BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 7 2007, pages 802–808
doi:10.1093/bioinformatics/btm017

*Sequence analysis*

# PROMALS: towards accurate multiple sequence alignments of distantly related proteins

Jimin Pei[1,*] and Nick V. Grishin[1,2]

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

Received on December 4, 2006; revised on January 12, 2007; accepted on January 17, 2007

Advance Access publication January 31, 2007

Associate Editor: Alex Bateman

**ABSTRACT**

**Motivation:** Accurate multiple sequence alignments are essential in protein structure modeling, functional prediction and efficient planning of experiments. Although the alignment problem has attracted considerable attention, preparation of high-quality alignments for distantly related sequences remains a difficult task.

**Results:** We developed PROMALS, a multiple alignment method that shows promising results for protein homologs with sequence identity below 10%, aligning close to half of the amino acid residues correctly on average. This is about three times more accurate than traditional pairwise sequence alignment methods. PROMALS algorithm derives its strength from several sources: (i) sequence database searches to retrieve additional homologs; (ii) accurate secondary structure prediction; (iii) a hidden Markov model that uses a novel combined scoring of amino acids and secondary structures; (iv) probabilistic consistency-based scoring applied to progressive alignment of profiles. Compared to the best alignment methods that do not use secondary structure prediction and database searches (e.g. MUMMALS, ProbCons and MAFFT), PROMALS is up to 30% more accurate, with improvement being most prominent for highly divergent homologs. Compared to SPEM and HHalign, which also employ database searches and secondary structure prediction, PROMALS shows an accuracy improvement of several percent.

**Availability:** The PROMALS web server is available at: http://prodata.swmed.edu/promals/

**Contact:** jpei@chop.swmed.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Multiple sequence alignments have broad applications in sequence similarity searches, structure modeling and phylogenetic analysis (Altschul *et al.*, 1997; Eddy, 1998; Ginalski and Rychlewski, 2003; Phillips *et al.*, 2000). They also aid in experimental design by revealing conserved residues with potential functional importance. A variety of alignment methods that rely on different algorithms and scoring functions have been developed (Edgar and Batzoglou, 2006). A rigorous method that aligns all sequences simultaneously (Lipman *et al.*, 1989) is computationally prohibitive for large sets of sequences. In contrast, a progressive method that aligns pairs of sequences and sequence groups along a tree is algorithmically simpler and much faster, requiring only $N-1$ steps of pairwise alignments for $N$ sequences. However, in progressive methods, alignment errors made at each step are propagated to subsequent steps. Many progressive methods use a scoring function called sum-of-pairs, i.e. a sum of amino acid substitution scores for pairs of amino acids between two positions (Edgar and Batzoglou, 2006; Thompson *et al.*, 1994). Such a scoring function yields reasonable alignment quality for closely related sequences (identity above 40%). However, alignment quality drops rapidly with decreasing sequence similarity (Thompson *et al.*, 1999).

Effective construction of multiple alignments with respect to accuracy and speed has been extensively researched in recent years. Refinement and consistency-based scoring are two major techniques to improve classical progressive methods. MUSCLE (Edgar, 2004) and MAFFT (Katoh *et al.*, 2005) represent two recent methods that use extensive refinement to correct errors made in progressive steps. They both implement sum-of-pairs scores, which are easy to compute and offer the advantage of great speed. In T-COFFEE (Notredame *et al.*, 2000), the scoring is derived by finding consistently aligned residue pairs in a library of pairwise alignments. Such consistency-based scoring functions can give better alignment quality than sum-of-pairs scores. Further improvement comes with a probabilistic treatment of consistency via pairwise hidden Markov models (HMMs), as first implemented in ProbCons (Do *et al.*, 2005). MUMMALS (Pei and Grishin, 2006) builds on the success of probabilistic consistency by introducing HMMs with more states that capture local structural information. Consistency transformation requires operations on sequence triplets, and therefore is computationally intensive. By aligning similar sequences with general substitution matrices and aligning divergent sequence groups with profile-based consistency, PCMA (Pei *et al.*, 2003) is able to achieve a balance between alignment accuracy and speed.

Even with refinement and consistency-based scoring, current methods still have difficulty in obtaining high-quality alignments when sequence identity drops below 20%. As homologous proteins can have very low sequence similarity while maintaining similar structures and functions

*To whom correspondence should be addressed.

