPRED secondary structure predictions (red: α-helix, blue: β-strand). A representative sequence and the immediate sequences below it with black names, if there are any, form a closely related group (determined by the option 'Identity threshold'). Sequences within each group are aligned in a fast way. The groups are aligned using profile consistency with enhanced information from database searches and secondary structure predictions.

sequences to be aligned in a fast and less accurate way, resulting in fewer representative groups subject to the time and memory-consuming steps of PSI-BLAST searches and profile consistency measure. This tradeoff generally leads to less computational time but lower alignment quality. If the number of pre-aligned groups is large (e.g. >100), PROMALS could run out of memory during the consistency measure step and generate an error message with the report of the number of pre-aligned groups in the second alignment stage. In this case, the user can lower the identity threshold (default 0.6) so that the number of sequence groups subject to consistency measure can be reduced. We also provide options for changing weights of amino acid scoring and predicted secondary structure scoring. The default values were determined by a large scale testing on divergent SCOP superfamily domains (20,21). Several parameters for running PSI-BLAST and processing PSI-BLAST alignments (used for generating amino acid profiles) are also provided, such as e-value cutoff, the number of PSI-BLAST iterations, identity cutoff to remove divergent hits, and the number of homologs kept for profile calculation.

Output of PROMALS results

The web server reports the resulting alignment in a standard CLUSTAL format. In addition, the server provides a colored alignment with information about sequence grouping, secondary structure predictions and positional conservation (Figure 2). Sequence grouping is reflected by the color of sequence names. Sequences with magenta names are representatives from pre-aligned groups. Sequences with black names immediately under a representative sequence belong to the same pre-aligned group as the representative sequence. For example, in Figure 2, 'Q7U096_MYCBO_208_376' and 'Q1TAV7_9MYCO_205_370' belong to the same pre-aligned group, and they are aligned in the fast alignment stage. Predicted secondary structures are shown for representative sequences (residues with red and blue fonts are predicted to be α-helices and β-strands, respectively). Above each alignment block, conserved positions are marked by their conservation indices (integer values from 0 to 9) calculated using our program AL2CO (22). The line beneath each alignment block shows consensus secondary structure predictions derived from predictions of individual representative sequences ('h': α-helix; 'e': β-strand). Such a coloring and labeling scheme provides additional information about the PROMALS alignment, and is helpful for further sequence and structural analysis of the target sequences. In addition to the alignments, the server also provides links to the original input sequences and intermediate results of PSI-BLAST alignments and PSI-PRED secondary structure predictions.

ACKNOWLEDGEMENTS

The authors would like to thank Lisa Kinch and Hong Zhang for testing the server and helpful comments. This work was supported in part by NIH grant GM67165 to NVG. Funding to pay the Open Access publication charges for this article was provided by Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Lipman,D.J., Altschul,S.F. and Kececioglu,J.D. (1989) A tool for multiple sequence alignment. *Proc. Natl Acad. Sci. USA*, **86**, 4412–4415.
2. 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.
3. Thompson,J.D., Plewniak,F. and Poch,O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.
4. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
5. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
6. 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.
7. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
8. 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.
9. Simossis,V.A. and Heringa,J. (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res.*, **33**, W289–W294.
10. Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.
11. O'Sullivan,O., Suhre,K., Abergel,C., Higgins,D.G. and Notredame,C. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.
12. Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related sequences. *Bioinformatics* doi: 10.1093/bioinformatics/btm017.
13. Zhou,H. and Zhou,Y. (2005) SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, **21**, 3615–3621.
14. Pei,J., Sadreyev,R. and Grishin,N.V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.
15. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
18. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
19. 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.
20. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
21. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
22. Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
23. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.