BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 18 2011, pages 2537–2545
doi:10.1093/bioinformatics/btr432

Structural bioinformatics
Advance Access publication July 26, 2011

Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling

Sheng Wang, Jian Peng and Jinbo Xu*
Toyota Technological Institute at Chicago, IL 60637, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Building an accurate alignment of a large set of distantly related protein structures is still very challenging.

Results: This article presents a novel method 3DCOMB that can generate a multiple structure alignment (MSA) with not only as many conserved cores as possible, but also high-quality pairwise alignments. 3DCOMB is unique in that it makes use of both local and global structure environments, combined by a statistical learning method, to accurately identify highly similar fragment blocks (HSFBs) among all proteins to be aligned. By extending the alignments of these HSFBs, 3DCOMB can quickly generate an accurate MSA without using progressive alignment. 3DCOMB significantly excels others in aligning distantly related proteins. 3DCOMB can also generate correct alignments for functionally similar regions among proteins of very different structures while many other MSA tools fail. 3DCOMB is useful for many real-world applications. In particular, it enables us to find out that there is still large improvement room for multiple template homology modeling while several other MSA tools fail to do so.

Availability: 3DCOMB is available at http://ttic.uchicago.edu/~jinbo/software.htm.

Contact: jinboxu@gmail.com

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 22, 2011; revised on June 30, 2011; accepted on July 16, 2011

1 INTRODUCTION

Multiple protein alignment has been extensively used for classification, analysis of evolutionary relationship, motif detection and structure/function prediction. When proteins are distantly related, sequence methods usually fail to yield accurate alignment. In contrast, structure methods, which exploit geometrical information, may still work well. As more protein structures are experimentally solved, multiple structure alignment (MSA) is becoming more useful and important. However, developing computational methods for accurate MSA, especially of a large set of distantly related protein structures, is still regarded as an open challenge. An MSA method consists of two critical components: a scoring function measuring of the quality of an MSA and an algorithm searching for the MSA optimizing the score.

Many MSA algorithms have been developed, such as MAMMOTH (Lupyan *et al.*, 2005), MATT (Menke *et al.*, 2008), MultiProt (Shatsky *et al.*, 2004), MUSTANG (Konagurthu *et al.*, 2006), POSA (Ye and Godzik, 2005) and SALIGN (Eswar *et al.*, 2008; Madhusudhan *et al.*, 2009). These methods can be broadly divided into two groups: '*horizontal-first*' or '*vertical-first*'. The former progressively merges pairwise alignments into an MSA, which may be suboptimal since pairwise alignment errors carry over to the final result. The vertical-first methods identify some *similar fragment blocks* (SFB) among proteins and then extend the SFB alignments to MSAs. The number of SFBs could grow exponentially with respect to the number of proteins, so these methods may have to examine a large number of SFBs not to miss the best MSA, which is usually computationally expensive. As such, the challenge facing a vertical-first method is to identify only those SFBs, which are very likely contained in the best MSA.

This article presents 3DCOMB, a novel vertical-first method for MSA. 3DCOMB distinguishes itself from others in its search algorithm and scoring function. 3DCOMB makes use of both local and global structure environments, combined by a novel machine learning method, to accurately identify *highly similar fragment blocks* (HSFBs), which are very likely contained in the best MSA. 3DCOMB searches for the best MSA starting from the alignment of an HSFB. *Local structure environment* is used to describe the structural difference between two short segments centered on two residues. *Global structure environment* is used to measure the similarity of two substructures centered at the two residues. Both features are combined by a probabilistic graphical model Conditional Random Fields (CRF) (Lafferty *et al.*, 2001) to determine if two residues shall be aligned or not. Existing methods [e.g. BLOMAPS (Wang and Zheng, 2009)] employ only local structure environment and sometimes fail to detect good HSFBs. They usually have to examine a large number of not-so-conserved fragment blocks in order to find the best MSA. In contrast, 3DCOMB can generate accurate MSAs from only very few HSFBs and thus, improve accuracy without too much computational time.

Many methods do not explicitly take into consideration the quality of pairwise alignments in building an MSA, so it may not necessarily contain high-quality pairwise alignments. 3DCOMB aims to generate an MSA with not only as many conserved cores as possible, but also accurate pairwise alignments. A core is a fully aligned column without any gaps, consisting of one residue from each input protein. 3DCOMB achieves this by employing a novel scoring function CORE-LEN $\times \overline{\text{TMscore}}$ for MSA where

*To whom correspondence should be addressed.

CORE-LEN is the number of cores in the MSA. TMscore (Zhang and Skolnick, 2004), ranging from 0 to 1, is a widely used measure for pairwise structure similarity. The higher the TMscore, the more similar two protein structures are. $\overline{\text{TMscore}}$ is the TMscore averaged over all the pairwise alignments and thus, analogous to 'sum-of-pairs' used for MSA. The strength of this scoring function will be detailed in Section 2.

Our tests show that 3DCOMB generates alignments with not only better CORE-LEN and $\overline{\text{TMscore}}$, but also smaller core RMSD although it is not explicitly optimized. 3DCOMB significantly excels others for the alignment of a large group of distantly related proteins. Although not specifically designed for this, 3DCOMB can also generate correct alignments for functionally similar sites among proteins not in the same superfamily while many other MSA tools fail. We also estimate the gap between 3DCOMB alignments and the best possible. When proteins are closely related, 3DCOMB alignments are almost the best possible. Otherwise they may still have a gap from the best possible. 3DCOMB also helps us find out that there is still large improvement room for homology modeling, but several other MSA tools fail to do so.

2 METHODS

This section contains (i) a machine learning method to detect HSFBs, using both local and global structure information; and (ii) a new scoring function and an optimization algorithm to search for the best MSA starting from the HSFBs.

2.1 HSFB

We use both local and global structure environments to determine how likely two residues should be aligned. To the best of our knowledge, no previous methods consider the global structure environment. The *local structure environment* of a residue consists of 10 structure segments of length $2k+1 (k=1,2,\dots,10)$ centered at this residue. Given two residues of two different proteins, the similarity of their local structure environments is measured by 10 TMscore values, each measuring the similarity of two structure segments of the same size. The *global structure environment* of a residue i is defined as follows. Let $N(i,d)$ denote all the residues in the same protein within distance cutoff d from i. The global structure environment of i (under a given distance cutoff d) consists of all the 5mer segments centered at the residues in $N(i,d)$. Given two residues of different proteins, to calculate the similarity of their global structure environments, we first generate a rigid body transformation for them by minimizing the RMSD of two center 5mer segments. Then we fix this transformation and run dynamic programming to align them by maximizing TMscore. We generate 10 such global structure environment features, using 10 distance cutoff values 6, 7, …, 15 Å.

Modeling pairwise structure alignment using CRF: CRFs are probabilistic graphical models that have been applied to protein secondary structure prediction (Wang *et al.*, 2010), protein conformation sampling (Zhao *et al.*, 2010a; Zhao *et al.*, 2010b), protein sequence alignment (Do *et al.*, 2006) and protein threading (Peng and Xu, 2010). Here, we use CRF to model protein structure alignment, which then is used to identify HSFBs among proteins under consideration.

