

Sequence analysis

SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structuresHongyi Zhou¹ and Yaoqi Zhou^{1,2,*}¹Department of Physiology and Biophysics, Howard Hughes Medical Institute Center for Single Molecule Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214, USA and²The Key Laboratory of Molecular Engineering of Polymers, Department of Macromolecular Science, Fudan University, Shanghai, China

Received on May 3, 2005; revised on July 5, 2005; accepted on July 12, 2005

Advance Access publication July 14, 2005

ABSTRACT

Motivation: Multiple sequence alignment is an essential part of bioinformatics tools for a genome-scale study of genes and their evolution relations. However, making an accurate alignment between remote homologs is challenging. Here, we develop a method, called SPEM, that aligns multiple sequences using pre-processed sequence profiles and predicted secondary structures for pairwise alignment, consistency-based scoring for refinement of the pairwise alignment and a progressive algorithm for final multiple alignment.

Results: The alignment accuracy of SPEM is compared with those of established methods such as ClustalW, T-Coffee, MUSCLE, ProbCons and PRALINE_{PSI} in easy (homologs) and hard (remote homologs) benchmarks. Results indicate that the average sum of pairwise alignment scores given by SPEM are 7–15% higher than those of the methods compared in aligning remote homologs (sequence identity <30%). Its accuracy for aligning homologs (sequence identity >30%) is statistically indistinguishable from those of the state-of-the-art techniques such as ProbCons or MUSCLE 6.0.

Availability: The SPEM server and its executables are available on <http://theory.med.buffalo.edu>

Contact: yqzhou@buffalo.edu

1 INTRODUCTION

Multiple sequence alignment is one of the most basic tasks in bioinformatics. It has been used for building phylogenetic trees (Saitou and Nei, 1987), locating conserved motifs and domains (Dayhoff *et al.*, 1978; Attwood, 2002) and predicting secondary (Rost *et al.*, 1994) and tertiary (Goebel *et al.*, 1994) structures. Three different algorithms have been developed (Notredame, 2002). The progressive algorithm (Hogeweg and Hesper, 1984) [used in, for example, Pileup (Devereux *et al.*, 1984), ClustalW (Thompson *et al.*, 1994), T-Coffee (Notredame *et al.*, 1998), MUSCLE (Edgar, 1994) and MAFFT (Katoh *et al.*, 2005)] starts with the alignment of two sequences and, then, adds other sequences one by one according to a predetermined order. Obviously, the outcome of the progressive algorithm strongly depends on the predetermined order. To avoid this problem an exact algorithm (Lipman *et al.*, 1989) was developed to align multiple sequences simultaneously. However, the computational

and memory requirement of this approach, even with the use of a divide and conquer algorithm (Stoye *et al.*, 1997), has limited its usage. More recent studies focused on iterative optimization [e.g. Praline (Heringa, 1999), IterAlign (Brocchieri and Karlin, 1998), Prrp (Gotoh, 1982), SAM (Hughes and Krogh, 1996), HMMER (Eddy, 1995), SAGA (Notredame and Higgins, 1996), AIMS (Wang and Li, 2004), MUSCLE (Edgar, 1994), ProbCons (Do *et al.*, 2005) and MAFFT (Katoh *et al.*, 2005)] and consistency-based scoring [such as DiAlign (Morgenstern *et al.*, 1996), ComAlign (Bucka-Lassen *et al.*, 1999) and T-Coffee (Notredame *et al.*, 1998)]. Many above-mentioned methods combined iterative optimization with either progressive algorithm and/or consistency-based scoring. An assessment on various iteration algorithms was made recently (Wallace *et al.*, 2005).

It is known for sometime that profile–profile alignment and incorporation of secondary and tertiary structure information improves pairwise sequence-to-structure alignment in fold recognition (Gribkov *et al.*, 1987; Fischer and Eisenberg, 1996; Rychlewski *et al.*, 2000; Xu and Xu, 2000; Koretke *et al.*, 2001; Skolnick and Kihara, 2001; Yona and Levitt, 2002; Zhou and Zhou, 2004, 2005a). It is only recently that the profile–profile alignment approach is used in multiple alignment [PCMA (Pei *et al.*, 2003), SATCHMO (Edgar and Sjölander, 2003) and PRALINE_{PSI} (Simossis *et al.*, 2005)]. In addition, a new method called 3D-Coffee (O’Sullivan *et al.*, 2004) incorporated structural information in multiple sequence alignment via the sequence–structure fold recognition method FUGUE (Shi *et al.*, 2001) and the structure–structure alignment methods SAP (Taylor and Orengo, 1989) and LSQman (Kabsch, 1978).

Recently, we have developed a set of fold-recognition methods called SP, SP² and SP³ (Zhou and Zhou, 2005a). In SP³, sequence profile, secondary structure profile and structure-derived sequence profile are employed to align a query sequence to a sequence with a known three-dimensional structure. SP³ is one of the best fully automatic servers for comparative modeling targets in a recently completed community-wide experiment on the critical assessment of techniques for protein structure prediction (CASP 6) (Zhou and Zhou, 2005b).

Unlike SP³, which relies on structural information of templates, SP² is a purely sequence-based method that uses sequence profiles and predicted secondary structures (or actual secondary structures if known). Thus, it is possible to employ SP² in multiple sequence

*To whom correspondence should be addressed.

