SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPUTATIONAL MODELS
PROTEIN FOLDING
COMPUTATIONAL PLATFORMS AND ENVIRONMENTS
COMPUTATIONAL BIOPHYSICS

Received
1 July 2013

Accepted
22 August 2013

Published
10 September 2013

Correspondence and requests for materials should be addressed to Y.Z. (zhng@umich.edu)

A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction

Renxiang Yan, Dong Xu, Jianyi Yang, Sara Walker & Yang Zhang

Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109.

Protein sequence alignment is essential for template-based protein structure prediction and function annotation. We collect 20 sequence alignment algorithms, 10 published and 10 newly developed, which cover all representative sequence- and profile-based alignment approaches. These algorithms are benchmarked on 538 non-redundant proteins for protein fold-recognition on a uniform template library. Results demonstrate dominant advantage of profile-profile based methods, which generate models with average TM-score 26.5% higher than sequence-profile methods and 49.8% higher than sequence-sequence alignment methods. There is no obvious difference in results between methods with profiles generated from PSI-BLAST PSSM matrix and hidden Markov models. Accuracy of profile-profile alignments can be further improved by 9.6% or 21.4% when predicted or native structure features are incorporated. Nevertheless, TM-scores from profile-profile methods including experimental structural features are still 37.1% lower than that from TM-align, demonstrating that the fold-recognition problem cannot be solved solely by improving accuracy of structure feature predictions.

Template-based modeling (TBM) is by far the only reliable approach to protein 3D structure prediction[1,2]. With rapid accumulation of experimental structures in the Protein Data Bank (PDB)[3], TBM plays an increasingly important role in protein structure determination and structure-based function annotation studies as more and more protein structures become available as putative templates. In fact, recent studies showed that the current PDB library has already approached completeness in structural space[4,5]. Nevertheless, only around 2/3 of targets can have the templates reliably identified by current threading (or fold-recognition) methods in genome-wide protein structure prediction[6–9]. A critical issue for protein template identification is the correct construction and scoring of the target-to-template alignments of amino acid sequences.

Early efforts on protein sequence alignments can be traced back to the 1970s when Needleman and Wunsch pioneered a global alignment algorithm for protein sequences via dynamic programming recursion[10]. Smith and Waterman extended the algorithm for identifying highly conserved subsequence motifs by local alignments[11]. However, dynamic programming is too slow for scanning large-scale sequence databases. Altschul, Lipman and coworkers developed FASTA and BLAST based on a heuristic search and extension of common sequence patterns (words) among the compared sequences, which significantly increases the speed of sequence alignment and database search[12,13]. Later, the authors extended BLAST to PSI-BLAST which improves the sensitivity of sequence-sequence alignments[14]. The key idea of PSI-BLAST is to generate multiple sequence alignments (MSAs) by iterative sequence database search, where a sequence profile in terms of a position-specific scoring matrix (PSSM) is constructed from the MSAs and used to enhance the accuracy of sequence alignment by sequence-profile comparisons.

The idea of sequence profiles has revolutionized the sequence alignment search and template-based protein structure prediction[15,16]. A variety of profile alignment based threading methods have been recently developed for efficient protein homologous template identification and structure prediction[17–21]; most of the methods rely on PSI-BLAST for MSA search and profile generations. The multiple sequence alignments and sequence profiles can also be created by hidden Markov models (HMMs), which are represented by a chain of match and insert/deletion nodes with the MSAs corresponding to the paths with the highest probabilities given by the product of amino acid emission and insertion/deletion probabilities[22]. Typical HMM-based threading algorithms include SAM[23] and HHsearch[24], where SAM is based on HMM-sequence alignments and HHsearch on HMM-HMM alignments.

In addition to sequence profiles, a variety of structure features have been recently introduced to improve the alignment accuracy. For example, secondary structure predictions from neural network training[25] are used by almost all contemporary threading/alignment programs[17,19,24]. Other structural characteristics, including residue-residue contacts, backbone torsion angles[26], solvation[27] and residue depth[19], are often exploited in protein threading approaches[17,21,28,29].

