BMC Bioinformatics

RESEARCH ARTICLE Open Access

Enhancing HMM-based protein profile-profile alignment with structural features and evolutionary coupling information

Xin Deng[1] and Jianlin Cheng[2*]

Abstract

Background: Protein sequence profile-profile alignment is an important approach to recognizing remote homologs and generating accurate pairwise alignments. It plays an important role in protein sequence database search, protein structure prediction, protein function prediction, and phylogenetic analysis.

Results: In this work, we integrate predicted solvent accessibility, torsion angles and evolutionary residue coupling information with the pairwise Hidden Markov Model (HMM) based profile alignment method to improve profile-profile alignments. The evaluation results demonstrate that adding predicted relative solvent accessibility and torsion angle information improves the accuracy of profile-profile alignments. The evolutionary residue coupling information is helpful in some cases, but its contribution to the improvement is not consistent.

Conclusion: Incorporating the new structural information such as predicted solvent accessibility and torsion angles into the profile-profile alignment is a useful way to improve pairwise profile-profile alignment methods.

Background

Pairwise protein sequence alignment methods have been essential tools for many important bioinformatics tasks, such as sequence database search, homology recognition, protein structure prediction and protein function prediction [1-5]. Following the development of global and local alignment methods of aligning two single sequences [6-8], profile-sequence alignment or profile-profile alignment methods such as PSI-BLAST, SAM [9], HMMer [10], HHsearch, HHsuite [4-6], which enrich two single sequences with their homologous sequences, has substantially improved both the sensitivity of recognizing remote homologs and the accuracy of aligning two protein sequences.

Due to their relatively high sensitivity in recognizing remote protein homologs, profile-profile alignment methods have become the default structural template identification method for many template-based protein structure modeling methods and servers [11-14]. For instance, HHsearch, one of top profile-profile alignment tools based on comparing the profile hidden Markov models (HMM) of two proteins, was used by almost all the template-based protein structure prediction methods tested during the last two Critical Assessment of Techniques for Protein Structure Prediction (CASP) [15,16]. The open source package HHsuite contains both the latest implementation of HHSearch that supports a full HMM-HMM alignment-based search on a HMM profile database and a very fast search tool HHblits [5] that reduces the number of unnecessary full HMM pairwise alignment in order to drastically improve its search speed. Moreover, the maximum accuracy (MAC) alignment algorithm is applied in HHsuite, but not in HHsearch. In this work, we aim to introduce new sources of information to improve profile-profile alignments with respect to both the original HHsearch package and the open source HHsuite package,

In order to more accurately align the structurally equivalent residues in a target protein and a template protein together, secondary structure information was incorporated into profile-profile sequence alignment methods, yielding the better sensitivity and accuracy [4,17]. Aiming to find the new source of information to further improve the sensitivity and accuracy of pairwise profile-profile alignment,

* Correspondence: chengji@missouri.edu
[2]Computer Science Department, Informatics Institute, C. Bond Life Science Center, University of Missouri-Columbia, Columbia, MO 65211, USA
Full list of author information is available at the end of the article

we examine the effectiveness of incorporating into profile-profile alignment methods some new features that have not been used in profile-profile alignments before, including protein solvent accessibility, torsion angles, and the evolutionary residue coupling information [18,19].

Specifically, we add the additional scoring terms for solvent accessibility, torsion angles, and evolutionary residue coupling information into the scoring function of HHsuite [5] in order to enhance the alignment process. According to our evaluation, adding solvent accessibility and torsion angles can improve the alignment accuracy, but incorporating the evolutionary residue coupling information is only useful in some cases.

Methods

We extended an existing profile-profile alignment method within the standard five-step alignment framework of HHsuite [5] shown in Figure 1, including discretization of profile columns, removal of very short or very dissimilar sequences, execution of Viterbi alignment and calculation of E-value and probability, realignment based on the maximum accuracy (MAC) algorithm, and retrieval of alignments by tracing-back. Different from HHsuite, our method applies solvent accessibility and torsion angle information to both the Viterbi alignment and the maximum accuracy alignment, and traces back with the aid of the evolutionary residue coupling information. In the following sections, we focus on describing how to incorporate the new features into the profile-profile method (i.e., HHsuite), while briefly introducing the necessary technical background.

Adding solvent accessibilities and torsion angles into the viterbi alignment

The score of aligning two columns in two protein profiles (namely a query profile q and a template profile t) in HHsuite was calculated according to Equation (1).

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

in which $q_i(a)$ and $t_j(a)$ denote the probability of amino acid at position i in the query profile and at position j the template profile, respectively, and $f(a)$ is the background frequency of residue a ($a \in \{1, 2, ..., 20\}$, representing 20 types of amino acids). The best alignment between two profile HMMs was obtained by maximizing the log-sum-odds score S_{LSO} according to Equation (2).

$$S_{LSO} = \sum_{k: X_k Y_k = MM} S_{aa}\left(q_{i(k)}, t_{j(k)}\right) + \log P_{tr} \quad (2)$$

where k denotes the index of columns that query HMM q aligned to template HMM t, $i(k)$ and $j(k)$ are the respective columns in q and t, P_{tr} is the product of all transition probabilities for the path through q and t. The latest version of HHsuite has included the secondary structure information into the calculation of the score. In this work, we further augment the calculation of the score by adding the terms to account for the solvent accessibility, and torsion angles.

The Viterbi dynamic program algorithm used five matrices S_{AB} (i.e., $AB \in \{MM, MI, IM, DG, GD\}$) representing matching different states (M: match, I: insertion,

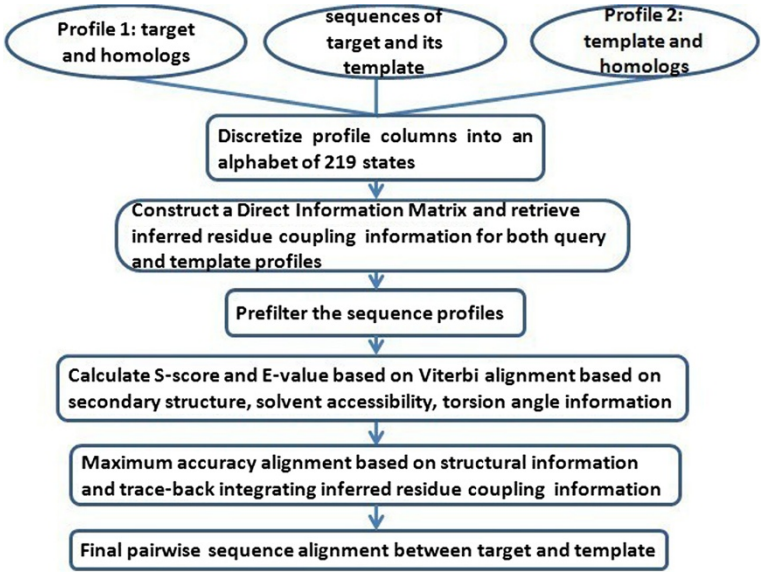

Figure 1 The workflow of the HMM-based profile-profile pairwise alignment.