 

(Murzin, 1998), aligning distantly related sequences is an important task. A recent trend in the multiple alignment field is to recruit various sources of sequence and structural information to improve alignment accuracy (Edgar and Batzoglou, 2006). Such sources include homologs detected in database searches (Katoh *et al*., 2005; Simossis and Heringa, 2005; Thompson *et al*., 2000), predicted secondary structure (Simossis and Heringa, 2005; Zhou and Zhou, 2005), and known 3D structures (O'Sullivan *et al*., 2004). Since additional homologs improve the quality of sequence profiles, and structural features such as secondary structure are generally more conserved than sequences, their usage can lead to improved alignment quality.

Here, we describe PROMALS, a multiple sequence alignment method that combines recent advances in computational approaches to tackle the difficult task of aligning divergent sequences. PROMALS improves probabilistic consistency-based scoring of profiles by utilizing predicted secondary structures and additional homologs found in database searches. To effectively combine these additional data, we developed and implemented a new hidden Markov model for profile-profile comparison, which scores both amino acid similarity and secondary structure similarity, and has local structure-dependent transition and emission probabilities. Like PCMA, PROMALS is made more computationally efficient by treating similar and divergent sequences with different alignment strategies. On several difficult data sets, we show that PROMALS gives the best alignment accuracy among leading methods such as SPEM, HHalign (Soding, 2005), MUMMALS, ProbCons and MAFFT.

## 2 METHODS

### 2.1 A hidden Markov model of profile–profile alignment

A classical pairwise HMM for aligning two sequences has three types of hidden states: a match state 'M' emitting a residue pair, an 'X' state emitting a residue in the first sequence and a 'Y' state emitting a residue in the second sequence (Durbin *et al*., 1998). 'X' and 'Y' states correspond to insertions or deletions in the two sequences. Our hidden Markov model for aligning two alignments (having profile representations) has the same architecture as a pairwise sequence HMM. In our model, an 'M' state emits a pair of positions instead of a pair of residues. For an 'X' or 'Y' state, a single position in the first alignment or in the second alignment is emitted, respectively. The emitted objects (observations) are amino acid frequency vectors and predicted secondary structure types.

We adopt a representation of amino acid sequence profile similar to the ones in PSI-BLAST (Altschul *et al*., 1997) and COMPASS (Sadreyev and Grishin, 2003). Two profile components are estimated for a position in an alignment: (i) effective frequencies of amino acids, and (ii) target frequencies of amino acids. The effective frequencies serve as the emitted objects (observations) in a position for the hidden Markov model. They are estimated from the position-specific independent counts (PSIC) of amino acids (Pei and Grishin, 2001; Sunyaev *et al*., 1999), which is a sequence-weighting scheme that corrects for biased similarities between sequences. If an amino acid is not present in a position, it has an effective frequency of zero. The target frequencies serve as the 'hidden' amino acid probabilistic generator for a position. The target frequencies are estimated from the effective frequencies, taking into account prior knowledge of amino acid substitution characteristics. The target frequency is a mixture (weighted average) between effective frequency and the pseudocount frequency (Altschul *et al*., 1997; Tatusov *et al*., 1994). Defined in this way, the target frequency of any amino acid, even if it is not present in a position, is always greater than zero. Details on derivation of the two profile components are in Supplementary Data.

For an 'M' state, the probability of emitting the observed amino acids for a position pair $(i, j)$ is the product of two probabilities: (i) the probability of generating the effective frequencies of position $i$ using the target frequencies of position $j$, and (ii) the probability of generating the effective frequencies of position $j$ using the target frequencies of position $i$. For an 'X' or 'Y' state, the probability of emitting the observed amino acids in a position $k$ is the probability of generating the effective frequencies of position $k$ using the background amino acid frequencies in insertion regions. Besides amino acids, an 'M' state also emits a pair of predicted secondary structures, and an 'X' or 'Y' state also emits a single predicted secondary structure. The emission probability in a hidden state ('M', 'X' or 'Y') is a weighted product of amino acid emission probability and secondary structure emission probability. The relative weights for the scoring terms of amino acids and predicted secondary structures have been optimized to increase the alignment accuracy of the training sequence pairs. Details on emission probability formulas, parameter estimation and the algorithm for aligning two profiles with optimal posterior probabilities of position matches are described in Supplementary Data.