Despite the extensive effort made in developing sequence alignment algorithms, little is known about the relative performance of the methods. In particular, a number of critical questions remain to be addressed; for example, what are the quantitative differences of sequence- versus profile-based or local- versus global-alignment methods on close- and distant-homology detections? As the two most often-used approaches, what are the strength and weakness of PSSM- and HMM-based profiles in sequence alignments? How much can we possibly gain in fold-recognition by developing the best structural feature prediction methods? The answer to these questions is of essential importance for guiding the uses within the biology community, as well as for leveraging future method development studies in the field.

As community-wide platforms, the CASP[30,31], CAFASP[32] and Livebench[33] experiments provided valuable opportunities for critical assessments of various threading methods. However, one limit of the assessments is due to the fact that predictors in the experiments usually exploit different template libraries, the construction of which can have important impact on the final modeling results. Meanwhile, the number of targets involved in the experiments is limited (~100) and unbalanced; most of the targets are closely homologous to the experimental structures, which are easy to be detected[34]. The problems have been partly addressed by several of previous studies that compared different sequence alignment methods on large sets of benchmark proteins[35–42]. For example, Park et al[41] and Madera and Gough[42] compared HMM- and PSSM-based profiles on the datasets collected from SCOP[43] and found that HMM-based profiles can detect more homologous relationships than PSSM-based profiles. Dunbrack and coworkers[35,37] examined different sequence alignment tools using structure alignments as the gold standard and found that sequence-profile alignments by PSI-BLAST are only slightly more accurate than sequence-sequence alignments by BLAST but PSI-BLAST achieves much longer alignments. Girshin and coworkers[36] evaluated the alignment methods in multiple reference-dependent/independent and global/local modes and showed that different aspects of evaluation reveal different properties of the methods. Barton and coworkers[38] developed a multiple-level benchmark suite to evaluate eight alignment methods and concluded that the majority of alignment improvements since 1985 were due to pair-score matrices rather than algorithmic refinements. Elofsson[40] compared different sequence alignment and threading algorithms and found that the alignment difference among different methods occurs mainly in the region of 15–20% sequence identity where secondary structure prediction and PSI-BLAST profiles are the major driven force of alignment improvements.

Despite the valuable insights revealed, most of the benchmark studies focused on a limited set of traditional sequence alignment algorithms and were performed nearly a decade ago. Many recent developments, e.g. structural feature integrations and HMM-HMM alignments which are important for protein structure prediction, are yet to be assessed. Meanwhile, the testing datasets used in these studies were mostly collected from the SCOP library and largely belong to the easy homology category (which represents a similar problem in the CASP experiments mentioned above), while the performance of the methods on detecting hard distant-homology templates, which are more challenging to the field, needs to be appropriately examined.

In this work, we aim to develop a comprehensive and balanced experiment to systematically examine the strength and weakness of various up-to-date sequence alignment methods. Ten publicly available methods and ten in-house methods specially designed for concept testing, which constitute a representative set of various alignment/threading approaches, were installed on the local computer cluster. These methods are tested on a large set of 538 proteins consisting of a balanced category distribution of difficulty (i.e. including similar number of easy, medium and hard protein targets), based on a uniform set of template structure libraries. We conducted a detailed analysis on the benchmark results to address a series of critical questions in sequence alignment and template-based protein structure prediction, which aim to provide insightful guidance for biological use and future method developments. All alignments and modeling data in this study can be downloaded at http://zhanglab.ccmb.med.umich.edu/publicdata/benchmark1.

Results

Dataset and template library. All the sequence alignment programs are benchmarked on the same set of 538 non-redundant proteins randomly collected from the PDB library[3]. These proteins have a pair-wise sequence identity less than 30% and length ranging from 34 to 804 residues. Proteins with broken chains or missing residues were not included. The sequences were divided into three categories: Easy, Medium and Hard targets, based on the consensus confidence score of the meta-threading LOMETS program[44], which consists of 9 protein threading programs (dPPAS, MUSTER, HHsearch-I, HHsearch-II, PPAS, PROSPECT, SAM, SPARKS and SP3). A target is defined as Easy if at least one strong template hit can be detected for the target by each program with the Z-score higher than the confidence cutoff; a target is defined as Hard if none of the threading programs has a strong template hit; otherwise, it is considered a Medium target. In total, the 538 proteins are selected to include a balanced category distribution of difficulty with 137 Easy, 177 Medium, and 224 Hard targets. Here, we have put more focus on the challenging targets by arbitrarily increasing the number of Medium and Hard proteins in our benchmark protein set, although a naturally collected sample from the PDB would have a much lower portion of Medium/Hard cases. A list of all the 538 proteins, together with the classification, are provided in http://zhanglab.ccmb.med.umich.edu/publicdata/benchmark1/protein_types.txt.