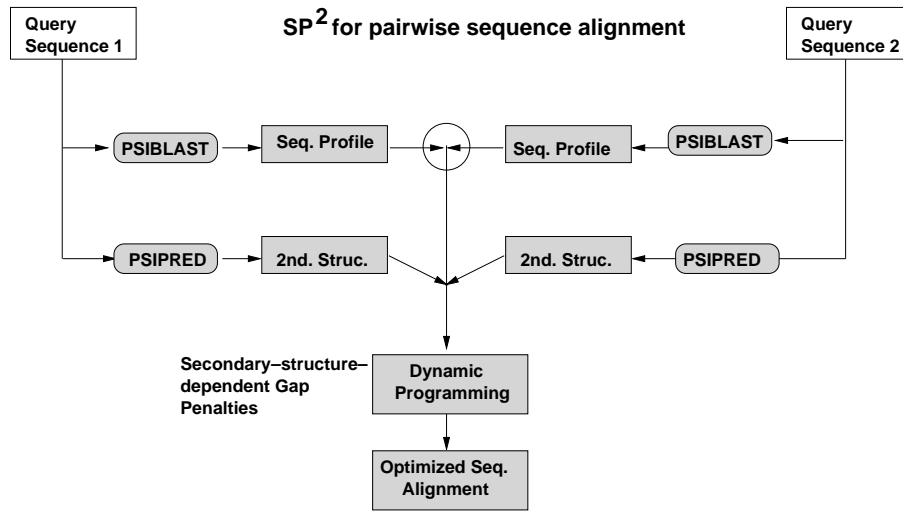


Fig. 1. The flow chart of SP² for pairwise sequence alignment.

alignment. The method, called SPEM (Sequence and secondary-structure Profiles Enhanced Multiple alignment), combines SP² with a consistency-based refinement for pairwise alignment and a progressive algorithm for multiple alignment. SPEM is tested on four benchmarks along with several leading methods. It is found that SPEM improves alignment of remote homologs over other leading methods while maintaining the accuracy of aligning homologs.

2 METHODS

2.1 SP² for pairwise sequence alignment

The method for SP² (Zhou and Zhou, 2005a) has been described elsewhere. Here, we give a brief summary for completeness. The algorithm of SP² for a pairwise sequence–sequence alignment is shown in Figure 1. The details are as follows.

First, the program PSIBLAST (Altschul *et al.*, 1997) is used to search homologous sequences of a query sequence from the NCBI non-redundant (NR) database (<ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz>). As in PSIPRED (Jones, 1999), the NR database was filtered to remove low-complexity regions, transmembrane regions and coiled-coil segments before being searched by PSIBLAST. This homolog search is conducted with an *E*-value cut-off of 0.001 and completed after three iterations. The homologous sequences found by PSIBLAST are then filtered by keeping only those sequences that have <98% identity with the query sequence and an *E*-value of <0.001. The filtered homologs are used to produce the sequence profile that characterizes evolutionary-derived probability of a residue type at a given query sequence position.

Second, PSIPRED (Jones, 1999) is used to predict the secondary structure of a query sequence. Three states (helix, strand and coil) are used for all secondary structures.

Third, two query sequences are aligned with a total matching score given by the following equation:

$$S(i, j) = -\frac{1}{2} \left[F_{q1}^{\text{seq}}(i) \cdot M_{q2}^{\text{seq}}(j) + F_{q2}^{\text{seq}}(j) \cdot M_{q1}^{\text{seq}}(i) \right] - w_{2\text{ndary}} \delta_{si,sj} + s_{\text{shift}}, \quad (1)$$

where $F_{qk}^{\text{seq}}(i)$ is the sequence-derived frequency profile of query sequence k (sequence profiles described above), $M_{qk}^{\text{seq}}(j)$ is the log odd profile (position-specific substitution matrix as in PSIPRED) of query sequence k produced by the above mentioned PSIBLAST search against the filtered NR database, s_{shift} is a to-be-determined constant shift, $w_{2\text{ndary}}$ is a weight parameter for

secondary structure profiles and $\delta_{si,sj}$ is a simple function of the secondary structure element si of the query sequence 1 at sequence position i and sj of the the query sequence 2 at sequence position j .

$$\delta_{si,sj} = \begin{cases} 1 & si = sj \\ -1 & si \neq sj. \end{cases}$$

Finally, the above-mentioned matching score is optimized by using a dynamic-programming alignment algorithm without penalty to end gaps (Needleman and Wunsch, 1970). A gap penalty that depends on secondary structures is employed. No gaps are allowed in helices or sheets (i.e. when $s_i = s_j = \alpha$ or $s_i = s_j = \beta$). The gap opening (w_0) and gap extension (w_1) penalties are applied to coil regions. However, no gap penalties are applied to the beginning and the end of sequences (i.e. no end-gap penalties). To avoid a possible trivial solution of aligning end gaps to whole sequences, a shift score, s_{shift} , is used (see, e.g., Wang and Dunbrack Jr (2004)). Alignment optimization is to minimize the total alignment score due to the negative signs in Equation (1).

The above procedure contains four unknown parameters (w_0 , w_1 , $w_{2\text{ndary}}$ and s_{shift}). In the original sequence-to-structure alignment method SP² these were obtained by optimizing the SP² performance on the ProSup sequence-to-structure alignment benchmark, where the reference alignment is from the ProSup structure-alignment program (Domingues *et al.*, 2000). The optimized parameters are: $w_0 = 7.8$, $w_1 = 0.18$, $w_{2\text{ndary}} = 0.73$ and $s_{\text{shift}} = -1.30$.