Let p, q denotes two input proteins and their associated structural features (i.e. local and global structure environment similarity scores). Let $X=\{M,I_p,I_q\}$ be a set of three possible alignment states. Meanwhile, M indicates that two residues are aligned and I_p and I_q indicate insertion at proteins p and q, respectively. Let $A=\{a_1,a_2,\dots,a_{NA}\}$ denote an alignment between p and q where $a_i \in X$ represents the state (or label) at position i and N_A denotes the length of this alignment. Our CRF model defines the conditional probability of an alignment A given p and q as follows:

$$P_\theta(A|p,q)=\frac{\exp\left(\sum_{i=1}^{N_A} F(A,p,q,i)\right)}{Z(p,q)} \tag{1}$$

$$Z(p,q)=\sum_{A'}\exp\left(\sum_{i=1}^{N_{A'}} F(A',p,q,i)\right) \tag{2}$$

and $\theta=\{\lambda_1,\lambda_2,\dots,\lambda_d\}$ is the model parameter and $F(A,p,q,i)$ is a function estimating the log-likelihood of an alignment at position i:

$$F(A,p,q,i)=\sum_k \lambda_k e_k(a_{i-1},a_i)+\sum_l \lambda_l v_l(a_i,p,q,i) \tag{3}$$

where $e_k(a_{i-1},a_i)$ and $v_l(a_i,p,q,i)$ are called edge and label feature functions, respectively. The edge features model the dependency of the state transition from alignment position $i-1$ to i. Here, we assume $e_k(a_{i-1}, a_i)$ is independent of protein features to make our formulation simple. $e_k(a_{i-1}, a_i)$ is equal to 1 if the transition $a_{i-1} \rightarrow a_i$ exists at position i; otherwise 0. We forbid transition from I_q to I_p, so there are in total eight state transitions. That is, k ranges from 1 to 8. The label features model the relationship between a_i and the local and global structure environment similarity scores at the alignment position i. There are in total 20 different label feature functions, each corresponding to one local or global structure environment similarity score. Therefore, l ranges from 1 to 20. To make the formulation simple, we assume an insertion state is independent of protein features, so modeling of insertions is implicitly taken into consideration in the edge feature functions.

We train the model parameters using a set of reference alignments taken from FSSP (Holm and Sander, 1994). In particular, we randomly selected 50 pairs for training and the other 50 pairs for test. The training and test sets have no overlap with our benchmarks. All the structural alignments (i.e. reference alignments) were generated by DALI (Holm and Sander, 1993) and each alignment has a DALI Z-score >8.0. The model parameters are initialized randomly to a value between 0 and 1 and the training process converges within ~100 iterations. Five-fold cross-validation is conducted to determine the regularization factor in the CRF model. Once the model parameters are determined, for a given protein pair p and q, we can generate their alignment by maximizing $P_\theta(A|p,q)$. This alignment may not be the best, but can be used as an initial alignment between p and q. Starting from these initial alignments, we can build very accurate pairwise alignments (Supplementary Material).

As shown in Supplementary Figure S1, the global features corresponding to distance cutoffs 7, 8, 14 and 15 Å have relatively large weight factors. This is consistent with the findings in da Silveira *et al.* (2009), which shows that 7 Å is the best cutoff for distance-based contact definition. Both 14 and 15 Å may be interpreted as the distance cutoffs for second-order contacts. To speed up, we exclude six local structure environment features and eight global structure environment features with small weight factors from our final CRF model. By using only six features, we will not lose much accuracy in identifying HSFBs. The remaining two global features correspond to global structure environments at radius 8.0 and 14.0 Å. The remaining four local features correspond to structure segments with lengths 9, 13, 17 and 21, respectively.

Detecting HSFBs using CRF: given two protein structures, we can calculate the marginal probability of two fragments being aligned using the forward–backward algorithm (Lafferty *et al.*, 2001). In this article, we consider only fragments of length $L=12$. Such a fragment is likely to cover at least one secondary structure segment. A slight change of L will not impact the final result much. This marginal probability is defined as the similarity score of two structure fragments. Given a short fragment F_1 in protein p_1 and another protein p_i, let F_i denote the fragment of the same size in p_i, which has the highest similarity score with F_1. All the fragments $F_1,F_2,\dots,F_M$ form an HSFB with F_1 being the pivot fragment. A protein of size N in total have $N-L+1$ HSFBs. Note that the 'highest similarity' relationship is asymmetric. That is, among all fragments in protein p_i, F_i is the most similar

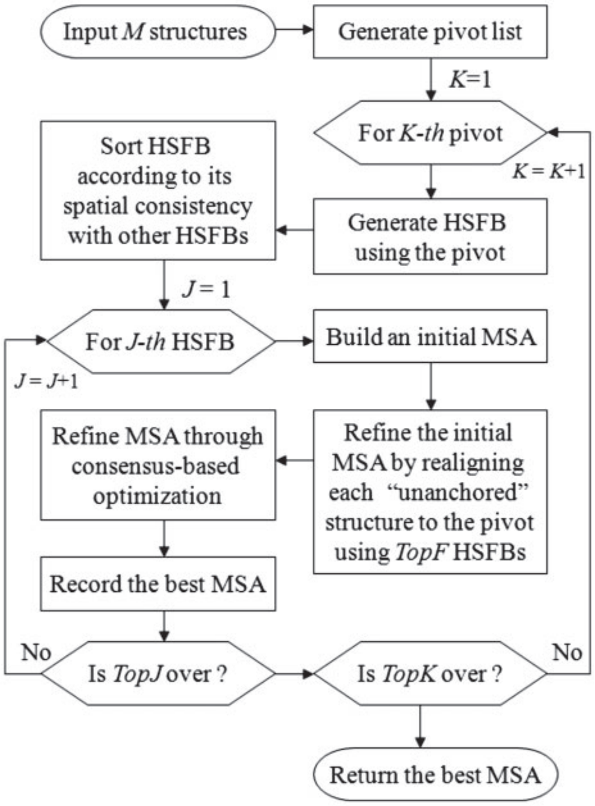
Input *M* structures
Generate pivot list
K=1
For *K-th* pivot
K = *K*+1
Sort HSFB according to its spatial consistency with other HSFBs
Generate HSFB using the pivot
J = 1
For *J-th* HSFB
Build an initial MSA
J = *J*+1
Refine MSA through consensus-based optimization
Refine the initial MSA by realigning each "unanchored" structure to the pivot using *TopF* HSFBs
Record the best MSA
No
Is *TopJ* over ?
Is *TopK* over ?
No
Return the best MSA

Fig. 1. The 3DCOMB algorithm overview.

one to F_1 may not imply that among all fragments of p_1, F_1 also is the most similar one to F_i. Therefore, given M proteins with lengths $N_1, N_2, \ldots, N_M$, there are in total $(\sum_{i=1}^{M} N_i) - \mathrm{ML} + M$ HSFBs.

