IMPROVED COMPUTATIONAL METHODS OF PROTEIN SEQUENCE ALIGNMENT, MODEL SELECTION AND TERTIARY STRUCTURE PREDICTION

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Computer Science

---

by

XIN DENG

Dr. Jianlin Cheng, Thesis Supervisor

DECEMBER 2013

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

IMPROVED COMPUTATIONAL METHODS OF PROTEIN ALIGNMENT, MODEL SELECTION AND TERTIARY STRUCTURE PREDICION

presented by XIN DENG,
a candidate for the degree of Doctor of Computer Science and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Jianlin Cheng

Dr. Dong Xu

Dr. Ye Duan

Dr. John C Walker

ACKNOWLEDGMENTS

I would never have been able to make some breakthroughs in my research or finish my dissertation without the guidance of my committee members, help from friends, and support from my parents.

I would like to express my deepest gratitude to my advisor, Dr.Cheng, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. He not only gave me his best help to explore my potential and develop me to be a good researcher, but also offered me many valuable opinions in the life philosophy. Particularly, he told us, "we need to earn everything not gain in this world". Thanks to his words, when I look back to my past five-year's PhD study and research, I gained some excited achievements, made my own contributions to the scientific society, and I did not waste my life.

I would like to thank Dr. Xu for leading the researches in the whole Computer Science Department and also having given much useful advice in my research. I would also like to thank Dr. Walker for giving me a chance in the cooperated project and leading me to a world of genes involved in abscission in Arabidopsis. Last, sincere thanks go to Dr. Duan, who patiently gave me advice in the qualify exam and was willing to participate in both my comprehensive exam and final defense.

I would also like to thank my parents. They were always supporting me and encouraging me with their best wishes. They cheered me up and stood by me through the good times and bad.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

Protein sequence and profile alignment 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.

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. 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. We also developed a novel profile-profile pairwise protein sequence alignment method based on pair HMM (Hidden Markov Model) by integrating the predicted secondary structure, solvent accessibility, torsion angle and evolutionary constraint information from the protein pairs. The evaluation showed that the secondary structure, relative solvent accessibility, torsion angle information significantly improved the alignment accuracy in comparison with the state of the art methods HHsearch and HHsuite. The evolutionary constraint information did help in some cases, especially the alignments of the proteins which are of short lengths, typically 100 to 500 residues. Furthermore, we

believe adopting evolutionary constraint information into the protein profile-profile pairwise alignment provides a useful point of view for the future improvement.

Protein Model selection is also a key step in protein tertiary structure prediction. We developed two SVM model quality assessment methods, taking either a query-single template pairwise alignment or a query-multi template alignment as input. The assessment results illustrated that such a novel, effective method may help improve the model selection, protein structure prediction and many other bioinformatics problems.

Based on the above methods, some in-house tools in our group, and other open public tools, we built up a MULTICOM conformation ensemble system of protein tertiary structure prediction. The system performed well in the CASP10 (Critical Assessment of Techniques for Protein Structure Prediction) competition.

Chapter 1

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

1.1 Introduction

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 tech-

niques 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 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, 11, 12, 13, 14, 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 [4, 5, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 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.

1.2 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.

1.2.1 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,...,x_{n1}), Y =(y_1,y_2,...,y_{n2}), where x_1,x_2,...,x_{n1} and y_1,y_2,...,y_{n2} are lists of the residues in X and Y, respectively. n1 is the length of sequence X, and n2 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 of all the possible global alignments of X and Y, and aln* $\in$ ALN 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:

$$
\begin{aligned}
&P(x_i \sim y_j \in aln^* \mid X, Y) = \sum_{aln^* \in ALN} P(aln \mid X, Y) I\{x_i \sim y_j \in aln\} \\
&I\{x_i \sim y_j \in aln\} = \begin{cases} 1 & \text{if } x_i \sim y_j \in aln \text{ true} \\ 0 & \text{otherwise} \end{cases} \\
&\qquad\qquad (1 \le x_i \le n_1, 1 \le y_j \le n_2)
\end{aligned} \tag{1.1}
$$

$P(x_i \sim y_j \in aln^* \mid X, Y)$ denotes the probability that is the true alignment aln^* . Thus, the posterior probability matrix is $n_1 \times n_2$ a collection of all the values $P(x_i \sim y_j \in aln^* \mid X, Y)(P(x_i \sim y_j \in aln^* \mid X, Y)$ for short) for $1 \le x_i \le n_1, 1 \le y_j \le n_2$. 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^1_{XY} and P^2_{XY}), respectively. The first kind of pairwise probability matrix P^1_{XY} 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, $F_Y(i, j)$ 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:

$$
\begin{aligned}
F_M(i,j) &= F(i-1, j-1) e^{W_1 \beta s(x_i, y_j) + W_2 SS(ss(x_i), ss(y_j)) + W_3 SA(sa(x_i), sa(y_j))} \\
F_Y(i,j) &= F_M(i, j-1) e^{\beta gap} + F_Y(i, j-1) e^{\beta ext} \\
F_X(i,j) &= F_M(i-1, j) e^{\beta gap} + F_X(i-1, j) e^{\beta ext} \\
F(i,j) &= F_M(i,j) + F_Y(i,j) + F_X(i,j)
\end{aligned} \tag{1.2}
$$

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 in protein Y according to the secondary structure similarity matrix SS, $SA(sa(x_i), sa(y_j))$ is the similarity score between the relative 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 MSAprobs, β 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:

$$SS = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

, where two identical secondary structures receive a score of 1 and different ones receive a score of 0. The 2 × 2 solvent accessibility similarity matrix SA contains the similarity scores of two kinds of relative solvent accessibilities (e, b) as follows:

$$SA = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

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 advance 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:

$$P^1(x_i \sim y_j) = \frac{F_M(i-1, j-1)F'_M(i+1, j+1)}{F} \times e^{W_1 \beta s(x_i, y_j) + W_2 SS(ss(x_i), ss(y_j)) + W_3 SA(sa(x_i), sa(y_j))} \tag{1.3}$$

where $F'_M(i, j)$ denotes the partition function of all the reverse alignments starting from the position (n1, n2) till position (i, j) with x_i aligned to y_j.

As in MSAProbs, 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.

$$P(x_i \sim y_j) = \sqrt{\frac{P^1(x_i \sim y_j)^2 + P^2(x_i \sim y_j)^2}{2}} \tag{1.4}$$

where $P^1(x_i \sim y_j)$ and $P^2(x_i \sim y_j)$ denote a posterior probability element in two kinds of posterior probability matrices (P^1_{XY} and P^2_{XY}), respectively.

1.2.2 Construction of pairwise distance matrices based on pairwise posterior probabilities and pairwise contact map 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:

$$AS(i,j) = max \begin{cases} AS(i-1, j-1) + P_{XY}(x_i \sim y_j) \\ AS(i-1, j) \\ AS(i, j-1) \end{cases} \tag{1.5}$$

$AS(n_1, n_2)$ 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 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,

$$x_1 x_2 \ldots - x_m \ldots x_p \ldots x_{n1}$$
$$y_1 - \ldots y_k y_{k+1} \ldots - \ldots y_{n2}$$

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

$$x_1 \ldots x_m \ldots\ldots x_{n1}$$
$$y_1 \ldots y_{k+1} \ldots\ldots y_{n2},$$

where 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:

$$CMap_X = \begin{bmatrix} x'_{11}x'_{12}\ldots\ldots x'_{1n} \\ x'_{21}x'_{22}\ldots\ldots x'_{2n} \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ x'_{n1}x'_{n2}\ldots\ldots x'_{nn} \end{bmatrix}$$
$$CMap_Y = \begin{bmatrix} y'_{11}y'_{12}\ldots\ldots y'_{1n} \\ y'_{21}y'_{22}\ldots\ldots y'_{2n} \\ \ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots \\ y'_{n1}y'_{n2}\ldots\ldots y'_{nn} \end{bmatrix} \quad (1.6)$$

x'_{ij} is the contact probability score between amino acid x'_i and y'_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

$CMap_Y$:

$$CMap_{XY} = CMap_X \times CMap_Y$$
$$= \begin{bmatrix} xy'_{11}xy'_{12}.....xy'_{1n} \\ xy'_{21}xy'_{22}.....xy'_{2n} \\ ..... \\ ..... \\ xy'_{n1}xy'_{n2}.....xy'_{nn} \end{bmatrix} \tag{1.7}$$

xy'_{ii} 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

$$\begin{aligned} &CMscore(X,Y) = \tfrac{1}{n^2}\sum_{i=1}^{n} CMap_{XY}(i,i) \\ &= \tfrac{1}{n^2}\sum_{i=1}^{n} xy'_{ii} = \tfrac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n} x'_{ij}y'_{ji} \end{aligned} \tag{1.8}$$

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

$$d(X,Y) = 1 - \frac{W_4 Optscore(X,Y)}{min\{n_1,n_2\}} - W_5 CMscore(X,Y) \tag{1.9}$$

where $W_4 + W_5 = 1$. The weights W4 and W5 are used to control the influence of sequences X and Y.

1.2.3 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):

$$d(W,Z) = \frac{d(W,X) \times Num(X) + d(W,Y) \times Num(Y)}{Num(X) + Num(Y)} \tag{1.10}$$

In which Num(X) is the number of leafs 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

$$P'_{XY} = \frac{1}{wN}((w_X + w_Y)P_{XY} + \sum_{Z \in S, Z \neq X,Y} w_Z P_{XZ} P_{ZY}) \tag{1.11}$$

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 wN is the sum of sequence weights in dataset S.

1.2.4 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 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].

1.3 Results and discussion

1.3.1 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, MSAProbs 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 se-quence 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 w_1 for amino acid similarity and weight w_2 for secondary structure information on BAliBASE 3.0 data sets. MSACompro worked wholly the best if the weight w_1 for amino acid similarity and the weight w_2 for secondary structure information were 0.4 and 0.5, respectively. Since the sum of w_1, w_2 and w_c is 1, we can deduce that wc is 0.1 if w_1 and w_2 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 w_1, w_2, and w_3 at their optimum values (0.4, 0.5, 0.1), or excluding that information by setting both w_2 and w_3 as 0. Evaluation results on BAliBASE 3.0 database were found to improve the most when w_c is 0.9 by integrating both secondary structure and relevant solvent accessibility informa-

tion. 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, comprehensive study on the effect of predicted structural information on the alignment accuracy. Consequently, the weightw_1, w_2, w_3 and w_c 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 1.1, 1.2, and 1.3** 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 for

Table 1.1: **Total SP scores on the full-length BAliBASE 3.0 subsets.** Bold denotes the highest scores. MSACompro yielded the highest SP scores on all the subsets except RV40. On some datasets such as RV11 and RV30, the improvement is substantial.

MSA tools	RV11	RV12	RV20	RV30	RV40	RV50
MSACompro	**73.14**	**94.84**	**93.30**	**87.16**	92.11	**91.41**
Clustalw	50.06	86.44	85.16	69.76	78.93	74.24
DIALIGN-TX	51.52	89.18	87.87	73.64	83.64	82.28
FSA	50.28	92.38	86.7	66.27	85.87	78.21
MAFFT	55.13	88.82	89.33	79.08	87.55	84.69
MSAProbs	68.18	94.65	92.81	83.19	**92.47**	90.76
MUSCLE	57.16	91.54	88.91	78.24	86.49	83.52
Opal	66.18	93.70	90.39	80.18	76.25	87.36
POA	37.96	83.19	85.28	69.18	78.22	71.49
Probalign	69.51	94.64	92.57	82.03	92.19	88.86
ProbCons	66.97	94.12	91.67	81.28	90.34	89.41
T-coffee	66.77	94.08	91.61	80.57	89.96	89.43

Table 1.2: **Total TC scores on the full-length BAliBASE 3.0 subsets.** Bold denotes the highest scores. MSACompro yielded the highest TC scores on all the 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	86.5	46.44	57.63	62.18	60.75
MUSCLE	31.79	80.39	35	38.6	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	53.6	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

Table 1.3: **Total TC scores on the full-length BAliBASE 3.0 subsets.** Bold denotes the highest scores. MSACompro yielded the highest TC scores on all the 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

multiple sequence alignment 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 1.4** 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 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 avert aligning unmatched regions, especially when the sequence similarity is unrecognizable.

Table 1.4: **Overall mean SP and TC scores on the SABmark 1.65.** Bold denotes the highest scores. The improvement of SP and TC scores on this data set is substantial.

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

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 OX-BENCH can be considered similar to BAliBASE core blocks. The mean SP and TC scores over the whole database are shown in **Table 1.5**. 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 1.6 and 1.7** 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 true experimental structures as input

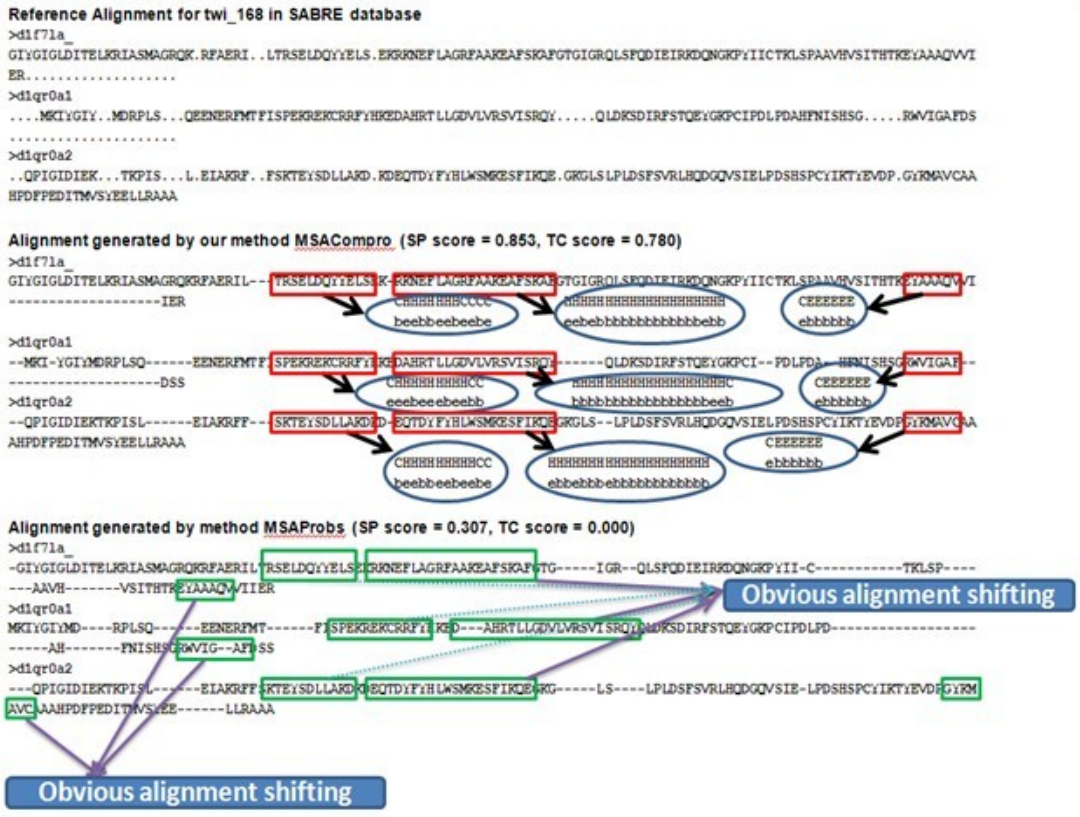


Figure 1.1: **an example in SABRE database comparing the alignments generated by our method and MSAProbs.**The reference alignment and resulting alignments generated by both methods are respectively shown in the figure. The correct alignment regions significantly improved by our MSACompro after taking 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.

Table 1.5: **Overall mean SP and TC scores on the OXBENCH.** Bold denotes the highest scores. The improvement of SP and TC scores on this data set is substantial.

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

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

Table 1.6: **Total SP scores of the tools which use the structural information on BAliBASE 3.0 subsets, SABmark data sets and OXBENCH data sets.** Bold denotes the highest scores.

MSA tools	RV11	RV12	RV20	RV30	RV40	RV50	BAli	SABmark	OX
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 1.7: **Total TC scores of the tools which use the structural information on BAliBASE 3.0 subsets, SABmark data sets and OXBENCH data sets.**Bold denotes the highest scores.

MSA tools	RV11	RV12	RV20	RV30	RV40	RV50	BAli	SABmark	OX
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**

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 1.8**. 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 MSA-Compro are mostly statistically lower than the other two alignment methods (PROMALS or PROM-ALS3D) 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,

Table 1.8: **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.** 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.

MSA tools / Score Type	Whole BAliBASE	SABmark	OXBENCH
Clustalw / SP score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
Clustalw / TC score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
DIALIGN-TX/ SP score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
DIALIGN-TX/ TC score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
FSA / SP score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
FSA / TC score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
MAFFET / SP score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
MAFFET / TC score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
MSAProbs / SP score	2.931×10^{-3}	2.2×10^{-16}	2.2×10^{-16}
MSAProbs / TC score	0.4839	2.2×10^{-16}	2.2×10^{-16}
MUSCLE / SP score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
MUSCLE / TC score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
Opal / SP score	3.384×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
Opal / TC score	2.15×10^{-14}	2.2×10^{-16}	2.2×10^{-16}
POA / SP score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
POA / TC score	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
Probalign / SP score	2.87×10^{-6}	2.2×10^{-16}	2.2×10^{-16}
Probalign / TC score	4.158×10^{-3}	2.2×10^{-16}	2.2×10^{-16}
ProbCons / SP score	2.16×10^{-15}	2.2×10^{-16}	2.2×10^{-16}
ProbCons / TC score	6.817×10^{-7}	2.2×10^{-16}	2.2×10^{-16}
T-coffee / SP score	1.225×10^{-14}	2.2×10^{-16}	2.2×10^{-16}
T-coffee / TC score	4.503×10^{-8}	2.2×10^{-16}	2.2×10^{-16}
MUMMALS / SP score	6.191×10^{-10}	2.2×10^{-16}	2.446×10^{-15}
MUMMALS / TC score	8.104×10^{-5}	2.2×10^{-16}	1.265×10^{-12}
PROMALS / SP score	$0.0116(-)$	$2.2 \times 10^{-16}(-)$	$0.0186(-)$
PROMALS / TC score	0.529	$2.2 \times 10^{-16}(-)$	$0.0274(-)$
PROMALS3D / SP score	$0.0149(-)$	$2.2 \times 10^{-16}(-)$	$2.2 \times 10^{-16}(-)$
PROMALS3D / TC score	$0.0078(-)$	$2.2 \times 10^{-16}(-)$	$2.2 \times 10^{-16}(-)$

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 because they need to generate extra information for alignment. We ran both PROMALS and MSACom-pro on the BAliBASE 3.0 database on an 4 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.

1.3.2 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 as 1, and setting the values of w_3 (weight for relative solvent accessibility) and wc (weight for contact map) to 0. The results for different w_2 values on the SABmark data sets are shown in **Table 1.9**. 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 w2 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 1.2 and 1.3** plot the SP and TC scores against weight values in **Table 1.9 and Table 1.10**, respectively.