Equation (1) for the sequence–sequence alignment uses a symmetric form for $F_{qk}^{\text{seq}}(i)$ and $M_{qk}^{\text{seq}}(j)$ with $k = 1$ and 2. This is slightly different from that used in the original SP² for the sequence–structure alignment (the alignment between a query sequence and a sequence with a known structure). The latter was from the asymmetric equation in the SP³ method (Zhou and Zhou, 2005a) that is built on two input sequences having different properties (one has a known structure and the other does not). The adoption of a symmetric version in Equation (1) is to avoid the dependence of the alignment result on the input order of sequences. We found that this symmetric score function gives essentially the same alignment accuracy on the ProSup benchmark (Domingues *et al.*, 2000) with the same optimized parameter values from the original sequence–structure alignment method SP². Thus, throughout this paper, we will use this optimized parameter set.

2.2 SPEM for multiple sequence alignment

The above-mentioned SP² algorithm for pairwise alignment is combined with a consistency-based scoring method for refining pairwise alignment and

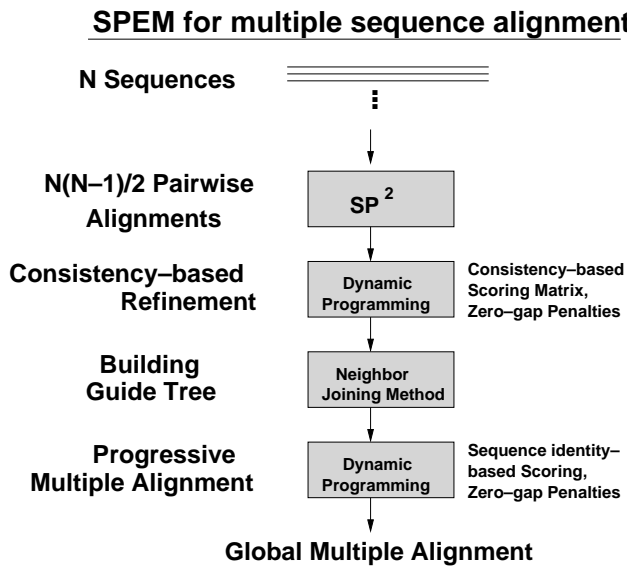


Fig. 2. The flow chart of SPEM for multiple sequence alignment.

a progressive algorithm for multiple sequence alignment. As illustrated in Figure 2, SPEM takes the following steps for multiple alignment.

- (1) *Pairwise alignment*: Given a set of N sequences, SP^2 is used to produce all $N(N-1)/2$ pairwise alignments.
- (2) *Pairwise alignment refinement*: The SP^2 pairwise alignment is refined by using a consistency score. For a given pair of sequences a and b , the consistency scoring matrix, $cons^{ab}(i, j) = -1$ if residue $a(i)$ is aligned with $b(j)$ by SP^2 , $cons^{ab}(i, j) = 0$ if otherwise. This matrix is then updated based on the alignment between a and the third sequence c and the alignment between b and c by SP^2 . If $c(k)$ is aligned with $a(m)$ and with $b(n)$, $cons^{ab}(m, n) = cons^{ab}(m, n) - 1$. The scoring matrix $cons^{ab}$ is updated by analyzing all third sequences. The final scoring matrix $cons^{ab}(i, j)$ is used to refine the alignment between sequences a and b by using the dynamic programming technique with a zero gap penalty. After this step, we obtained a refined set of $N(N-1)/2$ pairwise alignments and pairwise sequence identities. (Shift score and secondary-structure information are used only in SP^2 . They are not used in this refinement step.)
- (3) *Guide tree*: Neighbor joining method (Saitou and Nei, 1987) is used to construct the guide tree. The distance between two sequences is $1 - ID$ (ID denotes sequence identity). Each time, one joins the two nearest sequences (or sequence groups). The distance between two groups is the average distance of each pair between the two groups.
- (4) *Progressive multiple alignment*: Progressive multiple alignment is performed based on the guide tree and the refined $N(N-1)/2$ pairwise alignments and sequence identities obtained above. When two sequences (groups) are aligned, we first construct a scoring matrix $S(I, J)$ between position I in Group A and position J in Group B . The contribution of sequence a from group A and sequence b from Group B to $S(I, J)$ is $S^{ab}(I, J)$. $S^{ab}(I, J) = (-ID)$ if sequences a and b are aligned at the two positions after Step 2. $S^{ab}(I, J) = 0$ if otherwise. The scoring matrix between the two groups is the sum of matrices between all possible pairs of the two groups, i.e. $S(I, J) = \sum_{a \in A, b \in B} S^{ab}(I, J)$. Once $S(I, J)$ is obtained, a dynamic programming algorithm (without any gap penalty) is used to align the two sequences (groups) with scoring matrix $S(I, J)$. The multiple sequence alignment is completed when the guide tree reaches the root. This procedure produces a global alignment of all sequences.

This progressive multiple-alignment algorithm in the last step is very similar to that used in the T-Coffee method with two exceptions. First, a refined SP^2 pairwise alignment is used. Second, the consistency-scoring matrix is used only in pairwise refinement (Step 2) before the progressive multiple alignment (Step 4) but not in the progressive multiple alignment (Also see discussion).

It should be emphasized that no new parameter is introduced in combining SP^2 with the progressive algorithm.