D: deletion; G: Gap [4]) in two HMMs to maximize the augmented log-sum-of-odds score S_{LSO}. They are recursively calculated as:

$$S_{MM}(i,j) = S_{aa}(q_i, {}^{i}t_j) + w_{ss}S_{ss}(q_i,t_j) + w_{sa}S_{sa}(q_i,t_j) + w_{tors}S_{tors}(q_i,t_j) + \max\begin{cases} S_{MM}(i-1,j-1) + \log\left[q_{i-1}(M,M)t_{j-1}(M,M)\right] \\ S_{MI}(i-1,j-1) + \log\left[q_{i-1}(M,M)t_{j-1}(I,M)\right] \\ S_{IM}(i-1,j-1) + \log\left[q_{i-1}(I,M)t_{j-1}(M,M)\right] \\ S_{DG}(i-1,j-1) + \log\left[q_{i-1}(D,M)t_{j-1}(M,M)\right] \\ S_{GD}(i-1,j-1) + \log\left[q_{i-1}(M,M)t_{j-1}(D,M)\right] \end{cases} + S_{shift} \quad (3)$$

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

$$S_{MI}(i,j) = \max\begin{cases} S_{MM}(i-1,j) + \log\left[q_{i-1}(M,M)t_j(M,I)\right] \\ S_{MI}(i-1,j) + \log\left[q_{i-1}(M,M)t_j(I,I)\right] \end{cases} \quad (4)$$

$$S_{DG}(i,j) = \max\begin{cases} S_{MM}(i-1,j) + \log[q_{i-1}(M,D)] \\ S_{DG}(i-1,j) + \log[q_{i-1}(D,D)] \end{cases} \quad (5)$$

$S_{IM}(i,j)$ and $S_{GD}(i,j)$ are calculated similarly as $S_{MI}(i,j)$ and $S_{DG}(i,j)$.

The difference between Equation (3) above and the default one in HHsuite is that two new terms (S_{sa}, S_{tors}) were added to utilize the solvent accessibility and torsion angle information. In Equation (3), $S_{ss}(q_i, t_j)$ is the secondary structure score between column i in query HMM (q_i) and column j in template HMM (t_j), which was the same as the one originally used 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 the new terms introduced in this work. w_{ss}, w_{sa}, and 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 states. 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. $q_{i-1}(M,M)$ is the transition probability from state M at column *i-1* to next state M of in the query HMM, and $t_{j-1}(M,M)$ is the transition probability from state M at column *j-1* to next state M in the template HMM.

Here we denote this extension of the HHsuite method as HMMsato. HMMsato allows for scoring predicted (or known) solvent accessibilities of one protein against predicted (or known) ones of another protein. DSSP [20] is used to parse the true solvent accessibility of a protein if its tertiary structure is known. PSpro 2.0 [21] is used to predict the solvent accessibility of a protein. The solvent accessibility information can be automatically parsed or predicted in HMMsato, or alternatively provided by a user. The two types of solvent accessibilities (e: exposed, > = 25% of the maximum area of a residue is exposed; b: buried, < 25% of the maximum area of a residue is exposed) are employed. Assuming the predicted or true solvent accessibility states of the i$^{\mathrm{th}}$ residue (x_i) of the query protein and the j$^{\mathrm{th}}$ residue (y_j) of the template protein are $sa(x_i)$ and $sa(y_j)$, the solvent accessibility score between the two residues $S_{sa}(q_i, t_j)$ is defined as:

$$S_{sa}(q_i,t_j) = \boldsymbol{\delta}\left(sa(x_i), sa\left(y_j\right)\right) \quad (6)$$

The score is calculated by the kronecker-delta function $\delta(a, b)$, which equals to 1 if $a = b$, 0 otherwise.

Similarly as the solvent accessibility, the torsion angles including both phi angle (φ) and psi angle (ψ) can be automatically predicted by SPINE-X [22,23] or provided by a user. The range of both φ and ψ is (−180,180). Given the query sequence X and template sequence Y, the predicted phi angle and psi angle of the i-th residue x_i in the query are denoted as $\varphi(x_i)$ and $\psi(x_i)$, and those of the j-th residue y_j in the template as $\varphi(y_j)$ and $\psi(y_j)$. The torsion angle score $S_{tors}(q_i, t_j)$ between the two residues is calculated as:

$$_{Stors}(q_i,t_j) = 1 - \frac{\sqrt{0.5 * \left[\left(\boldsymbol{\varphi}(x_i) - \boldsymbol{\varphi}\left(y_j\right)\right)^2 + \left(\boldsymbol{\psi}(x_i) - \boldsymbol{\psi}\left(y_j\right)\right)^2\right]}}{180} \quad (7)$$

Realign the profiles by maximum accuracy alignment combining solvent accessibility and torsion angles

It has been shown that maximum accuracy (MAC) algorithm can generally create a more accurate alignment than the Viterbi algorithm, while the latter can generate better alignment scores, e-values and probabilities [5,24]. Consequently, the Viterbi algorithm is applied to compute e-values and scores, and the MAC algorithm is chosen to generate the final HMM-HMM pairwise alignment in HHsato by default.