Table 1.9: **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.**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.

W2	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SP	60.55	62.988	65.51	67.33	68.348	**68.698***	**68.465**	68.159	67.28	66.15	64.745

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 1.4**

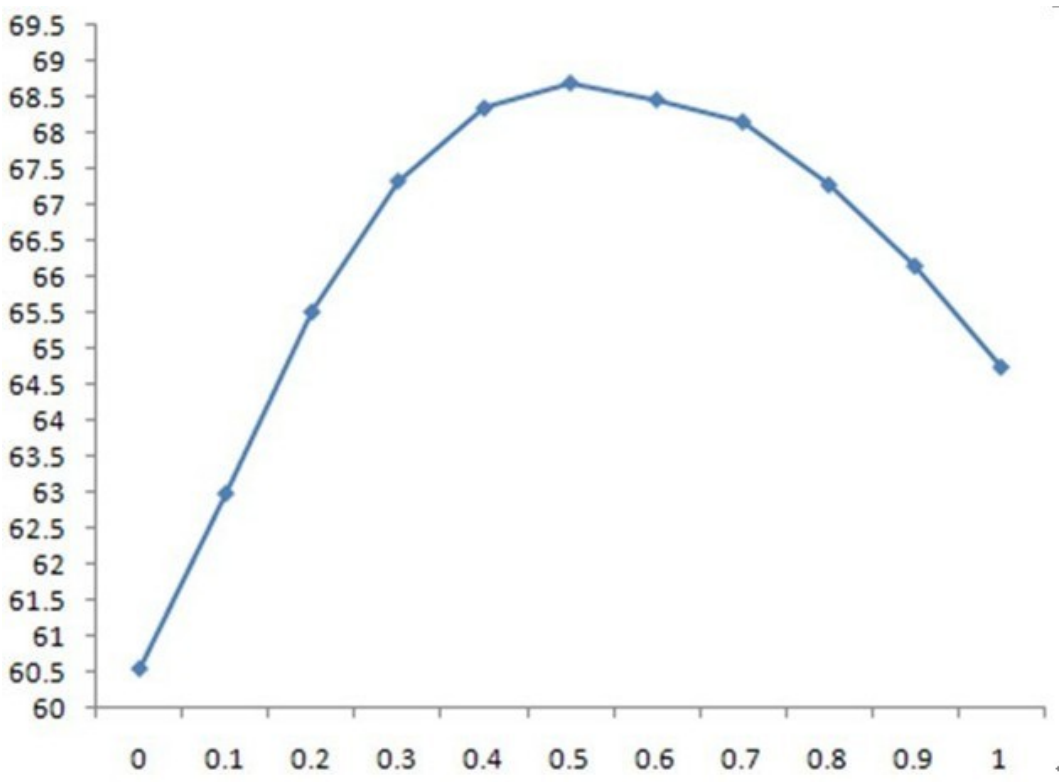


Figure 1.2: **the 2D plot of SP scores against w2 on the SABmark dataset.**

Table 1.10: **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.**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 49.10 when w2 is set to 0.5.

W2	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
TC	39.948	42.64	45.26	47.44	**48.75**	**49.005***	48.745	48.35	47.14	45.492	43.385

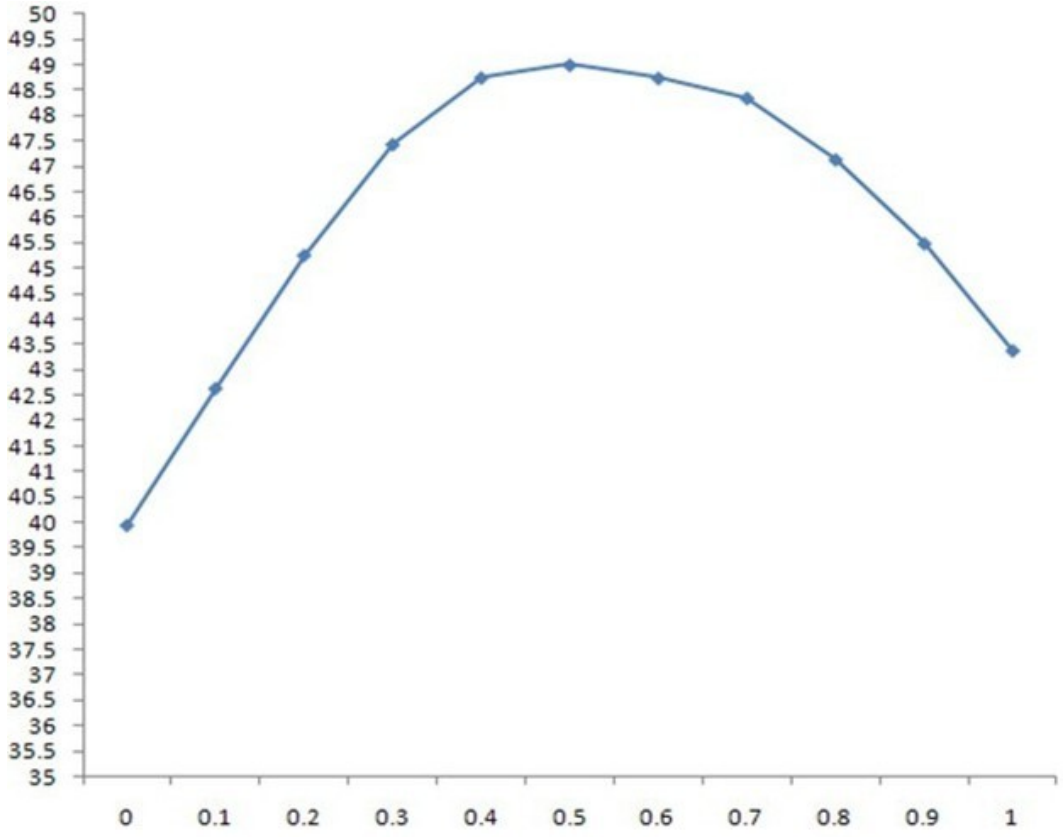


Figure 1.3: **the 2D plot of TC scores against w2 on the SABmark dataset.**

and 1.5, respectively. The highest SP and TC scores were achieved when w_3 was set to 0.5 or 0.6.

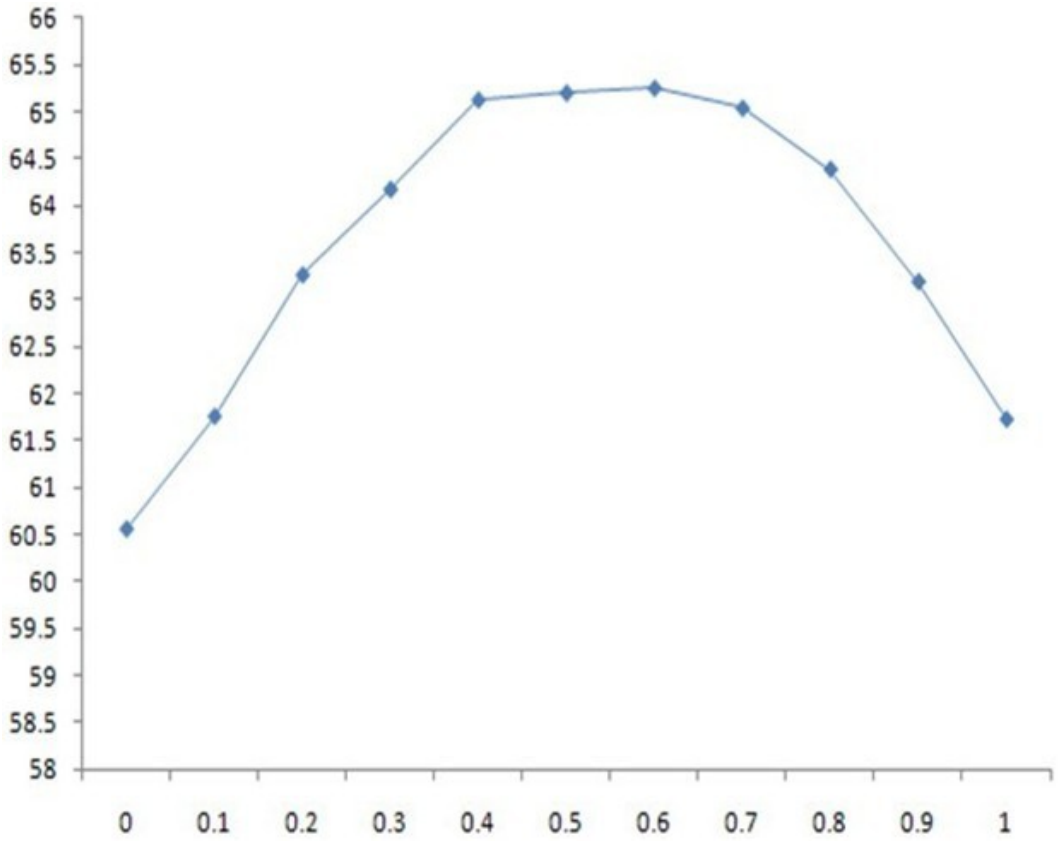


Figure 1.4: **the 2D plot of SP scores against w3 on the SABmark dataset.**

Table 1.11: **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.**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 65.25 when w3 is set to 0.6.

W3	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SP	60.55	61.75	63.26	64.17	65.12	**65.199**	**65.249***	65.037	64.388	63.188	61.72

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 wc 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,

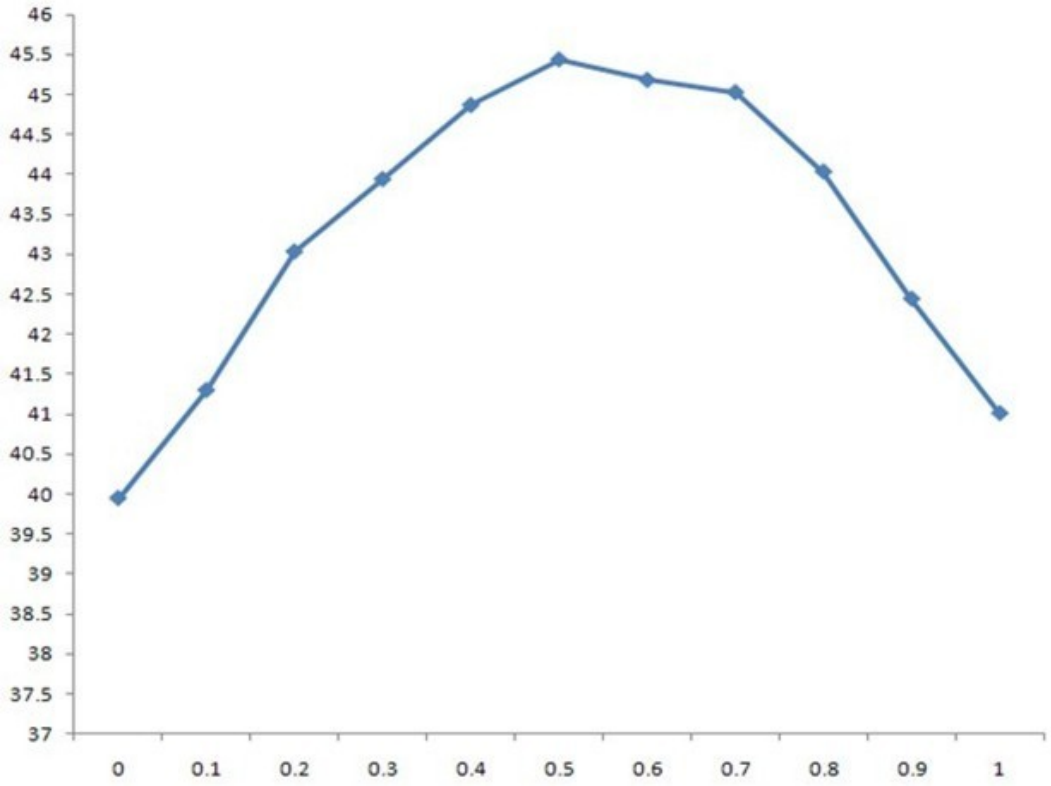


Figure 1.5: **the 2D plot of TC scores against w3 against the SABmark dataset.**

Table 1.12: **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.**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 45.25 when w3 is set to 0.6.

W3	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SP	39.948	41.3	43.04	43.94	44.87	**45.442***	**45.184**	45.03	44.038	42.447	41.012

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 wc 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.

Table 1.13: **SP scores for different weights for contact map on the BAliBASE3.0 database. Bold highlights the improved scores on each BAliBASE subset.**

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	0.946	0.945	0.946	0.945	0.946	0.945	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	0.863	0.864	0.864	**0.865**	0.817
RV40	0.928	0.926	0.926	0.924	0.923	0.924	0.924	**0.936**	**0.934**	**0.933**	0.927
RV50	0.909	0.908	**0.910**	**0.910**	**0.909**	**0.909**	**0.909**	0.907	0.907	0.908	0.886

Table 1.14: **TC scores for different weights for contact map on the BAliBASE3.0 database. Bold highlights the improved scores on each BAliBASE subset.**

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	**0.867**	**0.867**	**0.867**	0.865	0.858
RV20	0.482	0.479	0.473	0.460	0.457	0.462	0.453	0.453	0.457	0.453	0.419
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

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 combina-tions 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 to 0.5, and w_3 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 1.6 and 1.7**. We conducted similar experiments on BAliBASE 3.0 and OXBENCH and got the similar results (data not shown).

Table 1.15: **SP scores for different weight combinations (w1 - amino acid, w2 - secondary structure, w3 - solvent accessibility) on the SABmark 1.65 dataset.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.**

w2											
w1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	61.72	63.188	64.388	65.037	65.249	65.199	65.124	64.17	63.26	61.75	60.55
0.1	63.30	64.600	65.635	66.492	66.702	66.619	66.423	65.72	64.79	62.99	
0.2	64.76	66.055	67.161	67.598	68.104	67.831	67.469	66.78	65.51		
0.3	65.78	66.974	67.867	68.312	68.414	68.418	68.033	67.33			
0.4	66.42	67.531	68.251	68.743	69.016^{1}	68.920^{2}	68.3475				
0.5	66.85	67.907	68.4	68.859^{3}	68.9333	68.698					
0.6	66.84	67.911	68.544	68.560	68.465						
0.7	66.74	67.800	68.135	68.159							
0.8	66.39	67.119	67.282								
0.9	65.45	66.153									
1	64.75										

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

Table 1.16: **TC scores for different weight combinations (w1 - amino acid, w2 - secondary structure, w3 - solvent accessibility) on the SABmark 1.65 dataset.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.**

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	43.94	43.04	41.30	39.94
0.1	42.558	44.147	45.596	46.863	47.043	46.910	46.676	45.33	44.39	42.64	
0.2	43.915	45.678	47.270	47.927	48.619	48.080	47.584	47.00	45.26		
0.3	45.582	46.768	48.116	48.660	48.905	48.660	48.371	47.44			
0.4	46.104	47.340	48.473	48.889	49.508^{1}	49.159^{3}	48.754				
0.5	46.440	47.809	48.210	49.078	49.222^{2}	49.005					
0.6	46.577	47.619	48.487	48.797	48.745						
0.7	46.147	47.579	48.083	48.352							
0.8	45.714	46.898	47.142								
0.9	44.442	45.492									
1	43.385										

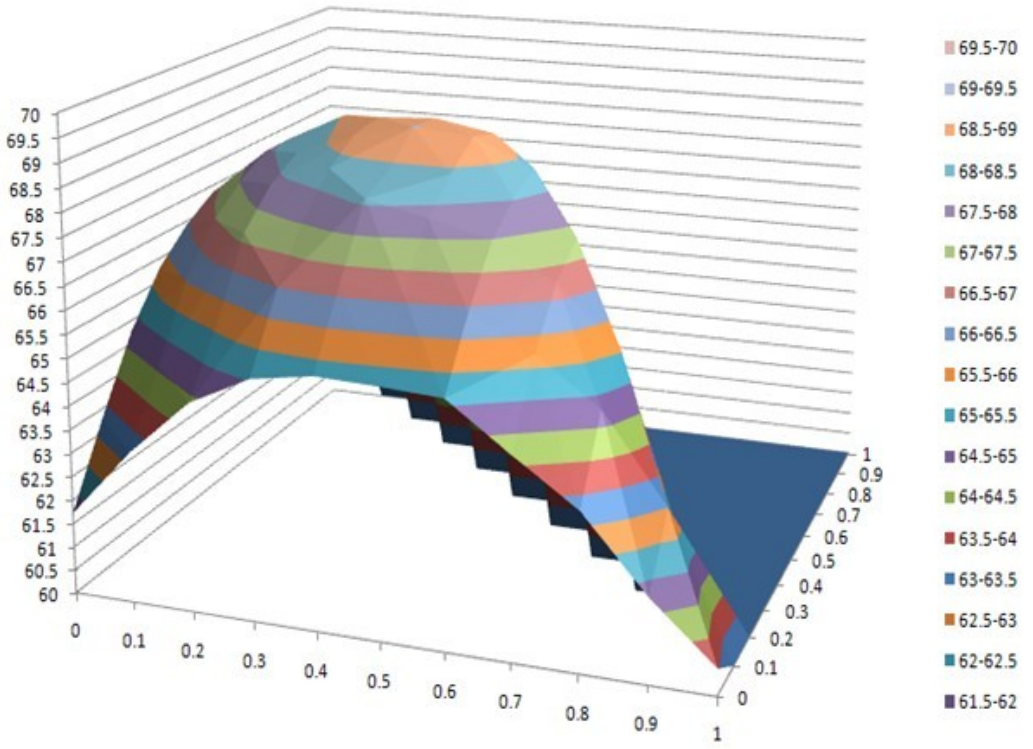


Figure 1.6: **3D plot of SP scores against secondary structure weight w2 and relative solvent accessibility weight w3.**

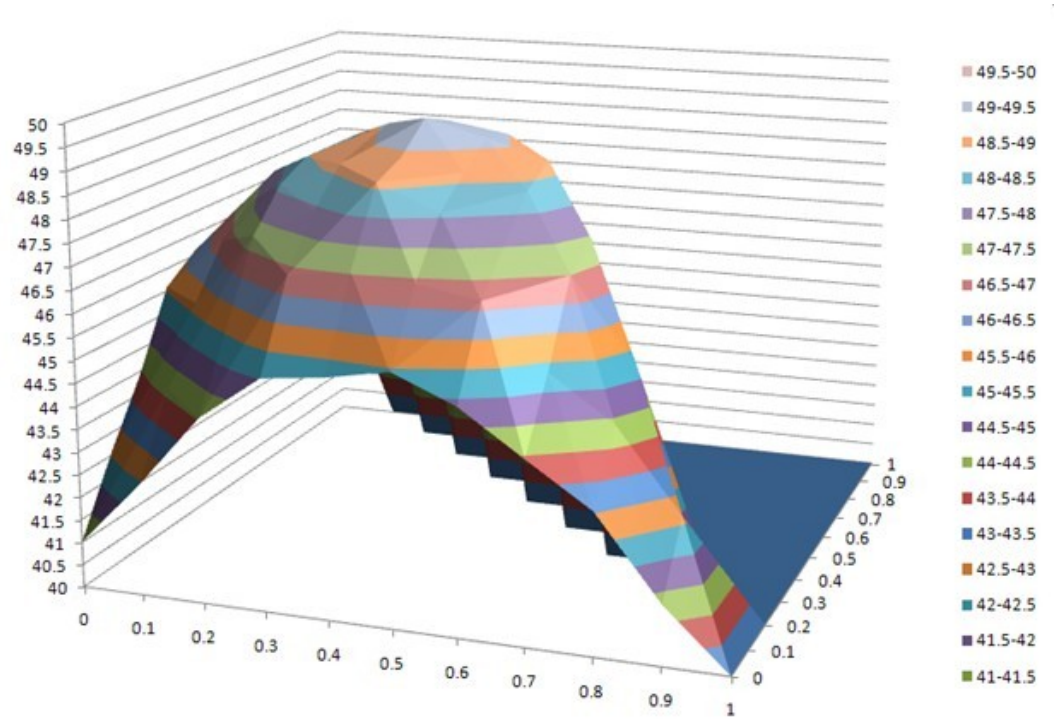


Figure 1.7: **3D plot of TC scores against secondary structure weight w2 and relative solvent accessibility weight w3.**

information, while keeping the w_1, w_2, and w_3 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 ($w_c = 0$) to maximum contact information ($w_c = 1$). The results show that the improvement caused by contact information seems not to be substantial and significant.

