MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts

Xin Deng¹ and Jianlin Cheng¹,²,³*

Abstract

Background: Multiple Sequence Alignment (MSA) is a basic tool for bioinformatics research and analysis. It has been used essentially in almost all bioinformatics tasks such as protein structure modeling, gene and protein function prediction, DNA motif recognition, and phylogenetic analysis. Therefore, improving the accuracy of multiple sequence alignment is important for advancing many bioinformatics fields.

Results: We designed and developed a new method, MSACompro, to synergistically incorporate predicted secondary structure, relative solvent accessibility, and residue-residue contact information into the currently most accurate posterior probability-based MSA methods to improve the accuracy of multiple sequence alignments. The method is different from the multiple sequence alignment methods (e.g. 3D-Coffee) that use the tertiary structure information of some sequences since the structural information of our method is fully predicted from sequences. To the best of our knowledge, applying predicted relative solvent accessibility and contact map to multiple sequence alignment is novel. The rigorous benchmarking of our method to the standard benchmarks (i.e. BAliBASE, SABmark and OXBENCH) clearly demonstrated that incorporating predicted protein structural information improves the multiple sequence alignment accuracy over the leading multiple protein sequence alignment tools without using this information, such as MSAProbs, ProbCons, Probalign, T-coffee, MAFFT and MUSCLE. And the performance of the method is comparable to the state-of-the-art method PROMALS of using structural features and additional homologous sequences by slightly lower scores.

Conclusion: MSACompro is an efficient and reliable multiple protein sequence alignment tool that can effectively incorporate predicted protein structural information into multiple sequence alignment. The software is available at http://sysbio.rnet.missouri.edu/multicom_toolbox/.

Background

Aligning multiple evolutionarily related protein sequences is a fundamental technique for studying protein function, structure, and evolution. Multiple sequence alignment methods are often an essential component for solving challenging bioinformatics problems such as protein function prediction, protein homology identification, protein structure prediction, protein interaction study, mutagenesis analysis, and phylogenetic tree construction. During the last thirty years or so, a number of methods and tools have been developed for multiple sequence alignment, which have made fundamental contributions to the development of the bioinformatics field.

State of the art multiple sequence alignment methods adapt some popular techniques to improve alignment accuracy, such as iterative alignment [1], progressive alignment [2], alignment based on profile hidden Markov models [3], and posterior alignment probability transformation [4,5]. Some alignment methods, such as 3D-Coffee [6] and PROMALS3D [7], use 3D structure information to improve multiple sequence alignment, which cannot be applied to the majority of protein sequences without tertiary structures. In order to overcome this problem, we have developed a method to incorporate secondary structure, relative solvent accessibility, and contact map information predicted from protein sequences into multiple alignment, which have made fundamental contributions to the development of the bioinformatics field.

* Correspondence: chengji@missouri.edu
1 Department of Computer Science, University of Missouri-Columbia, Columbia, MO 65211, USA
Full list of author information is available at the end of the article
sequence alignment. Predicted secondary structure information has been used to improve pairwise sequence alignment [8,9], but few attempts had been made to use predicted secondary structure information in multiple sequence alignment [10-15]. To the best of our knowledge, applying predicted relative solvent accessibility and residue-residue contact map to multiple sequence alignment is novel.

In order to use the predicted structural information to advance the state of the art of multiple sequence alignment, we first compared the existing multiple sequence alignment tools [16-31,4,5,32-37] on the standard benchmark data sets such as BAliBASE [38], SABmark [39] and OXBENCH [40], which showed that MAFFT [30], T-coffee [31], MSAProbs [4], and ProbCons [5] yielded the best performance. Then we developed MSACompro, a new multiple sequence alignment method, which effectively utilizes predicted secondary structure, relative solvent accessibility, and residue-residue contact map together with posterior alignment probabilities produced by both pair hidden Markov models and partition function as in MSAProbs [4]. The assessment results of MSACompro compared to the benchmark data sets from BAliBASE, SABmark and OXBENCH showed that incorporating predicted structural information has improved the accuracy of multiple sequence alignment over most existing tools without using structural features and sometimes the improvement is substantial.

**Method**

Following the general scheme in MSAProbs [4], MSACompro has five main steps: (1) compute the pairwise posterior alignment probability matrices based on both pair-HMM and partition function, considering the similarity in amino acids, secondary structure, and relative solvent accessibility; (2) generate the pairwise distance matrix from both the pairwise posterior probability matrices constructed in the first step and the pairwise contact map similarity matrices; (3) construct a guide tree based on pairwise distance matrix, and calculate sequence weights; (4) transform all the pairwise posterior matrices by a weighting scheme; (5) perform a progressive alignment by computing the profile-profile alignment from the probability matrices of all sequence pairs, and then an iterative alignment to refine the results from progressive alignment. Our method is different from MSAProbs in that it adds secondary structure and solvent accessibility information to the calculation of the posterior residue-residue alignment probabilities and computes the pairwise distance matrix with the help of predicted residue-residue contact information.

**Construction of pairwise posterior probability matrices based on amino acid sequence, secondary structure and solvent accessibility information**

For two protein sequences X and Y in a sequence group S to be aligned, we denote $X = (x_1, x_2, \ldots, x_{n_1})$, $Y = (y_1, y_2, \ldots, y_{n_2})$, where $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ are lists of the residues in X and Y, respectively. $n_1$ is the length of sequence X, and $n_2$ is the length of sequence Y. Suppose $x_i$ is the i-th amino acid in sequence X, and $y_j$ is the j-th amino acid in sequence Y. We let $aln$ denote a global alignment between X and Y, $ALN^*$ the set of all the possible global alignments of X and Y, and $aln^* \in ALN$ the true pairwise alignment of X and Y. The posterior probability that the i-th residue in X ($x_i$) is aligned to the j-th residue ($y_j$) in Y in $aln^*$ is defined as:

$$p(x_i \sim y_j | aln^*, X, Y) = \sum_{aln \in ALN} P(aln | X, Y) I(x_i \sim y_j \in aln)$$

(1)

$$I(x_i \sim y_j \in aln) = \begin{cases} 1, & \text{if } (x_i \sim y_j \in aln) \text{true} \\ 0, & \text{otherwise} \end{cases}$$

$P(aln | X, Y)$ denotes the probability that $aln$ is the true alignment $aln^*$: Thus, the posterior probability $n_1 \times n_2$ matrix $P_{XY}$ is a collection of all the values $p(x_i \sim y_j | aln^*, X, Y)$ (p(x_i \sim y_j) for short) for $1 \leq x_i \leq n_1, 1 \leq y_j \leq n_2$. The calculation process of the pairwise posterior probability matrix is described as follows.

