W30–W34 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 24 May 2008*
doi:10.1093/nar/gkn322

PROMALS3D web server for accurate multiple protein sequence and structure alignments

Jimin Pei[1,*], 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, TX 75390-9050, USA

Received January 31, 2008; Revised April 24, 2008; Accepted May 7, 2008

ABSTRACT

Multiple sequence alignments are essential in computational sequence and structural analysis, with applications in homology detection, structure modeling, function prediction and phylogenetic analysis. We report PROMALS3D web server for constructing alignments for multiple protein sequences and/or structures using information from available 3D structures, database homologs and predicted secondary structures. PROMALS3D shows higher alignment accuracy than a number of other advanced methods. Input of PROMALS3D web server can be FASTA format protein sequences, PDB format protein structures and/or user-defined alignment constraints. The output page provides alignments with several formats, including a colored alignment augmented with useful information about sequence grouping, predicted secondary structures and consensus sequences. Intermediate results of sequence and structural database searches are also available. The PROMALS3D web server is available at: http://prodata.swmed.edu/promals3d/.

INTRODUCTION

The quality of multiple sequence alignments directly affects their applications in structure modeling, similarity searches, function prediction and phylogenetic analysis. Constructing accurate multiple alignments for distantly related proteins remains a difficult task in computational biology. Aligning all sequences simultaneously by dynamic programing is not feasible for more than a few sequences (1). Therefore, many current programs use the heuristic progressive alignment technique, which reduces the problem of aligning multiple sequences to make a limited number of pairwise alignments. Although progressive methods can be very fast, errors made at early stages are not corrected later. Classic progressive methods based on general amino acid substitution matrices such as ClustalW (2) can give reasonable results for similar sequences, but fail to produce accurate alignments for divergent sequences (3). Refinement after progressive steps can correct alignment errors, as implemented in recent programs such as MAFFT (4) and MUSCLE (5). The consistency-based alignment strategy (6) derives a better scoring function than general substitution matrices before carrying out the progressive alignment steps. Using additional information from database homologs and known or predicted structures can lead to further improvement of alignment quality (4,7–10).

Our progressive method PROMALS (11) integrates advanced alignment techniques such as probabilistic consistency of profile–profile comparisons, and additional information from database homologs and predicted secondary structures. In PROMALS3D (12), alignment constraints from 3D structural comparisons are automatically derived and combined with constraints of PROMALS profile–profile alignments with secondary structures to derive consistency-based alignments. PROMALS3D has shown prominent improvements when 3D structures are available (Table 1).

Here we describe the PROMALS3D web server that constructs alignments for multiple protein sequences and/or structures. The output is a consensus alignment that brings together sequence and structural information about input proteins and their homologs. PROMALS3D server provides researchers a tool to produce high-quality alignments consistent with both sequences and structures in an automatic fashion. In addition to alignment construction, the server facilitates further analysis of target proteins by providing intermediate results of sequence and structural database searching, and presenting alignments with useful information about predicted secondary structures, sequence grouping and consensus sequences.

PROMALS3D MULTIPLE ALIGNMENT PROCEDURE

PROMALS3D (12) is a progressive method that clusters similar sequences and aligns them in a fast way, and uses more elaborate techniques to align the relatively divergent

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

clusters to each other. In the first alignment stage, PROMALS3D aligns similar sequences using a scoring function of weighted sum-of-pairs of BLOSUM62 (13) scores. The first stage is fast and results in a number of prealigned groups (clusters) that are relatively distant from each other. In the second alignment stage, one representative sequence is selected for each prealigned group. Representative sequences (also called targets or target sequences below) are subject to PSI-BLAST searches for additional homologs from UNIREF90 (14) database and to PSIPRED (15) secondary structure prediction. Then a hidden Markov model of profile–profile alignments with predicted secondary structure scoring is applied to pairs of representatives to derive sequence-based constraints. Structure-based constraints are derived from homologs with known structures (see details below) and are combined with sequence-based constraints to derive a probabilistic consistency scoring function (16). The representative sequences are progressively aligned using such a consistency scoring function, and the prealigned groups obtained in the first stage are merged into the alignment of representatives to form the final multiple sequence alignment.