2.3 Test sets and alignment accuracy assessment

SPEM is tested on four benchmarks: BAliBase 2.0 (Thompson *et al.*, 1999), SABmark 1.63 (Walle *et al.*, 2005), Prefab 4.0 (Edgar, 1994) and a HOMSTRAD dataset of remote homologs (March 10, 2005) (Mizuguchi *et al.*, 1998). Alignment accuracy is measured by the sum of pairwise alignment score (SPS)—the percentage of predicted pairwise alignment that is the same as those in the reference alignment. Another score, called column scores (CS), assesses the percentage of whole columns that are aligned correctly (Thompson *et al.*, 1999).

P -values are used to estimate the statistical significance of the difference in alignment accuracy between SPEM and other established methods. For two sets of data (x_1, x_2, \dots, x_n) and (y_1, y_2, \dots, y_n) , t -value (Press *et al.*, 1992) is defined as $t = \sqrt{n}D/S$, where $D = [\sum_{i=1}^n (x_i - y_i)/n]$ is the average difference between x and y , and $S = (\sum_{i=1}^n (x_i - y_i - D)^2/n - 1)$ is the standard deviation of the difference. The probability of difference $|d|$, which is greater than $|D|$, satisfies the equation

$$P(|d| > |D|) = 1 - I_{\frac{\nu}{\nu+2}}(\nu/2, 1/2),$$

where $\nu = n - 1$ is the degree of freedom and I is the incomplete β function. The probability, called P -value, if found to be 0.05 means that there is a 95% confidence about the found difference between the two sets of data. The smaller the P -value, the higher is the significance level of the difference.

To further analyze the performance of SPEM, we also tested the performance of SPEM when secondary-structure information is turned off. This is equivalent to a combination of the SP method for the sequence-to-structure alignment (Zhou and Zhou, 2005a) with the consistency-based refinement for pairwise alignment and a progressive algorithm for multiple alignment. The optimized parameter values from the SP method (Zhou and Zhou, 2005a) are used in this test. That is, $w_0 = 6.6$, $w_1 = 0.58$, $s_{\text{shift}} = -0.9$ and $w_{\text{2ndary}} = 0$.

3 RESULTS

3.1 Test set 1: BAliBase

BAliBase benchmark (Thompson *et al.*, 1999) contains five reference sets. Different sets were designed to test different aspects of alignment methods. Set 1 is made of approximately equidistant sequences; Set 2, a family with orphan sequences; Set 3, divergent families; Set 4, sequences with large N/C terminal insertions and Set 5, sequences with large internal insertions. Reference alignments from FSSP (Holm and Sander, 1994) and HOMSTRAD (Mizuguchi *et al.*, 1998) databases as well as manually constructed alignments from the literature are used. All reference alignments are manually-refined by BAliBase authors. The evaluation of multiple alignment results is also performed by using the evaluation programs supplied with the benchmark. The average pairwise sequence identity of this benchmark is 31.5%.

Table 1 compares the results given by SPEM with those by several other methods along with the P -values for the difference in alignment accuracy between SPEM and a given method for the overall results. The other methods are two popular methods ClustalW (Thompson *et al.*, 1994) and T-Coffee (Notredame *et al.*, 1998), a profile-profile alignment method PRALINE_{PSI} (Simossis *et al.*, 2005), MUSCLE 6.0 (Edgar, 1994) and the probabilistic consistency-based method

Table 1. Alignment accuracies given by several methods on the BALiBase benchmark for multiple sequence alignment

Method ^a	ClustalW	T-Coffee	MUSCLE 6.0	ProbCons	PRALINE _{PSI} ^b	SPEM ^c
Set 1 -CS ^d	78.3	80.0	84.7	83.9	83.9	83.9 (82.3)
(82) -SPS ^e	85.8	86.8	90.3	90.0	90.4	90.8 (89.4)
Set 2 -CS ^d	59.3	58.5	60.9	62.6	61.0	57.3 (52.4)
(23) -SPS ^e	93.3	93.9	94.4	94.5	94.0	93.4 (93.0)
Set 3 -CS ^d	48.1	54.8	61.9	63.1	55.8	56.9 (52.0)
(12) -SPS ^e	72.3	76.7	82.2	82.3	76.4	81.4 (78.1)
Set 4 -CS ^d	62.3	76.8	74.8	73.6	53.9	90.8 (87.8)
(12) -SPS ^e	83.4	92.1	91.8	90.9	79.9	97.4 (96.3)
Set 5 -CS ^d	63.4	86.1	92.1	91.7	68.6	92.3 (84.1)
(12) -SPS ^e	85.8	94.6	98.1	98.1	81.8	97.4 (91.7)
All (141) -CS ^d	70.0	74.6	78.7	78.4	73.9	78.6 (75.5)
<i>P</i> -value ^f	6.0×10^{-4}	0.067	0.95	0.95	—	—(0.014)
-SPS ^e	85.7	88.2	91.0	90.8	88.2	91.5 (89.8)
<i>P</i> -value ^f	8.1×10^{-6}	4.3×10^{-3}	0.54	0.16	—	—(7.4×10^{-3})

^aAll results for other established methods (except PRALINE_{PSI}) were done locally with their latest versions.

^bFrom Simossis *et al.* (2005), the *P*-value for the difference between this method and SPEM is not available due to lack of individual multiple-alignment results.

^cThis work. The number in parentheses is the result when secondary structure information is turned off in SP². That is, the SP alignment method (Zhou and Zhou, 2005a) rather than SP² method is used in SPEM.