Ranking HSFBs by spatial consistency: two HSFBs may be geometrically inconsistent with each other. That is, we cannot superimpose well the fragments in both HSFBs using a single set of rigid body transformations. We can calculate the degree to which two HSFBs are geometrically consistent and then rank all the HSFBs according to their consistency with others. The HSFB with the highest consistency score is very likely contained in the best MSA. We use a simple method to estimate the consistency score of one HSFB as follows. For each fragment in the HSFB, we generate a rigid body transformation by minimizing the RMSD between this fragment (Kabsch, 1976) and the pivot fragment. Let $B_1 = \{F_{11}, F_{12}, \ldots, F_{1M}\}$ denote the HSFB for which we want to calculate its consistency score and F_{11} is the pivot fragment. Let $T_i (i = 2, 3, \ldots, M)$ denote the rigid body transformations derived from superimposing F_{1i} to F_{11}. Let $B_2 = \{F_{21}, F_{22}, \ldots, F_{2M}\}$ denote another HSFB. For any fragment $F_{2i} (i = 2, 3, \ldots, M)$ in B_2, we superimpose F_{2i} to F_{21} using the transformation T_i and then calculate the RMSD between F_{2i} and F_{21}. If the distance is within 3 Å, we increase the consistency score of B_1 by 1, otherwise by 0.

2.2 Algorithm for MSA

Overview: as shown in Figure 1, 3DCOMB first generates a list of pivot structures. By default, this list contains all the input proteins, so *TopK* is equal to M. For each pivot structure, 3DCOMB uses the CRF model to generate HSFBs, which are ranked by their spatial consistency scores and only the *TopJ* with the highest scores are extended to initial MSAs. 3DCOMB identifies those 'unanchored' proteins which are not well aligned to the pivot. To improve an initial MSA, 3DCOMB conducts *TopF* trials to realign each of the 'unanchored' proteins to the pivot. Finally, 3DCOMB refine the whole MSA based on the consensus structure derived from the MSA.

Scoring function: a good MSA should have a large number of cores (i.e. CORE-LEN) and also a small core RMSD. A core is a fully aligned column, consisting of one residue from each input protein. In addition, pairwise alignments in an MSA should also be of high quality. It is challenging to develop an algorithm that can optimize these criteria simultaneously since sometimes they contradict with one another. For example, a large CORE-LEN usually leads to a large RMSD value. A simple solution is to fix one criterion and then optimize the others, e.g. maximizing CORE-LEN while restricting RMSD. This solution is not very flexible in that we have to determine RMSD in advance, not to mention that neither CORE-LEN nor RMSD is the best measure.

We use CORE-LEN × $\overline{\text{TMscore}}$ as the scoring function where $\overline{\text{TMscore}}$ of an MSA is defined as the average TMscore of all the pairwise alignments implied in the MSA. TMscore is widely used to measure the pairwise structure similarity and the quality of a protein model. It is defined as follows:

$$\text{TMscore}(p,q) = \frac{1}{L_S} \sum_{i}^{\text{Lali}} \frac{1}{1 + \left(d_i / d_0(L_S)\right)^2} \tag{4}$$

$$d_0(L_S) = 1.24 \times \sqrt[3]{L_S - 15} - 1.8 \tag{5}$$

Meanwhile, Lali is the alignment between protein p and q, L_S is the shorter protein length, d_i is the deviation of two aligned residues and d_0 is the length-related normalization term.

Our scoring function has the following features: (i) it leads to an MSA with not only a large number of conserved cores, but also high-quality pairwise alignments; (ii) the distance between two aligned residues is used as denominator, so it favors the aligned residue pairs within small distance and disfavors or even ignores those with large deviation. This enables us to detect even a small conserved region among proteins of very different structures. In contrast, RMSD for the whole alignment, even normalized by alignment length, is (Levitt and Gerstein, 1998; Siew *et al.*, 2000) greatly affected by a small number of badly aligned residue pairs and not sensitive in detecting small but conserved regions; (iii) TMscore is also better than the alignment length since the latter does not take into consideration the distance deviation of aligned residue pairs; (iv) TMscore is almost independent of the protein length (Zhang and Skolnick, 2004) since the distance at each position is normalized by a length-dependent factor. This is particularly useful for the alignment of a large set of proteins with very different lengths; (v) TMscore is not only good for alignment, but also very sensitive in detecting fold-level similarity. As reported in Xu and Zhang (2010), when TMscore >0.6, it is very likely (90% of chance) that two proteins have the same fold. When TMscore <0.4, it is very likely (90% of chance) that two proteins have different folds. When TMscore >0.5, there is 50% of chance that two proteins have similar folds.

Building an initial MSA from an HSFB: given an HSFB of M structures, we first generate a set of $M-1$ rigid body transformations by superimposing each fragment in the HSFB to the pivot fragment and minimizing the RMSD of these two fragments. Then we superimpose each structure to the pivot structure using the transformation generated from fragment superimposition, and run dynamic programming to generate an alignment of the two structures by maximizing the TMscore. Finally, we assemble the $M-1$ pairwise alignments into an initial MSA using the pivot structure as the anchor.

Adjustment of pairwise alignment: given the initial MSA, we may refine it by adjusting the pairwise alignment between each input structure and the pivot structure (Wang and Zheng, 2008). First, we calculate the TMscore of the pairwise alignment between each input structure and the pivot. If TMscore <0.5 (Xu and Zhang, 2010), this input structure is called 'unanchored'. We adjust the alignment between each unanchored structure and the pivot using rigid body transformations derived from other *TopF* HSFBs. In particular, for each top HSFB, let F_1 and F_2 denote the fragments in the HSFB belonging to the pivot and the unanchored structure, respectively. We realign the unanchored structure to the pivot structure using the rigid body transformation generated from minimizing the RMSD between F_1 and F_2. The pairwise alignment with the maximum TMscore is kept in the MSA.

Consensus-based MSA refinement: 3DCOMB refines an MSA by realigning each input structure to the consensus structure, which is constructed as follows. At each column of this MSA, we calculate the center of all the aligned residues (only Cα is considered). Second, we merge two neighbor columns into a single one if the following two conditions are satisfied: (i) the total number of aligned residues in these two columns is not more than the total number of input structures; and (ii) the distance between their two centers is <3.0 Å. We use 3.0 Å as the cutoff because in native protein structures, $>99\%$ of Cα–Cα virtual bonds are >3.0 Å. This merge procedure is repeated until no columns can be merged. The consensus structure consists of all the centers. This refinement procedure is repeated until a given number of iterations or the scoring function cannot be improved further.

2.3 3DCOMB time complexity

Let N denote the maximum length of the input M protein structures. We analyze the time complexity of 3DCOMB as follows:

Step 1: Generate HSFB: given one pivot protein and another one, it takes time $O(N^2)$ to generate protein features and run the forward–backward algorithm to calculate all the marginal probabilities. Afterwards, given one fragment F in the pivot, it takes time $O(N)$ to detect a fragment on another protein with the highest similarity score to F. There are M proteins, so the time complexity of HSFB generation for a given pivot protein is $O(MN^2)$.