### 2.2 PROMALS multiple sequence alignment procedure

PROMALS (PROfile Multiple Alignment with predicted Local Structure) is a progressive method (Fig. 1). The alignment order is set by a tree built using a $k$-mer count method (Edgar, 2004). Like PCMA (Pei *et al*., 2003) and MUMMALS (Pei and Grishin, 2006), PROMALS has two alignment stages for easy and difficult alignments. In the first stage, highly similar sequences are progressively aligned in a fast way with a weighted sum-of-pairs measure of BLOSUM62 scores (Henikoff and Henikoff, 1992) (step 2 in Fig. 1). If two neighboring groups on the tree have an average sequence identity higher than a certain threshold (default: 60%), they are aligned in this fast way. The result of the first alignment stage is a set of sequences or pre-aligned groups that are relatively divergent from each other. In the second alignment stage, one representative sequence (the longest one) is selected from each pre-aligned group. For each representative, PSI-BLAST is used to search for homologs from sequence database UNIREF90 (Wu *et al*., 2006) with three iterations and an E-value cutoff of 0.001. Hits with <20% identity to the query are removed and up to 300 hits are selected. The PSI-BLAST checkpoint file after three iterations is used to predict secondary structures by PSIPRED (Jones, 1999). 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 is obtained by forward and backward algorithms of the profile-profile HMM (see Supplementary Data for details). These matrices are used to calculate the probabilistic consistency scores as described in Do *et al*. (2005). The representatives are then aligned progressively according to the consistency-based scoring function, and the pre-aligned groups obtained in the first stage are merged to the multiple alignment of the representatives. Finally, gap placement is refined to make the gap patterns more realistic. For that, we define a core block as a set of consecutive positions with gap content less than 0.5 at each position. A highly gapped ('gappy') region is defined as a set of consecutive positions with gap contents no less than 0.5 at each position. A gappy region is either bound by two adjacent core blocks, or is at the start or the end of the alignment. If there are $l$ amino acid residues in a gappy segment, gap refinement introduces continuous gap characters in between the $[l/2]$th residue and the $(l-[l/2])$th residue, with the exceptions for any gappy segment in N- or C-terminus,

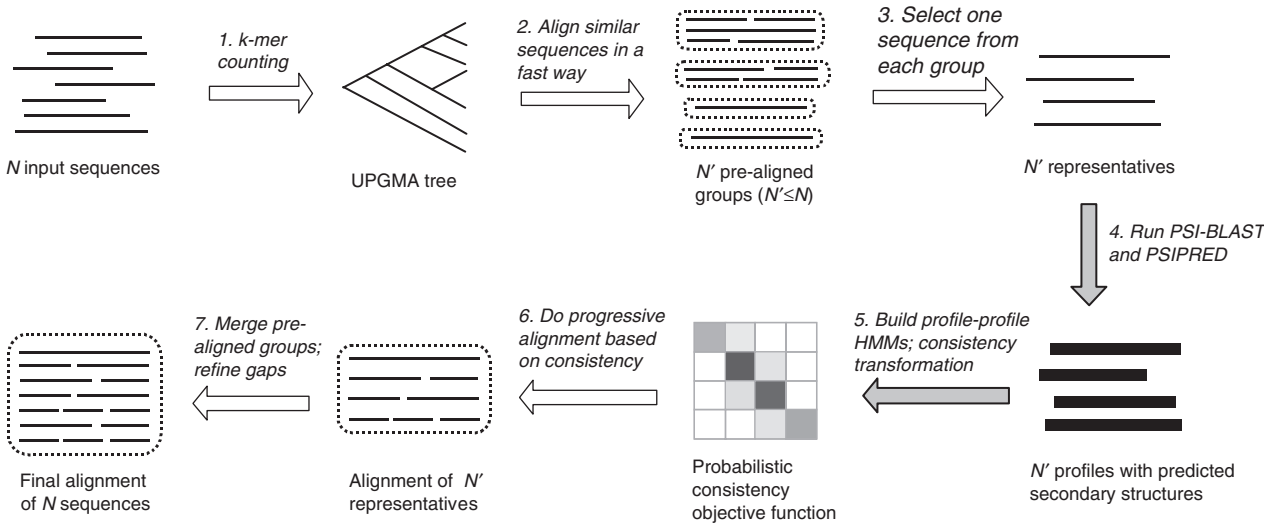


**Fig. 1.** PROMALS multiple sequence alignment procedure. The gray arrows indicate the two most time-consuming steps: running PSI-BLAST and PSIPRED (step 4) and profile consistency transformation (step 5).

where a single run of continuous gap characters is introduced at the sequence start or end.

### 2.3 Assessment of alignment methods

The following methods were tested: SPEM (Zhou and Zhou, 2005), HHalign (Soding, 2005), MUMMALS (Pei and Grishin, 2006), ProbCons (version 1.10) (Do *et al*., 2005), MAFFT (version 5.667) (Katoh *et al*., 2005), MUSCLE (version 3.52) (Edgar, 2004) and ClustalW (version 1.83) (Thompson *et al*., 1994). For MAFFT, we report two alignment options ('-linsi' and '-ginsi') that show the best results. HHalign is an enhanced version of HHsearch (Soding, 2005) that performs pairwise profile–profile alignment with predicted secondary structures (J. Soding, personal communication). Several parameters (score shift, secondary structure weight, pseudocount weight) of HHalign were selected that gave optimal performance on SCOP domain pairs with identity <20%.