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)
PROMALS3D	0.616	0.812	0.900
PROMALS	0.391	0.665	0.790
SPEM (10)	0.326	0.628	0.774
MUMMALS (24)	0.196	0.522	0.731
ProbCons (16)	0.166	0.485	0.716
MAFFT	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 (25) data sets (‘twi’ for ‘twilight zone’ set, ‘sup’ for ‘superfamilies’ set) and the PREFAB 4.0 (5) 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, PROMALS3D yields statistically higher accuracy than any other method (*P*-value <0.000001) according to Wilcoxon signed rank test.

In PROMALS3D, structural constraints are derived for representative sequences that have homologs with known structures. First, the program identifies homologs with 3D structures (homolog3D) for representative sequences. For each representative sequence, the profile of PSI-BLAST (stored as a checkpoint file) search against the UNIREF90 database is used to initiate a new PSI-BLAST search (one iteration, with -C option) against the SCOP40 domain database (17,18) that contains protein domain sequences with known structures. Only structural domains that pass certain similarity criteria (default: *e*-value <0.001 and sequence identity no <20%) are kept. Multiple homolog3Ds could be identified and used for one target sequence if it contains several distinct domains with known structures. Pairwise residue match constraints for two representative target sequences are derived from sequence-based target-to-homolog3D alignments and structure-based homolog3D-to-homolog3D alignments. For example, if residue *A* in target S1 is aligned to residue *B* in homolog3D T1, residue *B* in homolog3D T1 is aligned with residue *C* in homolog3D T2 according to a structure comparison program, and residue *C* in homolog3D T2 is aligned with residue *D* in target S2, then we deduce that residue *A* in sequence S1 is aligned with residue *D* in sequence S2, and this pair (*A*, *D*) is used as a structure-derived constraint (Figure 1). The alignment between a target sequence and its homolog3D can be the PSI-BLAST alignment, or they can be re-aligned by the profile–profile comparison routine used in PROMALS. The structure constraints among target sequences