1.4 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.

Table 1.17: **SP scores for different contact map weight wc 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. 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	0.729	**0.730**	0.728	0.726	0.726	0.726	0.727	0.725	**0.732**	**0.731**	0.722
RV12	0.947	**0.948**	0.947	**0.949**	**0.948**	**0.948**	**0.948**	**0.948**	**0.948**	**0.948**	0.945
RV20	0.934	0.933	0.932	**0.934**	0.934	0.934	0.933	0.9328	0.9332	0.933	**0.934**
RV30	0.876	**0.877**	**0.877**	**0.876**	0.873	0.873	0.873	0.873	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	0.908	**0.914**	**0.914**	0.871

Table 1.18: **TC scores for different contact map weight wc 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. 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	0.470	**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	0.625	0.624	0.625	**0.644**	0.612
RV50	0.601	0.595	**0.601**	**0.601**	0.596	0.596	0.586	0.625	0.636	**0.634**	0.55

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 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.

Chapter 2

New profile-profile pairwise protein sequence alignment by HMM-HMM comparison

2.1 Introduction

Homology searching and sequence alignment have drawn increasing attention in bioinformatics field, since many important bioinformatics tasks such as protein structure and function prediction depends crucially on sensitivity of homology sequence searching and accuracy of the resulting sequence alignment [57, 58, 59, 60, 9]. The development of profile-sequence alignment or profile-profile alignment methods such as PSI-BLAST, HHsearch, HHsuite [61, 60, 9] over sequence-sequence alignment methods has indicated that sequence profile can help improve the accuracy of alignment. This is led by the reason that a sequence profile built by a multiple alignment of homologous sequences can provide additional information compared to mere protein sequence information, increasing the sensitivity in recognizing the conserved positions among homologous sequences.

Moreover, profile-profile alignment methods have been widely used by many pro-

tein structure prediction servers [62, 63, 64, 9]. These servers stand among the best-performing methods in template-based structure prediction [65, 60, 66]. HHsearch, one of top profile-profile alignment tools, is a software suite for detecting remote homologues of proteins and generating profile-profile alignment for given query and template protein sequences based on HMM-HMM comparison [9]. Based on HHsearch, another alignment tool HHsuite is developed to enable fast, iterative sequence searches, as well as high-quality sequence alignments [60]. Here, we present HHpacom (HMM-HMM pairwise protein sequence alignment combining structural information and inferred residue pair coupling information), which extends HHsuite to enable fast and high-quality profile-profile pairwise alignment by integrating secondary structure, solvent accessibility, torsion angle and inferred residue pair coupling information.

2.2 Method

As shown in **Figure 2.1**, the work flow of our new method is as follows. Following the basic scheme in HHsuite [60], HHpacom performs four main steps: (1) Discretize profile columns into an alphabet of 219 states; (2) Prefilter the profiles by removing sequences with coverage of template or query less than a certain percent (default is 0, can be set by users) as well as those with sequence identity less than twenty percent [67]; (3) Perform Viterbi alignment based on the secondary structure, solvent accessibility and torsion angle information of the template and query sequences, calculate E-value and probability; (4) Realign using the maximum accuracy algorithm integrating secondary structure, solvent accessibility and torsion angle information, and trace-back the alignments with the help of inferred residue coupling information. Different from HHsuite, our method applies solvent accessibility and torsion angle information to both the Viterbi-alignment and the maximum accuracy alignment, and trace-back to gain the final alignment with the help of inferred residue pair coupling

information.

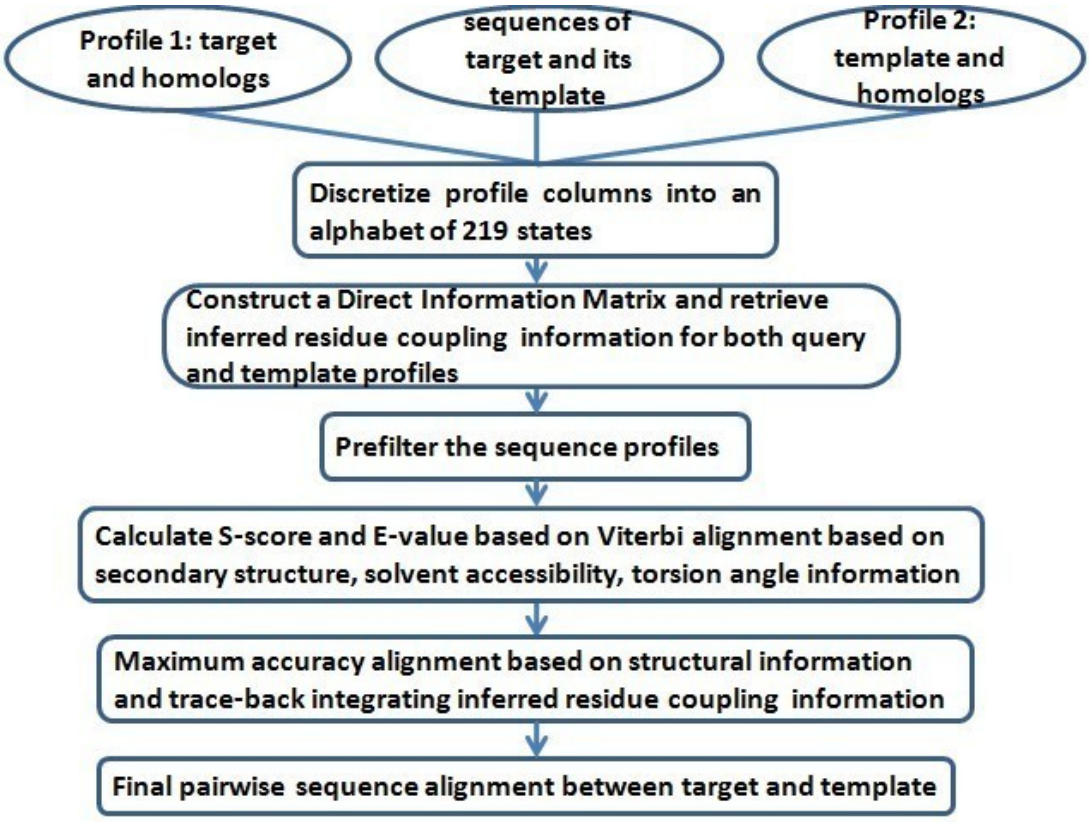


Figure 2.1: **The work flow of our new method HHpacom of profile-profile pairwise alignment.**

2.2.1 Discretize profile columns and prefilter

Refer to the discretizing and prefiltering scheme of hhsuite, both the query and template profile columns are discretized into an alphabet of 219 states, in which each one is a printable ASCII character. The column scores are first calculated for both query and template profiles, using equation (1) as follows:

$$S_{aa}(q_i, t_j) = log_2 \sum_{a=1}^{20} \frac{q_i(a)t_j(a)}{f(a)} \tag{2.1}$$

in which $q_i(a)$ and $t_j(a)$ denote the query profile at position i and the template profile at position j, respectively, and f(a) is the background frequency of residue a

($a \in 1, 2, \ldots, 20$,representing 20 types of amino acids).

Then the same prefilter process as in HHsuite is performed on both query and template profiles by removing sequences with coverage of template or query less than a certain percent (default is 0, can be set by users) as well as those with sequence identity less than twenty percent.

2.2.2 Viterbi alignment combining the structural information

As introduced in HHsearch [9],alignment between two profile HMMs is gained by maximizing the log-sum-odds score S_{LSO} as defined below:

$$S_{LSO} = \sum_{k:X_k Y_k = MM} S_{aa}(q_{i(k)}, t_{j(k)}) + logP_{tr} \tag{2.2}$$

where k denotes the number of columns that query HMM q aligned to template HMM t, and are the respective columns from q and t, P_{tr} is the product of all transition probabilities for the path through q and t.

Five matrices S_{AB} are used to calculate the log-sum-of-odds score S_{LSO} based on dynamical programming, and $AB \in \{MM, MI, IM, DG, GD\}$. They are recursively calculated as:

$$\begin{aligned} S_{MM}(i,j) &= S_{aa}(q_i, t_j) + w_{ss}S_{ss}(q_i, t_j + w_{sa}S_{sa}(q_i, t_j) + w_{tors}S_{tors}(q_i, t_j) \\ &+max \begin{cases} S_{MM}(i-1, j-1) + log[q_{i-1}(M,M)t_{j-1}(M,M)] \\ S_{MI}(i-1, j-1) + log[q_{i-1}(M,M)t_{j-1}(I,M)] \\ S_{IM}(i-1, j-1) + log[q_{i-1}(I,M)t_{j-1}(M,M)] \quad +S_{shift} \\ S_{DG}(i-1, j-1) + log[q_{i-1}(D,M)t_{j-1}(M,M)] \\ S_{GD}(i-1, j-1) + log[q_{i-1}(M,M)t_{j-1}(D,M)] \end{cases} \end{aligned} \tag{2.3}$$

$$(w_{ss}, w_{sa}, w_{tors} \in (0,1))$$

$$S_{MI}(i,j) = max \begin{cases} S_{MM}(i-1,j) + log[q_{i-1}(M,M)t_j(M,I)] \\ S_{MI}(i-1,j) + log[q_{i-1}(M,M)t_j(I,I)] \end{cases} \tag{2.4}$$

$$S_{DG}(i,j) = max \begin{cases} S_{MM}(i-1,j) + log[q_{i-1}(M,D)] \\ S_{DG}(i-1,j) + log[q_{i-1}(D,D)] \end{cases} \tag{2.5}$$

Besides, $S_{IM}(i,j)$ and $S_{GD}(i,j)$ are calculated similarly. In the formula [67], $S_{ss}(q_i, t_j)$ is the secondary structure score between column i in query HMM (q_i) and column j in template HMM (t_j), which is originally applied in HHsuite. $S_{sa}(q_i, t_j)$ is the solvent accessibility score between q_i and t_j , and $S_{tors}(q_i, t_j)$ is the torsion angle score between q_i and t_j, which are newly adopted in HHpacom. w_{ss}, w_{sa}, w_{tors} are weights for the secondary structure score, solvent accessibility score and torsion angle score respectively. S_{shift} is the score offset for match-match state. Three weights w_{ss}, w_{sa}, w_{tors} and shift score S_{shift} are set to 0.11, 0.72, 0.4 and -0.03 by default, and can be adjusted by users as well.

Secondary structure score $S_{ss}(q_i, t_j)$ is calculated the same way as in HHsuite. HHsuite is able to score a predicted secondary structure either against a predicted secondary structure or a known secondary structure [9].

HHpacom also allows scoring a predicted solvent accessibility either against a predicted solvent accessibility or a known solvent accessibility. DSSP [68] is used to parse the true solvent accessibility from the template sequence if the pdb file is known, and PSpro 2.0 [69] is used to predict the solvent accessibility from the query sequence and also the template sequence if true pdb file is not known. The solvent accessibility can be automatically parsed or predicted in HHpacom, or alternatively provided by a user. Two types of solvent accessibilities (e, b) are employed. Suppose query sequence X and template sequence Y are respectively denoted as X =(x_1,x_2,. . . ,x_{n1})

and Y =(y_1,y_2,. . . ,y_{n2}), where n_1 is the length of query sequence X as well as the query HMM, and n_2 is the length of template sequence Y as well as the template HMM. x_i is the i-th amino acid in sequence X, and y_j is the j-th amino acid in sequence Y. The corresponding predicted (by PSpro) or true (by DSSP) solvent accessibility states of x_i and y_j are $sa(x_i)$ and $sa(y_j)$. Accordingly, the solvent accessibility score $S_{sa}(q_i, t_j)$ is defined as:

$$S_{sa}(q_i, t_j) = \delta(sa(x_i), sa(y_j)) \tag{2.6}$$

The score is gained by kronecker-delta $\delta(a, b), whichequals 1 if a = b, 0 otherwise.$

Similarly, the torsion angles including both phi angle (ϕ)and psi angle (ψ) can be automatically predicted by SPINE-X [70, 71], or also provided by a user. The range of both ϕ and ψ is (-180,180). Given the query X and template Y, the predicted phi angle and psi angle of the i-th residue x_i in the query are denoted as $\phi(x_i)$ and $\psi(x_i)$, and those of in the template are and . Consequently, we came up with the formula (2.7) to calculate the torsion angle score $S_{tors}(q_i, t_j)$:

$$S_{tors}(q_i, t_j) = 1 - \frac{\sqrt{0.5 * [(\phi(x_i) - \phi(y_j))^2 + (\psi(x_i) - \psi(y_j))^2]}}{180} \tag{2.7}$$

2.2.3 Re-align by maximum accuracy alignment combining the structural information

It is found that maximum accuracy (MAC) algorithm can generally create a more accurate alignment than the Viterbi algorithm, while the latter can generate better scores, E-values and probabilities [60, 72]. Consequently, Viterbi algorithm is applied to compute E-values and scores, and MAC algorithm is chosen to achieve the final HMM-HMM pairwise alignment in HHpacom by default. Specifically speaking, re-align with MAC is chosen by default in HHpacom. If a user prefers not to re-align

by MAC, -norealign can be used in the command.

The maximum accuracy alignment, employed in [60, 72], creates the local alignment that maximizes the sum of probabilities for each residue pair to be correctly aligned minus a penalty (mact): $\sum i, j \in alignment[P({q_i}^M \sim {t_j})^M) - mact] \rightarrow max$, where $P({q_i}^M \sim {t_j}^M)$ represents the posterior probability of match state i in HMM q aligned to match state j in HMM t. With the parameter mact, users can control the alignment greediness, from nearly global, long (mact close to 0) to very precise, short alignments. Parameter mact is set to be 0.3501 in HHpacom, as in HHsuite. To find the best MAC alignment path, an optimal sub-alignment score matrix AS is calculated recursively using the posterior probability $P({q_i}^M \sim {t_j}^M)$ as substitution scores:

$$AS(i,j) = max \begin{cases} P({q_i}^M \sim {t_j}^M) - mact \\ AS(i-1, j-1) + P({q_i}^M \sim {t_j}^M) - mact \\ AS(i-1, j) - 0.5 * mact \\ AS(i, j-1) - 0.5 * mact \end{cases} \tag{2.8}$$

Here, Forward-Backward algorithm in local or global mode is applied to calculate the posterior probabilities $P({q_i}^M \sim {t_j}^M)$. Firstly, referring to [72], a newly introduced Pure Column Score $PS_{aa}(q_i, t_j)$ is defined as:

$$log_2 PS_{aa}(q_i, t_j) = S_{aa}(q_i, t_j) = log_2 \sum_{i=1)}^{20} \frac{q_i(a) t_j(a)}{f(a)} \tag{2.9}$$

Similar to HHsuite, we introduce Forward partition function $F_{MM}(i, j)$ and Backward partition function $B_{MM}(i, j)$, so that the posterior probability for pair state $({q_i}^M, {t_j}^M)$ to be part of an alignment between HMM q and HMM t is calculated as formula (2.10):

$$P({q_i}^M \sim {t_j}^M) = \frac{F_{MM}(i,j) B_{MM}(i,j)}{1 + \sum_{i,j} F_{MM}(i,j)} \tag{2.10}$$

However, different from HHsuite, we also integrate solvent accessibility and torsion angle information in the process of calculating both Forward and Backward partition function.

Five dynamic programming matrices F_{AB} are used to compute the Forward partition function F_{MM}, and $AB \in \{MM, MI, IM, DG, GD\}$. We initialize the top row and left column of the F_{MM} matrix to 0, and fill all the matrices recursively:

$$
\begin{aligned}
&F_{MM}(i,j) = PS_{aa}(q_i,t_j) * 2^{w_{ss}S_{ss}(q_i,t_j)} * 2^{w_{sa}S_{sa}(q_i,t_j)} * 2^{w_{tors}S_{tors}(q_i,t_j)} \\
&(pmin \\
&+F_{MM}(i-1,j-1)q_{i-1}(M,M)t_{j-1}(M,M) \\
&+F_{MI}(i-1,j-1)q_{i-1}(M,M)t_{j-1}(I,M) \\
&+F_{IM}(i-1,j-1)q_{i-1}(I,M)t_{j-1}(M,M) \\
&+F_{MM}(i-1,j-1)q_{i-1}(M,M)t_{j-1}(M,M) \\
&+F_{DG}(i-1,j-1)q_{i-1}(D,M)t_{j-1}(M,M) \\
&+F_{GD}(i-1,j-1)q_{i-1}(M,M)t_{j-1}(D,M) \\
&)
\end{aligned}
\tag{2.11}
$$

$$
\begin{aligned}
&F_{MI}(i,j) = F_{MM}(i-1,j)q_{i-1}(M,M)t_j(M,I) + \\
&F_{MI}(i-1,j)q_{i-1}(M,M)t_j(I,I)
\end{aligned}
$$

$$
\begin{aligned}
&f_{DG}(i,j) = F_{MM}(i-1,j)q_{i-1}(M,M)t_j(M,D) + \\
&f_{DG}(i-1,j)q_{i-1}(D,D)
\end{aligned}
$$

where pmin is 0 if HHpacom is in global alignment mode, 1 if in local alignment mode. In addition, $F_{IM}(i,j)$ and $F_{GD}(i,j)$ are calculated in a similar way. The calculation scheme of solvent accessibility score $S_{sa}(q_i,t_j)$ and torsion angle score $S_{tors}(q_i,t_j)$ is already introduced in Viterbi algorithm.