Step 2: Sort HSFB by spatial consistency: it takes $O(M)$ time to generate $M-1$ rigid body transformations to align the fragments in one HSFB. It takes $O(M)$ time to calculate the spatial consistency score of one HSFB with another HSFB. So the total time complexity for a given pivot protein is $O(MN^2)$.

Step 3: Building an initial MSA from an HSFB: it takes $O(M)$ time to generate $M-1$ rigid body transformations to align the $M-1$ fragments in one HSFB to the pivot fragment. It takes $O(N^2)$ time to align one structure to the pivot given a rigid body transformation. The total time complexity is $O(MN^2)$.

Step 4: Adjust pairwise alignments in an initial MSA: given one HSFB and the pivot protein, we need to conduct at most *TopF* adjustments to realign each of the unanchored proteins to the pivot. Each needs to run a dynamic programming algorithm with time $O(N^2)$, so the total time complexity is $O(TopF \times MN^2)$.

Step 5: Refine MSA by the consensus-based optimization: empirically, the length of the consensus structure is $<2N$, so it takes time $O(MN)$ to build a consensus structure from a given MSA. The column merge procedure takes time $O(MN)$ by using a 3D-hashing technique. It takes time $O(N^2)$ to align each structure to the consensus structure. At most 10 iterations are executed to refine the MSA, so the total time complexity is $O(MN^2)$.

The first two steps will be conducted for each of the *TopK*, resulting in a time complexity $O(TopK \times MN^2)$. The last two steps will be conducted for each of the *TopK* and *TopJ*, leading to $O(TopK \times TopJ \times TopF \times MN^2)$ time complexity. The overall time complexity is $O(TopK \times TopJ \times TopF \times MN^2)$.

3 RESULTS

3.1 3DCOMB alignment accuracy

The programs to be compared: we compare 3DCOMB with BLOMAPS (Wang and Zheng, 2009), MAMMOTH (Lupyan *et al.*, 2005), MAPSCI (Ilinkin *et al.*, 2010), MATT (Menke *et al.*, 2008), MultiProt (Shatsky *et al.*, 2004) and MUSTANG (Konagurthu *et al.*, 2006). Meanwhile, BLOMAPS, MultiProt and 3DCOMB are vertical-first algorithms while the other four are horizontal-first. We do not compare 3DCOMB with POSA because it has only a web server version and is not amendable to a large-scale test. 3DCOMB differs from BLOMAPS mainly in that 3DCOMB uses both local and global structure environments to identify HSFBs while BLOMAPS uses only local information. 3DCOMB also uses a better scoring function. 3DCOMB differs from MultiProt in both search algorithms and scoring functions.

Table 1. Alignment accuracy of seven MSA tools on three benchmarks HOMSTRAD, SABmark-sup and SABmark-twi

Method	CORE-LEN	RMSD	$\overline{\text{TMscore}}$
HOMSTRAD			
3DCOMB	170.58	2.00	0.800
MAPSCI	162.55	1.87	0.792
MAMMOTH	169.84	3.03	0.786
MATT	169.53	2.00	0.781
BLOMAPS	169.27	2.18	0.779
MUSTANG	169.49	2.66	0.765
MultiProt	140.82	1.33	0.649
SABmark-sup			
3DCOMB	106.66	2.59	0.655
MAPSCI	89.51	2.95	0.627
MAMMOTH	105.50	5.78	0.614
MATT	104.12	2.59	0.613
BLOMAPS	101.82	3.11	0.613
MUSTANG	103.86	4.20	0.583
MultiProt	68.70	1.61	0.404
SABmark-twi			
3DCOMB	71.63	3.02	0.526
MAPSCI	50.11	4.38	0.466
BLOMAPS	67.20	4.22	0.457
MATT	67.08	2.89	0.453
MAMMOTH	64.97	8.31	0.436
MUSTANG	66.89	5.10	0.422
MultiProt	36.38	1.75	0.259

The accuracy is measured by CORE-LEN, RMSD and $\overline{\text{TMscore}}$. The values in the table are averaged over an individual benchmark.

The benchmarks: we use three benchmarks: HOMSTRAD (Mizuguchi *et al.*, 1998), SABmark-sup and SABmark-twi (Van Walle *et al.*, 2005). HOMSTRAD contains 398 homologous protein families, each with at least three structures. SABmark-sup is the superfamily set in SABmark (version 1.65), containing 425 families with low to intermediate sequence identity. SABmark-twi represents the twilight set in SABmark, containing 209 families with low sequence identity. We apply three metrics CORE-LEN, core RMSD and $\overline{\text{TMscore}}$ to evaluating all the methods. The former two are also used by others such as MATT (Menke *et al.*, 2008). We normalize CORE-LEN by the length of the shortest protein in a group and denote it as CORE-LEN%.

Performance on HOMSTRAD: 3DCOMB obtains the largest average CORE-LEN and the third best average core RMSD (slightly larger than MultiProt and MAPSCI). Note that because MultiProt uses a very strict cutoff to determine if an aligned column is a core or not, it always obtains the smallest core RMSD and also very small CORE-LEN on all the benchmarks. As shown in Table 1, MAMMOTH and MUSTANG can generate alignments with CORE-LEN comparable to 3DCOMB, Matt and BLOMAPS, but much larger RMSD. The average $\overline{\text{TMscore}}$ achieved by 3DCOMB, BLOMAPS, MATT, MAMMOTH, MUSTANG, MAPSCI and MultiProt is 0.800, 0.779, 0.781, 0.786, 0.765, 0.792 and 0.649, respectively. 3DCOMB not only achieves the best average $\overline{\text{TMscore}}$, but also excels others on almost each individual structure group (Supplementary Fig. S2a).

Performance on SABmark-sup: 3DCOMB obtains the best average CORE-LEN and $\overline{\text{TMscore}}$. By core RMSD, 3DCOMB is second to only MultiProt, but MultiProt obtains a much smaller CORE-LEN. The average $\overline{\text{TMscore}}$ achieved by 3DCOMB, BLOMAPS, MATT, MAMMOTH, MUSTANG, MAPSCI and MultiProt is 0.655, 0.613, 0.613, 0.614, 0.583, 0.627 and 0.404, respectively. By $\overline{\text{TMscore}}$, 3DCOMB is ~4.5% better than the second best method MAPSCI and also outperforms others on almost each individual structure group (Supplementary Fig. S2b).

