Sequence analysis

SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures

Hongyi Zhou1 and Yaoqi Zhou1,2,*
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 and
2The 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 PRALINEPSI 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 (Brockheri and Karlin, 1998), Fpr (Gotoh, 1982), SAM (Hughey 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 (Gribskov 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 PRALINEPSI (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 SP1, 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. SP1 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 SP1, 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
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 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.

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 sj = sj = α or si = sj = β). The gap opening (w0) and gap extension (w1) 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, sshift, 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 (w0, w1, w2dary and sshift). 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: w0 = 7.8, w1 = 0.18, w2dary = 0.73 and sshift = −1.30.

Equation (1) for the sequence–sequence alignment uses a symmetric form for fseq(i) and fseq(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 SP² 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
SPEM for multiple sequence alignment

<table>
<thead>
<tr>
<th>N Sequences</th>
<th>N(N–1)/2 Pairwise Alignments</th>
<th>Consistency–based Refinement</th>
<th>Building Guide Tree</th>
<th>Progressive Multiple Alignment</th>
<th>Global Multiple Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dynamic Programming</td>
<td>Neighbor Joining Method</td>
<td>Dynamic Programming</td>
<td></td>
</tr>
</tbody>
</table>

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.

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). The scoring matrix S(I, J) is used to align the two sequences 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 (Waltz 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 that 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, ..., x_n) and (y_1, y_2, ..., 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 = \sqrt{\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 - \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{\beta}}\frac{1}{\sqrt{n}}\sqrt{n}\left(\frac{1}{\sqrt{2}}\right)

where \nu = n - 1 is the degree of freedom and I is the incomplete beta 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 = 0.5, w_2 = -0.9 and w_2n = 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 (Simossis et al., 2005), MUSCLE 6.0 (Edgar, 1994) and the probabilistic consistency-based method.
The accuracy of SPEM is similar to that of PRALINE PSI in Sets 1 and 2 but is not as accurate as SP2 in the SALIGN alignment benchmark (Zhou and Zhou, 2005a). The difference between the methods with and without secondary structure information is turned off and SPEM is a combination of the method SP with consistency-based refinement and a progressive-based method implemented in SPEM.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>ClustalW</th>
<th>T-Coffee</th>
<th>MUSCLE 6.0</th>
<th>ProbCons</th>
<th>PRALINEpS1</th>
<th>SPEM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1 -CSd</td>
<td>78.3</td>
<td>80.0</td>
<td>84.7</td>
<td>83.9</td>
<td>83.9</td>
<td>83.9</td>
</tr>
<tr>
<td>(82) -SPS3</td>
<td>85.8</td>
<td>86.8</td>
<td>90.3</td>
<td>90.4</td>
<td>90.4</td>
<td>89.4</td>
</tr>
<tr>
<td>Set 2 -CSd</td>
<td>59.3</td>
<td>58.9</td>
<td>60.9</td>
<td>61.6</td>
<td>61.0</td>
<td>61.0</td>
</tr>
<tr>
<td>(23) -SPS2</td>
<td>93.3</td>
<td>93.9</td>
<td>94.4</td>
<td>94.5</td>
<td>94.0</td>
<td>93.0</td>
</tr>
<tr>
<td>Set 3 -CSd</td>
<td>48.1</td>
<td>54.8</td>
<td>61.9</td>
<td>63.1</td>
<td>58.8</td>
<td>56.9</td>
</tr>
<tr>
<td>(12) -SPS2</td>
<td>72.3</td>
<td>76.7</td>
<td>82.2</td>
<td>82.3</td>
<td>76.4</td>
<td>78.1</td>
</tr>
<tr>
<td>Set 4 -CSd</td>
<td>62.3</td>
<td>76.8</td>
<td>74.8</td>
<td>73.6</td>
<td>53.9</td>
<td>59.0</td>
</tr>
<tr>
<td>(12) -SPS2</td>
<td>83.4</td>
<td>92.1</td>
<td>91.8</td>
<td>90.9</td>
<td>79.9</td>
<td>97.4</td>
</tr>
<tr>
<td>Set 5 -CSd</td>
<td>63.4</td>
<td>86.1</td>
<td>92.1</td>
<td>91.7</td>
<td>68.6</td>
<td>92.3</td>
</tr>
<tr>
<td>(12) -SPS2</td>
<td>85.8</td>
<td>94.6</td>
<td>98.1</td>
<td>98.1</td>
<td>81.8</td>
<td>97.4</td>
</tr>
<tr>
<td>All (141) -CSd</td>
<td>70.0</td>
<td>74.6</td>
<td>78.7</td>
<td>78.4</td>
<td>73.9</td>
<td>78.6</td>
</tr>
<tr>
<td>P-value3</td>
<td>6.0 × 10⁻¹⁴</td>
<td>0.067</td>
<td>0.95</td>
<td>0.95</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>-SPS4</td>
<td>85.7</td>
<td>88.2</td>
<td>91.0</td>
<td>90.8</td>
<td>88.2</td>
<td>91.5</td>
</tr>
<tr>
<td>P-value5</td>
<td>8.1 × 10⁻⁶</td>
<td>4.3 × 10⁻³</td>
<td>0.54</td>
<td>0.16</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

| All results for other established methods (except PRALINEpS1) were done locally with their latest versions. | All results for other established methods (except PRALINEpS1) were done locally with their latest versions. |
| From 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. | This 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. |
| The 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. | The 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. |

ProCons (Do et al., 2005). All these methods (except PRALINEpS1) are run locally with default settings. The results of PRALINEpS1 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 PRALINEpS1, the alignment accuracy of SPEM is similar to that of PRALINEpS1 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 ProCons 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).

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-based 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 BAiLiBase benchmark. Thus, it is a more challenging (‘harder’) benchmark than BAiLiBase. 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 BAiLiBase, 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), ProCons (Do et al., 2005) and SPEM are shown in Table 2. We did not obtain the results for PRALINEpS1 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 (ProCons), 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.
The alignment accuracy of SP2 is an indicator for the accuracy of SPEM multiple alignment. That is, the alignment accuracy of SP2 pairwise alignment is dictated by the accuracy of SPEM pairwise alignment (Table 2).

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 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 6.0, 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 significantly better than that of the other methods.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>ClustalW</th>
<th>T-Coffee</th>
<th>MUSCLE 6.0</th>
<th>ProbCons</th>
<th>SP2b</th>
<th>SPEMc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPS</td>
<td>61.7</td>
<td>69.2</td>
<td>69.6</td>
<td>70.5</td>
<td>77.0</td>
<td></td>
</tr>
<tr>
<td>P-valued</td>
<td>3.9 × 10^{-142}</td>
<td>2.4 × 10^{-21}</td>
<td>1.6 × 10^{-55}</td>
<td>4.5 × 10^{-36}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a All results for other established methods were done locally with their latest versions.

b Pairwise alignment by SP2, this work.

c Multiple alignment by SPEM, this work.

d The 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.
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 6.0, and a profile–profile multiple alignment method PRALINEPSI. 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 PRALINEPSI, 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 PRALINEPSI 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