As in MSAProbs, two different methods (a pair hidden Markov model and a partition function) are used to compute the pairwise posterior probability matrices ($P_{XY}^1$ and $P_{XY}^2$), respectively. The first kind of pairwise probability matrix $P_{XY}^1$ is calculated by a partition function $(F)$ of alignments based on dynamic programming. $F(i, j)$ denotes the probability of all partial global alignments of X and Y ending at position (i, j). $F_M(i, j)$ is the probability of all partial global alignments with $x_i$ aligned to $y_j$, $F_S(i, j)$, is the probability of all partial global alignments with $y_j$ aligned to a gap, and $F_X(i, j)$ is the probability of all partial global alignments with $x_i$ aligned to a gap. Accordingly, the partition function can be calculated recursively as follows:

$$F_M(i, j) = F(i - 1, j - 1)e^{\beta_{gap}} + F_Y(i, j - 1)e^{\beta_{ext}}$$

(2)

$$F_Y(i, j) = F_M(i, j - 1)e^{\beta_{gap}} + F_Y(i, j - 1)e^{\beta_{ext}}$$

$$F_X(i, j) = F_M(i - 1, j)e^{\beta_{gap}} + F_X(i - 1, j)e^{\beta_{ext}}$$
$F(i, j) = F_M(i, j) + F_Y(i, j) + F_X(i, j)$

Subject to the constraint $W_1 + W_2 + W_3 = 1$.

In the formula above, $s(x_i, y_j)$ is the amino acid similarity score between $x_i$ and $y_j$. One element of the substitution matrix $s$, $ss(ss(x_i), ss(y_j))$ is the similarity score between the secondary structure $(ss(x_i))$ of residue $x_i$ in protein X and that of residue $y_j$ in protein Y according to the secondary structure similarity matrix SS, $SA(sa(x_i), sa(y_j))$ is the similarity score between the solvent accessibility $(sa(x_i))$ of residue $x_i$ in protein X and that of residue $y_j$ in protein Y according to the solvent accessibility similarity matrix SA. $W_1$, $W_2$, $W_3$ are weights used to control the influence of the amino acid substitution score, secondary structure similarity score, and solvent accessibility similarity score. The secondary structure and solvent accessibility can be automatically predicted by SSpro/ACCpro [41] (http://sysbio.rnet.missouri.edu/multicom_toolbox/) using a multi-threading technique implemented in MSACompro, or alternatively be provided by a user. The values of the three weights are set to 0.4, 0.5, and 0.1 by default, and can be adjusted by users. The ensembles of bidirectional recurrent neural network architectures in ACCpro are used to discriminate between two different states of relative solvent accessibility, higher or lower than the accessibility cutoff - 25% of the total surface area of a residue [42], corresponding to e or b. As in MSAPros, $\beta$ is a parameter measuring the deviation between suboptimal and optimal alignments, $gap(gap \leq 0)$ is the gap open penalty, and $ext(ext \leq 0)$ is the gap extension penalty.

We used the Gonnet 160 matrix as a substitution matrix to generate the similarity scores between two amino acids in proteins [43]. The $3 \times 3$ secondary structure similarity matrix SS contains the similarity scores of three kinds of secondary structures (E, H, C) as follows:

$$SS = \begin{bmatrix} 100 \\ 010 \\ 001 \end{bmatrix}$$

where two identical secondary structures receive a score of 1 and different ones receive a score of 0.

The $2 \times 2$ solvent accessibility similarity matrix SA contains the similarity scores of two kinds of relative solvent accessibilities (e, b) as follows:

$$SA = \begin{bmatrix} 10 \\ 01 \end{bmatrix}$$

where two identical solvent accessibilities receive a score of 1 and different ones receive a 0. It is worth noting that we used the simple identity scoring matrix for secondary structure and solvent accessibility here. Employing more advanced scoring matrices defined in [44] may lead to further improvement. Each posterior residue-residue alignment probability element in the first kind of posterior probability matrix $(P^1_{XY})$ can be calculated from the partition function as:

$$p^1_{XY}(i, j) = \frac{F_M(i-1, j-1)F_M(i, j+1)}{e^{W_1\beta_{i,j} + W_2SS(sa(x_i), sa(y_j)) + W_3SA(sa(x_i), sa(y_j))}}$$

As in MSAPros, the second kind of pairwise probability matrix $(P^2_{XY})$ is calculated by a pair hidden Markov model (HMM) combining both Forward and Backward algorithm [4,5,45]. The pairwise probabilities can be generated under the guidance of pair HMM involving state emissions and transitions. $P^2_{XY}$ is only derived from protein sequences without using secondary structure and solvent accessibility, which is different from PROMALS [15] that lets HMM emit both amino acids and secondary structure alphabets.

The final posterior probability matrix $P_{XY}$ is calculated as the root mean square of the corresponding values in $P^1_{XY}$ and $P^2_{XY}$ as follows.

$$p_{XY}(i, j) = \sqrt{p^1_{XY}(i, j)^2 + p^2_{XY}(i, j)^2}$$

where $p^1_{XY}(i, j)$ and $p^2_{XY}(i, j)$ denote a posterior probability element in two kinds of posterior probability matrices $(P^1_{XY}$ and $P^2_{XY})$, respectively.

Construction of pairwise distance matrices based on pairwise posterior probabilities and pairwise contact scores

The posterior probability matrix $P_{XY}$ is used as a scoring function to generate a pairwise global alignment between sequences X and Y. The optimal global alignment score Opt(X,Y) of the global alignment is computed according to an optimal sub-alignment score matrix AS. The optimal sub-alignment score AS(i, j) denotes the score of the optimal sub-alignment ending at residues $i$ and $j$ in X and Y. The AS matrix is recursively calculated as:

$$AS(i, j) = \max \begin{cases} AS(i-1, j-1) + P_{XY}(i-1, j-1) \\ AS(i-1, j) \\ AS(i, j-1) \end{cases}$$

$AS(n_1, n_2)$ is the optimal score of the full global alignment between X and Y, which is denoted as OptScore(X,Y).

In addition to the optimal alignment score, we introduce a contact map score, CMscore(X, Y), for the optimal
pairwise alignment of X and Y, assuming that the spatially neighboring residues of two aligned residues should have a higher tendency to be aligned together. \(CMscore(X, Y)\) is calculated from the contact map correlation score matrix \(CMap_{XY}\) based on the residue-residue contact map matrices \(CMap_x\) and \(CMap_y\) of X and Y.

Assuming the optimal global alignment of X and Y is represented as,

\[
x_1x_2\ldots \ldots x_m\ldots \ldots x_p\ldots \ldots x_n
\]
\[
y'_1\ldots \ldots y'_{k+1}\ldots \ldots y'_{n_2}
\]

we can generate a new alignment after removing the pairs containing gaps:

\[
x'_1x'_2\ldots \ldots x'_m
\]
\[
y'_1y'_2\ldots \ldots y'_n
\]

, which can be denoted as

\[
x'_i, x'_{i+1}, \ldots , x'_{j-1}, x'_{j}, x'_{j+1}, \ldots , x'_{k}
\]
\[
y'_i, y'_{i+1}, \ldots , y'_{j-1}, y'_{j}, y'_{j+1}, \ldots , y'_{k}
\]

, where \(n\) is the length of the new alignment without gaps.