Performance on SABmark-twi: SABmark-twi is more challenging because it consists of mostly distantly related proteins. However, 3DCOMB excels others at an even larger margin. By CORE-LEN, 3DCOMB is 6.6% better than the second best method BLOMAPS and excels others on a majority of structure groups. By core RMSD, 3DCOMB is second to MultiProit and slightly to Matt. The $\overline{\text{TMscore}}$ obtained by 3DCOMB, BLOMAPS, MATT, MAMMOTH, MUSTANG, MAPSCI and MultiProt is 0.526, 0.457, 0.453, 0.436, 0.422, 0.466 and 0.259, respectively. By $\overline{\text{TMscore}}$, 3DCOMB outperforms the second best algorithm MAPSCI by 12.9% and also excels others on almost each individual structure group (Supplementary Fig. S2c).

Alignment accuracy versus core definition: different methods may use different distance cutoff values in determining if an aligned column is a core or not, so it is unfair to compare them simply in terms of CORE-LEN and core RMSD without specifying a uniform distance cutoff. To ensure a more fair comparison, we employ three cutoff values 4, 5 and 6 Å to determine if an aligned column is a core or not, respectively. That is, given a fully aligned column, we calculate all the pairwise distance of the aligned residues. If all the pairwise distances are within a given cutoff, this column is a core, otherwise not. As shown in Table 2, regardless of distance cutoffs, 3DCOMB consistently outperforms others in terms of CORE-LEN for distantly related proteins. In particular, 3DCOMB excels Matt in terms of both CORE-LEN and core RMSD. MAPSCI has similar performance as 3DCOMB on HOMSTRAD, but is much worse on the other two benchmarks.

Performance on large structure groups: 3DCOMB builds MSAs from only those HSFBs. Does 3DCOMB suffer from using too few similar fragment blocks (and thus, too few initial MSAs) in aligning a large structure group (with >15 structures) consisting of distantly related proteins? We examined the performance of 3DCOMB on all the large groups in the three benchmarks and found out that 3DCOMB does not have such an issue. Measured by $\overline{\text{TMscore}}$, 3DCOMB excels others especially on the groups containing distantly related proteins. The advantage of 3DCOMB over others is even larger on the large structure groups than that on all the structure groups. See Supplementary Tables S10, S11 and S12 for detailed results.

3.2 How much room is left for further improvement?

Given a set of M structures, how good is a given MSA in terms of a specific quality metric (e.g. TMscore)? Can we estimate the difference between this MSA and the best possible even if we do not know how to generate the best? This problem is important because if we can answer it, we may know how much room is left for further improvement. If the difference is quite small, it does not make much sense to further search for a better MSA. Here we estimate the (upper bound) quality, measured by $\overline{\text{TMscore}}$, of the best possible MSA using the average TMscore of all the pairwise alignments. However, it is challenging not to underestimate the TMscore of a protein pair. To handle this, we employ several tools such as TMalign, DALI, MATT and MAMMOTH to generate pairwise alignments and then choose the highest TMscore.

Figure 2 shows the real $\overline{\text{TMscore}}$ of the 3DCOMB alignments and their estimated $\overline{\text{TMscore}}$ difference from the best possible.

Table 2. Performance of seven MSA tools when three distance cutoff values 4.0, 5.0 and 6.0 Å are used to determine if an aligned column is a core or not

Method	4.0 Å	5.0 Å	6.0 Å
HOMSTRAD			
3DCOMB	141.98 (1.35)	152.38 (1.49)	159.09 (1.62)
BLOMAPS	137.92 (1.38)	149.40 (1.55)	156.92 (1.70)
MAMMOTH	136.40 (1.38)	146.59 (1.54)	153.49 (1.67)
MAPSCI	**143.97 (1.38)**	**154.30 (1.52)**	**159.35 (1.61)**
MATT	138.96 (1.40)	150.24 (1.56)	157.62 (1.69)
MultiProt	137.81 (1.28)	140.45 (1.32)	140.82 (1.33)
MUSTANG	130.86 (1.48)	142.93 (1.65)	150.81 (1.80)
SABmark-sup			
3DCOMB	**69.26 (1.57)**	**80.58 (1.79)**	**88.63 (1.98)**
BLOMAPS	65.40 (1.60)	76.92 (1.82)	85.41 (2.02)
MAMMOTH	62.96 (1.52)	72.37 (1.73)	78.95 (1.90)
MAPSCI	66.84 (1.49)	77.59 (1.74)	83.99 (1.92)
MATT	66.11 (1.62)	77.85 (1.87)	86.79 (2.06)
MultiProt	64.86 (1.51)	68.23 (1.59)	68.69 (1.61)
MUSTANG	53.76 (1.56)	64.45 (1.90)	72.82 (2.16)
SABmark-twi			
3DCOMB	**36.09 (1.67)**	**45.20 (1.95)**	**52.54 (2.20)**
BLOMAPS	34.10 (1.64)	42.90 (1.89)	49.85 (2.12)
MAMMOTH	27.69 (1.46)	34.26 (1.71)	38.65 (1.91)
MAPSCI	30.92 (1.53)	38.83 (1.83)	44.08 (2.08)
MATT	33.91 (1.67)	42.68 (1.97)	50.14 (2.25)
MultiProt	32.91 (1.63)	35.95 (1.74)	36.34 (1.76)
MUSTANG	22.13 (1.49)	30.08 (1.88)	36.63 (2.26)

The values outside and inside the parenthesis are CORE-LEN and RMSD, respectively. Bold fonts indicate the best values.

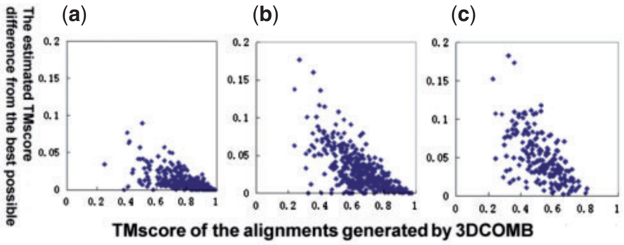

Fig. 2. MSA upper bound analysis on three datasets (**a**) HOMSTRAD, (**b**) SABmark-sup and (**c**) SABmark-twi. The X-axis is $\overline{\text{TMscore}}$ of the MSAs generated by 3DCOMB and the Y-axis is the estimated $\overline{\text{TMscore}}$ difference between the 3DCOMB alignments and the best possible.

Overall, the estimated $\overline{\text{TMscore}}$ difference decreases with respect to the similarity of input structures. That is, when the structures are quite similar, it is easier to generate an MSA consistent with all the pairwise alignments. Otherwise, it is much more challenging. When input structures are distantly related, an MSA consistent with all the pairwise alignments may not exist at all.

3.3 Specific examples

We have visually examined many structure groups in the three benchmarks and found out 3DCOMB generates significantly better alignments for many groups than others. It is very challenging to visualize the alignments, though. Here, we show three groups for which 3DCOMB generates correct alignments in functionally similar regions while others (especially the horizontal-first methods) fail. See the upplementary Material for more results and case studies.