The maximum accuracy algorithm [5,24] creates the local alignment that maximizes the sum of probabilities for each residue pair to be aligned minus a penalty (*mact*) (i.e., $argmax(\sum_{i,j \in alignment}\left[P\left(q_i^M \sim t_j^M\right) - mact\right])$), where $P\left(q_i^M \sim t_j^M\right)$ represents the posterior probability of the match state i in HMM q aligned to the match state j in HMM t. With the parameter *mact*, users can control the alignment greediness, from nearly global, long alignment ($mact = 0$) to very precise, short local alignments ($mact \approx 1$). The default value of *mact* is set to 0.3501 in HMMsato as in HHsuite. To find the best MAC alignment path, an optimal sub-alignment score

matrix AS is calculated recursively using the posterior probability $P\left(q_i^M \sim t_j^M\right)$ as substitution scores:

$$AS(i,j) = \max\begin{cases} P\left(q_i^M \sim t_j^M\right) - mact \\ AS(i-1,j-1) + P\left(q_i^M \sim t_j^M\right) - mact \\ AS(i-1,j) - 0.5 * mact \\ AS(i,j-1) - 0.5 * mact \end{cases} \tag{8}$$

Here, the Forward-Backward algorithm in local or global mode is applied to calculate the posterior probabilities $P\left(q_i^M \sim t_j^M\right)$. The Forward partition function $F_{MM}(i,j)$ and Backward partition function $B_{MM}(i,j)$ are introduced to calculate the posterior probability for pair state (q_i^M, t_j^M) according to Equation (9):

$$P\left(q_i^M \sim t_j^M\right) = \frac{F_{MM}(i,j)B_{MM}(i,j)}{1 + \sum_{i,j} F_{MM}(i,j)} \tag{9}$$

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

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

$$\begin{aligned} F_{MI}(i,j) = F_{MM}(i-1,j)q_{i-1}(M,M)t_j(M,I) + \\ F_{MI}(i-1,j)q_{i-1}(M,M)t_j(I,I) \end{aligned} \tag{10}$$

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

where $p\min$ controls the alignment model (0: global alignment mode, 1: local alignment mode). $F_{IM}(i,j)$ and $F_{GD}(i,j)$ are calculated similarly as $F_{MI}(i,j)$ and $F_{DG}(i,j)$. Solvent accessibility score $S_{sa}(q_i, t_j)$ and torsion angle score $S_{tors}(q_i, t_j)$ are calculated as in the Viterbi alignment.

In analogy to the Forward partition function, the Backward partition function matrix B_{MM} are calculated recursively as follows:

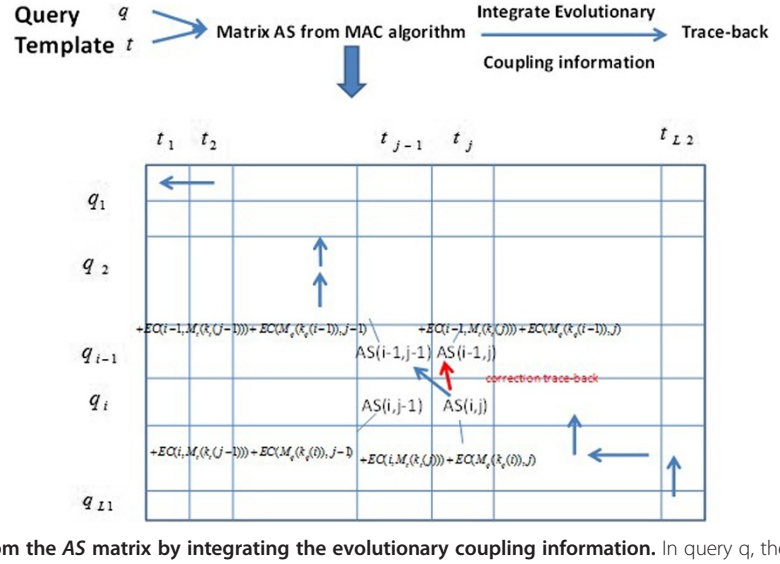

Figure 2 Tracing back from the *AS* matrix by integrating the evolutionary coupling information. In query q, the coupled position of *i* is $k_q(i)$, and that of *i-1* is $k_q(i-1)$. In template t, the coupled position of *j* is $k_t(j)$, and that of *j-1* is $k_t(j-1)$. $M_q(i)$ is the corresponding position in template t matched to position *i* in q during the original tracing-back. $M_t(j)$ is the corresponding position in query q matched to position *j* in t during the original tracing-back. Additional EC scores are added into the corresponding elements in the *AS* matrix as shown in the figure so that the correct tracing back is performed.

Table 1 The mean SP and TC scores of the pairwise alignments generated by HHsearch1.2, HHsuite and HMMsato on the CASP9 test data set consisting of 1,138 pairs of proteins

Method	Mean SP score	Mean TC score
HHsearch (without secondary structure information)	48.69	48.34
HHsearch (with secondary structure information)	50.00	49.65
HHsuite (without secondary structure information)	48.47	48.12
HHsuite (with secondary structure information)	49.76	49.41
HMMsato	**50.39**	**50.02**

Bold numbers are the highest scores.

$$
\begin{aligned}
&B_{MM}(i,j) = \\
&p\min \\
&+B_{MM}(i+1,j+1)PS_{aa}\left(q_{i+1},{}^{i+1i+1}t_{j+1}\right) \\
&\quad *2^{w_{ss}S_{ss}\left(q_{i+1},t_{j+1}\right)} * 2^{w_{sa}S_{sa}\left(q_{i+1},t_{j+1}\right)} \\
&\quad *2^{w_{tors}S_{tors}\left(q_{i+1},t_{j+1}\right)} q_i(M,M)t_j(M,M) \\
&+B_{GD}(i,j+1)t_j(M,D) \\
&+B_{IM}(i,j+1)q_i(M,I)t_j(M,M) \\
&+B_{DG}(i+1,j)q_i(M,D) \\
&+B_{MI}(i+1,j)q_i(M,M)t_j(M,I) \\
&B_{MI}(i,j) = B_{MM}(i+1,j+1)PS_{aa}\left(q_{i+1},{}^{i+1i+1}t_{j+1}\right) \\
&\quad *2^{w_{ss}S_{ss}\left(q_{i+1},t_{j+1}\right)} * 2^{w_{sa}S_{sa}\left(q_{i+1},t_{j+1}\right)} * 2^{w_{tors}S_{tors}\left(q_{i+1},t_{j+1}\right)} \\
&\quad *q_i(M,M)t_j(I,M) + B_{MI}(i+1,j)q_i(M,M)t_j(I,I)
\end{aligned}
\tag{11}
$$

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

Table 2 The average TM-scores and GDT-TS scores of the 3D models generated from the 1,127 pairwise test alignments produced by HHsearch1.2, HHsuite and HMMsato

Method	Average TM-score	Average GDT- TS score
HHsearch (without secondary structure information)	0.527	0.459
HHsearch (with secondary structure information)	0.548	0.479
HHsuite (without secondary structure information)	0.525	0.459
HHsuite (with secondary structure information)	0.543	0.476
HMMsato	**0.555**	**0.483**

Bold numbers are the highest scores.