For pairwise alignment tests, we used divergent SCOP superfamily domain pairs that were divided into three identity bins: below 10%, 10–15% and 15–20%. For multiple alignment tests, we added up to 24 homologs to each sequence in the testing cases of pairwise alignments. Details on construction of these testing data sets were given in our previous work (Pei and Grishin, 2006). Two large benchmark data sets compiled by other researchers were used as well. One is the SABmark database (version 1.65) (Van Walle *et al*., 2005), which contains two sets of multiple protein domains related at SCOP fold or superfamily level. The other is PREFAB database (version 4.0) (Edgar, 2004), which is based on structural alignments in FSSP database (Holm and Sander, 1998b) and homologous sequences from database searches. Reference-dependent alignment quality scores ($Q$-scores) were calculated using the built-in programs in SABmark and PREFAB packages. The $Q$-score is the number of correctly aligned residue pairs in the test alignment divided by the number of aligned residue pairs in the reference alignment. The value of the $Q$-score is between 0 and 1. Wilcoxon signed-ranks tests were performed to calculate the statistical significance of comparisons between alignment methods.

In addition to $Q$-score, we applied reference-independent evaluation of alignment quality to SCOP domain pairs, as described in our previous work (Pei and Grishin, 2006). We calculated several scores reflecting structural similarity of two SCOP domains compared according to aligned residues in a test alignment: DALI Z-score (Holm and Sander, 1998a), GDT-TS score (Zemla *et al*., 1999), TM-score (Zhang and Skolnick, 2004), 3D-score (Rychlewski *et al*., 2003) and two LiveBench contact scores (Rychlewski *et al*., 2003). These scores were scaled by taking into account self-comparison scores, random scores and alignment coverage (scaled scores are no larger than 1 and usually above 0). We also calculated two reference-independent sequence similarity scores: sequence identity and BLOSUM62 scores of aligned positions in a test alignment. These scores were also calculated for DaliLite (Holm and Sander, 1998a) structure-based alignments as a positive control.

## 3 RESULTS

PROMALS is a progressive multiple alignment method based on probabilistic consistency of profile-profile comparison, with enhanced profile information from homologs detected by PSI-BLAST and secondary structures predicted by PSIPRED (Fig. 1). SPEM and HHalign are comparable methods as they also use these two sources of extra data. While PROMALS and SPEM can align two or more sequences, HHalign performs only pairwise alignments. The other tested methods (MUMMALS, ProbCons, MAFFT, MUSCLE and ClustalW) are stand-alone multiple sequence methods that do not resort to other data sources or programs.

### 3.1 Reference-dependent evaluation of methods

*3.1.1 Tests on weakly similar SCOP domain pairs* We tested our profile-profile HMM on 1207 divergent SCOP domain pairs (Pei and Grishin, 2006) with <20% sequence identity (Table 1, first numbers in columns under 'SCOP'). The three methods that use extra data (PROMALS, SPEM and HHalign) produce substantially better results than stand-alone methods (MUMMALS, ProbCons, MAFFT, MUSCLE and ClustalW) that align a pair of sequences without using additional homologs or predicted secondary structures. For sequence

**Table 1.** Reference-dependent evaluation of alignment methods

| Method | SCOP[a] 0–10% (355) | SCOP[a] 10–15% (432) | SCOP[a] 15–20% (420) | SABmark-twi (209) | SABmark-sup (425) | PREFAB[c] (1682) |
|---|---|---|---|---|---|---|
| PROMALS | **0.435/0.457** | **0.612/0.619** | **0.761/0.772** | **0.391** | **0.665** | **0.790** |
| SPEM | 0.377/0.411 | 0.558/0.578 | 0.727/0.751 | 0.326 | 0.628 | 0.774 |
| HHalign[b] | 0.406/– | 0.567/– | 0.730/– | – | – | 0.787 |
| MUMMALS | 0.151/0.329 | 0.335/0.520 | 0.586/0.732 | 0.196 | 0.522 | 0.731 |
| ProbCons | 0.116/0.290 | 0.294/0.486 | 0.536/0.701 | 0.166 | 0.485 | 0.716 |
| MAFFT-linsi | 0.116/0.301 | 0.262/0.500 | 0.495/0.707 | 0.184 | 0.510 | 0.722 |
| MAFFT-ginsi | 0.116/0.308 | 0.265/0.497 | 0.496/0.714 | 0.176 | 0.495 | 0.715 |
| MUSCLE | 0.139/0.262 | 0.293/0.452 | 0.507/0.661 | 0.136 | 0.433 | 0.680 |
| ClustalW | 0.136/0.210 | 0.270/0.357 | 0.482/0.565 | 0.127 | 0.390 | 0.617 |