The existence of template structures in the library is a precondition for template identification. To eliminate potential bias of the alignment algorithms from the template structure library, we constructed the libraries of all threading programs using the same sequence identity cutoff updated to the same time stamp (by Jan, 2013). In fact, the template libraries for NW-align, SW-align, BLAST, PSI-BLAST, PSA, PPA, PPAS, dPPAS, MUSTER, SAM, PRC, PROSPECT, SPARKS, SP3 and FFAS are generated from the same set of non-redundant PDB proteins with a pair-wise sequence identify cutoff 70% (see http://zhanglab.ccmb.med.umich.edu/library/). The libraries for HHsearch-I and HHsearch-II are downloaded from ftp://toolkit.lmb.uni-muenchen.de, which has also a sequence identity cutoff of 70%. The size of these two libraries is about the same. The programs of all the tested methods are described in METHODS.

Summary of performance by individual alignment methods. Table 1 presents a summary of the 3D structural models, which are built by copying the framework of the highest ranked and the best in the top ten scoring templates based on the alignments generated by different alignment programs. The quality of alignments is generally measured by the root-mean-square deviation (RMSD) of the models (Columns 4–5), where BLAST, PSI-BLAST, PRC, FFAS and HHsearch programs have the lowest RMSD to the targets (~7–9 Å). However, the alignments by these programs tend to have a smaller number of residues aligned (i.e. lower alignment coverage, Columns 6–7), typically below 80%. Such short alignments can have a negative

Table 1 | Summary of template identification by different alignment methods

Methods[a]	TM-score[b] First	TM-score[b] Best in top10	RMSD (Å)[c] First	RMSD (Å)[c] Best in top10	Coverage[d] First	Coverage[d] Best in top10	CPU[e]
Profile-to-profile alignments							
MUSTER	0.435(0.449)	0.487(0.512)	10.3(15.2)	8.7(12.3)	0.875	0.875	27.0
HHsearch-II	0.429(0.449)	0.477(0.507)	9.5(21.6)	9.0(14.1)	0.767	0.820	13.0
dPPAS	0.426(0.438)	0.481(0.502)	9.6(20.6)	8.5(15.3)	0.819	0.844	17.0
PPAS	0.424(0.441)	0.473(0.499)	10.3(17.4)	8.9(13.4)	0.839	0.850	10.0
SP3	0.424(0.438)	0.476(0.499)	10.7(15.7)	9.1(12.3)	0.873	0.873	11.0
HHsearch-I	0.422(0.444)	0.472(0.502)	9.5(20.3)	9.0(14.7)	0.763	0.817	16.0
SPARKS	0.421(0.433)	0.469(0.493)	11.0(15.7)	9.4(12.1)	0.891	0.886	36.0
PROSPECT	0.418(0.428)	0.469(0.490)	11.5(13.3)	9.9(11.3)	0.914	0.903	15.0
PPA	0.397(0.413)	0.447(0.469)	10.9(17.5)	9.7(15.5)	0.844	0.851	25.0
FFAS	0.393(0.406)	0.444(0.465)	9.5(24.2)	8.6(18.9)	0.758	0.790	4.0
PRC	0.372(0.388)	0.417(0.442)	8.6(32.9)	8.0(24.3)	0.668	0.712	23.0
Average	**0.415(0.429)**	**0.465(0.5)**	**10.2(19.5)**	**9.0(17.7)**	**0.818**	**0.838**	**17.9**
Sequence-to-profile alignments							
SAM	0.344(0.358)	0.405(0.426)	10.6(27.5)	9.9(18.3)	0.717	0.778	8.0
PSA	0.338(0.333)	0.371(0.392)	12.9(17.5)	12.0(15.0)	0.870	0.873	9.0
PSI-BLAST	0.301(0.320)	0.344(0.369)	7.8(51.7)	7.4(42.1)	0.507	0.556	4.0
Average	**0.328(0.337)**	**0.373(0.4)**	**10.5(32.3)**	**9.8(25.1)**	**0.698**	**0.736**	**7.0**
Sequence-to-sequence alignments							
NW-align	0.321(0.336)	0.377(0.403)	12.7(21.7)	11.4(15.0)	0.849	0.866	5.0
SW-align	0.265(0.285)	0.324(0.348)	9.9(49.5)	9.2(35.7)	0.560	0.625	4.0
BLAST	0.246(0.263)	0.292(0.315)	8.5(59.7)	8.2(47.5)	0.470	0.529	0.1
Average	**0.277(0.295)**	**0.331(0.424)**	**10.4(43.7)**	**9.7(23.6)**	**0.626**	**0.673**	**3.0**
Other controls							
TM-align	0.661(0.664)	0.663(0.683)	3.1(7.5)	3.0(7.1)	0.856	0.846	90.0
MUSTER$^{SS + BTA + SA}$	0.482(0.511)	0.512(0.559)	8.0(14.1)	7.2(11.2)	0.797	0.800	26.0
MUSTER$^{SS + BTA}$	0.453(0.481)	0.493(0.536)	9.5(12.7)	8.1(11.3)	0.831	0.820	26.0
MUSTER^{SS}	0.447(0.474)	0.487(0.528)	9.7(12.7)	8.3(11.3)	0.839	0.830	26.0