^dCS: Column score, the percentage of whole columns that are aligned correctly.

^eSPS: sum-of-pairs score, the percentage of predicted pairwise alignments that are the same as those in the reference alignment.

^fThe *P*-value indicates the significance of the difference in alignment accuracies between SPEM and a given method for all datasets. A *P*-value of >0.05 indicates a non-significant difference.

ProbCons (Do *et al.*, 2005). All these methods (except PRALINE_{PSI}) are run locally with default settings. The results of PRALINE_{PSI} are from Simossis *et al.* (2005). SPEM outperforms ClustalW and T-Coffee for all sets (3–5% in SPS scores) except Set 2. For two profile-based methods SPEM and PRALINE_{PSI}, the alignment accuracy of SPEM is similar to that of PRALINE_{PSI} in Sets 1 and 2 but is significantly better than the latter in Sets 3, 4 and 5 (5–15% better in SPS scores). However, the overall accuracy of SPEM is statistically indistinguishable (based on *P*-values) from those of the iterative, consistency-based method ProbCons and the MUSCLE method. It is noted that the average pairwise alignment accuracy is high for all six methods (between 86% by ClustalW and 92% by SPEM). Thus, BALiBase can be viewed as an ‘easy’ benchmark. Indeed, its average pairwise sequence identity (31.5%) indicates that about half of all pair sequences are easily detectable homologs (sequence identity >30%).

To have a better understanding of the contribution of secondary structure information, Table 1 also shows the result when the secondary structure information is turned off and SPEM is a combination of the method SP with consistency-based refinement and a progressive algorithm. The difference between the methods with and without predicted secondary structures is significant. The difference is 3% in column score and 2% in SPS score. Similar magnitude of difference is observed between pairwise alignment accuracy given by SP and that by SP² in the SALIGN alignment benchmark (Zhou and Zhou, 2005a).

3.2 Test set 2: SABmark 1.63

SABmark (Walle *et al.*, 2005) (Sequence Alignment Benchmark) was designed to align the sequences that have low-to-intermediate sequence identities (the superfamily set) and very-low-to-low sequence identities (the twilight set) between each other. The average

pairwise sequence identity is 22.9% for the superfamily set and 16.9% for the twilight set, compared with 31.5% for the BALiBase benchmark. Thus, it is a more challenging (‘harder’) benchmark than BALiBase. It is also a larger one as it covers the entire known fold space (698 folds). Reference alignments are from consensus structural alignments by SOFI (Boutonnet *et al.*, 1995) and CE (Shindyalov and Bourne, 1998). Unlike BALiBase, which provides reference multiple alignments, SABmark supplies only reference pairwise alignments. As a result, only SPS score is evaluated. [SPS is called the developer score in SABmark. We omitted the modeler score (Walle *et al.*, 2005) since it yields essentially the same trend in relative accuracy among various methods.] The reference pairwise alignments permit us to evaluate the results of SP² in addition to those of SPEM. This makes it possible to examine the effect of consistency-based pairwise refinement of alignment and progressive multiple alignment.

Results on the SABmark 1.63 given by ClustalW (Thompson *et al.*, 1994), T-Coffee (Notredame *et al.*, 1998), MUSCLE 6.0 (Edgar, 1994), ProbCons (Do *et al.*, 2005) and SPEM are shown in Table 2. We did not obtain the results for PRALINE_{PSI} because it is not feasible to perform such a computationally intensive large-scale benchmark test by using a web server. For this ‘hard’ benchmark, the difference between previously developed methods is relatively small, whereas SPEM is 10.8% more accurate in the superfamily set and 14.7% more in the twilight set than the next best (ProbCons), according to SPS scores. This demonstrates the exceptional capability of SPEM to align remotely-related sequences. The change of pairwise alignment accuracy from SP² to SPEM is small (1%) but statistically significant according to the corresponding *P*-value. This indicates that the consistency-based scoring and progressive-based algorithm implemented in SPEM provide some additional noticeable improvements.

Table 2. Alignment accuracies based on SPS scores given by several methods on the SABmark 1.63 benchmark for multiple sequence alignment

Method ^a	ClustalW	T-Coffee	MUSCLE 6.0	ProbCons	SP ^{2b}	SPEM ^c
Superfamily (462)	49.9	54.8	54.8	56.2	65.7	67.0
Twilight (236)	21.9	27.4	26.3	29.2	43.5	43.9
All (698)	40.4	45.5	45.2	47.1	58.2	59.2
<i>P</i> -value ^d	5.6×10^{-112}	1.5×10^{-96}	2.3×10^{-85}	3.8×10^{-83}	8.3×10^{-6}	—

^aAll results for other established methods were done locally with their latest versions.

^bPairwise alignment by SP², this work.

^cMultiple alignment by SPEM, this work.

^dThe *P*-value indicates the significance of the difference in alignment accuracies between SPEM and a given method for all datasets. A *P*-value of >0.05 indicates a non-significant difference.

Table 3. Alignment accuracies based on SPS scores given by several methods on the PREFAB 4.0 benchmark (1682 families)

Method ^a	ClustalW	T-Coffee	MUSCLE 6.0	ProbCons	SP ^{2b}
SPS	61.7	69.2	69.6	70.5	77.0
<i>P</i> -value ^c	3.9×10^{-142}	2.4×10^{-21}	1.6×10^{-55}	4.5×10^{-16}	—

^aAll results for other established methods were done locally with their latest versions.

^bThis work. Due to the limitation of computational resource, the results for SPEM are not available.