Average *Q*-scores of three testing data sets of ASTRAL SCOP40 superfamily pairs, two SABmark data sets (twi—'twilight zone' set, sup— '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. The number of alignments in each testing data set is shown in parentheses. Identity ranges are shown for the three SCOP data sets. The first three methods use extra data from PSI-BLAST and PSIPRED. The other five are stand-alone methods. The option of MUMMALS (modeling secondary structure and solvent accessibility) is set to produce the best results on these data sets. 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.

[a]For tests on the SCOP data sets, there are two numbers in each cell separated by a slash. The first number is the average *Q*-score in pairwise alignment tests and the second number is the average *Q*-score in multiple alignment tests.

[b]HHalign only performs pairwise profile–profile alignments and does not construct multiple sequence alignments. Thus the values for SCOP multiple alignment tests and SABmark tests are not available.

[c]For PREFAB 4.0 data set, the scores of PROMALS, HHalign and SPEM are based on pairwise profile–profile alignments, while the scores for other methods are based on multiple alignments.

pairs with identity below 10%, the average *Q*-score of PROMALS (0.431) is almost three times higher than that of MUMMALS (0.156). For alignments with identity ranges 10–15% and 15–20%, PROMALS also gives substantial accuracy increases over MUMMALS of 0.272 and 0.176, respectively. PROMALS shows about 3–4% accuracy increases over SPEM and HHalign, suggesting that our profile-profile HMM utilizes homologs and predicted secondary structures in a better way.

We also tested the methods (except HHalign, which is a pairwise alignment program) on data sets of multiple sequences constructed by adding up to 48 homologs to each SCOP domain pair (Table 1, second numbers in columns under 'SCOP'). With multiple sequences, PROMALS and SPEM both show slight improvement (1–2% for PROMALS and 2–3% for SPEM) over their pairwise profile–profile alignments. PROMALS outperforms SPEM by ~2% on multiple sequences. With added homologs, stand-alone methods all yield better accuracies than pairwise sequence alignments, among which MUMMALS is the best method. PROMALS outperforms MUMMALS by 0.13, 0.1, and 0.05 for data sets with identities <10%, 10–15% and 15–20%, respectively.

*3.1.2 Tests on SABmark database* SABmark database (version 1.65) has two multiple alignment benchmark sets. The 'twilight zone' set contains 209 tests of SCOP (version 1.65) fold-level domains with very low similarity, and the 'superfamily' set contains 425 tests of SCOP superfamily-level domains with low to intermediate similarity. PROMALS achieves the best results among all methods for both sets. Its accuracy is ~6% and 4% higher than SPEM on 'twilight zone' set and 'superfamily' set, respectively. For the most difficult 'twilight zone' set, PROMALS doubles the accuracy of the best stand-alone method (MUMMALS). Nevertheless, only ~40% residues were correctly aligned on average by PROMALS for the 'twilight zone' set, suggesting that homology modeling of extremely divergent domains remains a difficult problem with regard to alignment quality.

*3.1.3 Tests on and PREFAB database* PREFAB 4.0 database consists of 1682 alignments averaging 45.2 sequences per alignment. Each alignment consists of two sequences with known structures and their homologs found by PSI-BLAST database searches. The reference structural alignment in each test is based on the consensus of FSSP (Holm and Sander, 1998b) and CE (Shindyalov and Bourne, 1998) alignments. We have used the performances of pairwise profile–profile alignments of PROMALS and SPEM as an indicator of their multiple alignment performances. The three methods that use additional data (PROMALS, SPEM and HHalign) give similar results, each with an average *Q*-score above 0.75. Their accuracies are higher than those on the two SCOP data sets with identity <15% and the two SABmark sets, suggesting that PREFAB 4.0 is an easier testing data set. PROMALS, SPEM and HHalign are more accurate than MUMMALS by 4–6%. PROMALS is statistically more accurate (*P*-value <0.000001) than SPEM and HHalign despite small differences in their average *Q*-scores. Results on PREFAB 4.0 confirm that alignment quality differences between methods become smaller on easier tests.