[a]Alignment methods as sorted by TM-score in each category.
[b]Average TM-score. Values in parentheses are for full-length models built by MODELLER. 'First' refers to the top-ranking model based on alignment score; 'Best in top10' to the best model of the highest TM-score among the top ten models with the highest alignment scores.
[c]RMSD to the native.
[d]Alignment coverage equals to the number of aligned residues divided by target length.
[e]Average CPU time in minutes, which consists of constructing profile and building of alignments in a HP DL1000h computer.

impact on the full-length structure construction by homology modeling since structure information is missed for a large portion of unaligned sequences. In fact, the full-length models by MODELLER[45] have a very high RMSD (>20 Å) for all these local alignment methods (see values in parentheses). Here, the full-length models ware generated by the script *model-default.py* in the MODELLER package. The modeling results from MODELLER are deterministic in the sense that more runs do not change the quality result of the final models.

In Columns 2–3, we also list the result of the alignment models on TM-score, which is defined to combine the alignment accuracy and coverage as a unique score[46]:

$$\mathrm{TM-score} = \frac{1}{L}\sum_{i=1}^{L_{ali}} \frac{1}{1+\left(\dfrac{d_i}{1.24(L-15)^{1/3}-1.8}\right)^2} \qquad (1)$$

where L is the length of target sequence, L_{ali} the number of aligned residues, and d_i the pair-wise distance of ith residue in the model and target after the optimal superposition. In this scoring function, the programs, which have a better balance of alignment coverage and RMSD, excel, including MUSTER, HHsearch and PPAS programs. The simple sequence-to-sequence based alignment algorithms generally have a lower TM-score.

Meanwhile, the local-to-local alignment based algorithms generally have a lower coverage and TM-score compared to the global-to-global alignment methods. A typical example is SW-align based on Smith-Waterman, which only identifies the highly conserved regions and has on average 56% residues aligned, while Needleman-Wunsch based NW-align uses the same parameter and scoring function but generates alignments with a much higher coverage (84.9%). Accordingly, the TM-score of NW-align is 21.1% higher than that of SW-align. The completeness of alignment searching also plays a role in final model determination. For instance, both BLAST and SW-align are local sequence alignments based on BLOSUM62 mutation scores. But BLAST searches are based on an incomplete heuristic word search algorithm, which has an average TM-score 7% lower than SW-align. BLAST is however 39 times faster than SW-align in our test.