In analogy to Forward partition function, Backward algorithm recursively computes Backward matrix B_{MM} from the bottom:

$$
\begin{aligned}
&B_{MM}(i,j) = \\
&pmin \\
&+B_{MM}(i+1,j+1)PS_{aa}(q_{i+1},t_{j+1}) * 2^{w_{ss}S_{ss}(q_{i+1},t_{j+1})} * 2^{w_{sa}S_{sa}(q_{i+1},t_{j+1})} * 2^{w_{tors}S_{tors}(q_{i+1},t_{j+1})} \\
&*q_i(M,M)t_j(M,M) \\
&+B_{GD}(i,j+1)t_j(M,D) \\
&+B_{IM}(i,j+1)q_i(M,I)t_j(M,M) \\
&+B_{DG}(i+1,j)q_i(M,D) \\
&+B_{MI}(i+1,j)q_i(M,M)t_j(M,I)
\end{aligned}
$$

$$
\begin{aligned}
&B_{MI}(i,j) = B_{MM}(i+1,j+1)PS_{aa}(q_{i+1},t_{j+1}) * 2^{w_{ss}S_{ss}(q_{i+1},t_{j+1})} * 2^{w_{sa}S_{sa}(q_{i+1},t_{j+1})}2^{w_{tors}S_{tors}(q_{i+1},t_{j+1})} \\
&*q_i(M,M)t_j(I,M) + B_{MI}(i+1,j)q_i(M,M)t_j(I,I)
\end{aligned}
$$

$$
\begin{aligned}
&B_{DG}(i,j) = B_{MM}(i+1,j+1)PS_{aa}(q_{i+1},t_{j+1}) * 2^{w_{ss}S_{ss}(q_{i+1},t_{j+1})} * 2^{w_{sa}S_{sa}(q_{i+1},t_{j+1})}2^{w_{tors}S_{tors}(q_{i+1},t_{j+1})} \\
&*q_i(M,M)t_j(M,M) + B_{DG}(i+1,j)q_i(D,D)
\end{aligned}
\tag{2.12}
$$

and similar calculation scheme for $B_{IM}(i,j)$ and $B_{GD}(i,j)$.

2.2.4 Trace-back maximum accuracy alignment based on inferred residue coupling information

Evolutionary sequence variation, namely inferred residue coupling information, has recently been employed to help improve the protein tertiary structure prediction [73, 74]. However, to our best knowledge, applying it to profile-profile pairwise alignment is still novel.

The Evolutionary Coupling (EC) stands for the correlation between two positions

or columns in a given multiple protein sequence alignment, in another word, a given protein profile. For a multiple sequence alignment/profile, an EC score matrix can represent the correlated scores of position pairs. The calculation of EC scores are based on the frequencies of amino acids in single columns and paired columns in a given profile, referring to [74]. Suppose a profile can be denoted as a N X L matrix $\{X_{im}\}$, where each row represents one of $m = 1, \ldots, N$ proteins and each column represents one of $i = 1, \ldots, L$ sequence positions. Each matrix element X_{im} can be either one of the twenty amino acid types or the gap, namely, it has q = 21 different values.

First, a weighting scheme is come up for the m-th sequence in the profile:

$$w_m = \sum_{n=1)}^{N} U(\delta(X_{im}, X_{in} - \theta L) \tag{2.13}$$

with U denoting the unit step function, $\theta(0 < \theta \leq 1)$ as sequence similarity threshold. So, the effective number of sequences in the profile after weighting can be $N_{ef} = \sum_{m=1}^{N} \frac{1}{w_m}$. θ is set as 0.28 in HHpacom.

Then, the frequency of amino acid type A in column i of the profile is defined as:

$$F_i(A) = \frac{1}{\lambda + N_{ef}}(\frac{\lambda}{q^2} + \sum_{m=1}^{N} \frac{1}{w_m}\delta(X_{im}, A)) \tag{2.14}$$

where kronecker-delta $\delta(a, b)$ equals 1 if a=b, 0 otherwise. Besides, λ is a pseudo-count variable, and is set as 0.5 as suggested in [74]. Similarly, the frequency of a amino acid pair A and B respectively in column i and j can be defined as:

$$F_{ij}(A, B) = \frac{1}{\lambda + N_{ef}}(\frac{\lambda}{q} + \sum_{m=1}^{N} \frac{1}{w_m}\delta(X_{im}, A)\delta(X_{jm}, B)) \tag{2.15}$$

Empirically, the frequency distribution of A and B between columns i and j is independent, namely, $F_{ij}(A, B) - F_i(A)F_j(B) \approx 0$.

Due to the computational efficiency in the empirical implementation, the Evolu-

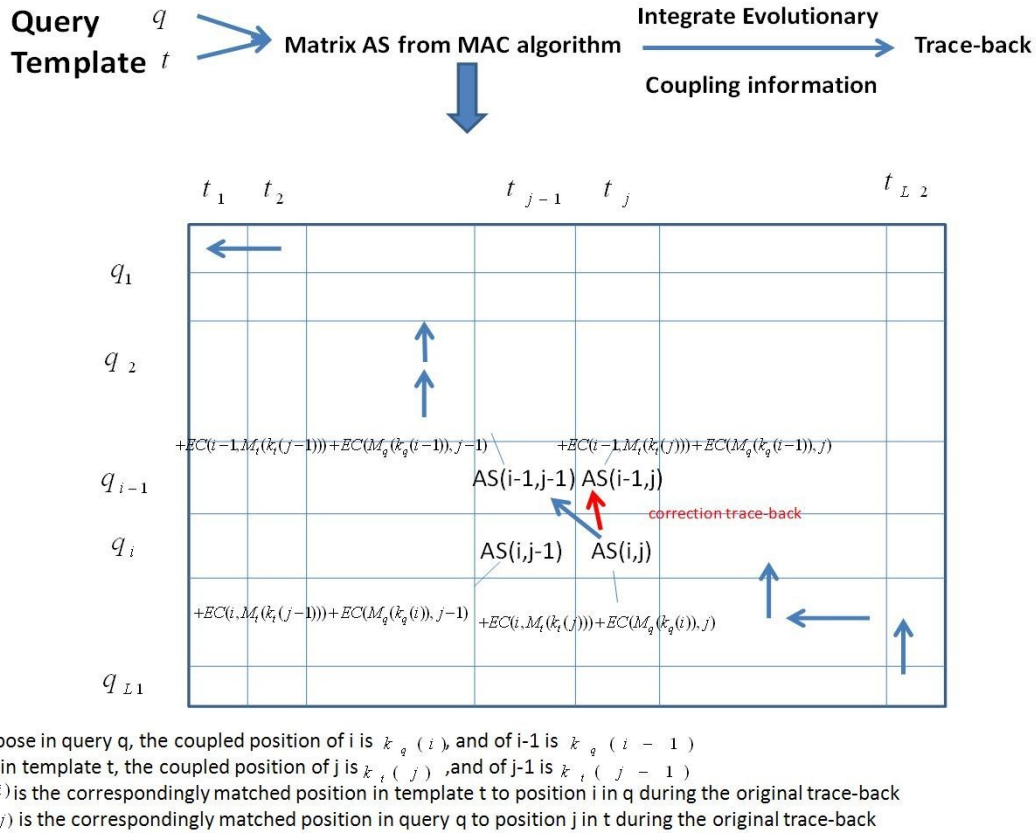
tionary Coupling matrix in HHpacom is computed through the mutual information (MI) instead of the direct information (DI) based on the global probability model as below [74]:

$$EC_{ij} = MI_{ij} = \sum_{X_i, X_j=1}^{q} F_{ij}(X_i, X_j) ln \frac{F_{ij}(X_i, X_j)}{F_i(X_i) F_j(X_j)} \tag{2.16}$$

With higher EC values corresponding to a stronger correlation between two columns in the given profile.

Based on the calculated EC value matrices for both the query and template profiles, top highly correlated paired positions (and the gaps between positions need to be more than five residues) with higher EC values for each profile are selected. The inferred evolutionary residue coupling information is then applied to check the counterpart pairs during the process of trace-backing through the sub-alignment score matrix AS from MAC algorithm. An application example is shown in **Figure 2.2**. Specifically, suppose in query q, the coupled positions of i and i-1 are $k_q(i)$ and $k_q(i-1)$, respectively, and in analogy, in template t, the coupled positions of j and j-1 are $k_t(j)$ and $k_t(j-1)$. Moreover, $M_q(i)$ denotes the correspondingly matched position in template t to position i in query q when tracing back the original AS matrix, $M_t(j)$ denotes the correspondingly matched position in query q to position j in template t when tracing back the original AS matrix, and w_{ec} is the weight for the evolutionary information. Then, after integrating the inferred evolutionary information, the modified AS scores for column pairs (i, j), (i, j-1), (i-1, j-1), and (i-1, j) are as below:

$$AS^{'}(i,j) = AS(i,j) + w_{ec}(EC(i, M_t(k_t(j))) + EC(M_q(k_q(i)), j))$$
$$AS^{'}(i,j-1) = AS(i,j-1) + w_{ec}(EC(i, M_t(k_t(j-1))) + EC(M_q(k_q(i)), j-1))$$
$$AS^{'}(i-1,j-1) = AS(i-1,j-1) + w_{ec}(EC(i-1, M_t(k_t(j-1))) + EC(M_q(k_q(i-1)), j-1))$$
$$AS^{'}(i-1,j) = AS(i-1,j) + w_{ec}(EC(i-1, M_t(k_t(j))) + EC(M_q(k_q(i-1)), j)) \quad (2.17)$$



Suppose in query q, the coupled position of i is $k_q(i)$, and of i-1 is $k_q(i-1)$
And in template t, the coupled position of j is $k_t(j)$,and of j-1 is $k_t(j-1)$
$M_q(i)$ is the correspondingly matched position in template t to position i in q during the original trace-back
$M_t(j)$ is the correspondingly matched position in query q to position j in t during the original trace-back

Figure 2.2: **Trace-back from AS by integrating the evolutionary coupling information.**

2.3 Evaluation and Results

2.3.1 Evaluation of HHpacom and other tools on CASP9 data

We compared pairwise alignment results by running our new method HHpacom, HHsearch1.2 [9] and HHsuite [60] against the testing data set in terms of two evaluation schemes: (1) we generated true or reference pairwise alignments by TMalign [75], and evaluated the alignment results from HHpacom, HHsearch1.2 and HHsuite in comparison with the true alignments 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 true alignment. The TC score is the number of correctly aligned columns in core blocks of the true alignment [76]. (2) 3D-models were obtained by MODELLER [77] based on the pairwise alignments generated by these methods. TM-scores and GDT-TS scores were calculated for the 3D-models generated by these methods to assess the quality of the predicted models based on the profile-profile alignments. GDT-TS score is the average coverage of the target sequence of the substructures with four different distance threshold 1, 2, 4, and 8 . TM-score is a variation of the Levitt-Gerstein (LG) score, also used to assess the similarity between the template and native structures. Both of their values lie between zero and one, with better templates having higher values [75].

To understand how various parameters of HHpacom affect the alignment accuracy, some experiments were carried out based on two algorithm changes: (1) combining relative solvent accessibility and torsion angle information in the process of calculating both Forward and Backward partition function; (2) Tracing-back maximum accuracy alignment based on inferred residue coupling information. To optimize the parameters, we divided 2621 pairs of which each contains a CASP9 target and its

single homolog released in CASP9 website into training and testing data sets. The training dataset is consisted of 1482 target-single template pairs generated from 60 CASP9 targets, and the testing dataset is consisted of 1138 pairs generated from 46 CASP9 targets. Firstly, we focused on the effect of solvent accessibility by testing different values of weight w_{sa} for solvent accessibility information ranging from 0 to 1, and discovered that HHpacom worked wholly the best on the training data if w_{sa} was set to 0.72. Then we focused on the effect of torsion angle by keeping w_{sa} as 0.72 and testing different values of weight w_{tors} for torsion angle information ranging from 0.1 to 1, and HHpacom were found to perform the best when w_{tors} were 0.4. Finally, we focused on the evolutionary constraint information by keeping w_{sa} and w_{tors} at their optimum values (0.72, 0.4), and found HHpacom was found to work the best when w_{ec} was 0.1. However, evolutionary constraint information did not help much to improve the alignment accuracy, yet we believe it provides a good direction in protein sequence alignment to some extent. More details are going to be discussed in the following section, A comprehensive study of the impact of the new information on the alignment accuracy. The weights w_{sa}, w_{tors} and w_{ec} are respectively set as 0.72, 0.4 and 0 in HHpacom by default, and w_{ec} can be a choice for users to adopt. All other tools were also evaluated under default parameters. Moreover, HHsearch and HHsuite were both evaluated with and without secondary structure information.

The overall mean SP and TC scores for the resulting pairwise alignment results generated by HHpacom and all the other tools against the whole testing data set are illustrated in **Table 2.1**. It is not hard to conclude integrating the new features in our method HHpacom helped improve the quality of the pairwise alignment.

MODELLER succeeded to generate models for 1127 out of 1138 resulting pairwise sequence alignments respectively by running HHsearch, HHsuite and HHpacom against the testing data set. The average TM-scores and GDT-TS scores of the 3D-models generated from the pairwise alignments by HHsearch1.2, HHsuite and our

Table 2.1: Overall mean SP and TC scores of the pairwise alignments generated by HHsearch1.2, HHsuite and HHpacom

Methods	Mean SP score	Mean TC score
HHsearch (without secondary structure information)	48.6936	48.3374
HHsearch (with secondary structure information)	50.0047	49.6520
HHsuite (without secondary structure information)	48.4668	48.1230
HHsuite (with secondary structure information)	49.7569	49.4051
HHpacom	50.3882	50.0237

new method are shown in **Table 2.2** as below. The same as the previous conclusion, HHpacom improved the alignment performance in comparison with HHsearch1.2 and HHsuite.

Table 2.2: The average TM-scores and GDT-TS scores of the 3D models generated from the pairwise alignments by running HHsearch1.2, HHsuite and HHpacom

Methods	Average TM-score	Average GDT- TS
HHsearch (without secondary structure information)	0.5268	0.4592
HHsearch (with secondary structure information)	0.5478	0.4787
HHsuite (without secondary structure information)	0.5250	0.4587
HHsuite (with secondary structure information)	0.5434	0.4757
HHpacom	0.5550	0.4828

In order to check if alignment score differences between HHpacom and the other pairwise alignment methods are statistically significant, we carried out the Wilcoxon matched-pair signed-rank test on both SP and TC scores of the methods on the testing data set. The p-values of alignment score differences between HHpacom and the other methods calculated by the Wilcoxon matched-pair signed-rank test are illustrated in **Table 2.3**. The alignment scores HHpacom are significantly higher than the other two tools without using secondary structure information yet not statistically higher than the other tools with using secondary structure information according to the significance threshold of 0.05.

Table 2.3: The statistical significance (p-values) of SP and TC score differences between HHpacom and the other tools on the testing data set

Tools	p-value of SP scores	p-value of TC scores
HHpacom – HHsearch (without ss information)	$1.078X10^{-6}$	3.414 X 10^{-7}
HHpacom – HHsearch (with ss information)	0.7538	0.8082
HHpacom – HHsuite (without ss information)	$1.724X10^{-6}$	1.515 X 10^{-9}
HHpacom –HHsuite (with ss information)	0.1535	0.1087

2.3.2 A comprehensive study of the impact of the new information on the alignment accuracy

To understand the effect of relative solvent accessibility, torsion angle and inferred residue coupling information on the accuracy of profile-profile pairwise sequence alignment, we tested their effects on alignments individually or in combination by adjusting the values of their weights.

I. Effect of solvent accessibility information

We studied the effect of solvent accessibility information by sorely adjusting the value of w_{sa} (weight for solvent accessibility information). The SP scores and TC scores of the resulting alignments on different w_{sa} by running HHpacom against the training data set are shown in **Table 2.4** as below. 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 solvent accessibility information always improves alignment accuracy over the baseline established without using solvent accessibility information (wsa = 0). The highest accuracy is achieved when wsa is set to 0.72. **Figure 2.3** and **Figure 2.4** show the 2D plot of the results. So, the best value of w_{sa} is 0.72.

II. Effect of torsion angle information

We also studied the effect of torsion angle information by sorely adjusting the value of w_{tors} (weight for torsion angle information) and keep w_{sa} as 0.72 at the best point. The SP scores and TC scores of the resulting alignments generated by HHpacom against the training data set on different w_{tors} are shown in **Table 2.5** as

Table 2.4: SP scores and TC scores on different value of wsa using HHpacom. Bold denotes the two best scores, and an extra superscript of star denotes the highest score

w_{sa}	0	0.1	0.2	0.3	0.4	0.5	0.6	0.61	0.62
SP score	40.89	41.58	41.82	41.92	42.06433	42.18476	42.23399	42.17889	42.20202
TC score	40.58	41.25	41.49	41.58	41.72758	41.85098	41.9029	41.84828	41.87121
0.63	0.64	0.65	0.66	0.67	0.68	0.69	0.7	0.71	0.72
42.19	42.22	42.22	42.23	42.23	42.24821	42.24356	42.28746	**42.29177**	42.3056*
41.86	41.89	41.89	41.90	41.90	41.91504	41.91301	41.95684	**41.96224**	41.97505*
0.73	0.74	0.75	0.76	0.77	0.78	0.79	0.8	0.9	1
42.27	42.29	42.27	42.28	42.27	42.2768	42.27121	42.24555	42.24032	42.20094
41.94	41.96	41.94	41.95	41.94	41.94268	41.93729	41.91105	41.91099	41.86649

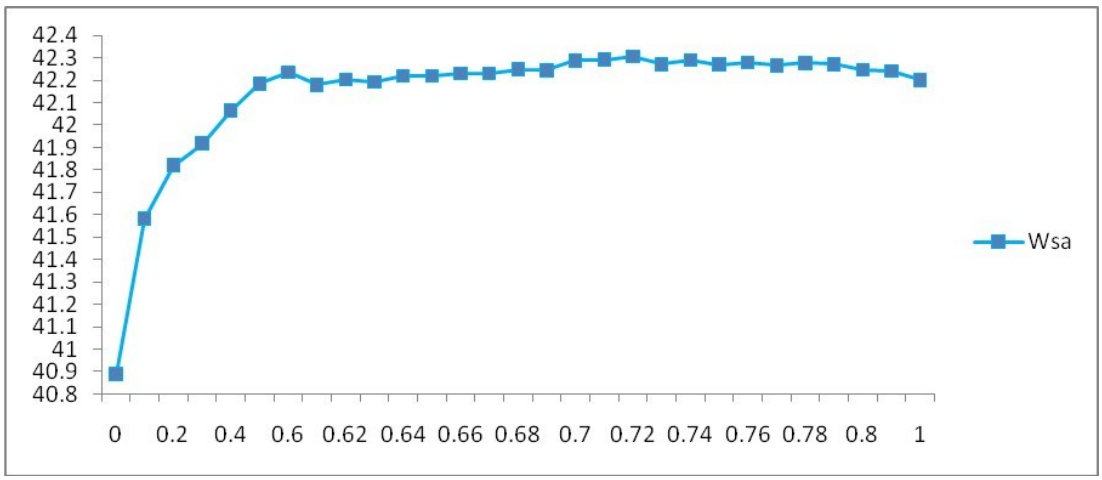


Figure 2.3: **The 2D plot of the SP scores on different w_{sa}.**

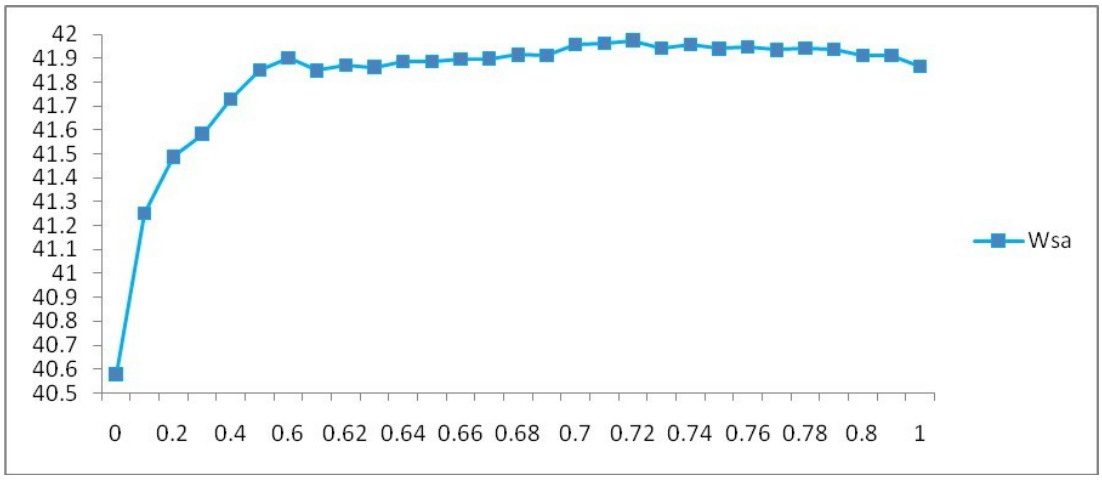


Figure 2.4: **The 2D plot of the TC scores on different w_{sa}.**

below. 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 torsion angle information also helps improve alignment accuracy. The highest accuracy is achieved when wtors is set to 0.4. **Figure 2.5** and **2.6** show the 2D plot of the results. So, the best value of w_{sa} is 0.4.