^cThe *P*-value indicates the significance of the difference in alignment accuracies between SPEM and a given method. A *P*-value of >0.05 indicates a non-significant difference.

3.3 Test set 3: PREFAB 4.0

PREFAB 4.0 contains 1682 families (Edgar, 1994). Each family contains only one pair of sequences with known structures. This pair of sequences are supplemented with total up to 50 homologous sequences obtained from the PSIBLAST. Alignment accuracy is evaluated based on reference structural alignment of the pair of sequences with known structures. The average pairwise sequence identity is 21.0%. Thus, this benchmark can be considered as a 'hard' benchmark. Each pair of structures are aligned using the CE aligner (Shindyalov and Bourne, 1998), and only those pairs for which FSSP (Holm and Sander, 1994) and CE agreed on 50 or more positions are retained.

Table 3 compares the results given by the pairwise alignment method SP² with four multiple alignment methods ClustalW, MUSCLE 6.0, T-Coffee and ProbCons. We did not carry out SPEM or PRALINE_{PSI} for this benchmark because it is too computationally demanding. For SPEM, the accuracy of SPEM multiple alignment is dictated by the accuracy of SP² pairwise alignment (Table 2). That is, the alignment accuracy of SP² is an indicator for the accuracy of SPEM. The average SPS score [called Qscore in PREFAB (Edgar, 1994)] is 61.7% for ClustalW, 69.2% for T-Coffee, 69.6% for MUSCLE 6.0, 70.5% for ProbCons and 77.0% for SP². The differences in alignment accuracy given by SP² and other methods are statistically significant because the *p*-values for the differences are very small.

3.4 Test set 4: HOMSTRAD

We also construct a test set based on the HOMSTRAD structural alignment dataset (dated March 10, 2005). In this dataset, the

structural alignments were performed using the programs MNY-FIT (Sutcliffe *et al.*, 1987), STAMP (Russell and Barton, 1992) and COMPARE (Sali and Blundell, 1990). These structure-based alignments are annotated with JOY and examined individually (Mizuguchi *et al.*, 1998). There are 75 families that contain a minimum of four sequences with average sequence identity $\leq 25\%$. This set with an average sequence identity of 18.7% further examines the ability of various methods to align remote homologs.

Table 4 compares the alignment accuracies given by several methods. The alignment accuracies given by ProbCons, T-Coffee, MUSCLE 6.0 and PRALINE_{PSI} are similar to each other, whereas SPEM is about 7% more accurate (in either SPS or CS) than the next best (ProbCons). The difference between SPEM and all other methods are statistically significant.

This small benchmark also permits us to investigate the effect of errors in predicted secondary structures on the accuracy of SPEM. To do this, we test a version of SPEM where the secondary structures used in SP² are obtained directly from known structures by using the program DSSP (Kabsch and Sander, 1983), rather than predicted by PSIPRED (Jones, 1999). The new version of the method (labeled SPEM-DSSP) uses the parameters of SPEM. The results are shown in Table 4. The average SPS score increases from 74.9 to 75.5% and the CS score increases from 55.7 to 56.9% after the exact secondary structures are used. However the differences are not statistically significant based on *P*-values. This limited change may be resulted from the error cancellation in predicted secondary structures between two query sequences and/or from the high accuracy ($\sim 80\%$) in secondary structure prediction by PSIPRED.

To further examine the dependence of different methods on the difficulty of benchmarks, we expand the 75 families of the HOMSTRAD benchmark to 233 families by including all alignments with >3 sequences. These families are binned in every 5% average pairwise sequence identity. More specifically, the bins are 10–15, 15–20, 20–25, 25–30, 30–35, 35–40, 40–45, 45–50, 50–55, 55–60 and 60–65%. The corresponding number of families in each bin is 11, 38, 26, 34, 36, 32, 22, 15, 19, 4 and 3, respectively. Here, we focus on families with sequence identities from 10 to 65% because there are only three families with sequence identities between 65–100%.

Figure 3 plots alignment accuracies (measured by SPS) as a function of average sequence identity given by SPEM, ProbCons, MUSCLE 6, T-Coffee and ClustalW. All methods have similar accuracy when the average sequence identity is high. The difference between various methods is clear when the average sequence identity is <30%. Below this point, the alignment accuracy of SPEM is

Table 4. Alignment accuracies based on SPS and CS scores given by several methods on the HOMSTRAD dataset of 75 families of remote homologs

Method ^a	ClustalW	T-Coffee	PRALINE _{PSI} ^b	MUSCLE 6.0	ProbCons	SPEM ^c	SPEM-DSSP ^c
SPS	60.7	64.9	66.0	67.7	68.0	74.9	75.5
<i>P</i> -value ^d	4.9×10^{-15}	6.7×10^{-13}	4.1×10^{-8}	1.4×10^{-9}	2.7×10^{-9}	—	0.33
CS	38.8	45.2	44.5	47.7	49.0	55.7	56.9
<i>P</i> -value ^d	1.9×10^{-12}	2.2×10^{-9}	9.6×10^{-7}	3.0×10^{-7}	4.3×10^{-6}	—	0.16

^aAll results for other established methods (except PRALINE_{PSI}) were done locally with their latest versions.

^bResults from the direct submission to the PRALINE_{PSI} server with default setting (<http://ibivu.cs.vu.nl/programs/pralinewww/>).

^cThis work. SPEM-DSSP is same as SPEM but with secondary structures extracted from known structures by DSSP.