From this alignment, we can construct two contact map matrices, \(CMap_x\) and \(CMap_y\), shown below:

\[
CMap_x = \begin{bmatrix}
x_{i1} & x_{i2} & \cdots & x_{in} \\
x_{i1}' & x_{i2}' & \cdots & x_{in}' \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1}' & x_{n2}' & \cdots & x_{nm}' \\
\end{bmatrix}
\]

(6)

\[
CMap_y = \begin{bmatrix}
y_{i1} & y_{i2} & \cdots & y_{in} \\
y_{i1}' & y_{i2}' & \cdots & y_{in}' \\
\vdots & \vdots & \ddots & \vdots \\
y_{n1}' & y_{n2}' & \cdots & y_{nm}' \\
\end{bmatrix}
\]

\(x'_{ij}\) is the contact probability score between amino acid \(x_i\) and \(x_j\) in protein sequence X, and \(y'_{ij}\) is the contact probability score between amino acid \(y_i\) and \(y_j\) in protein sequence Y. The residue-residue contact probabilities are predicted from the sequence by NNCon [46] (http://sysbio.rnet.missouri.edu/multicom_toolbox/). The contact map correlation score matrix \(CMap_{XY}\) is calculated as the multiplication of \(CMap_x\) and \(CMap_y\):

\[
CMap_{XY} = CMap_x \times CMap_y
\]

\[
\begin{bmatrix}
x'y_{11} & x'y_{12} & \cdots & x'y_{1n} \\
x'y_{21} & x'y_{22} & \cdots & x'y_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
x'y_{n1} & x'y_{n2} & \cdots & x'y_{nm} \\
\end{bmatrix}
\]

(7)

\(x'y_{ij}\) is the contact map score for an aligned residue pair (amino acid \(x_i\) in protein X and amino acid \(y_j\) in protein Y). The contact map score for the global alignment of two sequences X and Y is calculated as

\[
CMscore(X, Y) = \frac{1}{n^2} \sum_{i=1}^{n} CMap_{XY}(i, i)
\]

(8)

In practice, we only need to calculate the diagonal values in \(CMap_{XY}\).

Finally, we define the pairwise distance between sequences X and Y as

\[
d(X, Y) = 1 - \frac{W_4Optscore(X, Y)}{\min(n_1, n_2)} - W_5CMscore(X, Y)
\]

, where \(W_4 + W_5 = 1\). The weights \(W_4\) and \(W_5\) are used to control the influence of sequences X and Y.

Construction of guide tree and transformation of posterior probability

Akin to MSAProbs [4], a guide tree is constructed by the UPGMA method that uses the linear combinatorial strategy [47]. The distance between a new cluster Z formed by merging clusters X and Y, and another cluster W is calculated as (10):

\[
d(W, Z) = d(W, X) \times \frac{\text{Num}(X) + \text{Num}(Y)}{\text{Num}(X) + \text{Num}(Y)}
\]

(10)

In which \(\text{Num}(X)\) is the number of leaves in cluster X.

After the guide tree is constructed, sequences are weighted according to the schemes inferred in [4].

To reduce the bias of sampling similar sequences, we use a weighted scheme to transform the former posterior probability as

\[
P'_{XY} = \frac{1}{w_N}((w_x + w_y)P_{XY} + \sum_{Z \in S, Z \neq X, Y} w_x w_Z P_{ZX} P_{ZY})
\]

(11)

where \(w_x\) and \(w_y\) are, respectively, the weight of sequences X and Y, \(w_Z\) is the weight of a sequence Z other than X or Y in the given group of sequences, and \(w_N\) is the sum of sequence weights in dataset S.

Combination of progressive and iterative alignment

We first use the guide tree to generate a multiple sequence alignment by progressively aligning two clusters of the most similar sequences together. As in MSAProbs [4], we also apply a weighted profile-profile alignment to align two clusters of sequences. The sequence weights are the same as in the previous step. The posterior alignment probability matrix of two clusters/profiles is averaged from the probability matrices of all sequence pairs (X, Y), where x and y are from the
two different clusters. Formula (5) used to generate the global profile-profile alignment matrices is based on the posterior alignment probability matrices of the profiles. In order to further improve the alignment accuracy, we then use a randomized iterative alignment to refine the initial alignment. This randomized iterative refinement randomly partitions the given sequence group S into two separate groups, and performs a profile-profile alignment of the two groups. The iterative refinement can be completed after 10 iterations by default, or a fixed number of iterations set by users. Generally speaking, the final progressive alignment orders sequences along the guide tree from closely related to distantly related. To improve the alignment accuracy, a final iterative alignment is applied to refine the results from progressive alignment. In addition, a multi-thread technology based on OpenMP is also used to improve the efficiency of the program [48].

**Results and discussion**

**Evaluation of MSACompro and other tools on the standard benchmarks**

We tested MSACompro in comparison to three benchmarks: BALiBASE, SABmark and OXBENCH, and evaluated the alignment results in terms of sum-of-pairs (SP) score and true column (TC) score. The SP score is the number of correctly aligned pairs of residue in the test alignment divided by the total number of aligned pairs of residues in core blocks of the reference alignment [49]. The TC score is the number of correctly aligned columns in the test alignment divided by the total number of aligned columns in core blocks of the reference alignment [49]. We used the application bali_score provided by BALiBASE 3.0 to calculate these scores. We compared MSACompro to 11 other MSA tools which do not have access to the structural information, including ClustalW 2.0.12, DIALIGN-TX 1.0.2 [27], FSA 1.15.5, MAFFT 6.818, MSAPros 0.9.4, MUSCLE 3.8.31, Opal 0.2.0, POA 2, Probalign 1.3, Probcons and T-coffee 8.93. It is worth noting that a fair comparison between our method with these multiple sequence alignment methods without using structural features is not possible because these methods use less input information. So, the goal of comparison is to present the idea that structural information-based alignment may contain valuable information that is not available in sequence-based multiple sequence alignments and can therefore be a supplement to sequence-based alignments. And to make the evaluation more fair and comprehensive, we also compared MSACompro with four tools which use structural information, including MUMMALS 1.01 [14], PROMALS [15] and PROMALS3D [7].

To understand how various parameters of MSACompro affect alignment accuracy, some experiments were carried out to evaluate these variants based on two algorithm changes: (1) combining amino acids, secondary structure, and relative solvent accessibility information into the partition function calculation using respective weights for each of them; (2) computing the pairwise distance from both the pairwise posterior probability matrices and the pairwise contact map similarity matrices by introducing the weight wc for contact map information. To optimize the parameters, we used BALiBASE 3.0 data sets as training sets, and SABmark 1.65 and OXBENCH data sets as testing sets. Firstly, we focused on the effect of secondary structure and solvent accessibility information by testing different values of weight w1 for amino acid similarity and weight w2 for secondary structure information on BALiBASE 3.0 data sets. MSACompro worked wholly the best if the weight w1 for amino acid similarity and the weight w2 for secondary structure information were 0.4 and 0.5, respectively. Since the sum of w1, w2 and wc is 1, we can deduce that wc is 0.1 if w1 and w2 are 0.4 and 0.5. Then we focused on the effect of residue-residue contact map information under two different scenarios: using secondary structure and relevant solvent accessibility information by keeping the w1, w2, and wc at their optimum values (0.4, 0.5, 0.1), or excluding that information by setting both w2 and wc as 0. Evaluation results on BALiBASE 3.0 database were found to improve the most when wc is 0.9 by integrating both secondary structure and relevant solvent accessibility information. Additionally, to avoid over-fitting, we tested MSACompro against SABmark 1.65 and OXBENCH data sets using this set of parameters independently, and found that a significant improvement was also gained in comparison to other leading protein multiple sequence alignment tools. More details can be found in the next section, “A comprehensive study on the effect of predicted structural information on the alignment accuracy”. Consequently, the weights w1, w2, w3 and wc are respectively set at 0.4, 0.5, 0.1 and 0.9 in MSACompro by default. All other tools were also evaluated under default parameters.