SABmark-sup group 323 contains three Glutamine synthetase/guanido kinase proteins: d1f52a2 [SCOP (Murzin *et al.*, 1995) ID d.128.1.1, 368 AAs], d1m15a2 (d.128.1.2, 262 AAs) and d1qh4a2 (d.128.1.2, 279 AAs). PDB (Berman *et al.*, 2002) shows that 1f52 (Glutamine synthetase) and 1m15 (arginine kinase) bind to ADP. The three proteins share a common core consisting of two beta-alpha-beta2-alpha repeats and the six-beta sheets form an unsealed pocket where ATP/ADP binds. That is, the common core is the active site of the synthetase/kinase.

By the 3DCOMB alignment (Fig. 3a1), we may infer that the aligned unsealed pocket in 1qh4 possibly is its active site. This is confirmed by the prediction result from ConCavity (Capra *et al.*, 2009) in Figure 3a2, which predicts binding sites without using structure alignment. Two other vertical-first methods BLOMAPS and MultiProt generate reasonable but worse alignments than 3DCOMB (Supplementary Fig. S9). All the horizontal-first methods (MATT, MAPSCI, MUSTANG and MAMMOTH) generate incorrect alignment for 1f52a2 and the binding sites. This example demonstrates the vertical-first methods are better than horizontal-first methods in discovering conserved regions among proteins which are not very similar and also of very different sizes.

SABmark-twi group 124 contains six Lysozyme-like proteins: d153l_ (SCOP ID d.2.1.5, 185 AAs), d1dxja_ (d.2.1.1, 242 AAs), d1lw9a_ (d.2.1.3, 164 AAs), d1qgia_ (d.2.1.7, 259 AAs), d1qsaa2 (d.2.1.6, 168 AAs) and d3lzt_ (d.2.1.2, 129 AAs). Meanwhile, 153l, 3lzt and 1dxj are goose lysozyme, hen egg-white lysozyme and jack bean chitinase, respectively. Proteins 1lw9 and 1qgi are T4 lysozyme and chitosanase from *Bacillus circulans*. d1qsaa2 is Lytic Transglycosylase Slt70 from *Escherichia coli*, resembling goose-type lysozyme (van Asselt *et al.*, 1999). These proteins represent a superfamily of hydrolases arising from the divergent evolution of an ancient protein (Robertus *et al.*, 1998).

By SCOP, these proteins have 'alpha + beta motif for the active site region'. Although their sequence similarity is low, they have a conserved core containing two helices and a three-stranded sheet, which form the substrate binding and catalytic cleft (Monzingo *et al.*, 1996). Their active binding sites lie in the cleft between the upper and lower domains. Their two catalytic centers are located at the upper domain, one in front of the three-stranded sheet and the other in the N-terminal helix (Saito *et al.*, 1999).

As shown in Figure 3b1, 3DCOMB correctly aligns the functional sites described in Monzingo *et al.* (1996) (i.e. three-sheet plus two-helix in the upper domain). As shown in Figure 3b2, all the binding sites with predicted by ConCavity with high confidence are well aligned. 3DCOMB even correctly aligns the shortest protein 3lzt and the longest protein 1qgi, which are conserved only in the functional sites. 3DCOMB also indicates that there are several conserved helixes in the lower domain. As shown in Supplementary Figure S6, only 3DCOMB and BLOMAPS generate correct alignments for the functional sites while all the horizontal-first methods fail. This confirms the strength of vertical-first methods. However, another vertical-first method MultiProt fails to produce a correct alignment, maybe because MultiPort cannot identify good similar fragment blocks to build initial MSAs.

SABmark-twi group 118 contains six chelatase-like proteins: d1doza_ (SCOP ID c.92.1.1, 309 AAs), d1m1na_ (c.92.2.3, 477 AAs), d1m1nb_ (c.92.2.3, 522 AAs), d1n2za_ (c.92.2.2, 245 AAs), d1qgoa_ (c.92.1.2, 257 AAs) and d1toaa_ (c.92.2.2, 277 AAs). Meanwhile, d1doza_, d1qgoa_, d1n2za_ and d1toaa are ferrochelatase, cobalt chelatase, vitamin B12 binding protein and Zn^{2+} binding protein, respectively, and d1m1na_ and d1m1nb_ are the molybdenum iron (MoFe) protein of nitrogenase. It is very challenging to align these proteins since they are similar only at fold level and also of very different sizes. Even SCOP and CATH are inconsistent on them. CATH splits d1doza_, d1n2za_, d1qgoa_ and d1toaa_ into two domains, d1m1na_ into 3 and d1m1nb_ into 4 while SCOP treats all of them as single domain proteins.

These proteins have different (non-conserved) binding ligands, but the locations of the binding pockets and the domains surrounding the pockets are conserved. In particular, d1doza_, d1qgoa_, d1n2za_ and d1toaa_ have two structurally similar lobes (the synonym for 'domain', may refer to a smaller substructure unit), each being a Rossmann-like fold (Al-Karadaghi *et al.*, 1997). These two lobes interact with each other in a head-to-head manner (Lee *et al.*, 1999) and the active sites of these four proteins lie in a deep cleft between the two lobes (Schubert *et al.*, 1999). d1n2za_ and d1toaa_have a single long helix linking the two lobes, d1doza_ and d1qgoa do not. This unique helix is possibly adopted to limit the hinge motion associated with ligand exchange (Lee *et al.*, 1999). d1m1na_ (d1m1nb_) has three lobes (Borths *et al.*, 2002) and its αII (βII) and αIII (βIII) lobes are similar to those in d1n2za_ and d1toaa and also linked by a helix. Similar to others, the binding ligand of d1m1na_ is located at the interface between the αII and αIII lobes (Kim and Rees, 1992).

3DCOMB alignment is consistent with the above structural description and ConCavity prediction, as shown in Figure 3c1 and c2. Only 3DCOMB generates a correct alignment for this group while all the other methods fail (Supplementary Fig. S5). This result confirms that 3DCOMB not only has a good search algorithm, but also a good scoring function (since the other two vertical-first methods BLOMAPS and MultiProt also fail).

Comparison with a binding site alignment program MultiBind (Shulman-Peleg *et al.*, 2008): MultiBind needs to know binding site positions in order to generate alignments while 3DCOMB does not. MultiBind also runs very slowly and generates many alternative alignments. Many proteins in these examples have no binding site information in PDB. In order to run MultiBind, we assign binding