Table 3 The statistical significance (p-values) of SP and TC score differences between HMMsato and the other two tools on the test data set

Tools	p-value of SP scores	p-value of TC scores
HMMsato – HHsearch (without secondary structure information)	1.078×10^{-6}	3.414×10^{-7}
HMMsato – HHsearch (with secondary structure information)	0.7538	0.8082
HMMsato – HHsuite (without secondary structure information)	1.724×10^{-8}	1.515×10^{-9}
HMMsato – HHsuite (with secondary structure information)	0.1535	0.1087

$B_{IM}(i,j)$ and $B_{GD}(i,j)$ are calculated similarly as $B_{MI}(i,j)$ and $B_{DG}(i,j)$.

Trace back maximum accuracy alignments with the evolutionary residue coupling information

The Evolutionary Coupling (EC) stands for the correlation between two positions or columns in a multiple protein sequence alignment or a protein profile [19,20]. It has recently been employed to predict residue-residue contacts [18,19]. In order to improve profile-profile alignment with the evolutionary coupling information, we calculate the mutual information (MI) (one way of calculating EC value) for any two columns (i, j) of each profile according to Equation (12).

$$
EC_{ij} = MI_{ij} = \sum_{X_i,X_j=1}^{N} F_{ij}(X_i,X_j)\ln\frac{F_{ij}(X_i,X_j)}{F_i(X_i)F_j(X_j)} \tag{12}
$$

N is 21, standing for 20 amino acids plus gap. The joint probability of two residues X_i and X_j ($F_{ij}(X_i, X_j)$) and the probability of residue X_i ($F_i(X_i)$) are calculated in the same way as in [10]. However, EC_{ij} is calculated as the mutual information (MI) instead of the direct information (DI) based on the global probability model [19] in order to achieve the higher time efficiency. A higher EC value corresponds 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 position pairs with higher EC values for each profile are selected. The evolutionary residue coupling information is then applied to check the counterpart pairs during the process of tracing back through the sub-alignment score matrix AS (see Equation (8)) of the MAC alignment. Specifically, we denote the evolutionary coupled position for position i in query q as $k_q(i)$, and the coupled position of position j in template t as $k_t(j)$. Moreover, $M_q(i)$ denotes the position in template t matched with position i in query q when tracing back the original AS matrix, $M_t(j)$ denotes the position in query q matched with

Table 4 The SP scores and TC scores with different values of w_{sa} using HMMsato on the training data

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	42.06	42.18	42.23	42.18	42.20
TC score	40.58	41.25	41.49	41.58	41.73	41.85	41.90	41.85	41.87
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	42.25	42.24	42.29	**42.29**	**42.31***
41.86	41.89	41.89	41.90	41.90	41.92	41.91	41.96	**41.96**	**41.98***
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	42.28	42.27	42.25	42.24	42.20
41.94	41.96	41.94	41.95	41.94	41.94	41.94	41.91	41.91	41.87

Bold denotes the two best scores, and an extra superscript of star denotes the highest score.

position j in template t when tracing back the original AS matrix, and w_{ec} is the weight for the evolutionary coupling information. The new AS' matrix integrating the evolutionary coupling information is recalculated as follows during the track back process.

$$AS'(i,j) = AS(i,j) + w_{ec}(EC(i, M_t(k_t(j))) + EC(M_q(k_q(i)), j))$$

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

$$AS'(i-1,j-1) = AS(i-1,j-1) + w_{ec}(EC(i-1, M_t(k_t(j-1))) + EC(M_q(k_q(i-1)), j-1)) \quad (13)$$

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

Figure 2 illustrates an exampling of taking into account the evolutionary coupling information during the tracing back process to generate the final alignment.

Results and discussion

Evaluation data set and metric

We evaluated HMMsato along with HHSearch [4] and HHsuite on the alignments between 106 targets (queries) of the 9th Critical Assessment of Techniques for Protein Structure Prediction (CASP9) [15,16] and their homologous template proteins (templates) released at the CASP9's web site. The alignment data set has 2,621 pairs of query and template proteins. 1,483 pairs associated with 60 CASP9 targets were used as optimization data set to

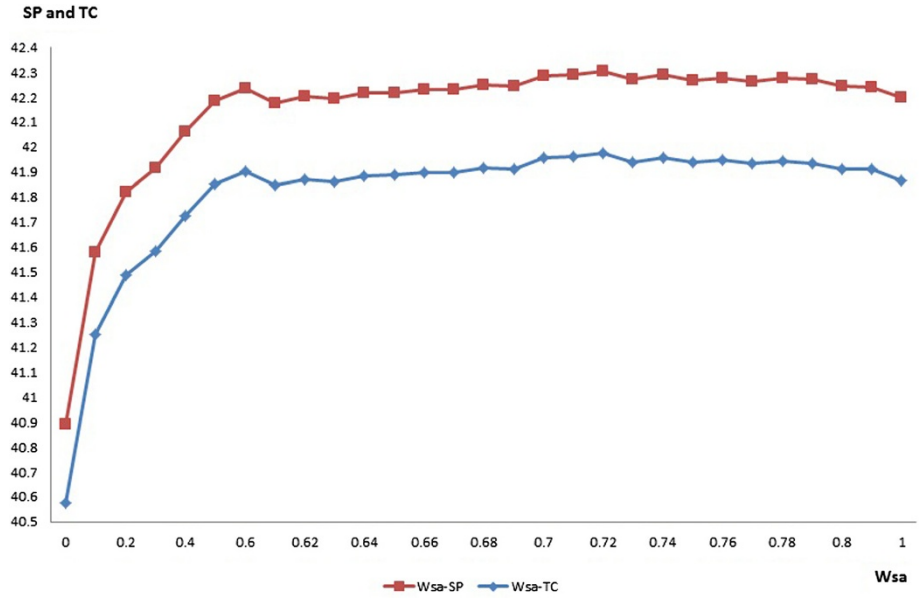

Figure 3 The plot of the SP and TC scores against different values of the weight of solvent accessibility (w_{sa}).