Firstly, we evaluated these methods on BALiBASE [16] - the most widely used multiple sequence alignment benchmark. The latest version, BALiBASE 3.0, contains 218 reference alignments, which are distributed into five reference sets. Reference set 1 is a set of equal-distant sequences, which are organized into two reference subsets, RV11 and RV12. RV11 contains sequences sharing >20% identity and RV12 contains sequences sharing 20% to 40% identity. Reference set 2 contains families with >40% identity and a significantly divergent orphan sequence that shares <20% identity with the rest of the family members. Reference set 3 contains families with >40% identity that share <20% identity between each two different sub-families. Reference set 4 is a set of
sequences with large N/C-terminal extensions. Reference set 5 is a set of sequences with large internal insertions. Tables 1, 2, and 3 report the mean SP scores and TC scores of MSACompro and the tools without using structural information for the six subsets and the whole database. All the scores in the tables are multiplied by 100, and the highest scores in each column are marked in bold. The results show that MSACompro received the highest SP and TC scores on the whole database and all the subsets except for the SP score for the subset RV40. In some cases, MSACompro's improvement was substantial.

Secondly, we evaluated MSACompro and other tools without the help of structural information on the SABmark database [4], which is a very challenging data set according to a comprehensive study [50]. SABmark is an automatically generated data set consisting of two sets. One set is from SOFI [51] and the other is from the ASTRAL database [52], which contains remote homologous sequences in twilight-zone or superfamily. Since some pairwise reference alignments in SABmark are not generally consistent with multiple alignments, a subset of SABmark, 1.65 called SABRE [53], has been widely used as a multiple sequence alignment benchmark database. SABRE was constructed by identifying mutually consistent columns (MCCs) in the pairwise reference structure alignment. MCCs are considered similar to BALiBASE core blocks. SABRE contains 423 out of 634 SABmark groups that have eight or more MCCs. Table 4 shows the overall mean SP and TC scores of the alignments. The mean SP and TC scores of MSACompro are 8.3 and 9.1 points higher than those of the second best-performer, MSAProbs, demonstrating that incorporating predicted structural features into multiple sequence alignments can substantially improve

### Table 2 Total TC scores on the full-length BALiBASE 3.0 subsets.

| MSA tools | RV11   | RV12   | RV20   | RV30   | RV40   | RV50   |
|-----------|--------|--------|--------|--------|--------|--------|
| MSACompro| 47.13  | 86.93  | 47.16  | 58.63  | 64.42  | 63.43  |
| Clustalw  | 22.74  | 71.30  | 21.98  | 25.63  | 39.55  | 30.75  |
| DIALIGN-TX| 26.53  | 75.23  | 30.49  | 36.83  | 44.82  | 46.56  |
| FSA      | 26.95  | 81.77  | 18.68  | 24.63  | 47.43  | 39.81  |
| MAFFT    | 28.05  | 74.36  | 32.85  | 41.07  | 47.51  | 49.31  |
| MSAProbs | 44.11  | 86.54  | 46.44  | 57.63  | 62.18  | 60.75  |
| MUSCLE   | 31.79  | 80.30  | 35.86  | 34.02  | 45.02  | 45.94  |
| Opal     | 41.97  | 84.05  | 34.61  | 42.03  | 51.35  | 50.06  |
| POA      | 15.26  | 63.84  | 23.34  | 26.73  | 33.67  | 27     |
| Probalign| 45.34  | 86.20  | 43.93  | 53.60  | 60.31  | 54.94  |
| ProbCons | 41.66  | 85.55  | 40.63  | 51.47  | 53.22  | 57.31  |
| T-coffee | 42.29  | 85.25  | 38.88  | 47     | 55.94  | 58.69  |

Bold denotes the highest scores. MSACompro yielded the highest TC scores on all the subsets.

### Table 3 Overall mean SP and TC scores on the full-length BALiBASE 3.0 subsets.

| MSA tools | Mean SP score | Mean TC score |
|-----------|---------------|---------------|
| MSACompro| 88.846        | 61.313        |
| Clustalw  | 74.980        | 37.161        |
| DIALIGN-TX| 78.48         | 44.10         |
| FSA      | 77.878        | 41.688        |
| MAFFT    | 81.112        | 46.028        |
| MSAProbs | 87.336        | 60.248        |
| MUSCLE   | 81.496        | 47.151        |
| Opal     | 82.030        | 51.789        |
| POA      | 71.795        | 33.165        |
| Probalign| 87.161        | 58.528        |
| ProbCons | 85.965        | 55.422        |
| T-coffee | 85.728        | 55.239        |

Bold denotes the highest scores. MSACompro has the highest mean SP and TC scores.

### Table 4 Overall mean SP and TC scores on the SABmark 1.65.

| MSA tools | Mean SP score | Mean TC score |
|-----------|---------------|---------------|
| MSACompro| 68.85         | 49.07         |
| Clustalw  | 52.18         | 31.17         |
| DIALIGN-TX| 50.49         | 29.66         |
| FSA      | 46.03         | 25.73         |
| MAFFT    | 51.99         | 31.72         |
| MSAProbs | 60.55         | 39.95         |
| MUSCLE   | 54.99         | 34.35         |
| Opal     | 58.28         | 37.84         |
| POA      | 38.28         | 19.02         |
| Probalign| 59.96         | 38.66         |
| ProbCons | 59.81         | 38.99         |
| T-coffee | 59.49         | 39.08         |

Bold denotes the highest scores. The improvement of SP and TC scores on this data set is substantial.
alignment accuracy for even remotely related homologous sequences. Figure 1 shows an example comparison between the alignments generated by our method, MSACompro, and MSAProbs from the SABRE database. The SP and TC scores significantly improved from 0.307 to 0.853 and 0 to 0.780, respectively. This case demonstrates that taking predicted structural information can help avoid aligning unmatched regions, especially when the sequence similarity is unrecognizable.

Thirdly, we also assessed all the tools without using the structural information on the OXBENCH database [54]. OXBENCH is also a popular benchmark database generated by the AMPs multiple alignment method from the 3Dee database of protein structural domains [55]. The conserved columns in OXBENCH can be considered similar to BAliBASE core blocks. The mean SP and TC scores over the whole database are shown in Table 5. The results show that MSACompro improves the alignment accuracy over all other methods.

Finally, we also compared the SP scores and TC scores of MSACompro and other tools which adopt the structural information on the six subsets of BAliBASE database, SABmark database, and OXBENCH database. Tables 6 and 7 demonstrate the SP and TC scores across the three databases. The results show that MSACompro gained the highest scores on three out of six subsets of BAliBASE and achieved the third highest scores on other data sets, which are lower than PROMALS3D that used

---

**Figure 1** An example in SABRE database comparing the alignments generated by our method and MSAProbs. The reference alignment and resulting alignments generated by both methods are respectively shown in the figure. The correct alignment regions significantly improved by our MSACompro after taking structural information are marked in red rectangles. In contrast, the corresponding incorrect alignment regions generated by MSAProbs are represented in green rectangles. The predicted secondary structure and solvent accessibility information for the correctly aligned regions are shown in circles.
true experimental structures as input and PROMALS that used both predicted secondary structures and additional homologous protein sequences found by PSI-BLAST search’s on a large protein sequence database [15]. Overall, MSACompro performed similarly as PROMALS, whereas the latter has an advantage on a remote homologous protein sequence data set SABmark since it directly incorporates additional homologous protein sequences to improve the alignment of remotely related target sequences during the progressive alignment process. Moreover, the accuracy of MSACompro on the BAliBASE 3.0 data sets seems to be higher than the published results of another alignment tool of using secondary structure information - DIALIGN-SEC [12], which was not directly tested in our experiment because it is only available as a web server other than a downloadable software package. Therefore, MSACompro is useful to improve the accuracy of multiple sequence alignment in general and particularly for most cases in reality where experimental structures are not available.