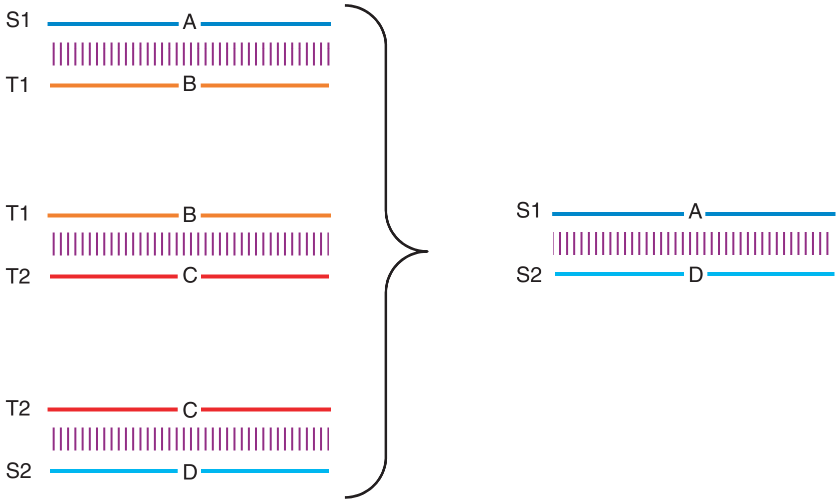
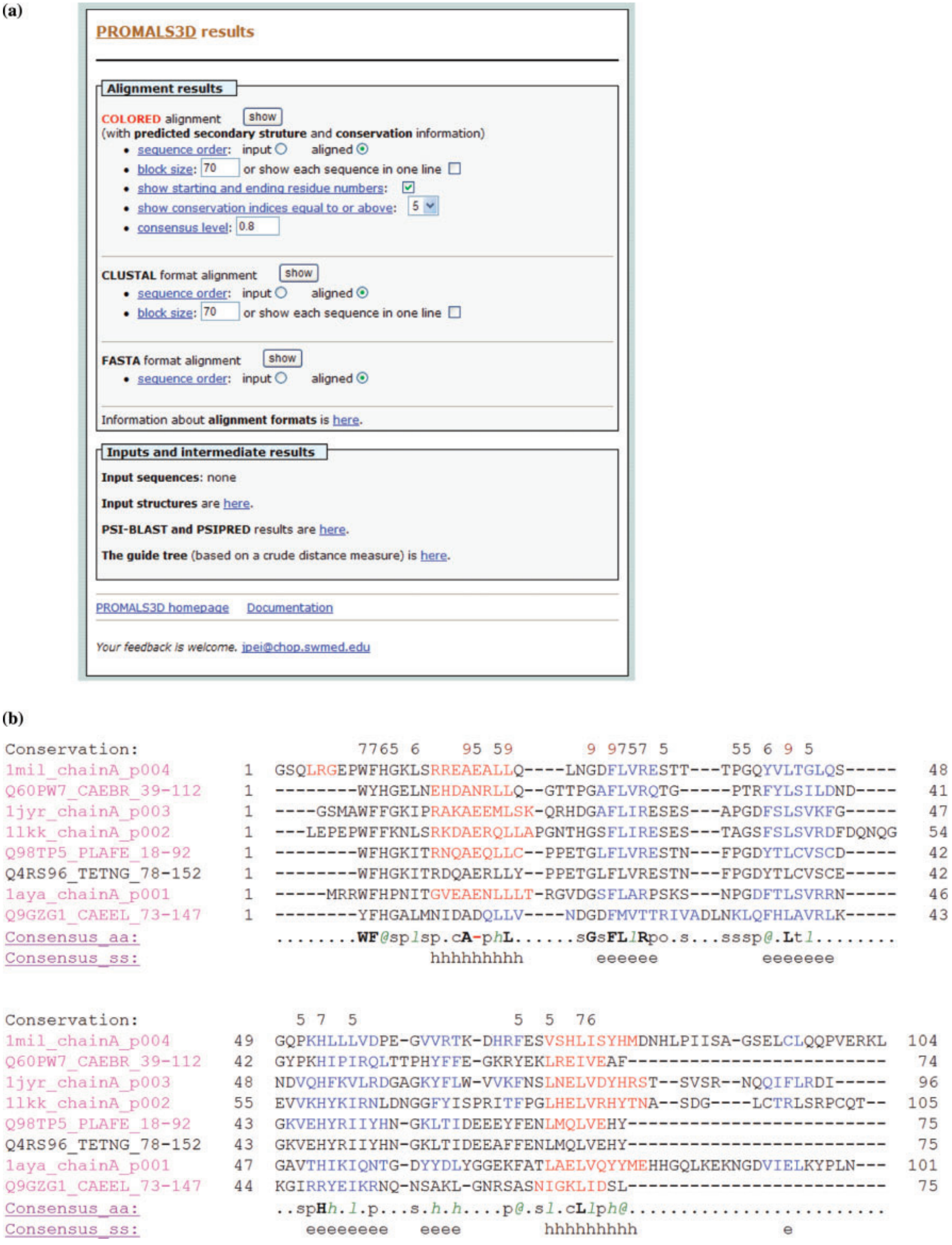


Figure 1. Deducing alignment constraints using homologs with 3D structures (homolog3Ds). S1 and S2 are two target sequences. T1 and T2 are their homolog3Ds. The alignment between two sequences S1 and S2 is deduced from two sequence-based sequence-to-homolog3D alignments and one structure-based homolog3D-to-homolog3D alignment. The three aligned residue pairs (A, B), (B, C) and (C, D) indicate that the pair (A, D) is aligned in the deduced alignment between two targets.



PROMALS3D results

Alignment results

COLORED alignment [show]
(with **predicted secondary struture** and **conservation** information)

- sequence order: input ○ aligned ●
- block size: 70 or show each sequence in one line ☐
- show starting and ending residue numbers: ☑
- show conservation indices equal to or above: 5
- consensus level: 0.8

CLUSTAL format alignment [show]

- sequence order: input ○ aligned ●
- block size: 70 or show each sequence in one line ☐

FASTA format alignment [show]

- sequence order: input ○ aligned ●

Information about **alignment formats** is here.

Inputs and intermediate results

Input sequences: none

Input structures are here.

PSI-BLAST and PSIPRED results are here.

The guide tree (based on a crude distance measure) is here.