Table 5 The SP scores and TC scores with different values of w_{tors} using HMMsato

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	42.47	42.47	42.47	42.49	42.50
TC score	41.98	41.99	42.02	42.12	42.14	42.14	42.14	42.16	42.16
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	**42.53***	**42.52**	42.49	42.50	42.50	42.51
42.17	42.17	42.17	42.18	**42.19***	**42.19**	42.15	42.16	42.17	42.17
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	42.50	42.46	42.45	42.40	42.46	42.40
42.17	42.16	42.17	42.17	42.17	42.13	42.12	42.07	42.13	42.07

Bold denotes the two best scores, and an extra superscript of star denotes the highest score.

optimize the parameters of HMMsato, and 1,138 pairs associated with the remaining 46 CASP9 targets were used to test the methods. The reference (presumably true) pairwise alignments of a query-template protein pair was generated by using TMalign [25] to align the tertiary (3D) structures of the two proteins together. The alignments generated by HMMsato and other tools were evaluated by three metrics, including sum-of-pairs (SP) score, true column (TC) score, and the quality of the tertiary structural models of the query proteins built from the alignments. The SP and TC scores are the two standard metrics for evaluating sequence alignment quality [26]. The quality of tertiary structural models indirectly assesses the quality of sequence alignments according to their effectiveness in guiding the construction of protein structural models.

The SP score is the number of correctly aligned pairs of residue in the predicted alignment divided by the total number of aligned pairs of residues in the core blocks (i.e., sequence alignment regions precisely determined by structural alignment of structurally equivalent residues in the structures of two proteins) of the true alignment [23]. The TC score is the number of correctly aligned

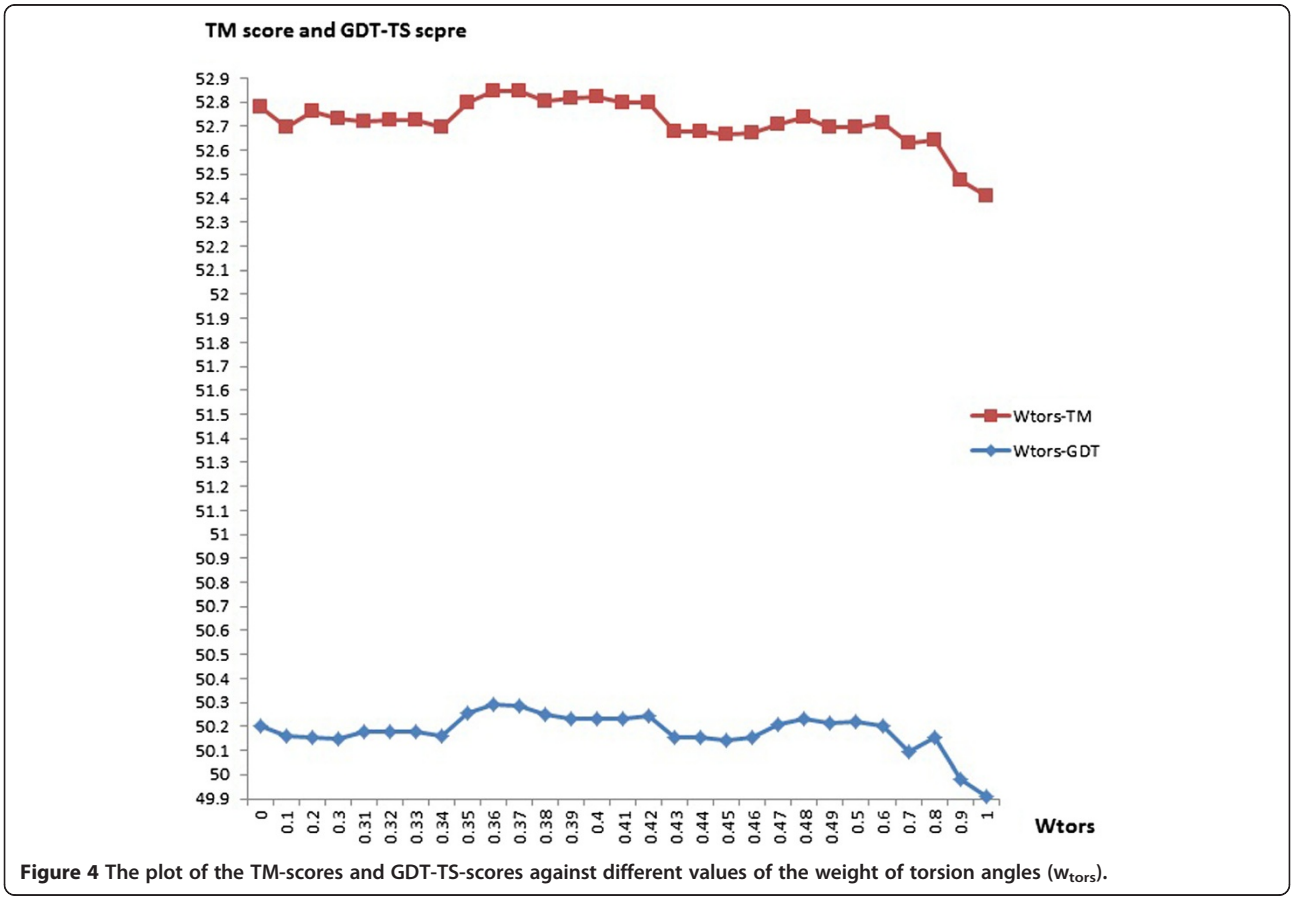

Figure 4 The plot of the TM-scores and GDT-TS-scores against different values of the weight of torsion angles (w_{tors}).

columns in the core blocks of the true alignment [27]. The 3D model of a query protein was produced by MODELLER [28] based on both the pairwise alignment generated by an alignment method and the known structure of the template protein in the alignment. We used TM-Score [29] to align a 3D model of a query protein against its true structure to generate TM-scores and GDT-TS scores [30] for the model in order to measure the quality of the alignment used to generate the model, assuming better alignments lead to better 3D models with higher TM-scores and GDT-TS scores. Both TM-score and GDT-TS score are in the range [0, 1] [31].