In order to check if alignment score differences between MSACompro and the other alignment methods are statistically significant, we carried out the Wilcoxon matched-pair signed-rank test [56] on both SP and TC scores of these methods on the three data sets. The p-values of alignment score differences calculated by the Wilcoxon matched-pair signed-rank test are reported in Table 8. Generally speaking, the alignment scores of MSACompro are significantly higher than all the alignment methods without using structural information and MUMMALS of using structural information in all but one case according to the significance threshold of 0.05. The exception is that MSACompro’s TC score is higher than MSAProbs on the BAliBASE, but not statistically significant. However, the alignment scores of MSACompro are mostly statistically lower than the other two alignment methods (PROMALS or PROMALS3D) of using predicted structural features, more homologous sequences, or tertiary structures.

In addition to alignment accuracy, alignment speed is also a factor to consider in time-critical applications. Because it is difficult to rigorously compare the speed of different methods due to the difference in implementation and inputs, we only report the roughly estimated running time of the different methods on BAliBASE based our empirical observations. The fastest methods are ClustalW, MAFFT, MUSCLE, and POA, which used less than one hour. The medium-speed methods that used a few hours to less than one day include FSA, Opal, Probalign, MSAProbs, ProbCons, T-coffee, MUMMALS, and DIALIGN-TX. The more time demanding methods are MSACompro, PROMALS, and PROMALS3D.

| MSA tools      | Mean SP score | Mean TC score |
|----------------|---------------|---------------|
| MSACompro      | 92.60         | 84.99         |
| Clustalw       | 89.45         | 80.19         |
| DIALIGN-TX     | 86.25         | 75.29         |
| FSA            | 86.47         | 75.79         |
| MAFFT          | 87.58         | 76.75         |
| MSAProbs       | 90.06         | 81.40         |
| MUSCLE         | 89.50         | 80.34         |
| Opal           | 89.38         | 79.77         |
| POA            | 82.19         | 68.40         |
| Probalign      | 89.97         | 81.39         |
| ProbCons       | 89.68         | 80.52         |
| T-coffee       | 89.56         | 80.27         |

Table 5: Overall mean SP and TC scores on the OXBENCH. Bold denotes highest scores.

| MSA tools      | RV11  | RV12  | RV20  | RV30  | RV40  | RV50  | Whole BAliBASE | SABmark | OXBENCH  |
|----------------|-------|-------|-------|-------|-------|-------|----------------|---------|----------|
| MSACompro      | 73.14 | 94.84 | 93.30 | 87.16 | 92.11 | 91.41 | 88.85          | 68.85   | 92.60    |
| MUMMALS        | 66.94 | 94.30 | 91.04 | 84.79 | 87.15 | 87.91 | 85.53          | 62.12   | 90.25    |
| PROMALS        | 79.08 | 93.55 | 93.31 | 88.30 | 89.80 | 90.27 | 89.00          | 77.40   | 93.76    |
| PROMALS3D      | 83.58 | 92.33 | 93.62 | 89.42 | 90.93 | 89.73 | 90.14          | 88.89   | 97.37    |

Table 6: Total SP scores of the tools which use the structural information on BAliBASE 3.0 subsets, SABmark data sets and OXBENCH data sets.

| MSA tools      | RV11  | RV12  | RV20  | RV30  | RV40  | RV50  | Whole BAliBASE | SABmark | OXBENCH  |
|----------------|-------|-------|-------|-------|-------|-------|----------------|---------|----------|
| MSACompro      | 47.13 | 86.93 | 47.16 | 58.63 | 64.42 | 63.43 | 61.31          | 49.07   | 84.99    |
| MUMMALS        | 41.61 | 83.98 | 42.83 | 49.40 | 48.55 | 52.88 | 53.85          | 41.96   | 81.43    |
| PROMALS        | 58.24 | 81.73 | 49.59 | 51.63 | 50.84 | 57.19 | 59.27          | 60.95   | 86.73    |
| PROMALS3D      | 66.71 | 79.30 | 55.95 | 61.07 | 51.67 | 54.38 | 62.16          | 80.22   | 93.25    |

Table 7: Total TC scores of the tools which use the structural information on BAliBASE 3.0 subsets, SABmark data sets and OXBENCH data sets.

| MSA tools      | RV11  | RV12  | RV20  | RV30  | RV40  | RV50  | Whole BAliBASE | SABmark | OXBENCH  |
|----------------|-------|-------|-------|-------|-------|-------|----------------|---------|----------|
| MSACompro      | 47.13 | 86.93 | 47.16 | 58.63 | 64.42 | 63.43 | 61.31          | 49.07   | 84.99    |
| MUMMALS        | 41.61 | 83.98 | 42.83 | 49.40 | 48.55 | 52.88 | 53.85          | 41.96   | 81.43    |
| PROMALS        | 58.24 | 81.73 | 49.59 | 51.63 | 50.84 | 57.19 | 59.27          | 60.95   | 86.73    |
| PROMALS3D      | 66.71 | 79.30 | 55.95 | 61.07 | 51.67 | 54.38 | 62.16          | 80.22   | 93.25    |
because they need to generate extra information for alignment. We ran both PROMALS and MSACompro on the BAliBASE 3.0 database on an eight-core (i.e. 32 CPU cores) Linux server to calculate their running time. It took about 4 days and 6 hours for PROMALS to run on the whole BAliBASE 3.0 data sets, and about 9 hours and 13 minutes for MSACompro to run on the same data sets. MSACompro was faster because it used a multiple-threading implementation to call SSpro/ACCpro to predict secondary structure and solvent accessibility in parallel. Out of about 9 hours and 13 minutes, about four hours and 17 minutes were used by MSACompro to align sequences if secondary structure and solvent accessibility information was provided. However, if only one CPU core is used, it took around 6 days and 14 hours for SSpro and ACCpro called by MSACompro to predict secondary structure and solvent accessibility information alone, which is time-consuming. Therefore, MSACompro will be slower than PROMALS if it runs a single CPU core, but faster on multiple (> = 3) CPU cores. As for PROMALS3D, it used about 9 days to extract tertiary structure information and make alignments.

A comprehensive study of the effect of predicted structural information on the alignment accuracy

To understand the impact of predicted secondary structure, relative solvent accessibility, and contact map on the accuracy of multiple sequence alignment, we tested their effects on alignments individually or in combination by adjusting the values of their weights used in the partition function (i.e. for secondary structure and solvent accessibility) or in the distance calculation (i.e. for contact map).

I. Effect of secondary structure information

We studied the effect of secondary structure information by adjusting the values of $w_1$ (weight for amino acid sequence information) and $w_2$ (weight for secondary structure information), the sum of which was kept constant.