## 3.2 Reference-independent evaluation of methods

On our data sets of 1207 SCOP domain pairs with identity below 20%, we evaluated alignment quality using reference-independent scores that reflect the similarity between two structures compared according to aligned residue pairs in the test alignment (Pei and Grishin, 2006). These structural

**Table 2.** Reference-independent evaluation on 1207 representative SCOP40 domain pairs with identity <20%

| Method | Structural similarity | | | | | | Sequence similarity | |
|---|---|---|---|---|---|---|---|---|
| | DALI Z-score | GDT-TS | TM-score | 3D-score | LBcona | LBconb | Identity | BLOSUM62 |
| PROMALS | **0.1562**[a] | **0.3079**[a] | **0.3675**[a] | **0.3097**[a] | **0.2692**[a] | **0.3527**[a] | 0.0868 | 0.1555 |
| SPEM | 0.1400 | 0.2886 | 0.3451 | 0.2893 | 0.2521 | 0.3319 | 0.0992 | 0.1724 |
| HHalign | 0.1334 | 0.2914 | 0.3488 | 0.2907 | 0.2469 | 0.3263 | 0.0874 | 0.1535 |
| MUMMALS | 0.1231 | 0.2570 | 0.3070 | 0.2563 | 0.2240 | 0.2909 | 0.0932 | 0.1651 |
| ProbCons | 0.1003 | 0.2324 | 0.2767 | 0.2307 | 0.2060 | 0.2670 | 0.0983 | 0.1719 |
| MAFFT-linsi | 0.1135 | 0.2485 | 0.2982 | 0.2467 | 0.2143 | 0.2820 | 0.0923 | 0.1632 |
| MAFFT-ginsi | 0.1126 | 0.2454 | 0.2960 | 0.2429 | 0.2152 | 0.2803 | 0.0972 | 0.1725 |
| MUSCLE | 0.0980 | 0.2297 | 0.2777 | 0.2266 | 0.1941 | 0.2535 | 0.0939 | 0.1686 |
| ClustalW | 0.0723 | 0.1916 | 0.2318 | 0.1876 | 0.1551 | 0.2030 | 0.0733 | 0.1344 |
| DaliLite | 0.4206 | 0.4936 | 0.5571 | 0.5289 | 0.4087 | 0.5110 | **0.0697**[b] | **0.1268**[b] |

The first three methods use extra data given by PSI-BLAST and PSIPRED. The last method (DaliLite) produces alignments based on comparison of known 3D structures. The other five are stand-alone methods. All sequence-based methods except HHalign construct multiple sequence alignments for target domain pairs with up to 48 homologs. HHalign constructs pairwise profile–profile alignments. Scores are calculated for pairwise alignments of target domain pairs extracted from multiple sequence alignments.

[a]PROMAL yields statistically higher structure-similarity scores (in bold) than other sequence alignment methods ($P$-value < 0.000001) according to Wilcoxon signed rank test.

[b]DaliLite structure-based sequence alignments have the lowest average sequence similarity scores (in bold).

similarity scores are DALI Z-score, TM-score, GDT-TS score, 3D-score, and two LiveBench contact scores (Table 2). Consistent with reference-dependent evaluation, PROMALS produces significantly higher average structural similarity scores than other methods. Used as a positive control, structural alignment method DaliLite yields higher structural similarity scores than any sequence-based alignment method (Table 2). Interestingly, DaliLite alignments have the lowest reference-independent sequence similarity scores (sequence identity and BLOSUM62 scores). PROMALS also shows lower sequence similarity scores than several other sequence-based methods. These observations suggest that for distantly related sequences (sequence identity <20%), sequence similarity scores, such as identity or BLOSUM62, may not correlate with alignment quality measured by 3D structural comparison, and maximization of these scores may not improve structural models based on sequence alignments.

### 3.3 Pairwise comparisons of alignment methods

To gain further understanding of the differences between alignment methods, we compared their performance on individual domain pairs from the SCOP sets (identity <20%). Table 3 shows the number of pairs, for which one method performs better than another method by a relatively large margin of 0.1 or more (measured by scaled TM-score or $Q$-score, both scores are between 0 and 1). Although PROMALS clearly leads by a large margin, it does not offer the best alignment in each and every case. For example, PROMALS gives a TM-score increase of 0.1 or more over SPEM on 197 alignments, while producing significantly inferior alignments for 109 pairs. Even stand-alone methods (MUMMALS, ProbCons, MAFFT, MUSCLE and ClustalW) outperform PROMALS by a TM-score of 0.1 or more on a small number of pairs (~5%, i.e. 49–67 out of 1207 alignments). These comparisons suggest that alignments constructed by different methods can vary much for divergent sequences, and a method with an overall inferior performance is capable of generating better alignments in some cases. Careful inspection of alignments produced by several programs could help improve alignment quality for divergent sequences.