^dThe *P*-value indicates the significance of the difference in alignment accuracies between SPEM and a given method. A *P*-value of >0.05 indicates a non-significant difference.

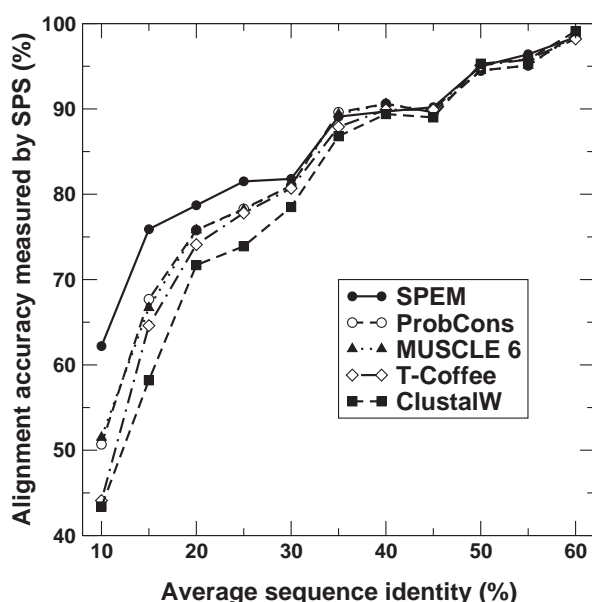


Fig. 3. Alignment accuracies (measured by SPS) as a function of average sequence identity given by methods SPEM, ProbCons, MUSCLE 6.0, T-Coffee and ClustalW, shown as labeled. Each point is represented by the lower bound of sequence identity at each bin.

consistently better than those of other methods. (PRALINE_{PSI} is not performed on this extended benchmark due to computational requirement. One expects that its performance will be similar to those of MUSCLE 6.0 or ProbCons as suggested by results in Tables 1 and 4).

4 DISCUSSION

In this paper, we combine a recently developed profile–profile and secondary-structure enhanced alignment method with a progressive algorithm for multiple sequence alignment. The method, called SPEM, provides a significant improvement in aligning remote homologs when compared with the state-of-the-art techniques such as ClustalW, T-Coffee, ProbCons, MUSCLE and a profile–profile multiple alignment method PRALINE_{PSI}. Meanwhile, it also provides an excellent alignment for homologs (statistically indistinguishable from ProbCons and MUSCLE in the BAliBase benchmark). Profile–profile alignment method is the main source for improving alignment of remote homologs. The use of predicted secondary structures also

contributes to the accuracy of SPEM. Predicted secondary structures are responsible for improving alignment accuracy by an additional 0.4–5.7% in SPS depending on specific testing sets in the BAliBase benchmark (Table 1). Employing exact secondary structures makes minor changes to alignment accuracy.

SPEM, as PRALINE_{PSI}, is more time-consuming than T-Coffee, ProbCons and MUSCLE. It spends most of its computing time in calculating sequence profiles using PSIBLAST. However, the required computing time is affordable. A multiple sequence alignment of 50 sequences between 100 and 200 residues long takes about several hours on a single-processor PC. The gain in accuracy for aligning remote-related sequences significantly outweighs the increase in computing time. SPEM improves over PRALINE_{PSI} in two benchmarks tested. This may be due to the difference in how the profile–profile alignment is made (Wang and Dunbrack Jr, 2004).

Table 2 indicates that SPEM provides a small but significant improvement over the pairwise alignment from SP². We have also tested the combination of T-Coffee with the SP² pairwise alignment. The combination also yields a similar improvement over the input pairwise alignment but with a longer computing time. In addition, we implemented a procedure for refining multiple alignment by randomly partitioning all into two sets for 100 times (Do *et al.*, 2005). This increased the computational time significantly but improved little with respect to alignment accuracy. However, there are many other iterative algorithms (Taylor and Brown, 1999; Hughey and Krogh, 1996; Edgar, 1994; Katoh *et al.*, 2005; Heringa, 1999; Brocchieri and Karlin, 1998; Gotoh, 1982; Eddy, 1995; Notredame and Higgins, 1996; Wang and Li, 2004) that are potentially useful for refining the pairwise alignment from SP². A recent evaluation of iterative alignment algorithms indicates that iterative algorithms can improve over some methods but not others (Wallace *et al.*, 2005). Further study in this area is required.

Another way to improve the accuracy of SPEM, as 3D-Coffee (O’Sullivan *et al.*, 2004), is to take advantage of structural information if one or more sequences to be aligned in multiple alignment have known structures. This can be achieved by combination of SP², SP³ and structure–structure alignment programs. [For a recent assessment of various structure–alignment techniques, see Kolodny *et al.* (2005).] SP³ will be used to align a sequence of unknown structure with a sequence of known structure. Benchmark tests suggested that SP³ increases alignment accuracy by an additional 2–3% over SP² (Zhou and Zhou, 2005a). Clearly, as more structures are known, more accurate the multiple alignment will be. We shall defer this to future studies.

ACKNOWLEDGEMENTS

The authors gratefully thank the authors who made their programs and databases available for comparison. This work was supported by NIH (R01 GM 966049 and R01 GM 068530), a grant from HHMI to SUNY Buffalo, and by the Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo. Y. Z. is also partially supported by a two-base grant (No. 20340420391) from the national science foundation of China.