Optimization of weights for the solvent accessibility, torsion angles and evolutionary coupling information

We estimated the weights of the solvent accessibility, torsion angles and evolutionary residue coupling information on the training alignments step by step. Firstly, we found the best weight value (w_{sa} = 0.72) for solvent accessibility. Then, we identified the best weight value (w_{tors} = 0.4) for torsion angles while keeping the weight for solvent accessibility fixed. Finally, we found the best parameter value (w_{ec} = 0.1) for the evolutionary residue coupling information by keeping w_{sa} and w_{tors} at their optimum values. HHsearch and HHsuite were both evaluated with and without secondary structure information. The default parameter values were used with HHsearch and HHsuite.

Comparison of HMMsato, HHSearch, and HHsuite on the test data set

The mean SP and TC scores for the pairwise alignment results generated by HMMsato, HHSearch and HHsuite for 1,138 protein pairs are reported in Table 1. The mean SP score and the mean TC score of HMMsato are 50.39 and 50.02 respectively, higher than HHsearch and HHsuite with or without secondary structure information. The average TM-scores and GDT-TS scores of the 3D models successfully generated from 1,127 out of 1,138 alignments by MODELLER were listed in Table 2. The average TM-score and GDT-TS score of the models generated from the HMMsato alignments are 0.555 and 0.483, respectively, better than those of HHSearch and HHsuite. Furthermore, we carried out the Wilcoxon matched-pair signed-rank test on both SP and TC scores of the three methods on the

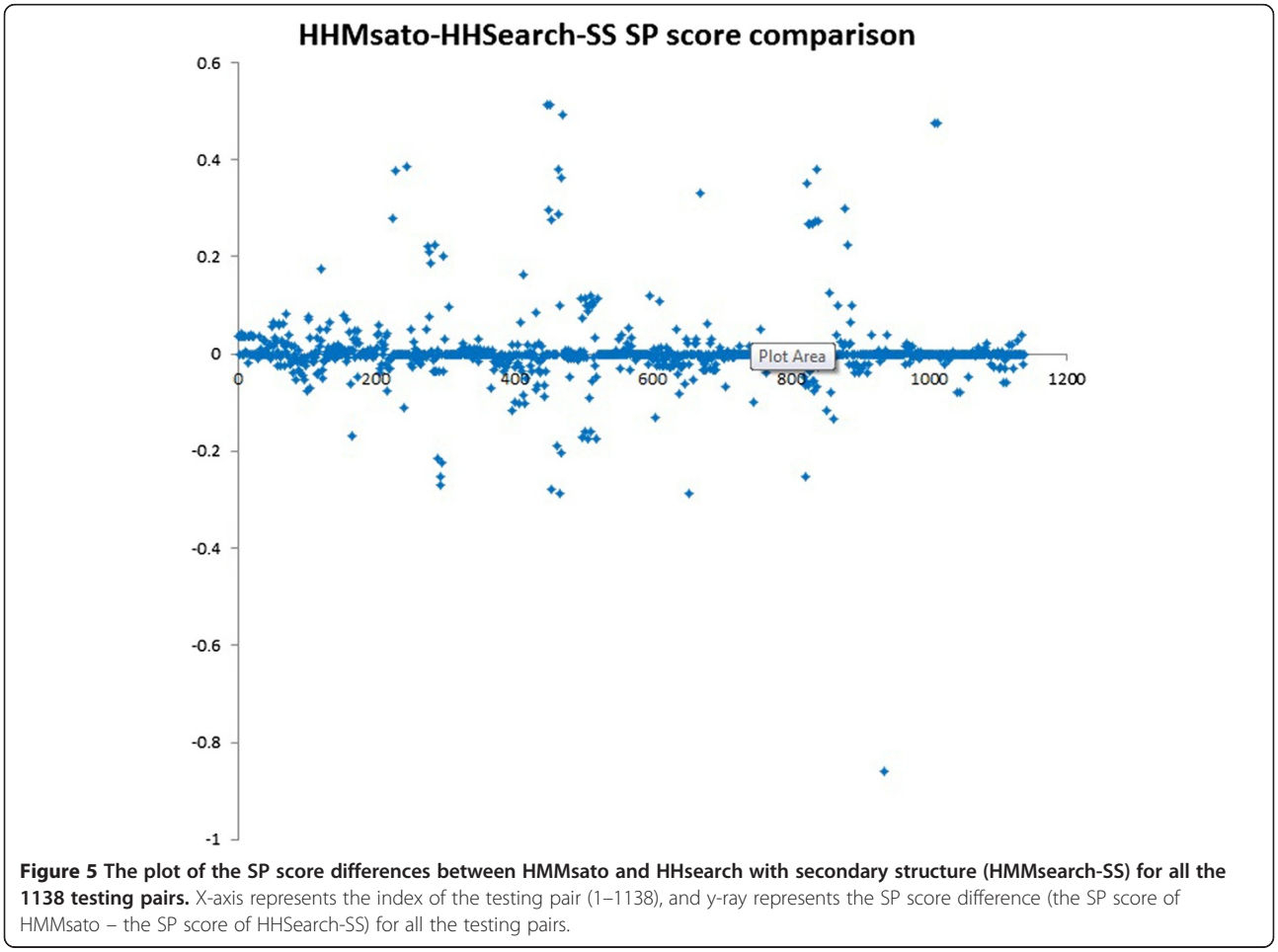

Figure 5 The plot of the SP score differences between HMMsato and HHsearch with secondary structure (HMMsearch-SS) for all the 1138 testing pairs. X-axis represents the index of the testing pair (1–1138), and y-ray represents the SP score difference (the SP score of HMMsato – the SP score of HHSearch-SS) for all the testing pairs.

test data set. The p-values of alignment score differences between HMMsato and the other methods calculated by the Wilcoxon matched-pair signed-rank test are reported in Table 3.

Impact of solvent accessibility, torsion angles and evolutionary coupling information on the alignment accuracy

We studied the effect of the solvent accessibility information by solely adjusting the value of its weight w_{sa}. The SP scores and TC scores of the alignments generated by HMMsato with different w_{sa} values on the training data set are shown in Table 4. The results show that incorporating the solvent accessibility information always improves alignment accuracy in comparison with the baseline not using solvent accessibility information ($w_{sa} = 0$). The highest accuracy is achieved when w_{sa} is set to 0.72. Figure 3 shows the plot of SP scores/TC scores against the different values of $w_{sa.}$ Red curve represents the SP scores and blue represents the TC scores.