## 4 DISCUSSION

Judging by its performance, PROMALS is a definite advance compared to our previous alignment programs MUMMALS (Pei and Grishin, 2006). MUMMALS derives probabilistic consistency from pairwise HMMs with built-in local structural information (secondary structure and/or solvent accessibility), and shows slight but significant improvement (a few percent) over other stand-alone methods such as ProbCons (Do *et al.*, 2005) and MAFFT (Katoh *et al.*, 2005). However, since no additional homologs are used, the local structure prediction implicitly performed by MUMMALS is of low accuracy compared to advanced methods such as PSIPRED (Jones, 1999). In contrast, PROMALS incorporates database searches and more accurate secondary structure prediction, and derives probabilistic consistency from profile–profile HMMs. Moreover, the HMM in PROMALS has a two-track structure (Karchin *et al.*, 2003) that treats both amino acids and predicted secondary structures as emitted objects, while MUMMALS HMMs only emit amino acids. Owing to additional data sources and the advanced profile–profile HMM, PROMALS shows significant improvement over MUMMALS and other stand-alone methods, especially for highly divergent sequences.

The HMM in PROMALS adopts a numerical representation of sequence profile (see Supplementary Data for details) that successfully works in other profile-sequence or profile–profile

**Table 3.** Pairwise comparisons among alignment methods on 1207 SCOP domain pairs with identity <20%

| | PROMALS | SPEM | HHalign | MUMMALS | ProbCons | MAFFT-linsi | MAFFT-ginsi | MUSCLE | ClustalW |
|---|---|---|---|---|---|---|---|---|---|
| PROMALS | – | **109/196** | **76/179** | **67/340** | **44/458** | **67/398** | **61/374** | **60/464** | **49/650** |
| SPEM | **199/81** | – | 140/148 | 108/281 | 71/389 | 98/324 | 99/326 | 82/400 | 43/574 |
| HHalign | **265/84** | 196/121 | – | 77/254 | 49/368 | 73/288 | 78/301 | 66/393 | 53/571 |
| MUMMALS | **685/286** | 648/305 | 627/333 | – | 38/169 | 111/138 | 82/128 | 82/227 | 59/431 |
| ProbCons | **726/263** | 693/277 | 674/303 | 201/62 | – | 172/80 | 162/76 | 162/169 | 110/336 |
| MAFFT-linsi | **718/276** | 680/295 | 662/325 | 239/128 | 133/188 | – | 85/98 | 93/196 | 60/387 |
| MAFFT-ginsi | **714/271** | 676/284 | 664/313 | 199/117 | 113/184 | 111/132 | – | 90/185 | 67/395 |
| MUSCLE | **783/239** | 741/255 | 727/279 | 401/83 | 302/138 | 295/110 | 327/106 | – | 75/288 |
| ClustalW | **858/193** | 840/209 | 819/228 | 649/55 | 559/103 | 585/70 | 600/76 | 449/100 | – |

Each off-diagonal cell has two numbers separated by a slash. The first number is the number of pairs where the alignment score of the method listed to the left is inferior to that of the method listed above (in a column) by 0.1 or more. The second number is the number of pairs where the score of the method listed to the left is better than that of the method listed above by 0.1 or more. The alignment quality scores used for comparison in the lower triangle and the upper triangle are *Q*-scores and weighted and scaled TM-scores, respectively. These scores are calculated based on results of multiple sequence alignments (target domain pairs plus up to 48 added homologs), with the exception of HHalign alignments, which are pairwise profile–profile alignments. Comparisons of PROMALS with other methods are highlighted in bold.

alignment methods such as PSI-BLAST (Altschul *et al.*, 1997) and COMPASS (Sadreyev and Grishin, 2003). A recent comprehensive study also supported the effectiveness of this profile–profile scoring scheme (Wang and Dunbrack, 2004). To adequately use predicted secondary structures, we not only score them as emitted objects, but also use transition and emission probabilities that are dependent on predicted secondary structure types (Supplementary Data). Unlike HHalign, which treats each alignment as a classical profile HMM (Eddy, 1998), our HMM has a simpler structure similar to the classical 3-state pairwise HMM (Durbin *et al.*, 1998). SPEM (Zhou and Zhou, 2005) does not use HMMs, but applies an empirical profile–profile alignment method ($SP^2$) that identifies the optimal alignment path. In contrast, the HMM in PROMALS allows estimation of posterior probabilities of matches between positions. As a result, PROMALS has a probabilistic treatment of consistency similar to the one in ProbCons and MUMMALS, while simple consistency measures are used in SPEM, T-COFFEE (Notredame *et al.*, 2000) and PCMA (Pei *et al.*, 2003). PROMALS performs significantly better than SPEM and HHalign on difficult tests, suggesting the advantages of our profile–profile comparison scheme.