Table 8 The statistical significance (i.e. p-values) of SP and TC alignment score differences between MSACompro and the other tools on three benchmark data sets.

| MSA tools/Score Type | Whole BAliBASE | SABmark | OXBENCH |
|----------------------|----------------|---------|---------|
| Clustalw/SP score    | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| Clustalw/TC score    | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| DIALIGN-TX/SP score  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| DIALIGN-TX/TC score  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| FSA/SP score         | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| FSA/TC score         | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| MAFFT/SP score       | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| MAFFT/TC score       | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| MSAProbs/SP score    | 2.931 x 10^{-3}  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| MSAProbs/TC score    | 0.4839           | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| MUSCLE/SP score      | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| MUSCLE/TC score      | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| Opal/SP score        | 3.84 x 10^{-16}  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| Opal/TC score        | 2.15 x 10^{-14}  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| POA/SP score         | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| POA/TC score         | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| Probalign/SP score   | 2.87 x 10^{-6}   | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| Probalign/TC score   | 4.158 x 10^{-3}  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| ProbCons/SP score    | 2.16 x 10^{-15}  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| ProbCons/TC score    | 6.817 x 10^{-7}  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| T-coffee/SP score    | 1.225 x 10^{-14} | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| T-coffee/TC score    | 4.503 x 10^{-8}  | < 2.2 x 10^{-16} | < 2.2 x 10^{-16} |
| MUMMALS/SP score     | 6.191 x 10^{-10} | < 2.2 x 10^{-16} | 2.446 x 10^{-15} |
| MUMMALS/TC score     | 8.104 x 10^{-5}  | < 2.2 x 10^{-16} | 1.265 x 10^{-12} |
| PROMALS/SP score     | 0.0116 (-)       | < 2.2 x 10^{-16} | (-) 0.0186 (-) |
| PROMALS/TC score     | 0.529 (-)        | < 2.2 x 10^{-16} | (-) 0.0274 (-) |
| PROMALS3D/SP score   | 0.0049 (-)       | < 2.2 x 10^{-16} | (-) 2.2 x 10^{-16} |
| PROMALS3D/TC score   | 0.0078 (-)       | < 2.2 x 10^{-16} | (-) 2.2 x 10^{-16} |

The p-values were calculated using the Wilcoxon matched-pair signed-rank test. All the p-values except for ones denoted by "(-)" are for hypothesis testing that MSACompro has higher alignment scores than the other methods. The p-values denoted by "(-)" are for hypothesis testing that MSACompro has lower alignment scores than the other methods.
as 1, and setting the values of $w_3$ (weight for relative solvent accessibility) and $w_c$ (weight for contact map) to 0. The results for different $w_2$ values on the SABmark data sets are shown in Table 9. The highest score is denoted in bold and by a superscript of star, and the second highest is denoted in bold. The results show that incorporating secondary structure information always improves alignment accuracy over the baseline established without using secondary structure information ($w_2 = 0$). The highest accuracy is achieved when $w_2$ is set to .5, at which point the score is 8 points greater than the baseline. $w_2 = 1$ means that only secondary structure is used to calculate the posterior alignment probability in the partition function (i.e. equation set (2)), but amino acid sequence similarity is still used to calculate the other posterior alignment probability by the pair Hidden Markov Models. Figures 2 and 3 plot the SP and TC scores against weight values in Table 9 and Table 10, respectively.

**II. Effect of relative solvent accessibility information**

Similarly, we studied the effect of relative solvent accessibility on the SABmark by adjusting the values of $w_1$ and $w_3$ and setting the values of $w_2$ and $w_c$ to 0. The SP and TC scores with respect to different weight values are shown in Tables 11 and 12, respectively. The scores are also plotted against the weights in Figures 4 and 5, respectively. The highest SP and TC scores were achieved when $w_3$ was set to 0.5 or 0.6.

**III. Effect of residue-residue contact map information**

We investigated the effect of contact map information on the BALIBASE 3.0 data set by adjusting $w_c$ and setting $w_2$ and $w_3$ to 0. We used NNcon to successfully predict the contact maps for subset RV11, RV30, 42 out of 44 alignments in RV12, 38 out of 40 in RV20, 33 out of 46 in RV40, and 14 out of 16 in RV50. We tested the MSACompro method against this data with contact predictions. Tables 13 and 14 show the SP and TC scores for different $w_c$ values on the subsets of the BALIBASE dataset. The results show that using contact information improved the alignment accuracy on some, but not all, subsets.

**IV. Effect of combining secondary structure and solvent accessibility information**

We adjusted the values of $w_1$ (weight for amino acid), $w_2$ (weight for secondary structure) and $w_3$ (weight for relative solvent accessibility) simultaneously to investigate the effect of using secondary structure and relative solvent accessibility together. SP and TC scores on different parameter combinations are shown in Tables 15 and 16. The highest score is denoted in bold and by a superscript of 1, the second in bold and by a superscript of 2, and the third in bold and by a superscript of 3. The results show that the highest scores are achieved when $w_1$ ranges from 0.4 to 0.5, $w_2$ from 0.4

### Table 9 SP scores for different weights of secondary structures on the SABmark benchmark. Bold denotes the two best scores, and an extra superscript of star denotes the highest score.

| $w_2$ | 0  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|-------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|---|
| SP    | 60.553 | 62.988 | 65.514 | 67.333 | 68.348 | **68.698** | 68.465 | 68.159 | 67.282 | 66.153 | 64.745 |

The results show that using secondary structure information (i.e. $w_2 > 0$) always increases the alignment scores over without using it (i.e. $w_2 = 0$). MSACompro yielded the highest accuracy score of ~68.70 when $w_3$ is set to 0.5.
to 0.5, and w3 from 0.1 to 0.2. Also, using both secondary structure and solvent accessibility improves alignment accuracy over using either one. The best alignment score, which uses both secondary structure and solvent accessibility, is >8 points higher than the baseline approach, which does not use them. The changes of SP scores and TC scores with respect to the weights are visualized by the 3D plots in Figures 6 and 7. We conducted similar experiments on BALiBASE 3.0 and OXBENCH and got the similar results (data not shown).

V. Effect of using contact map information together with secondary structure and solvent accessibility information

In order to study whether or not contact information can be used effectively with secondary structure and solvent accessibility, we adjusted the weight wc for contact information, while keeping the w1, w2, and w3 at their optimum values (0.4, 0.5, and 0.1 respectively). Tables 17 and 18 report the SP and TC scores on the BALiBASE 3.0 data set for different wc values from no contact information (wc = 0) to maximum contact information (wc = 1). The results show that the improvement caused by contact information seems not to be substantial and significant.

Table 10 TC scores for different weights of secondary structures on the SABmark benchmark. Bold denotes the two best scores, and an extra superscript of star denotes the highest score.

| w2  | 0   | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| TC  | 39.948 | 42.643 | 45.262 | 47.442 | 48.754 | 49.005* | 48.745 | 48.352 | 47.142 | 45.4023 | 43.385 |

The results show that using secondary structure information (i.e. w2 > 0) always increases the alignment scores over without using it (i.e. w2 = 0). MSACompro yielded the highest accuracy score of ~68.70 when w2 is set to 0.5.

Table 11 SP scores for different weights of relative solvent accessibility on the SABmark benchmark. Bold denotes the two best scores, and an extra superscript of star denotes the highest score.

| w3  | 0   | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SP  | 60.553 | 61.753 | 63.260 | 64.171 | 65.124 | 65.199 | 65.249 | 65.037 | 64.388 | 63.1882 | 61.723 |

The results show that using relative solvent accessibility information (i.e. w3 > 0) always increases the alignment scores over without using it (i.e. w3 = 0). MSACompro yielded the highest accuracy score of ~68.70 when w3 is set to 0.6.