Table 2.5: SP scores and TC scores on different value of w_{tors} using HHpacom. Bold denotes the two best scores, and an extra superscript of star denotes the highest score

w_{tors}	0	0.1	0.2	0.3	0.31	0.32	0.33	0.34	0.35
SP score	42.31	42.32	42.35	42.45	42.47431	42.46999	42.47262	42.49	42.49616
TC score	41.98	41.99	42.02	42.12	42.14093	42.13621	42.14093	42.16	42.162511
0.36	0.37	0.38	0.39	0.4	0.41	0.42	0.43	0.44	0.45
42.50	42.51	42.50	42.51	42.52771*	**42.52218**	42.48584	42.49636	42.50	42.50755
42.17	42.17	42.17	42.18	42.19285*	**42.18881**	42.15172	42.16251	42.17	42.17195
0.46	0.47	0.48	0.49	0.5	0.6	0.7	0.8	0.9	1
42.51	42.50	42.50	42.50	42.50486	42.46116	42.45314	42.40175	42.46	42.40128
42.17	42.16	42.17	42.17	42.1733	42.12744	42.1207	42.07148	42.13	42.07417

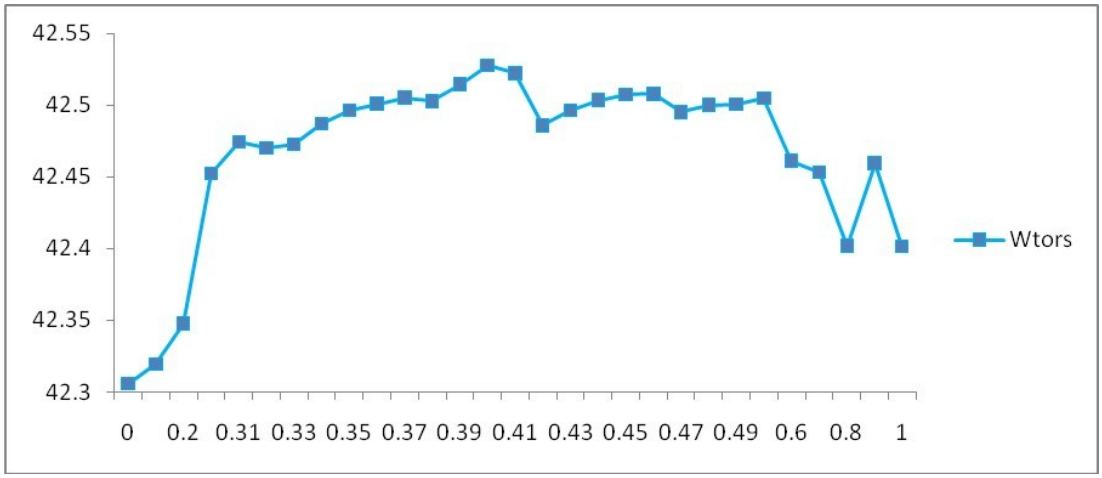


Figure 2.5: **The 2D plot of the TC scores on different w_{tors}.**

III. Case study on the effect of inferred residue coupling information

We also studied the effect of inferred residue coupling information in a similar way. HHpacom worked the best when w_{ec} was 0.1. However, evolutionary constraint information did not help much to improve the alignment accuracy on the training data

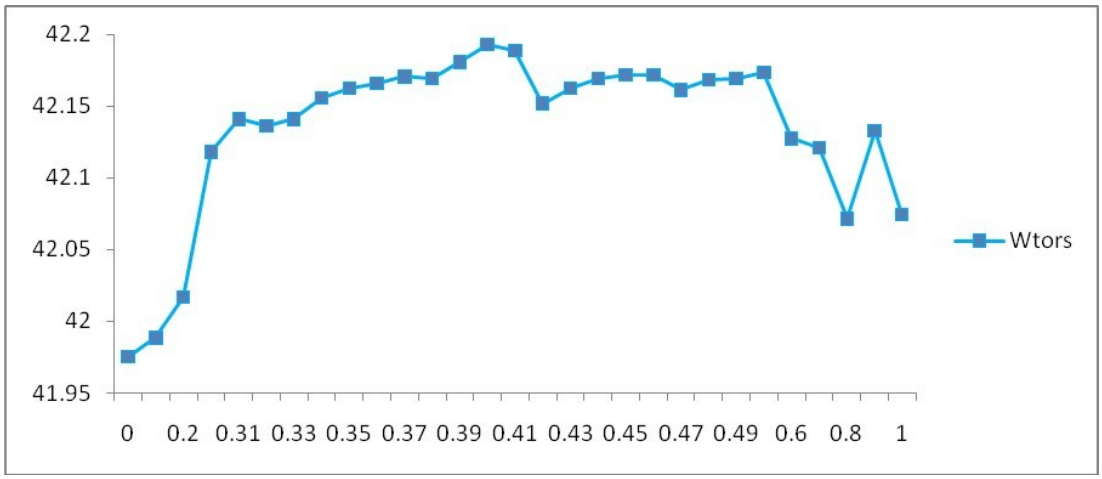
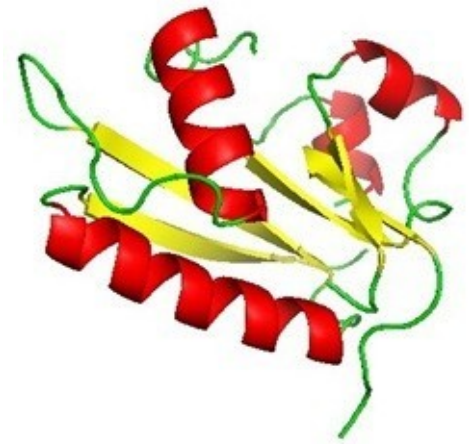


Figure 2.6: **The 2D plot of the GDT-TS scores on different w_{tors}.**

set, since the average SP score and TC score are respectively 42.5171 and 42.1821 by running HHpacom integrating the evolutionary constraint information. Specifically speaking, the alignment quality increased in 57 out of the total 1483 resulting pairwise alignments from the training data set, stayed the same in 1363, yet decreased in 61. Moreover, we also run HHpacom by setting w_{ec} as 0.1 against the testing data set, and the alignment quality increased in 59 out of the total 1138 resulting pairwise alignments, stayed the same in 1024, yet decreased in 55. After taking a deep look at the data, we discovered that integrating inferred residue coupling information mostly helped improve the alignments of the proteins which are of short lengths, typically 100 to 500 residues. **Figure 2.7** illustrated an example of a native 3D structure of T0606, a 3D model structure predicted based on the pairwise sequence alignment by running HHpacom integrating the evolutionary constraint information on the pair of T0606 and its homolog 2nooA, and a 3D model structure predicted based on the pairwise sequence alignment by running HHpacom without using the evolutionary constraint information. The evolutionary constraint information helped increasing the TM-score of the predicted model from 0.5965 to 0.6035. Consequently, in some cases, the evolutionary constraint information could help improve the quality of the pairwise alignment.



The native structure of T0606

The predicted model based on the pairwise alignment generated by HHpacom without using the evolutionary constraint information on the pair of T0606 and 2nooA

TM score of the model: 0.5965
SP score of the alignment: 0.592

The predicted model based on the pairwise alignment generated by HHpacom integrating the evolutionary constraint information on the pair of T0606 and 2nooA

TM score of the model: 0.6035
SP score of the alignment: 0.602

Figure 2.7: **An example of the case study.**

2.4 Conclusion

In this work, we designed a new method to incorporate relative solvent accessibility, torsion angle information and inferred residue pair information into profile-profile pairwise protein sequence alignment. Our experiments on the CASP9 data set showed that the method improved pairwise sequence alignment accuracy over HHsearch and HHsuite. However, the inferred residue pair information did not improve much on the CASP9 data set, yet our case study of the effect of the inferred residue pair information provided an useful view sight for the future direction in the pairwise sequence alignment.

Chapter 3

Predicting Protein Model Quality from Sequence Alignment by Support Vector Machines

3.1 Abstract

3.1.1 Background

Assessing the quality of a protein structural model is essential for protein structure prediction. Here, we developed a Support Vector Machine (SVM) method to predict the quality score (GDT-TS score) of a protein structure model from the features extracted from the sequence alignment used to generate the model.

3.1.2 Results

We developed a Support Vector Machine (SVM) model quality assessment method, taking either a query-single-template pairwise alignment or a query-multi-template alignment as input. For the pairwise alignment scheme, the input features fed into the

SVM predictor include the normalized e-value of the given alignment, the percentage of identical residue pairs in the alignment, the percentage of residues of the query aligned with those of the template, and the sum of the BLOSUM scores of all aligned residues divided by the length of the aligned positions. Similarly, for the multiple-alignment scheme, the input features include the percentage of the residues of the target sequence aligned with those in one or more templates, the percentage of aligned residues of the target sequence that are the same as that of any one template, the average BLOSUM score of aligned residues, and the average Gonnet160 score of aligned residues. A SVM regression predictor was trained on the training data to predict the GDT-TS scores of the models from the input features. The root mean square error (RMSE) and the absolute mean error (ABS) between predicted and real GDT-TS scores were calculated to evaluate the performance. A five-fold cross validation was applied to select the best parameter values based on the average RMSE and ABS on the five folds. The RMSE and ABS of the optimized SVM predictor on the testing data were close to 0.1.

3.1.3 Conclusions

The good performance of the SVM and sequence alignment based predictor indicates that integrating sequence alignment features with a SVM is effective for protein model quality assessment.

3.2 Background

The knowledge of protein three-dimensional (3D) structures is vitally important for biomedical research, such as protein function analysis, mutagenesis experiments, and rational drug design. Although the X-ray crystallography technique can determine protein 3D structures with high resolution, they are still time-consuming, expensive,

and cannot be readily applied to the proteins that cannot be successfully crystallized, including most membrane proteins. The nuclear magnetic resonance (NMR) is a powerful tool that can determine the 3D structures of membrane proteins of small and medium size in solutions [78, 79, 80], but it is also time-consuming and costly. In order to acquire the protein structural information at a large scale and in a timely manner, high-throughput, fast computational protein structure prediction methods, such as homology modeling (e.g. [81, 82]), need to be used. Since the accuracy of predicted protein structures depend on the relatedness of homologous structural templates and the correctness of sequence alignment [81], assessing the quality of protein structural models is important for controlling and analysing the quality of the predicted models.

Thus, protein model quality assessment plays a profound role in protein structure prediction and related applications [83]. Accurate quality assessment of protein models can help rank a pool of candidate models predicted for a given query protein. A number of model quality assessment methods and tools, such as ModelEvaluator [84], APOLLO [84], QMEAN [85], have been developed. These methods evaluate the quality of models based on the structural information extracted from protein models, without considering the source information (e.g., sequence alignment, homologous template structure) used to generate the models. The quality assessment methods without utilizing the source information may be considered a black box approach, while those considering the source information [86] is a white box approach [87]. Since the factors of largely determining the quality of a model such as the sequence similarity between a query protein and a homologous template structure are generally available in the template-based protein structure prediction (e.g., homology modelling and fold recognition), the white box approach can take advantage of the information to improve model quality assessment.

There are a few standard scores to assess the similarity between the model struc-

ture and the native target structure, such as TM score and GDT-TS score [88, 89] TM-score is to calculate the similarity of topologies of two protein structures. It can be exploited to quantitatively access the quality of protein structure predictions relative to native. GDT Total Score (GDT-TS) is the average percentage of residues in the model whose position is within 1.0, 2.0, 4.0, 8.0 α-carbon distance with that of their counterparts in the native structure after four optimal superimpositions [90]. Neither TM score nor GDT-TS score can directly tell the accuracy of a target-template alignment, yet given a few alignments consisted of the same target and different one or more than one templates, they can effectively measure which alignment can generate a model that is more structurally similar to the target. Consequently, if we succeed to predict these scores, our prediction method can be an effective alignment-based model selector, which plays a significant role in protein structure prediction, function prediction, and other essential bioinformatics tasks.

Here, extending from our previous model quality assessment method based on a query-single-template alignment [91], we designed and developed a support vector machine (SVM) [92] and alignment-based model quality assessment method, taking either a query-single template pairwise alignment or a query-multi template alignment as input to predict the GDT-TS score of a model generated from the input alignment. The method can be applied to select the protein models based on the query-template alignments used to generate the models in the widely used template-based protein modelling process.

3.3 Methods

Figure 3.1 shows the workflow of the SVM model quality assessment method based on the features extracted from the query-single template pairwise alignment employed to generate the model. The input features provided to the SVM predictor include the

logarithm of e-value of the given query-template alignment, the percent of identical residue pairs in aligned positions, the percent of residues of the query that are aligned with a residue in the template, and the average of BLOSUM [93] scores of all aligned residue pairs. The input feature vectors in the training data set were extracted from 245 pairwise protein sequence alignments generated for 50 CASP9 targets by PSI-BLAST [61]. The output score of each input feature vector was the real GDT-TS [94] score of the model generated from the corresponding pairwise alignment. The real GDT-TS score is the structural similarity score between a model and its corresponding native structure calculated by the TM-score program [89]. This data was used to train a SVM regression predictor equipped with a Gaussian radial basis kernel (RBF) to predict the GDT-TS scores of models from the input features. The SVM-Light software package [95] was employed to carry out the training and testing experiments. Three parameters of the SVM including the epsilon width of the regression tube (w), the margin option (c) and the gamma in the RBF kernel (g) were tuned during the training process. The root mean square error (RMSE) and the absolute mean error (ABS) between the predicted and real GDT-TS scores were used as the evaluation scheme to optimize the parameter values. Three standard cross-validation methods are commonly adopted to check the effectiveness of a predictor: independent dataset test, K-fold cross-validation, and jackknife test [96]. Here, we utilized the five-fold cross validation approach as many other SVM based prediction methods due to the computational efficiency. Specifically, many rounds of five-fold cross validations were applied to the training data to select the best parameter values of w from 0.5, 0.2, 0.1, 0.05, 0.02, and 0.01, and c from 2.0, 1.0, 0.5, 0.1, 0.05, and 0.01, and g from 0.5, 0.3, 0.2, 0.1, 0.05, 0.01, 0.005, and 0.001 in order to reduce the average ABS and RMSE on all the five folds. The set of parameter values with the lowest RMSE and ABS was selected.

Similarly, **Figure 3.2** shows the workflow of the SVM model quality assessment

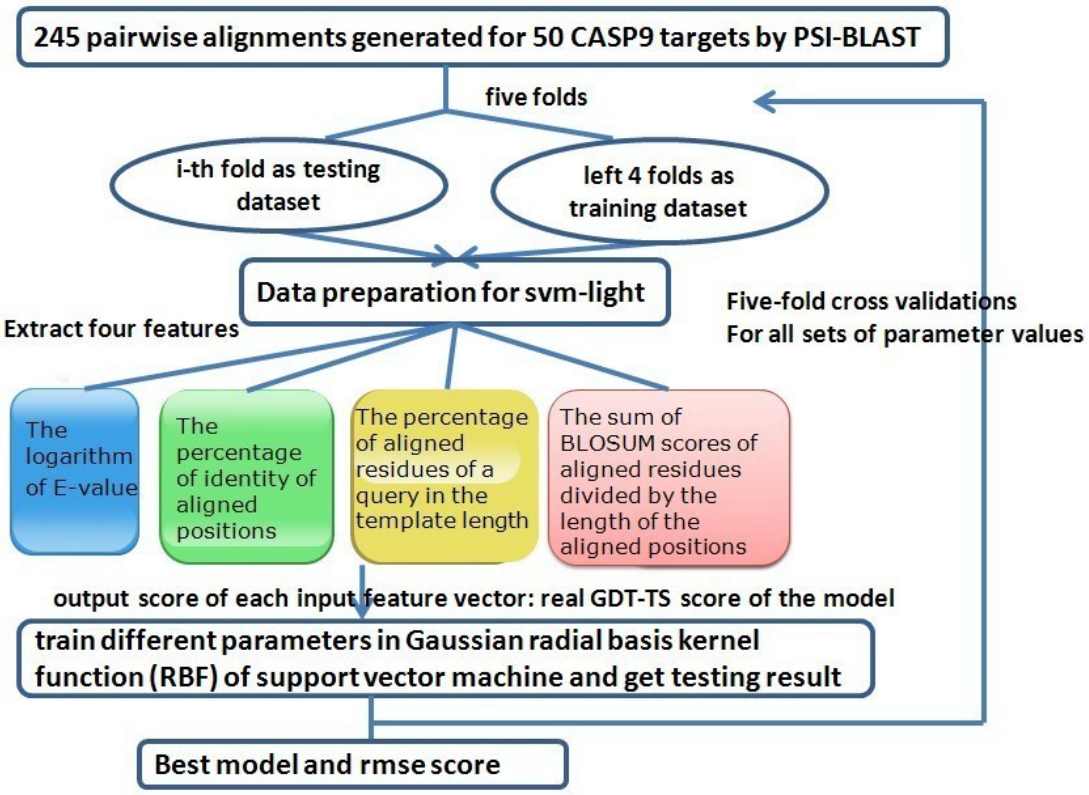


Figure 3.1: The workflow of the pairwise alignment based SVM model quality prediction method

method based on the features extracted from the query-multi template alignment employed to generate the model. Slightly different from the pairwise alignment based scheme above, the input features include the percentage of the residues of the target sequence aligned with those in one or more templates, the percentage of identical residues of the target sequence that are the same as that of any one template, the average BLOSUM score of aligned residues, and the average Gonnet160 score of aligned residues. Specifically, as for the average BLOSUM score, if a residue of the target is aligned in those from multiple templates, the pair BLOSUM score between the residue of the target and that of the template ranked higher in the alignment file (e.g., more significant) is counted. Consequently, the average BLOSUM score associated with all aligned residues of the target sequence was calculated as one feature. The average Gonnet 160 score of all aligned residues is calculated in a similar way. The input feature vectors in the training data set were extracted from 4850 multiple protein sequence alignments generated for 60 CASP9 targets by many different alignment tools, such as BLAST, PSI-BLAST [61], HHSearch [8], SAM [97], SPEM [98] and the output score of each input feature vector was the real GDT-TS score of the model generated from the corresponding multiple alignment. Many rounds of ten-fold cross validations were applied to the training data to select the best parameter values of w from 0.1, 0.08, 0.06, 0.05, 0.02, and 0.01, and g from 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.01, 0.005, and 0.001, and c from 2.0, 1.0, 0.5, 0.1, 0.05, and 0.01.