Although TM-score aims to balance the accuracy and coverage of alignments, it still favors algorithms that have a higher coverage, since including additional residues in the alignments always has a positive contribution to TM-score according to Eq. 1, although the contribution is small if the added residues from templates are far away from the target. To examine the effect of such bias, we constructed full-length models of the targets based on the alignments, using the widely-used comparative modeling tool MODELLER[45]. Although TM-score is now normalized by the same length of the target sequence, the TM-score ranking of full-length models is largely consistent with that of the original alignments, except for some small but notable variations. Taking the top hits as an example, the original alignments by HHsearch-I have a lower TM-score than those by dPPAS (0.422 vs. 0.426) due to the low coverage of the sequences (76.3% vs. 81.9%). After full-length modeling, the TM-score of HHsearch-I becomes higher than that of dPPAS (0.444 vs. 0.438) and several other related algorithms (e.g. PPAS and SP3). Here, the more precise alignments by HHsearch-I in the aligned regions have probably introduced some restraint/guidance to

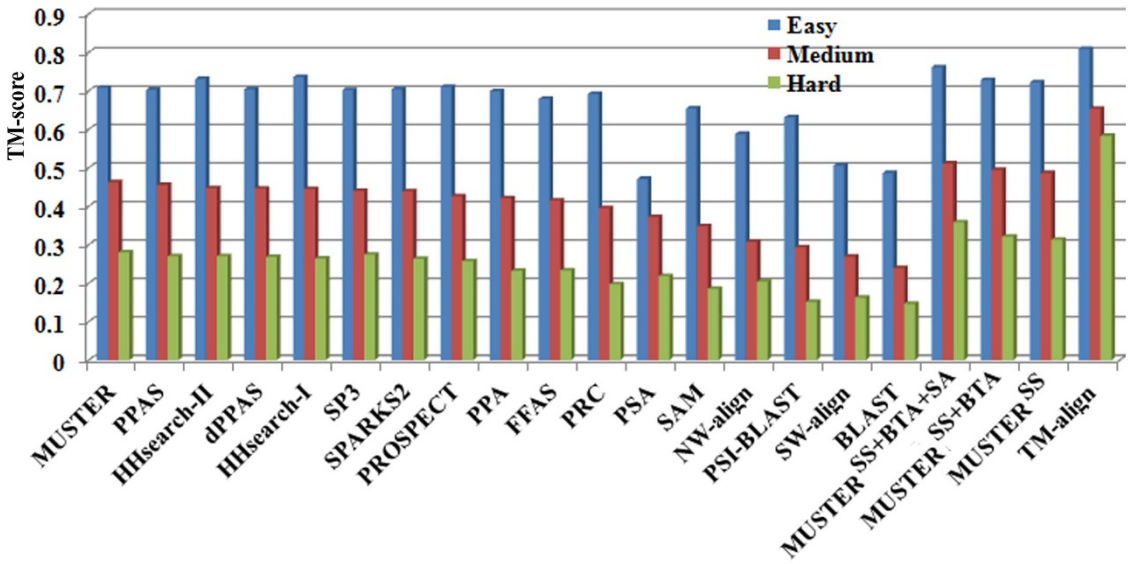

Figure 1 | TM-score histogram of the top hits identified by different algorithms in Easy, Medium and Hard categories.

the structure modeling of the unaligned regions, e.g. through bond-length and change connectivity, which have resulted in models of a higher overall TM-score.

Performance of sequence alignment programs in different target categories. The performance of different alignment programs varies with the difficulty of the targets, i.e. the evolutionary distance between target and template proteins. If we use the target structure as a probe to search through the PDB library by TM-align[47], the average TM-scores of aligned regions of the best structural templates in the three categories of Easy, Medium and Hard are 0.779, 0.666 and 0.586, respectively, after excluding homologous templates with a sequence identity $>$ 30%. This data on one hand sets up an upper-bar for template identifications by fold-recognition; on the other hand, it demonstrates that the target category as defined by the LOMETS prediction is largely consistent with the actual difficulty of the template identification for the targets.

In supplementary Tables S1, S2, and S3, we summarize the results of different programs on the Easy, Medium and Hard targets, respectively. Figure 1 is the histogram of the average TM-score achieved by different programs. As shown in Table S1, HHsearch programs generate the highest TM-score in the Easy targets. MUSTER and other structure-assisted alignment methods (dPPAS, SP3 etc) generally outperform the HHsearch programs in the Medium and Hard targets. This data demonstrates the usefulness of structure-based features in detecting the distant homologous templates.