PROMALS3D homepage Documentation

Your feedback is welcome. jpei@chop.swmed.edu

```
Conservation:                     7765 6    95 59       9 9757 5      55 6 9 5
1mil_chainA_p004       1  GSQLRGEPWFHGKLSRREAEALLQ----LNGDFLVRESTT---TPGQYVLTGLQS-----   48
Q60PW7_CAEBR_39-112    1  --------WYHGELNEHDANRLLQ--GTTPGAFLVRQTG-----PTRFYLSILDND----   41
1jyr_chainA_p003       1  ----GSMAWFFGKIPRAKAEEMLSK-QRHDGAFLIRESES---APGDFSLSVKFG-----   47
1lkk_chainA_p002       1  ---LEPEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESES---TAGSFSLSVRDFDQNQG   54
Q98TP5_PLAFE_18-92     1  --------WFHGKITRNQAEQLLC--PPETGLFLVRESTN---FPGDYTLCVSCD-----   42
Q4RS96_TETNG_78-152    1  --------WFHGKITRDQAERLLY--PPETGLFLVRESTN---FPGDYTLCVSCE-----   42
1aya_chainA_p001       1  -----MRRWFHPNITGVEAENLLLT-RGVDGSFLARPSKS---NPGDFTLSVRRN-----   46
Q9GZG1_CAEEL_73-147    1  --------YFHGALMNIDADQLLV----NDGDFMVTTRIVADLNKLQFHLAVRLK-----   43
Consensus_aa:             ........WF@splsp.cA-phL......sGsFLlRpo.s...sssp@.Ltl........
Consensus_ss:                            hhhhhhhhh       eeeeee          eeeeeee

Conservation:               5 7  5               5  5  76
1mil_chainA_p004      49  GQPKHLLLVDPE-GVVRTK-DHRFESVSHLISYHMDNHLPIISA-GSELCLQQPVERKL  104
Q60PW7_CAEBR_39-112   42  GYPKHIPIRQLTTPHYFFE-GKRYEKLREIVEAF-------------------------   74
1jyr_chainA_p003      48  NDVQHFKVLRDGAGKYFLW-VVKFNSLNELVDYHRST--SVSR--NQQIFLRDI-----   96
1lkk_chainA_p002      55  EVVKHYKIRNLDNGGFYISPRITFPGLHELVRHYTNA--SDG----LCTRLSRPCQT--  105
Q98TP5_PLAFE_18-92    43  GKVEHYRIIYHN-GKLTIDEEEYFENLMQLVEHY-------------------------   75
Q4RS96_TETNG_78-152   43  GKVEHYRIIYHN-GKLTIDEEAFFENLMQLVEHY-------------------------   75
1aya_chainA_p001      47  GAVTHIKIQNTG-DYYDLYGGEKFATLAELVQYYMEHHGQLKEKNGDVIELKYPLN---  101
Q9GZG1_CAEEL_73-147   44  KGIRRYEIKRNQ-NSAKL-GNRSASNIGKLIDSL-------------------------   75
Consensus_aa:             ..spHh.l.p...s.h.h....p@.sl.cLlph@..........................
Consensus_ss:                eeeeeeee   eeee        hhhhhhhhh              e
```

Figure 2. (**a**) Output page of the PROMALS3D server. This page allows access of PROMALS3D alignments in three formats with various adjustable parameters. Links to inputs and intermediate results are also provided. (**b**) An example of colored alignment. This alignment is for input of four SH2 domain sequences selected from PFAM database (23) and four SH2 structures (pdb ids are 1aya, 1jyr, 1lkk and 1mil). The first line in each alignment block begins with 'Conservation:' and shows conservation index numbers for conserved positions. The line in each block beginning with 'Consensus_ss:' shows the consensus secondary structure predictions ('h': α-helix; 'e': β-strand). The line in each block beginning with 'Consensus_aa' shows consensus amino acids. If the weighted frequency of certain type of residues is above a certain threshold, the consensus symbol of that type is displayed. Symbols are provided for the following types: conserved amino acid residues: bold and uppercase letters; aliphatic residues (I, V, L): *l*; aromatic residues (Y, H, W, F): @; hydrophobic residues (W, F, Y, M, L, I, V, A, C, T, H): *h*; alcohol residues (S, T): o; polar residues (D, E, H, K, N, Q, R, S, T): p; tiny residues (A, G, C, S): t; small residues (A, G, C, S, V, N, D, T, P): s; bulky residues (E, F, I, K, L, M, Q, R, W, Y): b; positively charged residues (K, R, H): +; negatively charged residues (D, E): −; charged (D, E, K, R, H): c. 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 and they are aligned in the first stage.

are combined with those constraints derived from profile–profile comparisons in the original PROMALS to deduce a consistency-based scoring function that integrates database sequence profiles, predicted secondary structures and 3D structural information. We used an empirical weight ratio of 1.5 (can be modified in server) for structure constraints relative to the sequence constraints of profile–profile comparison in the original PROMALS.