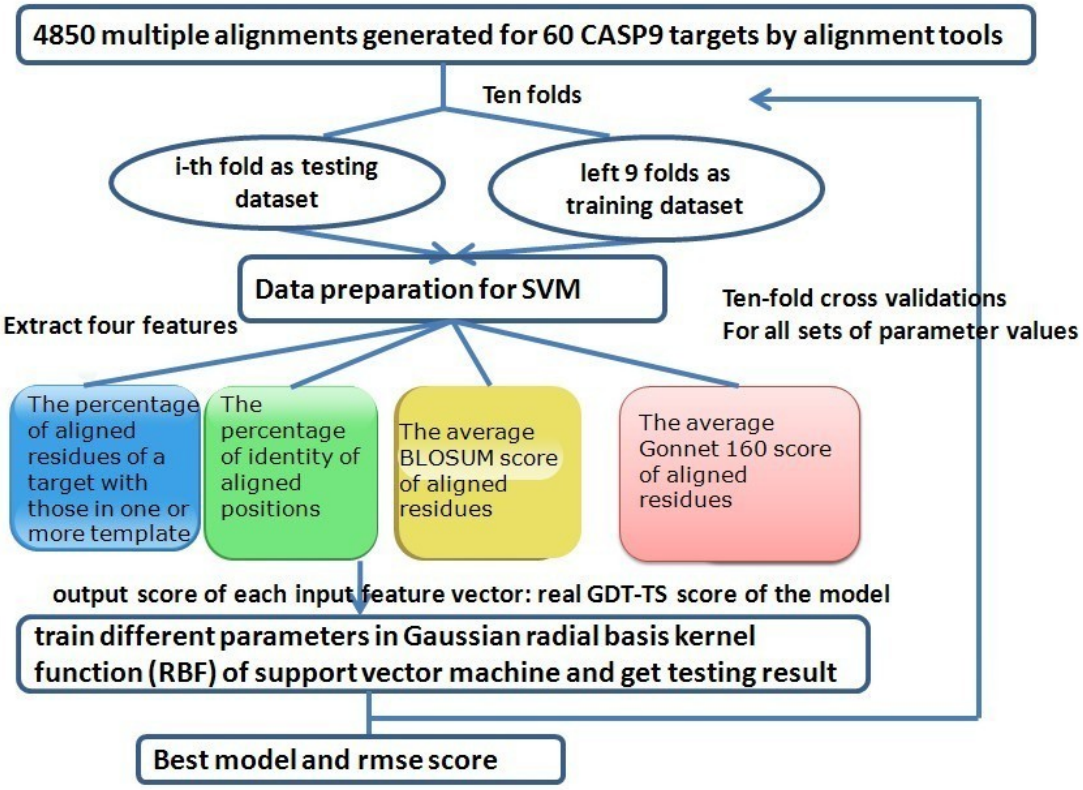


Figure 3.2: The workflow of the multiple alignment based SVM model quality prediction method

3.4 Results

3.4.1 Evaluation of the pairwise alignment based SVM model quality assessment method

The global average RMSE and ABS of the SVM trained with the best set of parameter values (w, c, g) = (0.02, 1.0, 0.5) on the five-fold training data set were 0.083 and 0.061, respectively. The trained pairwise alignment based SVM predictor was applied to predict the GDT-TS scores of models of 46 CASP9 targets not used in training from the input features extracted from the corresponding 225 PSI-BLAST alignments. The RMSE and ABS were respectively 0.098 and 0.073, illustrating that the predicted GDT-TS scores are close to the real ones. The RMSE and ABS of the trained SVM with the best parameter set on each fold of the training data as well as the testing data set are shown in **Table 3.1**.

Table 3.1: The RMSE and ABS of the trained pairwise sequence alignment based SVM with the best parameter set on each fold of the training data as well as the testing data set

The data set	RMSE	ABS
Fold 1 of the training data	0.0868	0.0606
Fold 2 of the training data	0.0923	0.0674
Fold 3 of the training data	0.0821	0.0631
Fold 4 of the training data	0.0771	0.0557
Fold 5 of the training data	0.0783	0.0566
Test data	0.0978	0.0734

Moreover, we used the predicted model quality scores to rank the models of 46 CASP9 targets [87]. The total real GDT-TS score of the top 1 models selected by the SVM predictor for these targets was compared with that of the top 1 models selected according to the e-values (i.e. significance) of the PSI-BLAST alignments and that of the top 1 models selected by APOLLO [85], a black box quality assessment tool using a pairwise model comparison approach. The total GDT-TS score of the models selected by the SVM predictor is 20.95, which is higher than 20.10 of the pure e-value based

model selection method, as well as 19.53 of APOLLO [85]. The t-test and Wilcox-test were respectively performed in order to calculate the p-values on the scores of our SVM predictor and the e-value based model selection method, as well as the ones on the scores of our predictor and APOLLO. The p-values are illustrated in **Table 3.2**. The results suggest the SVM predictor based on pairwise alignments performed significantly better than the e-value based predictor and APOLLO according to the standard p-value threshold (i.e. 0.05). Moreover, the Pearsons correlation coefficient score between the predicted and true GDT-TS scores on the testing data set is 0.913, indicating that the predicted and true scores are highly linearly correlated. The results demonstrate that integrating alignment e-value with other features by SVM can improve the accuracy of ranking models over the nave e-value based model ranking method and a state-of-art black-box model evaluation method (i.e. APOLLO).

Table 3.2: the p-values on the scores of our SVM predictor and the e-value based model selection method, and the ones on the scores of our predictor and APOLLO based on t-test and Wilcox-test

p-value	t-test	Wilcox-test
Our predictor/ e-value based method	0.044	0.042
Our predictor/ APOLLO	0.044	0.016

3.4.2 Evaluation of the multiple alignment based SVM model quality assessment method

The global average RMSE and ABS of the SVM trained with the best set of parameter values (w, c, g) = (0.1, 2.0, 0.05) on the ten-fold training data set were 0.185 and 0.149, respectively. The trained SVM predictor was applied to predict the GDT-TS scores of models of 47 CASP9 targets generated from 3809 multiple protein sequence alignments that were not used in training. The RMSE and ABS were respectively 0.176 and 0.142. This error is higher than that of the pairwise alignment-based predictor tested on models generated from PSI-BLAST alignments alone in the previous

experiment, probably due to the higher diversity in alignments and model quality in this experiment. However, the advantage of this SVM predictor is that it can be applied to the alignments generated from any alignment methods and does not require an alignment e-value as input, which varies from one alignment method to another. The RMSE and ABS of the trained SVM predictor with the best parameter values on each fold of the training data as well as the test data set are shown in **Table 3.3**.

Table 3.3: The RMSE and ABS of the trained multiple sequence alignment based SVM with the best parameter set on each fold of the training data as well as the testing data set

The data set	RMSE	ABS
Fold 1 of the training data	0.2057	0.1678
Fold 2 of the training data	0.1516	0.1238
Fold 3 of the trainingdata	0.1746	0.1393
Fold 4 of the training data	0.1538	0.1226
Fold 5 of the training data	0.1677	0.1383
Fold 6 of the training data	0.1692	0.1348
Fold 7 of the training data	0.1900	0.1487
Fold 8 of the training data	0.2330	0.1873
Fold 9 of the training data	0.2287	0.1939
Fold 10 of the training data	0.1721	0.1377
Test data	0.1764	0.1423

We also used the predicted model quality scores to rank the models of 47 CASP9 targets in the testing data [11]. The total real GDT-TS score of the top 1 models selected by the multiple alignment based SVM predictor for these targets was compared with that of the top 1 models selected by APOLLO. The total GDT-TS score of the top models selected by the multiple-alignment based SVM predictor is 22.59, which is lower than 25.26 of APOLLO. The lower performance of this multiple sequence alignment based SVM predictor is probably due to the lack of the alignment e-value feature used in the pairwise alignment based SVM predictor. Thus, one direction of improving multiple sequence alignment-based method is to include some features similar to the e-value of measuring the significance of alignments. And despite the lower performance of the current implementation of the multiple sequence alignment

based SVM predictor, it is likely complementary with the black-box model quality assessment methods like APOLLO because it used completely different features in prediction. And compared to the pairwise model comparison method like APOLLO that needs a pool of models of a protein as input, the alignment-based model quality assessment methods can be applied to assess the quality of one single model. Furthermore, the Pearsons correlation coefficient score between the predicted and true GDT-TS scores on the testing data set is 0.969, indicating that the predicted and true model quality scores are highly linearly correlated.

3.5 Conclusions

In this work, we designed and developed a SVM protein model quality prediction method, taking either a pairwise sequence alignment or a multiple-sequence alignment as input. The evaluation results showed that integrating pure sequence alignment features with a SVM is an effective approach to protein model quality assessment. The new method can be integrated with template-based protein modelling methods to rank and select models. Since user-friendly and publicly accessible web-servers are important for making bioinformatics methods available to the community [25], we will make the model quality assessment methods developed in this work available as a easy-to-use web service for the community in the future.

Chapter 4

A Protein Tertiary Structure Prediction Pipeline

4.1 Introduction

Nowadays, there have been large quantities of efforts from biologists on the protein tertiary structure prediction. Among those efforts and outcome developed methods, the most common ways are ab-initio protein modeling and comparative protein modeling. Ab-initio protein modeling aims at building three-dimensional protein models based on some characters such as physical features and so on for the query protein. Comparative protein modeling, in contrast, generates a model in terms of the alignment between the target protein and its homologous template and the template structure. However, few efforts have ever been made on protein structure prediction by using protein similarity network especially structural similarity network so far. As we know, some protein homology identification algorithms based on protein sequence similarity network have been developed for detecting evolutionary, structural or functional relationships [99, 100, 101, 102, 103]. Earlier algorithms mainly focused on the individual edges of the network and performed local search through the protein

sequence similarity network [99, 100, 101]. Improved from them, protein homology ranking algorithms based on the global structure of the protein sequence similarity network have been developed, such as RANKPROP algorithm [102, 103]. These methods, in practical, build up a good foundation for protein structure prediction, and we can take them as an important preliminary step. Taking advantage of global protein network searching, we also adopt a network-based inference algorithm similar to RANKPROP to rank the proteins in terms of the similarity to the query and pick up the top ten hits as the templates for tertiary structure prediction of a query protein. However, different from RANKPROP, we build up a structure similarity-based network. Furthermore, other steps including sequence/profile alignment, model generation, model quality assessment and model selection are carried out following the template identification in our protein tertiary structure prediction pipeline.

4.2 Methods

4.2.1 Overview of the prediction pipeline

The protein tertiary structure prediction pipeline consists of four major components. The template identification component accepts an input query sequence and searches it against a non-redundant protein sequence database based on our proposed network-based fold recognition method. A set of top ranked templates and the query protein are fed into the query-template alignment component. This component returns a set of query-single template pairwise alignments and query-multi template alignments. The query-template alignments and template structures are fed into model generation tools (model generator) to sample conformations for the query. The model generators usually produce a number of models, which are then evaluated by the model quality assessment component. The model quality assessment tools assign a global quality

score to each model measuring its overall quality (e.g. overall similarity between the model and the known native structure) and a local quality score to each residue predicting its deviation compared with the native structure. At the end, the models with the best predicted qualities are released from the system as the final predictions.

4.2.2 Fold recognition using protein structural similarity network

In this component, we used TM-align [75, 104] on 12,775 proteins in PDB database to construct all-versus-all weighted protein structure similarity based network. This network represents the degree of similarity between each protein pair by assigning weights to each edge. Then our proposed Netprop algorithm starts from the pre-computed protein similarity network. It firstly performs similarity network weighting and normalization, and helps the query protein find the target subnets which contain hits in the structure similarity based network, and calculate the extracted global information for all nodes in the target subnets by propagating link information outward from the query. As a result, all the proteins in the subnets are ranked in terms of the amount of link information they received from the query. Last, we choose the top ten protein from the structure similarity based subnets as the templates to predict the tertiary structure for the query. The work flow is illustrated in **Figure 4.1**.

I. Data preparation

My labmate and I calculated the sequence similarity between each pair of proteins of 32,227 proteins in the Protein Data Bank (PDB) [105](CASP1 ∼8). If the sequence similarity between one pair of proteins was higher than 0.3, we deleted one of them from this protein set and reserved the other. After that, 12,775 proteins were retained. And then we calculated the TM-Score between each pair of proteins using TM-Align [75]. If the TM-Score was higher than 0.5 between one pair of proteins, an edge was generated between them. Consequently, we performed our novel homology ranking

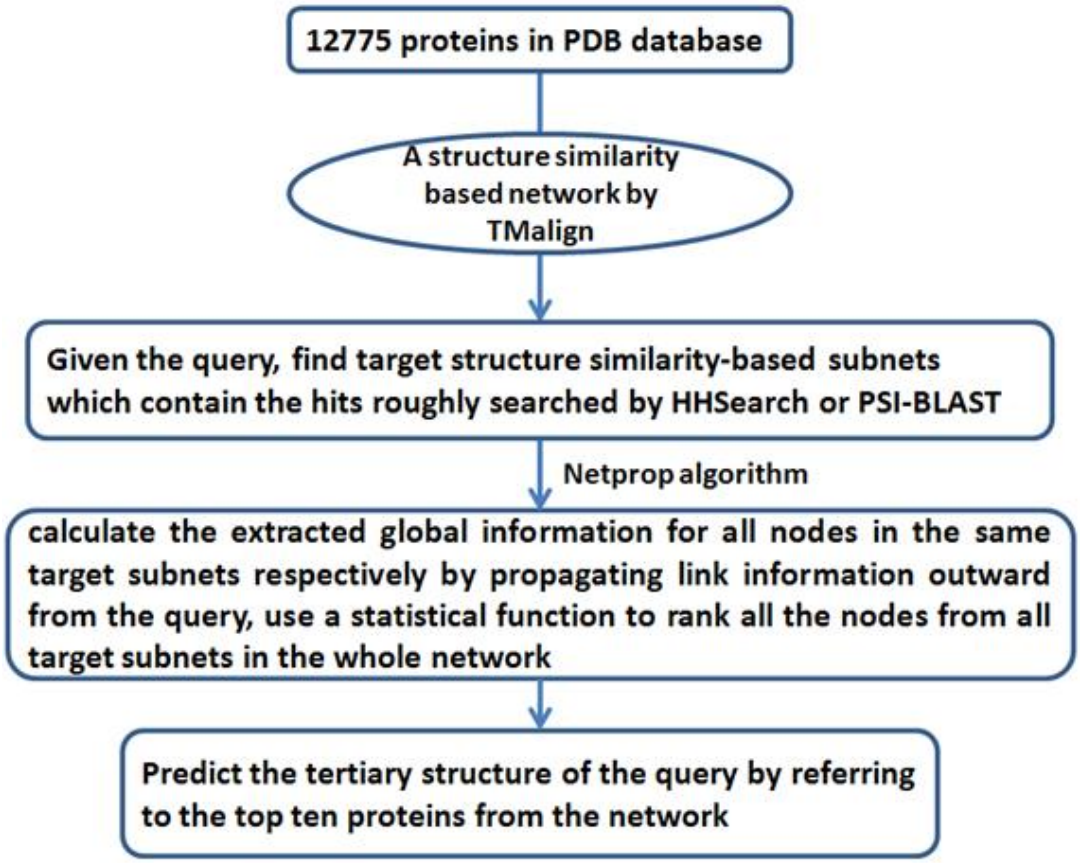


Figure 4.1: The work flow of our new methodology of network-based protein structure prediction

algorithm Netprop based on the pre-computed network.

II. Similarity network weighting and normalization

For the pre-computed structural similarity network, we can use TM-score to represent the degree of similarity between protein pairs, namely $w_{ij} = TM_{ij}$. TM-score is order-independent, which means $TM_{ij} = TM_{ji}$, and $w_{ij} = w_{ji}$. TM-score can assess the similarity of topologies of two protein structures [75, 104, 105]. TM-score lies between 0 and 1, and the higher the score is, the closer the two structures are. So, when building the structure similarity based network, edges are only included for TM-scores higher than threshold 0.5. Symmetric un-normalized weighting matrices W can be constructed based on the above step. Based on these un-normalized weights, we define that a diagonal matrix S in which S(i, i) is the sum of row i in W , and create a normalized similarity matrix by transformation function $W' = S^{-\frac{1}{2}}WS^{-\frac{1}{2}}(w'_{ij} = w_{ij}/\sqrt{S(i,i)S(j,j)})$ which constructs a symmetric matrix with row sums no higher than 1 [106]. This weighting and normalization scheme is different from Rankprop [102, 103].

III. Protein homology ranking by information retrieval in the similarity sub-networks

Given the query protein, we search the hits against all the 12775 proteins we used to build the similarity networks by HHSearch [9] or PSI-BLAST [61]. Then the structure-based sub-networks which contain the hits are considered as the target subnets. Based on the all-versus-all normalized similarity matrices for the chosen subnets, we operate protein homology ranking for the query respectively by information retrieval in each similarity subnet.

Similar to Rankprop, our approach also adopts a diffusion technique [107] which is closely related to spreading activation networks [108, 109]. First, initial activation scores are assigned to all the nodes in the subnet representing each target protein's similarity to the query. Then the activation score at each node is iteratively replaced

by a function of the weighted sum of the scores between query and all its incoming nodes. If the hits are found by HHSearch or PSI-BLAST, we denote the initial activation scores X_i^0 as the probability between the query and node i by HHSearch or PSI-BLAST, scores for other nodes in the target subnet are assigned as zero. We set X^0 as the initial column vector of activation scores, correspondingly, X^t can be taken as the column vector at iteration t of the diffusion process. The diffusion function is $X_i^{t+1} = X_i^0 + \alpha W'' X_i^t$ if X_i is not query and $X_i^{t+1} = 1$ otherwise. is the normalized weighting matrices of the subnet, and α is a constant parameter to control the rate of diffusion. We set 20 for iteration number for each query, which makes the diffusion process close to convergence. As a result, all the proteins in the subnet are ranked in terms of activation scores they gain by the diffusion process.