Specificity of alignment programs. Except for the accuracy of the template alignments (or sensitivity), the specificity of the alignments (i.e. the correlation of the scoring function and the accuracy of the final alignments) is another important measurement of the alignment algorithms, as this correlation essentially decides how the results can be used in the comparative structural modeling and function annotations.

In Figure 2, we present the TM-score data of the highest ranked alignment models versus the alignment scores by the 20 alignment programs. Here, we tried both the default alignment score of the programs and the Z-score (defined as the difference between the raw alignment score and average in units of standard deviation), and chose the one with the highest correlation to the TM-score of the final models to present in the plot. As expected, positive correlations are observed for all the alignment programs, with PPAS, SAM and MUSTER having the highest Pearson correlation coefficients (0.789, 0.787, and 0.782, respectively). The NW-align and BLAST programs have the lowest correlation coefficient because a number of targets have a high alignment score but with low quality (TM-score $<$ 0.5), indicating a low specificity of the programs.

We also mark in Figure 2 an alignment score cut-off that minimizes the false positive rate, FPR = FP/(FP + TN), and the false negative rate, FNR = FN/(TP + FN), where a model of TM-score $>$ 0.5 is defined as a positive hit that has the correct fold[48]. The score cut-offs, false positive and false negative rates are listed in Table 2. The programs with alignment score that are calibrated by the statistics of random samples, including PSI-BLAST, SAM and FFAS, have the lowest FPR + FNR values, i.e. the highest specificity, based on this measurement. Meanwhile, the Easy and Hard targets are clearly grouped in the right-up and left-bottom regions in Figure 2 for all programs, demonstrating the dependence of the performance of the alignment algorithms on the evolutionary distance of target and templates.

Profile-based alignments versus sequence-based alignments. Depending on whether the homologous sequences are included in the target-template alignments, the alignment methods can be grouped into the three general categories of sequence-to-sequence alignment (including NW-align, SW-align and BLAST), sequence-to-profile alignment (PSI-BLAST, SAM and PSA), and profile-to-profile alignment (PRC, HHsearch-I, HHsearch-II, PPA, PPAS, dPPAS, MUSTER, PROSPECT, SPARKS, SP3 and FFAS). Since the sequence profiles derived from multiple sequence alignment of protein families contain important information of conserved/diverged locations along the sequences, the profile-based alignments can generally generate more accurate target-template alignments than that made by single sequence-based alignments[16,49].

Such insight is also observed in our data analysis. As shown in Table 1 (rows highlighted in bold), the average TM-score obtained by the sequence-to-profile based methods is 18.4% higher than the TM-score from the sequence-to-sequence based methods. Similarly, the TM-score from profile-to-profile alignment methods is 49.8% higher than that of sequence-to-sequence based methods. These increases in TM-score are not only due to the higher coverage of alignments (81.8% vs. 62.6%), but also the enhanced accuracy of alignments as the average RMSD is reduced slightly in the profile-profile methods from 10.4 to 10.2 Å. This tendency is also seen in Tables S1–3 where