Since PROMALS relies on PSI-BLAST and PSIPRED to collect additional homologs and predicted secondary structures, the speed of PROMALS is considerably slower than that of stand-alone progressive methods. Our strategy for improving speed is to use different algorithms for easy and difficult alignments (Pei *et al.*, 2003). By aligning highly similar sequences in a fast way, the number of sequences subject to the time-consuming steps (running PSI-BLAST, PSIPRED and consistency transformation) could be substantially reduced. For example, for 1207 SCOP domain pairs with up to 48 added homologs, the average number of sequences in an alignment is 41.6. After PROMALS aligns similar sequences with identity above 60% in the first stage, only ~24 sequences on average require database searches, secondary structure prediction, and consistency transformation. For these tests, the median CPU time of PROMALS is ~30 min per alignment, as compared to 67 min for SPEM (on Redhat Enterprise Linux 3, AMD Opteron 2.0 GHz). The stand-alone methods (MUMMALS, PROBCONS, MAFFT, MUSCLE and ClustalW) are much faster, all with a median CPU time <1 min.

As in our previous work (Pei and Grishin, 2006), we demonstrated the effectiveness of reference-independent evaluation of alignment quality in this study. First, we observed a good correlation between reference-dependent and reference-independent evaluations, suggesting that it may not be necessary to spend significant efforts on development of reference alignment databases. Second, reference-independent techniques solve the problem of reference alignment ambiguity, which becomes significant when similarity is low. Third, reference-independent evaluation helps answer general questions such as whether alignments can be further improved for sequences with low similarity, and whether such improvements will help structure modeling. For several structural similarity measures (GDT-TS, 3Dscore, TM-score, LB contact scores), the ratio between the average score of PROMALS sequence-based alignment and the average score of DaliLite structure-based alignment is ~0.6 on domain pairs with <20% sequence identity (Table 2), suggesting that we are still 40% below what can be achieved with structures in hand. Notably, for these divergent sequences, DaliLite structural alignments have lower sequence similarity scores (identity and BLOSUM62 scores) than alignments produced by any sequence method, suggesting that scoring functions based only on amino acid sequence similarity may not be suitable for aligning divergent sequences for the purpose of homology modeling. This observation further justifies the use of alternative scoring schemes, such as the ones that recruit structural information.

## ACKNOWLEDGEMENTS

We would like to thank Bong-Hyun Kim for the reference-independent evaluation routine, and Johannes Soding for providing the HHalign program. We would like to thank Lisa Kinch, Ruslan Sadreyev and James Wrabl for critical reading

of the manuscript and helpful comments. This work was supported in part by NIH grant GM67165 to NVG.

*Conflict of Interest*: none declared.

## REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Do,C.B. *et al*. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res*., **15**, 330–340.

Durbin,R. *et al*. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.

Edgar,R.C. and Batzoglou,S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol*, **16**, 368–373.

Ginalski,K. and Rychlewski,L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res*., **31**, 3291–3292.

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

Holm,L. and Sander,C. (1998a) Dictionary of recurrent domains in protein structures. *Proteins*, **33**, 88–96.

Holm,L. and Sander,C. (1998b) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res*., **26**, 316–319.

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

Karchin,R. *et al*. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins*, **51**, 504–514.

Katoh,K. *et al*. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*., **33**, 511–518.

Lipman,D.J. *et al*. (1989) A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA*, **86**, 4412–4415.

Murzin,A.G. (1998) How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol*., **8**, 380–387.

Notredame,C. *et al*. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*., **302**, 205–217.

O'Sullivan,O. *et al*. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol*., **340**, 385–395.

Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.

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.

Pei,J. *et al*. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.

Phillips,A. *et al*. (2000) Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol*., **16**, 317–330.

Rychlewski,L. *et al*. (2003) LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **53** (Suppl. 6), 542–547.

Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol*., **326**, 317–336.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*., **11**, 739–747.

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**, W289–294.

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Sunyaev,S.R. *et al*. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng*., **12**, 387–394.

Tatusov,R.L. *et al*. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. USA*, **91**, 12091–12095.

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

Thompson,J.D. *et al*. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res*., **28**, 2919–2926.

Van Walle,I. *et al*. (2005) SABmark—a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.

Wang,G. and Dunbrack,R.L., Jr. (2004) Scoring profile-to-profile sequence alignments. *Protein Sci*., **13**, 1612–1626.

Wu,C.H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–191.

Zemla,A. *et al*. (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, (Suppl. 3), 22–29.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhou,H. and Zhou,Y. (2005) SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, **21**, 3615–3621.