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.
- Attwood,T.K. (2002) The PRINTS database: a resource for identification of protein families. *Brief Bioinformatics*, **3**, 252–263.
- Boutonnet,N.S. *et al.* (1995) Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. *Protein Eng.*, **8**, 647–662.
- Brocchieri,L. and Karlin,S. (1998) Asymmetric-iterated multiple alignment of protein sequences. *J. Mol. Biol.*, **276**, 249–264.
- Bucka-Lassen,K. *et al.* (1999) Combining many multiple alignments in one improved alignment. *Bioinformatics*, **15**, 122–130.
- Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In *Atlas of Proteins Sequences and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5, pp. 345–352.
- Devereux,J. *et al.* (1984) GCG package. *Nucleic Acids Res.*, **22**, 387–395.
- Do,C.B. *et al.* (2005) Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
- Domingues,F.S. *et al.* (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.*, **297**, 1003–1013.
- Eddy,S.R. (1995) Multiple alignment using hidden Markov models. In *Third International Conference on Intelligent Systems for Molecular Biology (ISMB)*. AAAI Press, Cambridge, England, Menlo Park, CA.
- Edgar,R.C. (1994) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
- Edgar,R.C. and Sjölander,K. (2003) SATCHMO: sequence alignment and tree construction using hidden Markov models. *Bioinformatics*, **19**, 1404–1411.
- Fischer,D. and Eisenberg,D. (1996) Protein fold recognition using sequence-derived predictions. *Protein Sci.*, **5**, 947–955.
- Goebel,U. *et al.* (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.
- Gotoh,O. (1982) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinements as assessed by reference to structural alignments. *J. Mol. Biol.*, **264**, 823–838.
- Gribskov,M. *et al.* (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
- Heringa,J. (1999) Two strategies for sequence comparison: profile-preprocessed and secondary-structure-induced multiple alignment. *Comput. Chem.*, **23**, 341–364.
- Hogeweg,P. and Hesper,B. (1984) The alignment of sets of sequences and the construction of phylogenetic trees. An integrated method. *J. Mol. Evol.*, **20**, 175–186.
- Holm,L. and Sander,C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, **22**, 3600–3609.
- Hughey,R. and Krogh,A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biol. Sci.*, **12**, 95–107.
- Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
- Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **34**, 827–828.
- Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
- Kato,H. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
- Kolodny,R. *et al.* (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.
- Koretke,K.K. *et al.* (2001) Fold recognition from sequence comparisons. *Proteins, Suppl. 5*, 68–75.
- Lipman,D.J. *et al.* (1989) A tool for multiple sequence alignment. *Proc. Natl Acad. Sci. USA*, **86**, 4412–4415.
- Mizuguchi,K. *et al.* (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
- Morgenstern,B. *et al.* (1996) Multiple DNA and protein sequence based on segment-to-segment comparison. *Proc. Natl Acad. Sci. USA*, **93**, 12098–12103.
- Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
- Notredame,C. (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, **3**, 131–144.
- Notredame,C. and Higgins,D.G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.*, **4**, 1515–1524.
- Notredame,C. *et al.* (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics*, **14**, 407–422.
- O’Sullivan,O. *et al.* (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, **340**, 385–395.
- Pei,J. *et al.* (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.
- Press,W.H. *et al.* (1992) *Numerical Recipes: The Art of Scientific Computing*, 2nd edn. Cambridge University Press, Cambridge, UK.
- Rost,B. *et al.* (1994) PHD—an automatic server for protein secondary structure prediction. *Comput. Appl. Biosci.*, **10**, 53–60.
- Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison. *Proteins*, **14**, 309–323.
- Rychlewski,L. *et al.* (2000) Comparison of sequence profiles. strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
- Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
- Sali,A. and Blundell,T.L. (1990) Definition of general topological equivalence in protein structures. a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428.
- Shi,J. *et al.* (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
- Shindyalov,I.N. and Bourne,P. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
- Simossis,V.A. *et al.* (2005) Homolog-extended sequence alignment. *Nucleic Acids Res.*, **33**, 816–824.
- Skolnick,J. and Kihara,D. (2001) Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins*, **42**, 319–331.
- Stoye,J. *et al.* (1997) DCA: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. *Comput. Appl. Biosci.*, **13**, 625–626.
- Sutcliffe,M.J. *et al.* (1987) *Protein Eng.*, **1**, 377–384.
- Taylor,W.R. and Brown,N.P. (1999) Iterated sequence databank search methods. *Comput. Chem.*, **23**, 365–385.
- Taylor,W.R. and Orengo,C.A. (1989) Protein structure alignment. *J. Mol. Biol.*, **128**, 1–22.
- Thompson,J. *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**, 4673–4690.
- Thompson,J.D. *et al.* (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.
- Wallace,I.M. *et al.* (2005) Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics*, **21**, 1408–1414.
- Walle,I.V. *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.
- Wang,Y. and Li,K.B. (2004) An adaptive and iterative algorithm for refining multiple sequence alignment. *Comput. Biol. Chem.*, **28**, 141–148.
- Xu,Y. and Xu,D. (2000) Protein threading using PROSPECT: design and evaluation. *Proteins*, **40**, 343–354.
- Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
- Zhou,H. and Zhou,Y. (2004) Single-body knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins*, **55**, 1005–1013.
- Zhou,H. and Zhou,Y. (2005a) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*, **58**, 321–328.
- Zhou,H. and Zhou,Y. (2005b) SPARKS 2 and SP³ servers in CASP 6. *Proteins (CASP Suppl. Issue)*, (in press).