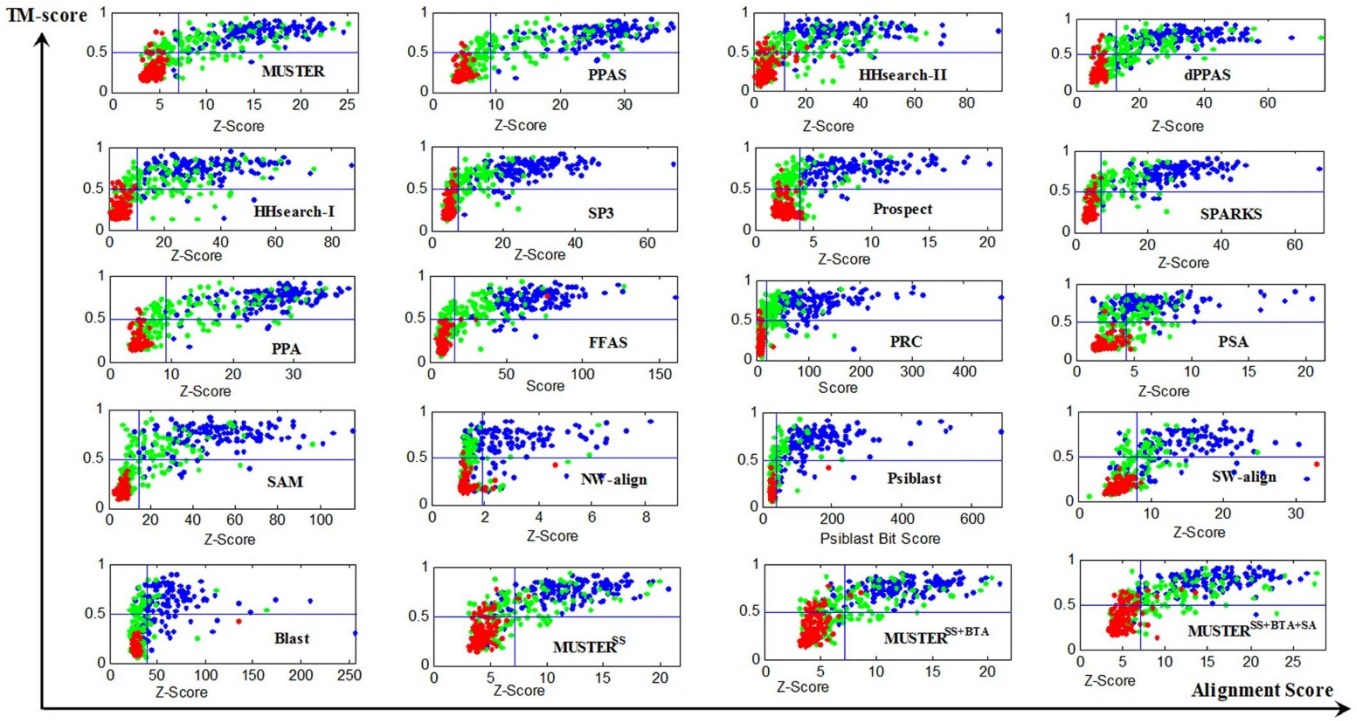

Figure 2 | TM-score of full-length models of 20 threading methods on 538 non-homologous proteins versus the alignment scores. Easy, Medium and Hard targets are colored blue, green and red, respectively. PSI-BLAST, BLAST and PRC use bit score and others use z-score to score the alignments.

the targets were categorized into different groups of Easy, Medium, and Hard, demonstrating that the profile-based alignments enhance both close and distant homology identifications.

Two types of sequence profiles, PSSM and HMM, are often exploited in various alignment methods. Given a MSA, the PSSM profile is designed to account for the estimated frequency of amino acids at each position, while the HMM profile accounts for both amino acid frequency and position-specific probabilities for insertion and deletion. Although the HMMs seem to contain additional gap information from MSAs, there is no obvious difference between the HMM-based (e.g. HHsearch-I and –II) and PSSM-based alignment algorithms (e.g. PPAS), in terms of the TM-score of the alignment models (Table 1). However, HMM-based methods did generate higher TM-scores than PSSM-based methods in the Easy targets (Table S1). Additionally, the HMM-based methods have generally a lower RMSD and lower coverage of alignments, indicating that the HMM method is more sensitive in detecting local structural motifs and scaffolds.

Meanwhile, there are a number of targets that have the correct templates identified by either HMM- or PSSM-based methods (but not both), demonstrating that these two types of methods are complementary to each other. This complementarity from multiple alignment algorithms is essential to the success of meta-server based structure modeling approaches[44,50].

Table 2 | Score cutoffs and false positive and negative rates of different programs

Methods*	Cutoff	FPR	FNR	FPR + FNR
PSI-BLAST	50.4	0.093	0.094	0.187
SAM	14.5	0.129	0.099	0.229
FFAS	12.9	0.170	0.100	0.270
PPA	7.8	0.126	0.158	0.284
SP3	6.5	0.117	0.175	0.292
SPARKS	6.4	0.111	0.194	0.306
SW-align	8.0	0.162	0.149	0.311
PRC	20.8	0.110	0.205	0.316
BLAST	35.7	0.149	0.169	0.318
PPAS	6.9	0.186	0.145	0.331
dPPAS	13.2	0.113	0.233	0.346
HHsearch-I	8.1	0.172	0.179	0.351
MUSTER	6.2	0.147	0.205	0.353
HHsearch-II	9.3	0.16	0.200	0.360
PROSPECT	4.2	0.075	0.317	0.392
PSA	4.1	0.128	0.400	0.528
NW-align	1.5	0.464	0.160	0.625

*Methods sorted by sum of false positive rate (FPR) and false negative rate (FNR).