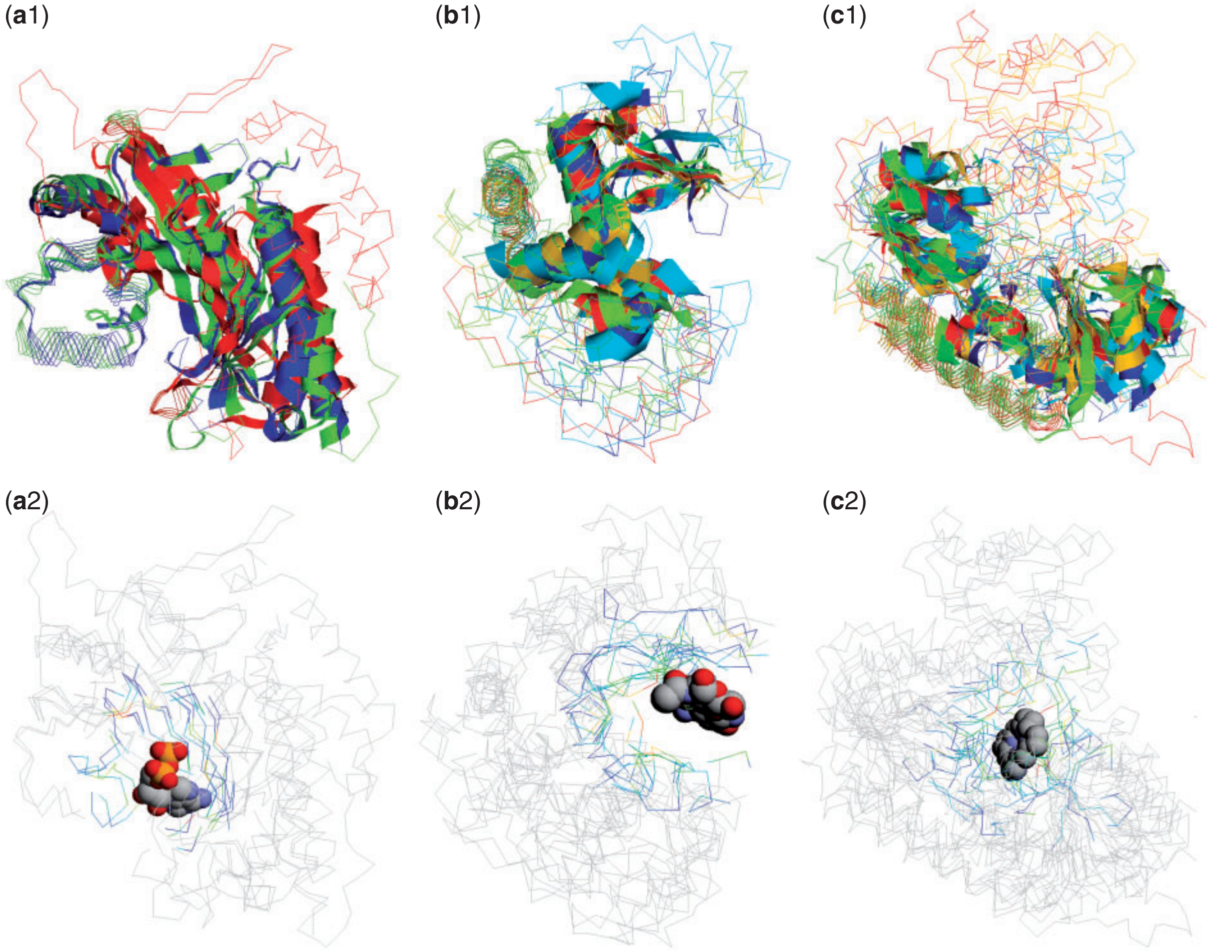

Fig. 3. The pictures in (**a1**), (**b1**) and (**c1**) show the 3DCOMB alignments of SABmark-sup 323, SABmark-twi 124 and SABmark-twi 118, respectively. Proteins are displayed in different colors. The full-core positions (where all proteins are aligned) are displayed in ribbon while the partial-core positions (where >50% of proteins are aligned) in strand. The pictures in (**a2**), (**b2**) and (**c2**) show the ligand positions of the three structure groups (in spacefill form) and ConCavity predictions with confidence in colors. Color close to blue indicates low confidence while close to red high, while gray indicates no predictions at all. (**a1**) The proteins in blue, green and red are d1qh4a2, d1m15a2 and d1f52a2, respectively. (**a2**) The ligand ADP is from 1m15. (**b1**) The proteins in blue, cyan, dark green, light green, yellow and red are d1lw9a_, d1qgia_, d153l__, d1dxja_, d3lzt__ and d1qsaa2, respectively. (**b2**) The ligand GlcN6 is from 1qgi. (**c1**) The proteins in blue, cyan, dark green, light green, yellow and red are d1qgoa_, d1doza_, d1toaa_, d1n2za_, d1m1na_ and d1m1nb_, respectively. (**c2**) The ligand heme is from 1doz.

sites to them based upon 3DCOMB alignments and ConCavity predictions. Even so, for these examples MultiBind cannot generate correct alignments at functionally similar regions.

3.4 Implications to homology modeling

Multiple template modeling has been used to enhance homology modeling (Cheng, 2008; Joo *et al.*, 2007; Peng and Xu, 2011). An interesting question to ask is how much improvement room is left for this method in terms of alignment accuracy? To answer this, we conduct an experiment using 47 CASP9 (the 9th Critical Assessment of Structure Prediction) test targets, all of them have at least two good templates. See Supplementary Material for a list of the targets and templates. We use seven MSA tools to build an MSA for a target and its templates, assuming that the native structures of the target and templates are known. We also used RaptorX (Peng and Xu, 2011), one of the best threading programs, to generate an alignment between the target and its templates, without using the native structure for the target. All the alignments are fed into MODELLER to generate 3D models for the targets. As shown in Supplementary Figure S11 and Table S14, the 3DCOMB models excel the RaptorX models. That is, there may be still improvement room for multiple template modeling. However, the 3D models derived from other MSA tools are not better than the RaptorX models. This shows that in order to identify the limitation of multiple template modeling, a good MSA tool is critical. Otherwise we may reach very different or even opposite conclusions.

3.5 3DCOMB running time analysis

We tested 3DCOMB on an Ubuntu Linux PC with 2 GB RAM and Intel®Core™2 Quad CPU T5600 @1.83 GHz. The input structures are sorted according to the length in ascending order. We use the $TopK = \text{All}$, $TopJ = 1$, $TopF = 5$ as our default parameters to run 3DCOMB on SABmark-twi. 3DCOMB, MATT and MUSTANG have running times of 43 706, 45 328 and 50 728 s, respectively. 3DCOMB yields better accuracy than MATT and MUSTANG. It takes MultiProt and MAMMOTH 39 642 and 1279 s, respectively, to run this benchmark, but they have much worse alignment accuracy than 3DCOMB and MATT. BLOMAPS and MAPSCI are very fast, taking only 780 and 240 s. The performance of 3DCOMB depends on three parameters *TopK*, *TopJ* and *TopF*, which can be further set smaller to reduce running time without losing much alignment accuracy. Supplementary Table S13 shows the 3DCOMB running time and accuracy on SABmark-twi with respect to different parameter s.

4 DISCUSSION AND FUTURE WORK

This article presents a novel method 3DCOMB for MSA. By using a probabilistic model to combine both local and global structure information, we can accurately identify the most conserved short fragment blocks among proteins to be aligned. These conserved fragment blocks are very likely contained in the best MSA, so 3DCOMB can quickly extend them to the best MSA. This article also introduces a novel scoring function to generate an MSA with a large number of cores and also high-quality pairwise alignments.