IV. Protein homology ranking in the whole networks

Based on the previous step, we gained the activation scores for all the nodes in the target subnets of structural similarity networks. The conceptual structure of the whole network is illustrated in **Figure 4.2**.

However, we only calculate the activation scores for those nodes in the sphere of each separate sub-network, so those scores can be just taken as local activation scores. We propose a statistical transformation to convert these local activation scores for all the nodes to global scores, otherwise it will be a bias if we directly compare the local activation scores in the sphere of the whole network. Suppose in each subnet S_i, the initial local activation scores we gained from last step are $K_{i1}, K_{i2}, \ldots\ldots, K_{in}$ for all the nodes $p_{i1}, p_{i2}, \ldots\ldots, p_{in}$. We use $P(p_{ij}|S_i), P(p_{ij}|S)$ to denote the possibility that protein p_{ij} ranks the first in S_i and S respectively. Our aim is to generate transformed global activation scores $K'_{ij}(1 \leq i \leq k, 1 \leq j \leq n_i)$. K_{ij} is proportional to $P(p_{ij}|S_i)$, and K'_{ij} is proportional to $P(p_{ij}|S)$.

First, we try to gain $P(p_{ij}|S)$ according to $P(p_{ij}|S_i)$.

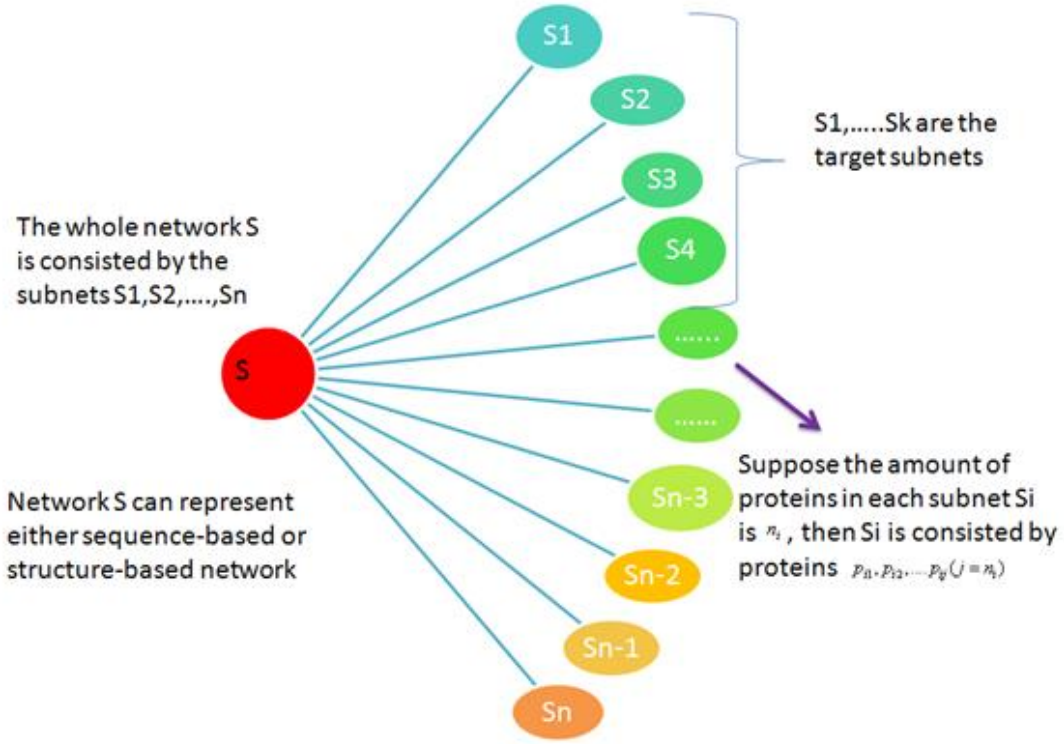


Figure 4.2: The conceptual structure of the whole network

$$
\begin{aligned}
P(p_{ij}|S) &= \textstyle\sum_{l=1}^{n} P(p_{ij}, S_l|S) \\
&= \textstyle\sum_{l=1}^{n} P(p_{ij}|S_l, S) \bullet P(S_l|S) \\
&= \textstyle\sum_{l=1}^{n} P(p_{ij}|S_l) \bullet P(S_l|S)
\end{aligned} \tag{4.0}
$$

Since $p_{ij} \in S_i, P(p_{ij}|S_l) = 0$ if $l \neq i$, namely:

$$
P(p_{ij}|S) = P(p_{ij}|S_i) \bullet P(S_l|S) = P(p_{ij}|S_i) \bullet \frac{n_i}{n_1 + n_2 + \ldots\ldots + n_n} \tag{4.0}
$$

From $K_{ij} \propto P(p_{ij}|S_i), K'_{ij} \propto P(p_{ij}|S)$, we can get global activation scores $K'_{ij} (1 \leq k, 1 \leq j \leq n_i)$ by transformation function:$K'_{ij} = K_{ij} \bullet \frac{n_i}{n_1+n_2+\ldots\ldots+n_n}$. Then, we can rank all the nodes by global activation scores in either sequence-based or structure-based network, and choose the top targets.

Ten top proteins are chosen from the structure-based network in terms of the final global activation scores. Then we used these hits with known structure to predict the tertiary structure of the query.

V. Analysis on the pre-computed network

The structure-based network was built including 2,214 sub networks after computing the TM-Score for all pairs of proteins. The largest sub network was comprised of 7,494 proteins and 199,736 edges. **Figure 4.3-4.5** showed some statistics for this structure-based network.

Figure 4.3 illustrated the frequency of the number of proteins for sub networks. The x-axis means the number of proteins in sub networks. The y-axis means the number of sub networks which have the same number of proteins. Most of sub networks have no more than 100 proteins.

Figure 4.4 showed the frequency of the number of edges for sub networks. The x-axis means the number of edges in sub networks. The y-axis means the number of sub networks which have the same number of edges. Most of sub networks have no more than 1,000 edges.

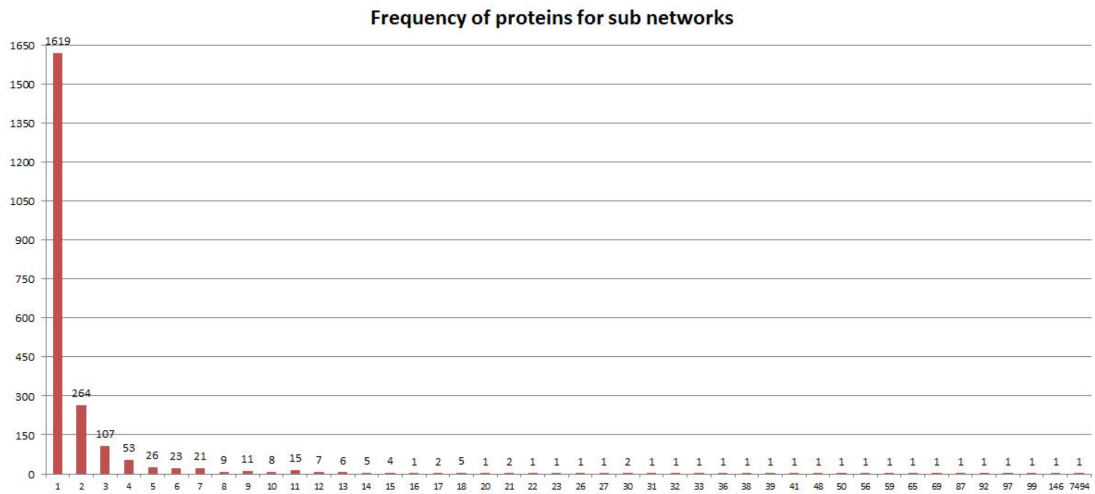


Figure 4.3: The frequency of proteins in sub-nets

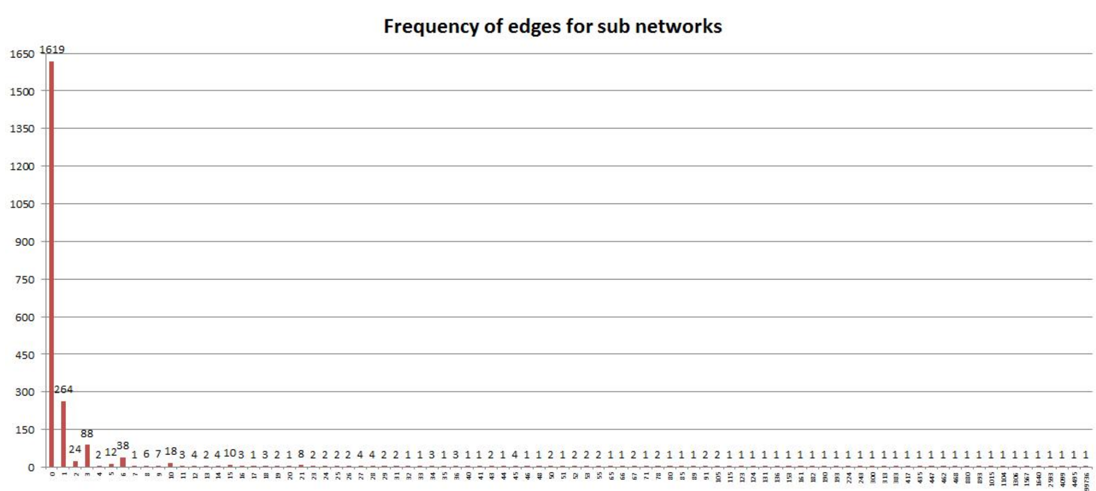


Figure 4.4: frequency of the number of edges for subnets

Figure 4.5 showed the frequency of the number of edges for proteins. The x-axis means the number of edges in proteins. The y-axis means the number of proteins which have the same number of edges. Most of proteins have no more than 300 edges.

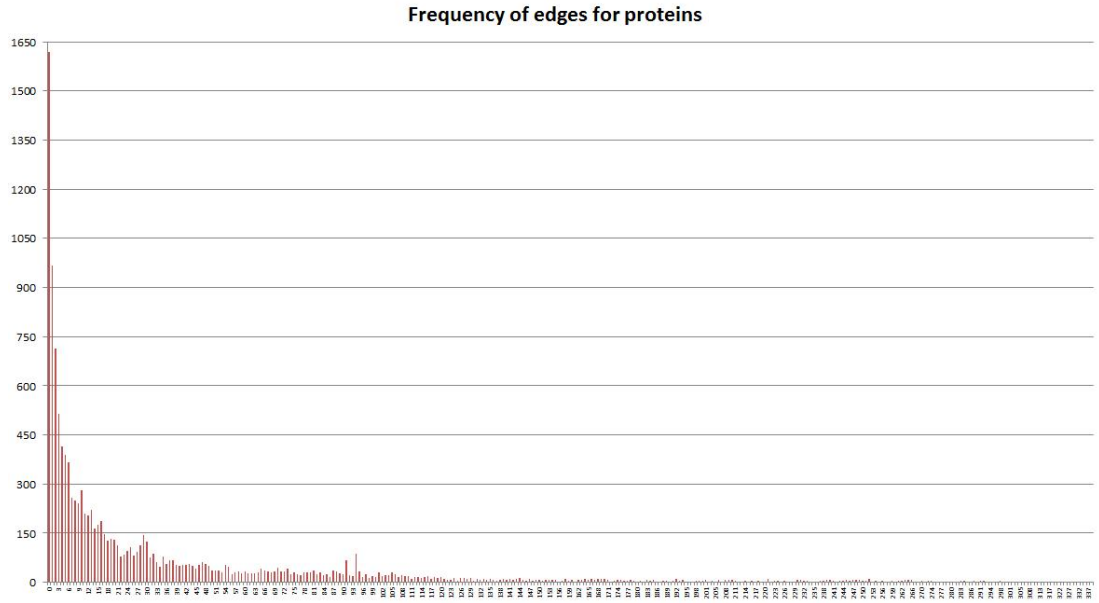


Figure 4.5: The frequency of the number of edges for proteins

We used Cytoscape [110] to visualize the structure-based network of proteins. **Figure 4.6** showed the visualization of the second largest sub network.

In total, the SBN of proteins contains 2,214 disconnected sub-graphs (each one has no edges connecting to any other sub-graphs); and most of the sub-graphs have less than 100 nodes (proteins). (A) is the second largest SBN sub-graph of proteins that has 146 nodes and 1,640 edges. (B) is an enlarged partial view of the second largest SBN, in which protein 2F5KA is a hub. The graph shown in **Figure 4.7** is an actual example of the small sub-graphs in SBN.

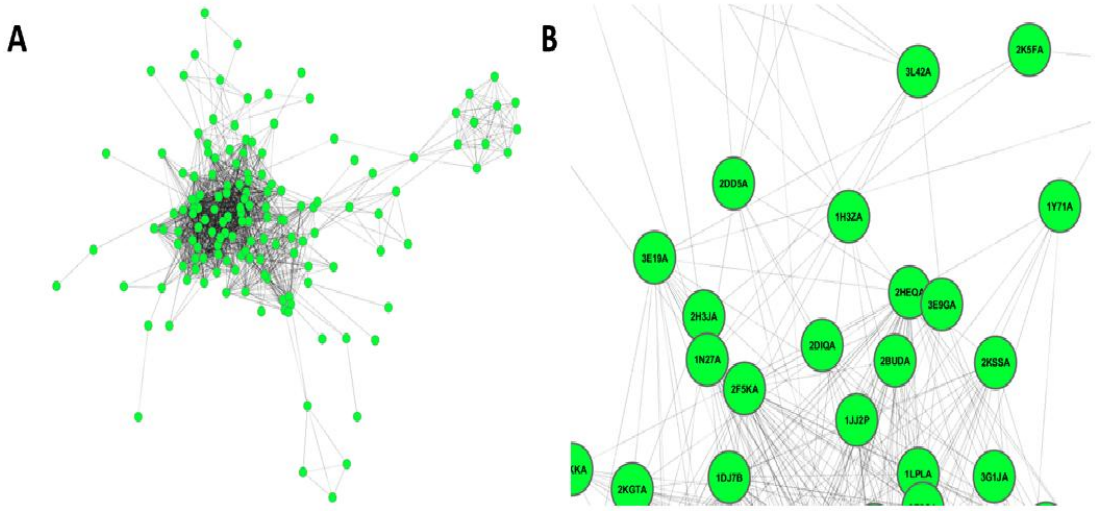


Figure 4.6: The frequency of the number of edges for proteins

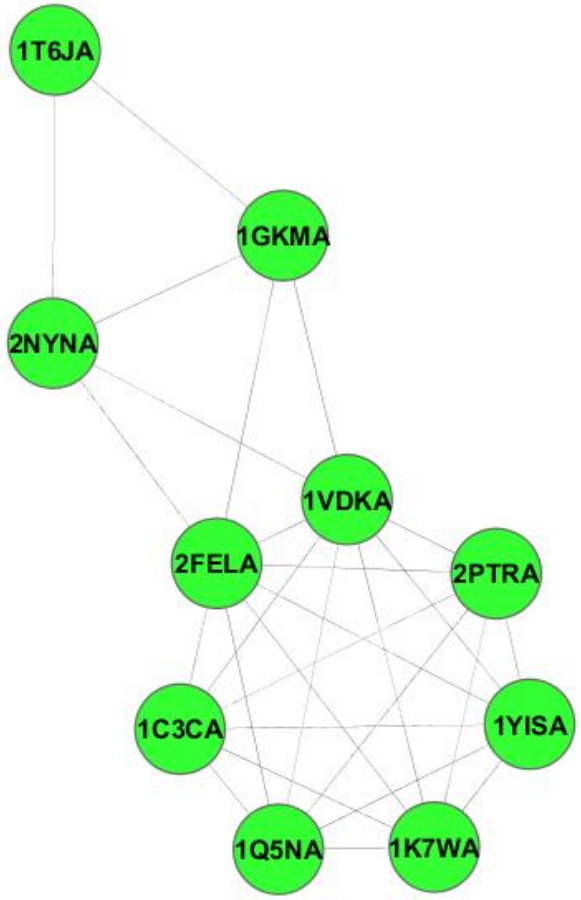


Figure 4.7: The frequency of the number of edges for proteins

4.2.3 Query-template alignment

For the new fold recognition component in our predition pipeline, a list of top ten hits is returned. Multiple sequence alignments and profile-proflie pairwse alignments are respectively performed afterwards. Specifically, for each template in the list, its structure is first aligned with that of each of the remaining templates using TM-align [75], and those with a high structural similarity score (i.e. GDT-TS score higher than 0.5) are selected into the template group with the given template as the seed. Consequently, ten different template groups are generated and the multiple sequence alignments are carried out for each template group and the query sequence recursively by MSACompro introduced in Chapter 1. Similarity, profile-profile pairwise alignments are carried out recursively between the query and each template by the old version of our pairwise alignment method introduced in Chapter 2. The latest version of our pairwise alignment method could be adopted in our new tertiary structure prediction pipeline.

4.2.4 Model generation

In the model generation component, for the simple target, two template-based model generator tools (Modeller [77] and our in-house model generator) are adopted to generate models based on the query-template alignments and the corresponding template structures. In contrast, for the hard target, a template-free model generator, Rosetta [111] is utilized to generate models for the target in addition to the two template-based model generator tools.

4.2.5 Model quality assessment

As model assessment is very challenging and none of the current methods can consistently select the best model, three model quality assessment methods (single-model

approach, model pairwise comparison approach (APOLLO) [85], and the SVM based model quality assessment aproach introduced in Chapter 3) are employed to assess the quality of the models in this layer. The single-model method (i.e. ModelEvaluator [84]) assigns an absolute quality score (e.g. GDT-TS score, the expected similarity between the model and the native structure) to each model by comparing the secondary structure, solvent accessibility, contact map, and beta-sheet topology of the model with that predicted from the query sequence[41, 46]. This method is generally effective at discriminating good models from poor models. The pairwise comparison method (APOLLO) compares a model against all other models using a structure alignment tool (e.g. TM-score [88] and calculates their similarity in terms of GDT-TS score, TM-score, and MaxSub score. The average similarity between a model and all other models is used as the predicted quality of the model. Note that the accuracy of the pairwise comparison method is input dependent (i.e. it works well only if the size of the model pool is large enough and the largest group of similar models in the pool are of good quality). At the end of this component, all models in the pool have been ranked by the average quality scores predicted by these three methods. Five top models are selected as the final models predicted by our pipeline.

Chapter 5

Summary and concluding remarks

Protein sequence alignment has played an essential role in the Bioinformatics field. An accurate sequence alignment among a protein target and its templates can successfully lead to generate a better quality model for the query protein.

In the multiple sequence alignment work, we designed a new algorithm MSAComPro 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 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.