We studied the effect of torsion angles on alignments by solely adjusting the value of w_{tors} (weight for torsion angle information) while keeping w_{sa} as 0.72. The SP scores and TC scores of the alignments generated by HMMsato with different w_{tors} values on the training data set are shown in Table 5. The results show that incorporating the torsion angle information also helps improve alignment accuracy. The highest accuracy is achieved when w_{tors} is set to 0.4. Figures 4 shows the TM-scores and GDT-TS scores of the 3D models constructed from the alignments generated by HMMsato with both torsion angles and solvent accessibility with respect to different w_{tors} values.

The effect of evolutionary residue coupling information on alignment accuracy

We studied the effect of the evolutionary residue coupling information on alignment accuracy in a similar way. HMMsato worked the best when w_{ec} was 0.1. However, the evolutionary coupling information did not improve the overall alignment accuracy on the training data set, probably due to lack of a large number of diverse sequences in many cases required by the evolutionary coupling calculation to obtain the sufficient discriminative power. Specifically speaking, the alignment quality increased in 57 alignments, stayed the same in 1363 alignments, but decreased in 61 alignments. Similarly, on the test data set, the alignment quality increased in 59 alignments, stayed the same in 1024 alignments, but decreased in 55 alignments. Generally speaking, the evolutionary coupling information contributed to the improvement of alignment accuracy in some cases, but its effect was rather inconsistent.

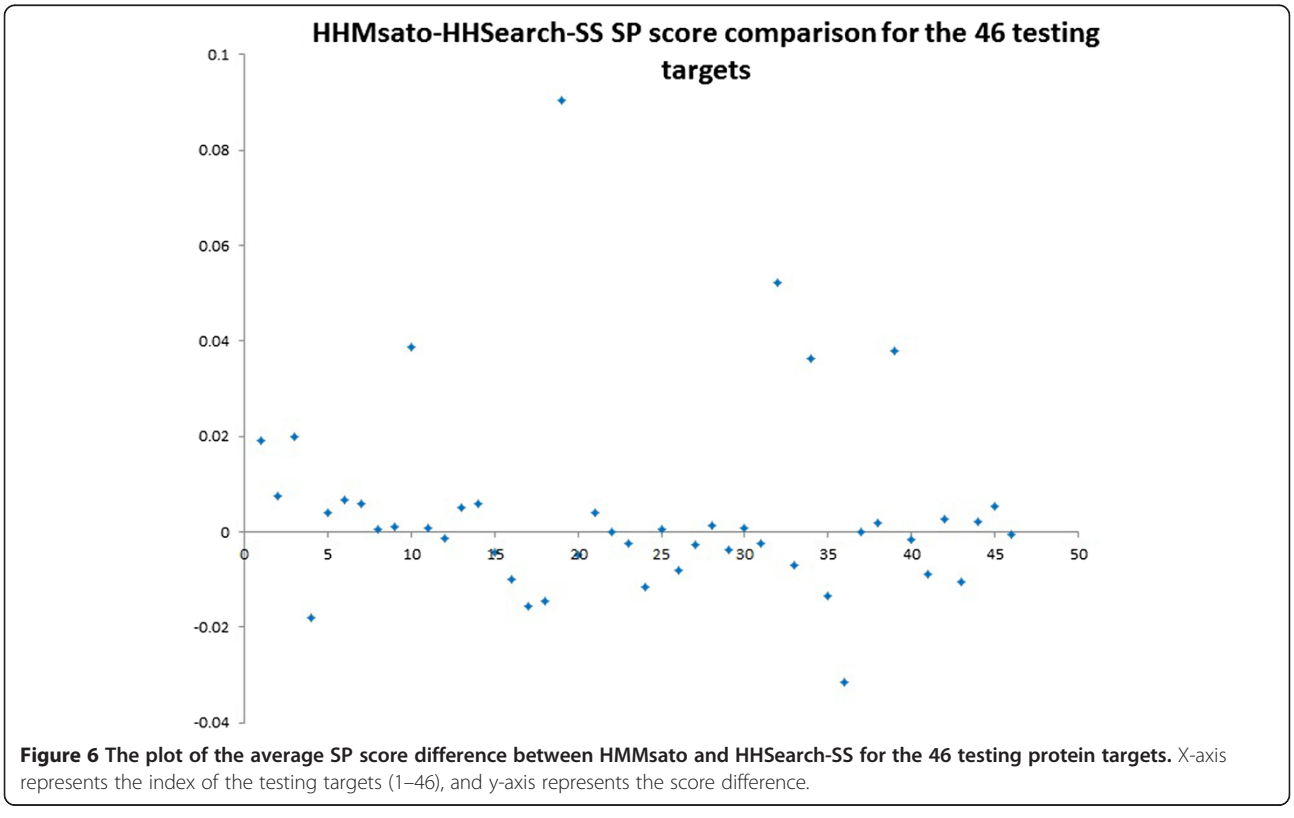

Figure 6 The plot of the average SP score difference between HMMsato and HHSearch-SS for the 46 testing protein targets. X-axis represents the index of the testing targets (1–46), and y-axis represents the score difference.

Comparison of HMMsato and HHSearch with secondary structure information on the test data set

We studied the SP score differences between HMMsato and HHSearch with secondary structure for all the 1138 testing pairs. The plot of the SP score difference (SP score of HMMsato minus SP score of HHSearch) for these pairs is shown in Figure 5. Similarly, the plot of the average SP score difference between HMMsato and HHSearch-SS for the 46 testing protein targets is shown in Figure 6. X-axis represents the index of the testing targets (1–46), and y-axis represents the score difference. Specifically, the alignment quality increased for 24 targets, stayed the same for 2 targets, but decreased for 20 targets. We found that HMMsato often improved the alignment quality for proteins of length ranging from 70 to 450 residues.