We have compared 3DCOMB with BLOMAPS, MAMMOTH, MAPSCI, MATT, MultiProt and MUSTANG on the popular benchmarks. Both the MSAs and the pairwise alignments generated by 3DCOMB are of high quality. In particular, 3DCOMB shows significant advantages over others in aligning a large set of distantly related proteins and their functionally similar regions. 3DCOMB also has a very reasonable running time and can scale well up to a large structure group.

Currently, 3DCOMB uses only geometrical information. We may improve 3DCOMB by including sequence information and the physical–chemical properties of amino acids. Sequence similarity measure can be used to explicitly model evolutionary relationship. Other geometrical information such as Voronoi tessellations (Birzele *et al.*, 2007) and Conformational Letter (CLE) (Zheng, 2008a; Zheng, 2008b; Zheng and Liu, 2005) can also be used to further improve HSFB identification.

3DCOMB will be useful for many real-world applications and sometimes can produce results dramatically different from other tools. For example, 3DCOMB indicates that there may be still large improvement room for multiple template modeling while other MSA tools fail to do so. Coupled with tools like ConCavity (Capra *et al.*, 2009), 3DCOMB can be potentially used for binding site prediction.

Funding: National Institutes of Health (grant R01GM089753 to J.X.); National Science Foundation (grant DBI-0960390 to J.X).

Conflict of Interest: none declared.

REFERENCES

Al-Karadaghi,S. *et al.* (1997) Crystal structure of ferrochelatase: the terminal enzyme in heme biosynthesis. *Structure*, **5**, 1501–1510.

Berman,H. *et al.* (2002) The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.

Birzele,F. *et al.* (2007) Vorolign–fast structural alignment using Voronoi contacts. *Bioinformatics*, **23**, e205–e211.

Borths,E.L. *et al.* (2002) The structure of Escherichia coli BtuF and binding to its cognate ATP binding cassette transporter. *Proc. Natl Acad. Sci. USA*, **99**, 16642.

Capra,J.A. *et al.* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.

Cheng,J. (2008) A multi-template combination algorithm for protein comparative modeling. *BMC Struct. Biol.*, **8**, 18.

da Silveira, C. *et al.* (2009) Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins Struct. Funct. Bioinformatics*, **74**, 727–743.

Do,C. *et al.* (2006) CONTRAlign: discriminative training for protein sequence alignment. In *RECOMB/Lecture Notes in Computer Science*, Vol. 3909, Springer, Venice, Italy, pp. 160–174.

Eswar,N. *et al.* (2008) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **426**, 145.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–128.

Holm,L. and Sander,C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, **22**, 3600–3609.

Ilinkin,I. *et al.* (2010) Multiple structure alignment and consensus identification for proteins. *BMC Bioinformatics*, **11**, 71.

Joo,K. *et al.* (2007) High accuracy template based modeling by global optimization. *Proteins*, **69** (Suppl. 8), 83–89.

Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **32**, 922–923.

Kim,J. and Rees,D. (1992) Structural models for the metal centers in the nitrogenase molybdenum-iron protein. *Science*, **257**, 1677.

Konagurthu,A.S. *et al.* (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.

Lafferty,J. *et al.* (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Citeseer*, 282–289.

Lee,Y.H. *et al.* (1999) Treponema pallidum TroA is a periplasmic zinc-binding protein with a helical backbone. *Nat. Struct. Mol. Biol.*, **6**, 628–633.

Levitt,M. and Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.

Lupyan,D. *et al.* (2005) A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, **21**, 3255–3263.

Madhusudhan,M.S. *et al.* (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng. Des. Sel.*, **22**, 569–574.

Menke,M. *et al.* (2008) Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput. Biol.*, **4**, e10.

Mizuguchi,K. *et al.* (1998) HOMSTRAD: a database of protein structure alignments for homologous families *Protein Sci.*, **7**, 2469–2471.

Monzingo,A.F. *et al.* (1996) Chitinases, chitosanases, and lysozymes can be divided into procaryotic and eucaryotic families sharing a conserved core. *Nat. Struct. Biol.*, **3**, 133–140.

Murzin,A. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Peng,J. and Xu,J. (2009) Boosting protein threading accuracy. In *RECOMB/Lecture Notes in Computer Science*, Vol. 5541, Springer, Tucson, AZ, pp. 31–45.

Peng,J. and Xu,J. (2010) Low-homology protein threading. *Bioinformatics*, **26**, i294.

Peng,J. and Xu,J. (2011) A multiple template approach to protein threading. *Proteins*, **79**, 1930–1959.

Robertus,J.D. *et al.* (1998) Structural analysis shows five glycohydrolase families diverged from a common ancestor. *J. Exp. Zool.*, **282**, 127–132.

Saito,J. *et al.* (1999) Crystal structure of chitosanase from Bacillus circulans MH-K1 at 1.6-Å resolution and its substrate recognition mechanism *J. Biol. Chem.*, **274**, 30818.

Schubert,H.L. *et al.* (1999) Common chelatase design in the branched tetrapyrrole pathways of heme and anaerobic cobalamin synthesis. *Biochemistry*, **38**, 10660–10669.

Shatsky,M. *et al.* (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.

Shulman-Peleg,A. *et al.* (2008) MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Res.*, **36**, W260.

Siew,N. *et al.* (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

van Asselt,E.J. *et al.* (1999) High resolution crystal structures of the Escherichia coli lytic transglycosylase slt70 and its complex with a peptidoglycan fragment1. *J. Mol. Biol.*, **291**, 877–898.

Van Walle,I. *et al.* (2005) SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.

Wang,S. and Zheng,W.M. (2008) CLePAPS: fast pair alignment of protein structures based on conformational letters. *J. Bioinform. Comput. Biol.*, **6**, 347–366.

Wang,S. and Zheng,W. (2009) Fast multiple alignment of protein structures using conformational letter blocks. *Open Bioinformatics J.*, **3**, 69–83.

Wang,Z. *et al.* (2010) Protein 8-class secondary structure prediction using conditional neural fields. *BIBM 2010*.

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

Ye,Y. and Godzik,A. (2005) Multiple flexible structure alignment using partial order graphs. *Bioinformatics*, **21**, 2362.

Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.

Zhao,F. *et al.* (2010a) A probabilistic and continuous model of protein conformational space for template-free modeling. *J. Comput. Biol.*, **17**, 783–798.

Zhao,F. *et al.* (2010b) Fragment-free approach to protein folding using conditional neural fields. *Bioinformatics*, **26**, i310.

Zheng,W. (2008a) Protein conformational alphabets. In Roswell,L.B. (ed.) *Protein Conformations: New Research*. Nova Science Publishers, New York, pp. 1–49.

Zheng,W. (2008b) The use of a conformational alphabet for fast alignment of protein structures. *Bioinformatics Res. Appl.*, 331–342.

Zheng,W. and Liu,X. (2005) *A Protein Structural Alphabet and its Substitution Matrix CLESUM*. Transactions on Computational Systems Biology II.: Springer, Berlin.