PROMALS3D WEB SERVER

The PROMALS3D web server is available at: http://prodata.swmed.edu/promals3d/.

Input

Users can input or upload protein sequences, structures or user-defined alignment constraints. The sequences should be in FASTA format and identical sequence names are not allowed. The structures should be in PDB format. In addition to uploading bulky structural files, users can also specify just the PDB ids and chain ids. The output is a multiple alignment of input sequences and sequences extracted from input structures. A name can be entered to identify the submitted job. It is also recommended that the user provides an email address to receive alignment results, as PROMALS3D can take a considerable amount of time (several hours) to finish for a large number of divergent sequences, mainly due to the time-consuming steps of running PSI-BLAST searches and calculating the profile-based consistency scoring function.

Alignment parameters

A number of alignment parameters 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 sequences to be aligned in a fast but less accurate way, resulting in fewer representative groups subject to the time- and memory-consuming steps of PSI-BLAST searches, structural comparisons and profile consistency measure. This tradeoff generally leads to less memory usage and computational time but potentially lower alignment quality. If the number of prealigned groups is large (e.g. >60), PROMALS3D could run out of memory during the consistency measure step. Therefore, if the number of prealigned groups is above a threshold (currently 60), the server automatically adjusts the identity threshold to keep the number of prealigned groups to a fixed number (currently 60). This automatic adjustment allows PROMALS3D to run for up to several thousand input sequences.

We also provide options for changing weights of sequence-based constraints and structural-based constraints, and the weights of amino acid scoring and predicted secondary structure scoring in profile–profile alignments. Parameters for running PSI-BLAST and processing PSI-BLAST alignments (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. For structure alignments of input structures or homologs with 3D structures, we provide options of using any combination of three structural comparison programs: DaliLite (19), FAST (20) and TM-align (21).

Output

We designed an output page that facilitates analysis of alignments (Figure 2a). Three alignment formats can be accessed in this page: CLUSTAL format, FASTA format and a colored alignment format. Sequences in the alignment can be displayed in aligned order or input order. In a colored alignment, useful information about sequence grouping, secondary structure predictions, positional conservation and consensus sequences (Figure 2b) is reported. Sequence grouping is reflected by the color of sequence names if sequences are in aligned order. Sequences with magenta names are representatives from prealigned groups. Sequences with black names immediately under a representative sequence belong to the same prealigned group as the representative sequence. 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 two lines beneath each alignment block show the consensus amino acid sequence (with symbols explained in Figure 2 legend) and consensus secondary structure predictions ('h': α-helix; 'e': β-strand). Such a coloring and labeling scheme is helpful for further sequence and structural analysis of input sequences and structures. In addition to the alignments, the server also provides links to the original input sequences and structures and intermediate results such as the guide tree, PSI-BLAST alignments, detected homologs with 3D structures, PSIPRED secondary structure predictions. Superimposed coordinates of input structures are also available.

ACKNOWLEDGEMENTS

We would like to thank Lisa Kinch, Kira Makarova, Andrei Osterman 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. 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.
5. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
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. 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.
8. 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.
9. 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.
10. Zhou,H. and Zhou,Y. (2005) SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, **21**, 3615–3621.
11. Pei,J. and Grishin,N.V. (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, **23**, 802–808.
12. Pei,J., Kim,B.H. and Grishin,N.V. (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.
13. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
14. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
15. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
16. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
17. 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.
18. 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.
19. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
20. Zhu,J. and Weng,Z. (2005) FAST: a novel protein structure alignment algorithm. *Proteins*, **58**, 618–627.
21. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
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., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
24. 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.
25. 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.