Conclusion

We designed a method to incorporate relative solvent accessibility, torsion angles and evolutionary residue coupling information into HMM-based pairwise profile-profile protein alignments. Our experiments on the large CASP9 alignment data set showed that utilizing solvent accessibility and torsion angles improved the accuracy of HMM-based pairwise profile-profile alignments. However, the effect of the evolutionary residue coupling information on alignments is less consistent according to our current experimental setting, even though it may still be a valuable source of information to explore in the future. Particularly, we will use the latest method (i.e., direct information) of calculating evolutionary coupling information to guide the profile alignment process. Furthermore, we will carry out more extensive search of optimal weights for solvent accessibility, torsion angle, secondary structure, and evolutionary coupling information to improve alignment accuracy.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JC and XD designed the project. XD implemented and tested the method. XD and JC wrote the manuscript. XD and JC read and approved the manuscript.

Acknowledgements

The work was partially supported by a NIH R01 grant (R01GM093123) to JC.

Author details

[1]LexisNexis | Risk Solutions | Healthcare, Orlando, FL 32811, USA. [2]Computer Science Department, Informatics Institute, C. Bond Life Science Center, University of Missouri-Columbia, Columbia, MO 65211, USA.

Received: 7 January 2014 Accepted: 17 July 2014
Published: 25 July 2014

References

1. Kinch LN, Wrabl JO, Krishna S, Majumdar I, Sadreyev RI, Qi Y, Pei J, Cheng H, Grishin NV: **CASP5 assessment of fold recognition target predictions.** *Proteins: Structure, Function, and Bioinformatics* 2003, **53**(S6):395–409.
2. Bork P, Koonin EV: **Predicting functions from protein sequences—where are the bottlenecks?** *Nat Genet* 1998, **18**(4):313–318.
3. Henn-Sax M, Höcker B, Wilmanns M, Sterner R: **Divergent evolution of (βα) 8-barrel enzymes.** *Biol Chem* 2001, **382**(9):1315–1320.
4. Söding J: **Protein homology detection by HMM–HMM comparison.** *Bioinformatics* 2005, **21**(7):951–960.
5. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.** *Nat Methods* 2011, **9**:173–175.
6. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.
7. Mott R: **Smith–Waterman algorithm.** *eLS* 2005. http://onlinelibrary.wiley.com/doi/10.1038/npg.els.0005263/abstract.
8. Holmes I, Durbin R: **Dynamic programming alignment accuracy.** *J Comput Biol* 1998, **5**(3):493–504.
9. Hughey R, Karplus K, Krogh A: *SAM: Sequence alignment and modeling software system. Technical Report UCSC-CRL-99-11.* Santa Cruz, CA 95604: Baskin Center for Computer Engineering and Science, University of California; 2003.
10. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res* 2011, **39**(suppl 2):W29–W37.
11. Ginalski K, Pas J, Wyrwicz LS, Von Grotthuss M, Bujnicki JM, Rychlewski L: **ORFeus: detection of distant homology using sequence profiles and predicted secondary structure.** *Nucleic Acids Res* 2003, **31**(13):3804–3807.
12. Tang CL, Xie L, Koh IYY, Posy S, Alexov E, Honig B: **On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles.** *J Mol Biol* 2003, **334**(5):1043–1062.
13. Tomii K, Akiyama Y: **FORTE: a profile–profile comparison tool for protein fold recognition.** *Bioinformatics* 2004, **20**(4):594–595.
14. Söding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction.** *Nucleic Acids Res* 2005, **33**(suppl 2):W244–W248.
15. Kryshtafovych A, Fidelis K, Moult J: **CASP9 results compared to those of previous CASP experiments.** *Proteins: Structure, Function, and Bioinformatics* 2011, **79**(S10):196–207.
16. Kryshtafovych A, Fidelis K, Moult J: **CASP10 results compared to those of previous CASP experiments.** *Proteins: Structure, Function, and Bioinformatics* 2013, **82**(S2):164–174.
17. Hildebrand A, Remmert M, Biegert A, Söding J: **Fast and accurate automatic structure prediction with HHpred.** *Proteins: Structure, Function, and Bioinformatics* 2009, **77**(S9):128–132.
18. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C: **Protein 3D structure computed from evolutionary sequence variation.** *PLoS One* 2011, **6**(12):e28766.
19. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS: **Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing.** *Cell* 2012, **149**(7):1607–1621.
20. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features.** *Biopolymers* 1983, **22**(12):2577–2637.
21. Cheng J, Li J, Wang Z, Eickholt J, Deng X: **The MULTICOM toolbox for protein structure prediction.** *BMC Bioinformatics* 2012, **13**(1):65.
22. Faraggi E, Yang Y, Zhang S, Zhou Y: **Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction.** *Structure* 2009, **17**(11):1515–1527.
23. Zhang W, Liu S, Zhou Y: **SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model.** *PLoS One* 2008, **3**(6):e2325.
24. Biegert A, Söding J: **De novo identification of highly diverged protein repeats by probabilistic consistency.** *Bioinformatics* 2008, **24**(6):807–814.
25. Zhang Y, Skolnick J: **TM-align: a protein structure alignment algorithm based on the TM-score.** *Nucleic Acids Res* 2005, **33**(7):2302–2309.
26. Thompson JD, Koehl P, Ripp R, Poch O: **BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark.** *Proteins: Structure, Function, and Bioinformatics* 2005, **61**(1):127–136.
27. Deng X, Cheng J: **MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts.** *BMC Bioinformatics* 2011, **12**:472.

28. Eswar N, Webb B, Marti-Renom MA, Madhusudhan M, Eramian D, Shen M-y, Pieper U, Sali A: **Comparative Protein Structure Modeling Using Modeller.** *Curr Protoc Bioinformatics* 2006. **15**(5.6):5.6.1–5.6.30.
29. Xu J, Zhang Y: **How significant is a protein structure similarity with TM-score = 0.5?** *Bioinformatics* 2010, **26**(7):889–895.
30. Zemla A, Venclovas Č, Moult J, Fidelis K: **Processing and analysis of CASP3 protein structure predictions.** *Proteins: Structure, Function, and Bioinformatics* 1999, **37**(S3):22–29.
31. Zhang Y, Skolnick J: **Scoring function for automated assessment of protein structure template quality.** *Proteins: Structure, Function, and Bioinformatics* 2004, **57**(4):702–710.

doi:10.1186/1471-2105-15-252
Cite this article as: Deng and Cheng: Enhancing HMM-based protein profile-profile alignment with structural features and evolutionary coupling information. *BMC Bioinformatics* 2014 **15**:252.