Table 12 TC scores for different weights of relative solvent accessibility on the SABmark benchmark. Bold denotes the two best scores, and an extra superscript of star denotes the highest score.

| w3  | 0   | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| TC  | 39.948 | 41.300 | 43.035 | 43.943 | 44.870 | 45.442* | 45.184 | 45.031 | 44.0383 | 42.4471 | 41.012 |

The results show that using relative solvent accessibility information (i.e. w3 > 0) always increases the alignment scores over without using it (i.e. w3 = 0). MSACompro yielded the highest accuracy score of ~68.70 when w3 is set to 0.5.
Table 13 SP scores for different weights for contact map on the BAliBASE3.0 database. Red color highlights the improved scores on each BAliBASE subset. Bold denotes the increased scores.

| subset | wc | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|--------|----|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|---|
| RV11   |    | 0.6829 | 0.686 | 0.686 | 0.684 | 0.684 | 0.683 | 0.687 | 0.684 | 0.687 | 0.687 | 0.668 |
| RV12   |    | 0.9461 | 0.946 | 0.946 | 0.945 | 0.945 | 0.946 | 0.946 | 0.945 | 0.946 | 0.946 | 0.944 |
| RV20   |    | 0.9297 | 0.927 | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 | 0.926 | 0.927 | 0.924 |
| RV30   |    | 0.865 | 0.865 | 0.864 | 0.864 | 0.864 | 0.863 | 0.864 | 0.864 | 0.864 | 0.865 | 0.817 |
| RV40   |    | 0.928 | 0.926 | 0.926 | 0.924 | 0.924 | 0.924 | 0.924 | 0.924 | 0.936 | 0.934 | 0.933 | 0.927 |
| RV50   |    | 0.9091 | 0.908 | 0.910 | 0.910 | 0.909 | 0.909 | 0.909 | 0.907 | 0.907 | 0.908 | 0.886 |

Table 14 TC scores for different weights for contact map on the BAliBASE 3.0 database. Red highlights the improved scores on each BAliBASE subset. Bold denotes the increased scores.

| subset | wc | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|--------|----|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|---|
| RV11   |    | 0.441 | 0.445 | 0.445 | 0.444 | 0.444 | 0.444 | 0.447 | 0.447 | 0.448 | 0.451 | 0.417 |
| RV12   |    | 0.8669 | 0.865 | 0.866 | 0.866 | 0.866 | 0.866 | 0.866 | 0.867 | 0.867 | 0.867 | 0.858 |
| RV20   |    | 0.482 | 0.479 | 0.473 | 0.460 | 0.457 | 0.462 | 0.453 | 0.453 | 0.453 | 0.457 | 0.439 |
| RV30   |    | 0.607 | 0.605 | 0.594 | 0.594 | 0.592 | 0.592 | 0.591 | 0.591 | 0.593 | 0.592 | 0.415 |
| RV40   |    | 0.67 | 0.667 | 0.667 | 0.661 | 0.659 | 0.662 | 0.662 | 0.682 | 0.682 | 0.681 | 0.642 |
| RV50   |    | 0.625 | 0.621 | 0.634 | 0.633 | 0.629 | 0.628 | 0.631 | 0.615 | 0.615 | 0.603 | 0.556 |

Table 15 SP scores for different weight combinations (w1 - amino acid, w2 - secondary structure, w3 - solvent accessibility) on the SABmark 1.65 dataset.

| w2 \ w1 | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|--------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|---|
| 0      | 61.723 | 63.188 | 64.388 | 65.037 | 65.249 | 65.199 | 65.124 | 64.171 | 63.260 | 61.753 | 60.553 |
| 0.1    | 63.303 | 64.600 | 65.635 | 66.492 | 66.702 | 66.619 | 66.423 | 65.717 | 64.790 | 62.988 |
| 0.2    | 64.759 | 66.055 | 67.161 | 67.598 | 68.104 | 67.831 | 67.469 | 66.775 | 65.514 |
| 0.3    | 65.781 | 66.974 | 67.867 | 68.312 | 68.414 | 68.418 | 68.033 | 67.333 |
| 0.4    | 66.424 | 67.531 | 68.251 | 68.743 | 69.016 | 68.920 | 68.347 |
| 0.5    | 66.847 | 67.907 | 68.4 | 68.859 | 68.933 | 68.698 |
| 0.6    | 66.843 | 67.911 | 68.544 | 68.560 | 68.465 |
| 0.7    | 66.739 | 67.800 | 68.135 | 68.159 |
| 0.8    | 66.389 | 67.119 | 67.282 |
| 0.9    | 65.445 | 66.153 |
| 1      | 64.745 |

Bold denotes the top 3 highest scores. The highest score is indicated by a superscript of 1, the second highest by a superscript of 2, and the third highest by a superscript of 3. The table only shows the values of w1 and w2 because w3 can be inferred by 1 - w1 - w2.

Table 16 TC scores for different weight combinations (w1 - amino acid, w2 - secondary structure, w3 - solvent accessibility) on the SABmark 1.65 dataset.

| w2 \ w1 | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|--------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|---|
| 0      | 41.012 | 42.447 | 44.038 | 45.031 | 45.184 | 45.442 | 44.870 | 43.943 | 43.035 | 41.300 | 39.948 |
| 0.1    | 42.558 | 44.147 | 45.596 | 46.863 | 47.043 | 46.910 | 46.676 | 45.333 | 44.900 | 42.643 |
| 0.2    | 43.915 | 45.678 | 47.270 | 47.927 | 48.619 | 48.080 | 47.584 | 47.002 | 45.262 |
| 0.3    | 45.582 | 46.768 | 48.116 | 48.660 | 48.905 | 48.660 | 48.371 | 47.442 |
| 0.4    | 46.104 | 47.340 | 48.473 | 48.889 | 49.508 | 49.1589 | 48.754 |
| 0.5    | 46.440 | 47.809 | 48.210 | 49.078 | 49.222 | 49.005 |
| 0.6    | 46.577 | 47.619 | 48.487 | 48.797 | 48.745 |
Conclusion
In this work, we designed a new method to incorporate predicted secondary structure, relative solvent accessibility, and residue-residue contact information into multiple protein sequence alignment. Our experiments on three standard benchmarks showed that the method improved multiple sequence alignment accuracy over most existing methods without using secondary structure and solvent accessibility information. However, the performance of the method is comparable to PROMALS and PROMALS3D by slightly lower scores on some subsets and behind it by a large margin on SABMARK probably because these two methods used homologous sequences or tertiary structure information in addition to secondary structure information. Since multiple sequence alignment is often a crucial step for bioinformatics analysis, this new method may help

Table 16 TC scores scores for different weight combinations (w₁ - amino acid, w₂ - secondary structure, w₃ - solvent accessibility) on the SABmark 1.65 dataset. (Continued)

|   | 0.7  | 0.8  | 0.9  | 1.0  |
|---|------|------|------|------|
| w₁ | 46.147 | 47.759 | 48.083 | 48.352 |
| w₂ | 45.714 | 46.898 | 47.142 |      |
| w₃ | 44.424 | 45.492 |      |      |

Bold denotes the top 3 highest scores. The highest score is indicated by a superscript of 1, the second highest by a superscript of 2, and the third by a superscript of 3. The table only shows the values of w₁ and w₂ because w₃ can be inferred by 1 - w₁ - w₂.