Besides multiple sequence alignment, profile-profile pairwise sequence alignment is another critical point to improve the protein tertiary structure prediction, since it has the advantage of employing the information from the templates of the query protein. In the profile-profile pairwise alignment work, we developed HHpacom (HMM-HMM pairwise protein sequence alignment combining structural information and inferred residue pair coupling information), which extends HHsuite to enable fast and high-quality profile-profile pairwise alignment by integrating secondary structure, solvent accessibility, torsion angle and inferred residue pair coupling information. To optimize the parameters of HHpacom, we divided 2621 pairs of which each contains a CASP9 target and its single homolog released in CASP9 website into training and testing data sets. The training dataset is consisted of 1482 target-single template pairs generated from 60 CASP9 targets, and the testing dataset is consisted of 1138 pairs generated from 46 CASP9 targets. Two evaluation schemes were carried on for the assessment: (1) we generated true or reference pairwise alignments by TMalign, and calculated the SP score and TC score for the pairwise alignments generated by HHpacom, HHsearch and HHsuite; (2) 3D-models were obtained by MODELLER based on the pairwise alignments generated by these methods. TM-scores and GDT-TS scores were calculated for the 3D-models respectively. The evaluation results showed that the method improved pairwise sequence alignment accuracy over HHsearch and HHsuite by incoporating the solvent accessibility and torsion angle information, and the accuracy significantly improved in comparision with both HHsearch and HHsuite without applying secondary structure information. However, the inferred residue pair information did not improve much on the CASP9 data set, yet our case study of the effect of the inferred residue pair information provided a useful view sight for the future direction in the profile-profile pairwise sequence alignment. Furthermore, there are a few potential ways to improve our profile-profile pairwise alignment method in the future work: (1) More states of solvent accessibility may be considered: eg. three

states including exposed, buried, or intermediate. Or real values of solvent accessibilities may be predicted, and then the similarity scores between residue pairs could be calculated according to certain thresholds. (2) It is found that mutual information (MI) may bring some noises in the calculation of residue coupling information, while direct information (DI) is able to discover effective residue couplings from a global maximum entropy model [74]. Consequently, we may use a maximum entropy model to infer residue pair couplings, so as to improve the alignment quality.

Given different sequence alignments among a given target protein and its single or multiple templates, how to effectively select the top alignments so as to generate better-quality models for the target is the next key step. Consequently, we implemented two SVM protein model quality prediction methods, taking either a pairwise alignment or a multiple alignment as input. The evaluation results showed that integrating pure sequence alignment features with a SVM is effective, convenient and cheap for protein model quality assessment. In addition, we believe that there is still a large space to improve such a method. In the future work, we may improve the method by adding more features from the 3-D models of the template sequences in the sequence alignments into the Support Vector Machine.

Last, we developed a protein tertiary structure prediction pipeline. Some components such as sequence alignment, profile-profile alignment and SVM based model quality assessment were built into our group's MULTICOM tertiary structure prediction system. The automatic evaluation of MULTICOM on MODEL 1 released by CASP10 website illustrated our performance in Protein Tertiary Structure Prediction was ranked number four among all the participants.

Moreover, here is a list of my main publications during my PhD study:

[1] X. Deng and J. Cheng. (2013) New profile-profile pairwise protein sequence alignment by HMM-HMM comparison (under submission).

[2] X. Deng and J. Cheng. (2013) Predicting Protein Model Quality from Sequence

Alignments by Support Vector Machines (under submission).

[3] J. Li, X. Deng, J. Eickholt, J. Cheng. (2013) Designing and Benchmarking the MULTICOM Protein Structure Prediction System. BMC Structural Biology. 13:2.

[4] B. Adhikari, X. Deng, J. Li, D. Bhattacharya, and J. Cheng. (2013) A Contact-assisted Approach to Protein Structure Prediction and Its Assessment in CASP10. AAAI workshop.

[5] M. Zhu, X. Deng, T. Joshi, D. Xu, G. Stacey, J. Cheng. (2012) Reconstructing Differentially Co-expressed Gene Modules and Regulatory Networks of Soybean Cells. BMC Genomics, 13:434.

[6] J. Cheng, J. Li, Z. Wang, J. Eickholt, and X. Deng. (2012) The MULTICOM Toolbox for Protein Structure Prediction. BMC Bioinformatics, 13:65.

[7] J. Cheng, J. Eickholt, Z. Wang, and X. Deng. (2012) Recursive Protein Modeling: a Divide and Conquer Strategy for Protein Structure Prediction and its Case Study in CASP9. ournal of Bioinformatics and Computational Biology, vol. 10, no. 3.

[8] X. Deng and J. Cheng. (2011) MSACompro: Protein Multiple Sequence Alignment Using Predicted Secondary Structure, Solvent Accessibility, and Residue-Residue Contacts. BMC Bioinformatics, 12:472.

[9] X. Deng, J. Eickholt and J. Cheng. (2012) A Comprehensive Overview of Computational Protein Disorder Prediction Methods. Molecular BioSystems, 8(1):114-121.

[10] J. Eickholt, X. Deng and J. Cheng. (2011) DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. BMC Bioinformatics, February.

[11] X. Deng, J. Eickholt and J. Cheng. (2009) PreDisorder: Ab Initio Sequence-based Prediction of Protein Disordered Regions. BMC Bioinformatics, 10:436.

Bibliography

[1] Barton GJ. and Sternberg MJ. A strategy for the rapid multiple alignment of protein sequences. confidence levels from tertiary structure comparisons. *J. Mol.Biol*, 198:327–337, 1987.

[2] Feng DF and Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol*, 25:351–361, 1987.

[3] Krogh A. et al. Hidden markov models in computational biology: applications to protein modeling. *J. Mol.Biol*, 235:1503–1531, 1994.

[4] Liu YC, Schmidt B, Douglas LM. Msaprobs: multiple sequence alignment based on pair hidden markov models and partition function posterior probabilities. *Bioinformatics*, 26(16):1958–1964, 2010.

[5] Do CB, et al. Probcons: probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15:330–340, 2005.

[6] Poirot O, Suhre K, Abergel C, Eamonn OT and Notredame C. 3dcoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Research*, 32:37–40, 2004.

[7] Pei J, Kim B, and Grishin NV. Promals3d: a tool for multiple sequence and structure alignment. *Nucleic Acids Research*, 36(7):W244–W248, 2008.

[8] Soding J, Biegert A, and Lupas AN. The hhpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33:2295–2300, 2005.

[9] Soding J. Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21:951–960, 2005.

[10] Heringa J. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *J. Comput. Biol*, 23:341–364, 2006.

[11] Kim NK, Xie J. Protein multiple alignment incorporating primary and secondary structure information. *Comput. Chem*, 13:75–88, 1999.

[12] Amarendran RS, Suvrat H, Rasmus S, Peter M, Eduardo C, and Burkhard M. Dialign-tx and multiple protein alignment using secondary structure information at gobics. *Nucleic Acids Research*, 38(suppl 2):W19–W22, 2010.

[13] Zhou HY, Zhou YQ. Spem: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, 21:3615–3621, 2005.

[14] Pei J, Grishin NV. Mummals: multiple sequence alignment improved by using hidden markov models with local structural information. *Nucleic Acids Res*, 34(16):4364–4374, 2006.

[15] Pei J, Grishin NV. Promals: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, 23:802–808, 2007.

[16] Brudno M, Steinkamp R and Morgenstern B. The chaos/dialign www server for multiple alignment of genomic sequences. *Nucl. Acids Res*, 32 (Supplement 2):W41.

[17] Larkin M, et al. Clustal w and clustal x version 2.0s. *Bioinformatics*, 23(21):2947–2948, 2007.

[18] Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, 31:3497–3500, 2003.

[19] Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with clustal x. *Trends Biochem Sci*, 23:403–405, 1998.

[20] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The clustal x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, 25:4876–4882, 1997.

[21] Higgins DG, Thompson JD, Gibson TJ. Using clustal for multiple sequence alignments. *Methods Enzymol*, 266:383–402, 1996.

[22] Thompson JD, Higgins DG, Gibson TJ. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–4680, 1994.

[23] Higgins DG. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–4680, 1994.

[24] Higgins DG, Bleasby AJ, Fuchs R. Clustal v: improved software for multiple sequence alignment. *Comput. Appl. Biosci*, 8:189–191, 1992.

[25] Higgins DG. and Sharp PM. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73:237–244, 1988.

[26] Bailey TL and Noble WS. Searching for statistically significant regulatory modules. *Bioinformatics*, 19 (Suppl. 2), 2003.

[27] Amarendran RS, Kaufmann M, Morgenstern B. Dialign-tx: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for Molecular Biology*, 3:6, 2008.

[28] Amarendran RS, Jan WM, Kaufmann M, Morgenstern B. Dialign-t: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, 6:66, 2005.

[29] Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. Fast statistical alignment. *PLoS Computational Biology*, 5:e1000392, 2009.

[30] Katoh K, Misawa K, Kuma K, Miyata T. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*, 30(14):3059–66, 2002.

[31] Notredame C, Higgins D, Heringa J. T-coffee: A novel method for multiple sequence alignments. *JMB*, 302:205–217, 2000.

[32] Brudno M, Do CB, Cooper G, Michael FK, Davydov E, EricDG, Sidow A, and Batzoglou S. Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. *Genome Research*, 2003.

[33] Edgar RC. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–97, 2004.

[34] Edgar RC. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113, 2004.

[35] Chikkagoudar S, Roshan U and Livesay DR. eprobalign: generation and manipulation of multiple sequence alignments using partition function posterior probabilities. *Nucleic Acids Research*, 35:W675–W677, 2007.

[36] Sze SH, Lu Y, and Yang Q. A polynomial time solvable formulation of multiple sequence alignment. *Journal of Computational Biology*, 13:309–319, 2006.

[37] Roshan U, Livesay DR. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, 22(22):2715–21, 2006.

[38] Thompson JD, Koehl P, Ripp R, Poch O. Balibase 3.0: Latest developments of the multiple sequence alignment benchmark. *Proteins: Structure, Function, and Bioinformatics*, 61:127–136, 2005.

[39] Walle V, et al. Align-mca new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, 20:1428–1435, 2004.

[40] Raghava GP. et al. Oxbench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, 4:47, 2003.

[41] Cheng J, Randall A, Sweredoski M, Baldi P. Scratch: a protein structure and structural feature prediction server. *Web Server Issue*, 33:72–76, 2005.

[42] Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 47:142–153, 2002.

[43] Gonnet GH, Cohen MA and Benner SA. Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445, 1992.

[44] Kawabata T and Nishikawa K. Protein structure comparison using the markov transition model of evolution. *Proteins*, 41:108–122, 2000.

[45] Durbin R. et al. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. *Cambridge University Press*.

[46] Tegge AN, Wang Z, Eickholt J, and Cheng J. Nncon: Improved protein contact map prediction using 2d-recursive neural networks. *Nucleic Acids Research*, 37:w515–w518, 2009.

[47] Sneath PHA. and Sokal RP. Numerical taxonomy. *Freeman*, 1973.

[48] Openmp tutorial. *[https://computing.llnl.gov/tutorials/openMP].*

[49] Thompson JD, Frederic P and Olivier P. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, 27:2682–2690, 1999.

[50] Walle V, et al. Align-mca new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, 20:1428–1435, 2004.

[51] Boutonnet NS, et al. Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. *Protein Eng*, 8:647–662, 1995.

[52] Brenner SE, et al. The astral compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28:254–256, 2000.

[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*, 4:7, 2003.

[55] Poirot O, Suhre K, Abergel C, Eamonn OT and Notredame C. 3dcoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Research*, 32:37–40, 2004.

[56] Wilcoxon F. Probability tables for individual comparisons by ranking methods. *Biometrics*, 3:119–122, 1947.

[57] P Bork, and EV Koonin,. Predicting functions from protein sequenceswhere are the bottlenecks? *Nature genetics*, 18:313–318, 1998.

[58] M Henn-Sax, et al. Divergent evolution of (a)8-barrel enzymes. *Biological chemistry*, 382:1315–1320, 2001.

[59] LN. Kinch, et al. Casp5 assessment of fold recognition target predictions. *Proteins: Structure, Function, and Bioinformatics*, 53:395–409, 2003.

[60] M Remmert, et al. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 2011.

[61] SF Altschul, et al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25:3389–3402, 1997.

[62] K Ginalski, et al. Orfeus: detection of distant homology using sequence profiles and predicted secondary structure. *Biological chemistry*, 382:1315–1320, 2001.

[63] CL Tang, et al. On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *Journal of molecular biology*, 334:1043–1062, 2003.

[64] K Tomii, and Y Akiyama. Forte: a profileprofile comparison tool for protein fold recognition. *Bioinformatics*, 20:594–595, 2004.

[65] et al. D Fischer. Cafasp3: the third critical assessment of fully automated structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 53:503–516, 2003.

[66] L Rychlewski, D Fischer, and A Elofsson. Livebench-6: Large-scale automated evaluation of protein structure prediction servers. *Proteins: Structure, Function, and Bioinformatics*, 53:542–547, 2003.

[67] et al. D Keedy. The other 90template-based and high-accuracy models. *Proteins: Structure, Function, and Bioinformatics*, 2009.

[68] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[69] J Cheng, et al. The multicom toolbox for protein structure prediction. *BMC Bioinformatics*, 13:65, 2013.

[70] E Faraggi, et al. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure*, 17:1515–1527, 2009.

[71] W Zhang, S Liu, and Y Zhou. Sp5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. *PLoS One*, 3:e2325, 2008.

[72] A Biegert, and J Sding. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, 24:807–814, 2008.

[73] TA Hopf, et al. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 2012.

[74] DS Marks, et al. Protein 3d structure computed from evolutionary sequence variation. *PLoS One*, 6:e28766, 2011.

[75] Y Zhang, and J Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33:2302–2309, 2005.

[76] X Deng, and J Cheng. Msacompro: Protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts. *BMC Bioinformatics*, 12:472, 2011.

[77] N Eswar, et al. Comparative protein structure modeling using modeller. *current protocols in bioinformatic*, pages 5.6. 1–5.6. 30, 2006.

[78] Berardi MJ, Chou JJ. Structure and mechanism of the m2 proton channel of influenza a virus. *Nature*, 451(7178):591–595, 2008.

[79] Schnell JR, Shih WM, Harrison SC, Chou JJ. Mitochondrial uncoupling protein 2 structure determined by nmr molecular fragment searching. *Nature*, 476(7358):109–113, 2011.

[80] OuYang B, Xie S, Berardi MJ, Zhao X, Dev J, Yu W, Sun B, Chou JJ. Unusual architecture of the p7 channel from hepatitis c virus. *Nature*, 2013.

[81] Chou K-C. Structural bioinformatics and its impact to biomedical science. *Current medicinal chemistry*, 11(16):2105–2134, 2004.

[82] Chou K-C. Coupling interaction between thromboxane a2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. *Journal of proteome research*, 4(5):1681–1686, 2005.

[83] Lundstrm J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: A neural-networkbased consensus predictor that improves fold recognition. *Protein Science*, 10(11):2354–2362, 2001.

[84] Z Wang, AN Tegge, J Cheng. Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 75(3):638–647, 2006.

[85] Z Wang, J Eickholt, J Cheng. Apollo: a quality assessment service for single and multiple protein models. *Bioinformatics*, 27(12):1715–1716, 2011.

[86] Chen H, Kihara D. Estimating quality of template-based protein models by alignment stability. *Proteins: Structure, Function, and Bioinformatics*, 71(3)):1255–1274, 2008.

[87] J Li, X Deng, J Cheng. Designing and benchmarking the multicom protein structure prediction system. *BMC structural biology*, 13(1):1–14, 2013.

[88] Y Zhang, and J Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[89] J Xu, and Y Zhang. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895, 2010.

[90] A Zemla. Lga: a method for finding 3d similarities in protein structures. *Nucleic acids research*, 31(13):3370–3374, 2003.

[91] J Cheng,J Li, Z Wang, J Eickholt, X Deng. The multicom toolbox for protein structure prediction. *BMC Bioinformatics*, 13(65), 2012.

[92] C Cortes ,V Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[93] S Henikoff ,JG Henikoff . Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(2):10915–10919, 1992.

[94] Contreras-Moreira B, Ezkurdia I, Tress M, Valencia A. Empirical limits for template-based protein structure prediction: the casp5 example. *FEBS letters*, 579(5):1203–1207, 2005.

[95] Joachims T. Making large scale svm learning practical. 1999.

[96] Chou K-C, Zhang C-T . Prediction of protein structural classes. *Critical reviews in biochemistry and molecular biology*, 30(4):275–349, 1995.

[97] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[98] H Zhou, Y Zhou. Spem: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, 21(18):3615–3621, 2005.

[99] Smith TF, Waterman MS. Identification of common molecular subsequences. *J mol Biol*, 147(1):195–197, 1981.

[100] Pearson WR. Rapid and sensitive sequence comparison with fastp and fasta. *Methods in enzymology*, 183:63–98, 1990.

[101] Altschul S. Basic local alignment search tool. *J Mol Biol*, 215:403–410, 1990.

[102] Weston J, Elisseeff A, Zhou D, Leslie CS, Noble WS. Protein ranking: from local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6559, 2004.

[103] Melvin I, Weston J, Leslie C, Noble WS. Rankprop: a web server for protein remote homology detection. *Bioinformatics*, 25(1):121, 2009.

[104] Y Zhang, and J Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[105] Bernstein FC. The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, 112:535–542, 1977.

[106] Vanunu O, Sharan R. A propagation based algorithm for inferring genedisease associations. *Citeseer*, 2008.

[107] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions. *In*, page 912, 2003.

[108] Anderson JR. The architecture of cognition. *Cambridge: MA*, 1983.

[109] Shrager J, Hogg T, Huberman BA. Observation of phase transitions in spreading activation networks. *Science*, 236(4805):1092, 1987.

[110] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[111] Randall A, Baldi P. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, 487:545–74, 2011.

VITA

I was born in Wuhan, a capital city of Hubei Province in China. As the only child in my family, I was well cultivated and highly expected by my mom, a professor in a chinese university and also my dad, a government officer. In 2007, I finished my four-year's undergraduate study in Wuhan University, a top 10 University in China, with my bachelor degree in Computer Science. The same year, I was admitted without any entrance exam to the Computer Science Department in Wuhan University to go on my graduate study because of a high GPA in the whole department. However, I received a research assistantship from the Computer Science Department in University of Missouri-Columbia in 2008. Consequently, I decided to quit the master program in Wuhan University and start my PhD study in University of Missouri-Columbia in 2008. During the five-year's PhD training process, I have attended many conferences, participated in many valuable projects, published about 10 papers and submitted another 1-2 papers. Now, I am going to get my PhD degree in Computer Science from University of Missouri-Columbia.

I am excited to start a new life after five-year's hardwork in University of Missouri. I am going to be a research scientist in a healthcare information company in Orlando. I am still single, so I wish I could have my own family in the near future. I want to be a successful scientist in industry, and balance both my personal life and career well. Moreover, wherever I go or work in the future, I am always part of MIZZOU.