Figure 6 3D plot of SP scores against secondary structure weight w₂ and relative solvent accessibility weight w₃.
Figure 7 3D plot of TC scores against secondary structure weight $w_2$ and relative solvent accessibility weight $w_3$.

Table 17 SP scores for different contact map weight $w_c$ on the BAliBASE3.0 database while keeping the weights for amino acid, secondary structure, solvent accessibility to 0.4, 0.5, and 0.1, respectively.

| subset | the weight $w_c$ | 0   | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1   |
|--------|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| RV11   | 0.729           | 0.730| 0.728| 0.726| 0.726| 0.726| 0.727| 0.72547| 0.732| 0.731| 0.722|
| RV12   | 0.947           | 0.948| 0.947| 0.949| 0.948| 0.948| 0.948| 0.94855| 0.948| 0.948| 0.945|
| RV20   | 0.934           | 0.933| 0.932| 0.934| 0.934| 0.934| 0.933| 0.93282| 0.933| 0.933| 0.934|
| RV30   | 0.876           | 0.877| 0.877| 0.876| 0.873| 0.873| 0.873| 0.87287| 0.873| 0.872| 0.846|
| RV40   | 0.909           | 0.908| 0.909| 0.909| 0.909| 0.909| 0.909| 0.909| 0.909| 0.921| 0.913|
| RV50   | 0.911           | 0.910| 0.911| 0.909| 0.909| 0.908| 0.902| 0.90807| 0.914| 0.914| 0.871|

Bold denotes the increased scores.

Table 18 TC scores for different contact map weight $w_c$ on the BAliBASE3.0 database while keeping the weights for amino acid, secondary structure, solvent accessibility to 0.4, 0.5, and 0.1, respectively.

| subset | the weight $w_c$ | 0   | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1   |
|--------|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| RV11   | 0.470           | 0.472| 0.471| 0.469| 0.468| 0.468| 0.468| 0.468| 0.475| 0.471| 0.450|
| RV12   | 0.870           | 0.870| 0.869| 0.872| 0.872| 0.871| 0.871| 0.872| 0.870| 0.869| 0.863|
| RV20   | 0.481           | 0.465| 0.460| 0.478| 0.478| 0.477| 0.477| 0.472| 0.471| 0.472| 0.468|
| RV30   | 0.609           | 0.591| 0.590| 0.588| 0.589| 0.588| 0.588| 0.587| 0.589| 0.586| 0.434|
| RV40   | 0.628           | 0.626| 0.624| 0.625| 0.625| 0.625| 0.624| 0.624| 0.6249| 0.644| 0.6124|
| RV50   | 0.601           | 0.595| 0.60071| 0.601| 0.596| 0.596| 0.586| 0.625| 0.63643| 0.634| 0.55|

Bold denotes the increased scores.
improve the solutions to many bioinformatics problems such as protein sequence analysis, protein structure prediction, protein function prediction, protein interaction analysis, protein mutagenesis and protein engineering.

Acknowledgements
The work was partially supported by a NIH R01 grant (no. 1R01GM093123) to JC. We thank Angela Zhang for English editing.

Author details
1Department of Computer Science, University of Missouri-Columbia, Columbia, MO 65211, USA. 2Informatics Institute, University of Missouri-Columbia, Columbia, MO 65211, USA. 3C. Bond Life Science Center, University of Missouri-Columbia, Columbia, MO 65211, USA.

Authors’ contributions
JC and XD designed the algorithm. XD implemented the algorithm and carried out the experiments. XD and JC analyzed the data. XD and JC wrote the manuscript. XD and JC approved it.

Received: 16 April 2011 Accepted: 14 December 2011

Published: 14 December 2011

References
1. Barton GJ, Sternberg MJ: A strategy for the rapid multiple alignment of protein sequences, confidence levels from tertiary structure comparisons. J Mol Biol 1987, 198:327-337.
2. Feng DF, Doolittle RF: Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 1987, 25:351-361.
3. Krogh A, et al: Hidden markov models in computational biology: applications to protein modeling. J Mol Biol 1994, 235:1503-1513.
4. Liu YC, Schmidt B, Douglas LM: MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. Bioinformatics 2010, 26(16):1958-1964.
5. Do CB, et al: ProCons: probabilistic consistency-based multiple sequence alignment. Genome Res 2005, 15:330-340.
6. Parot O, Suhe K, Abergel C, Sammon OT, Notredame C: 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. Nucleic Acids Research 2004, 32:37-40.
7. Pei J, Kim B, Grishin NV: PROMALS3D: a tool for multiple sequence and structure alignment. Nucleic Acids Research 2008, 36(7):2295-2300.
8. Söding J, Biegert A, Lupas AN: The HHpred interactive server for protein homology detection by HMM-HMM comparison. Bioinformatics 2005, 21:951-960.
9. Heringa J: Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. Comput Chem 1999, 23:341-364.
10. Kim NK, Xie J: Protein multiple alignment incorporating primary and secondary structure information. J Comput Biol 2006, 13:75-88.
11. Amarendran RS, Kaufmann M, Morgenstern B: DIALIGN-TX and multiple protein alignment using secondary structure information at GOBICS. Nucleic Acids Research 2010, 38.Suppl 2(W19):W21-W22.
12. Zhou HY, Zhou YQ: SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. Bioinformatics 2005, 21:3615-3621.
13. Pei J, Grishin NV: MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic Acids Res 2006, 34(16):4364-4374.
14. Pei J, Grishin NV: PROMALS: towards accurate multiple sequence alignment of distantly related proteins. Bioinformatics 2007, 23:802-808.
15. Brudno M, Steinkamp R, Morgenstern B: The CHAOS/DIALIGN www server for multiple alignment of genomic sequences. Nuclei Acids Res 32(Supplement 2):W41.
16. Larkin M, et al: Clustal W and Clustal X version 2.0. Bioinformatics 2007, 23(21):2947-2948.
45. Durbin R, et al: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press Cambridge, UK; 1998.
46. Tegge AN, Wang Z, Eickholt J, Cheng J: NNcon: Improved Protein Contact Map Prediction Using 2D-Recursive Neural Networks. Nucleic Acids Research 2009, 37:w515-w518.
47. Sneath PHA, Sokal RP: Numerical taxonomy. Freeman San Francisco, USA; 1973.
48. OpenMP tutorial. [https://computing.llnl.gov/tutorials/openMP].
49. Thompson JD, Frederic P, Olivier P: A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Research 1999, 27:2682-2690.
50. Walle V, et al: Align-m-a new algorithm for multiple alignment of highly divergent sequences. Bioinformatics 2004, 20:1428-1435.
51. Boutonnet NS, et al: Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. Protein Eng 1995, 8:647-662.
52. Brenner SE, et al: The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000, 28:254-256.
53. Edgar RC. [http://www.drive5.com/bench].
54. Raghava GP, et al: OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics 2003, 4:47.
55. Poirot O, Suhre K, Abergel C, Eamonn OT, Notredame C: 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. Nucleic Acids Research 2004, 32:37-40.
56. Wilcoxon F: Probability tables for individual comparisons by ranking methods. Biometrics 1947, 3:119-122.