How much space is left for improvement by structural feature prediction? The performance of profile alignments could be further improved by incorporating structural information. One example is secondary structure comparison, which has been used by almost all contemporary alignment/threading methods to guide the target-template alignments. As a quantitative test of the impact of secondary structure information on alignment accuracy, we developed two sequence profile-profile based methods, PPA and PPAS, where the only difference is that PPAS contains a secondary structure match in the scoring function but PPA doesn't (see Eqs. 3 and 4 in METHODS). As a result, the inclusion of secondary structure prediction increases the TM-score of PPA by 6.8%. MUSTER is another typical profile-profile alignment based algorithm that incorporates multiple composite structure features in the alignments, including secondary structure, residue depth, solvent accessibility, and backbone torsion angle predictions. These features result in a TM-score increase of 9.6% compared to the PPA method, corresponding to a p-value $< 10^{-14}$ in paired student t-test.

The performance of the structure feature assisted algorithms relies on the accuracy of structure feature predictions for the target sequence. In our test on the 538 non-homologous proteins, the average Q3 score (three structure states per residue overall accuracy) for PSSpred and PSI-pred is 83.1% and 80.7%, respectively; the mean absolute errors in ψ and φ angle predictions are 28° and 41°, respectively; and the Pearson correlation coefficient correlation between

predicted and actual solvent accessibility is 0.678. Incorrect assignments of the structure features can compromise the performance of MUSTER. In fact, we observed a number of cases where the TM-score of the alignments by MUSTER, which considers additional structural features, is lower than that of PPAS.

In order to explore the potential of the alignment improvement obtained by considering structural features, we implemented MUSTER using the native structure features derived from the target structures, where the weighting parameters were re-optimized in a separate test of 100 proteins. As shown in Table 1, the average TM-score of the full-length models from MUSTER alignments showed a gradual increase from 0.449 to 0.511, when we exploited more native structure features from secondary structures ($MUSTER^{SS}$), backbone torsion angle ($MUSTER^{SS+BTA}$), and solvent accessibility ($MUSTER^{SS+BTA+SA}$). This change corresponds to an overall increase of 13.8% in the average TM-score.

In Figure 3, we present an illustrative example from the PP7 bacteriophage coat protein (PDB ID: 2qudA), which has the secondary structures arranged as β1-β2-β3-β4-β5-β6-α1-α2 from the N- to C-terminals. The PSI-pred method has however mis-assigned most secondary structure elements in 12S-90M (see '*' in Figure 3A), which resulted in the first two beta-strands (7T-11S and 15A-25T) in the target mis-aligned to the coiled regions in the template 1qbeB by MUSTER. This alignment has a TM-score = 0.607. When using the actual secondary structure assignment, the $MUSTER^{SS}$ program correctly matches the two beta strands of the target with the strands on the template. Based on the same template, the correction of the secondary structure comparison increases the TM-score of the model to 0.7 in this example.

Despite the significant increase in alignment accuracy brought by the integration of structure features, the quality of the alignments using the best structure features from the native is still far from the best templates detected by structural alignments, i.e. the average TM-score by TM-align is 37.1% higher than that by $MUSTER^{SS + BTA + SA}$ (Table 1), which demonstrates considerable room for further alignment improvement. The gap is relatively small in Easy targets (7.4%) according to the data in Table S1, which indicates that the current state-of-the-art alignment methods generate nearly optimal alignments for close homology targets. But for the Medium and Hard targets, the gaps become highly significant, which correspond to a TM-score difference of 35.6% and 79.2%, respectively (Tables S2 and S3). Apparently, such gaps cannot be filled by solely improving the structure feature prediction methods, and a completely different alignment system based on novel scoring and alignment schemes might be required.

Discussion

We developed a comprehensive experiment to systematically examine the strength and weakness of 20 representative sequence alignment methods for template-based protein structure prediction. The data analysis demonstrates the dominant advantage of profile-profile based alignment methods in protein template identification, which

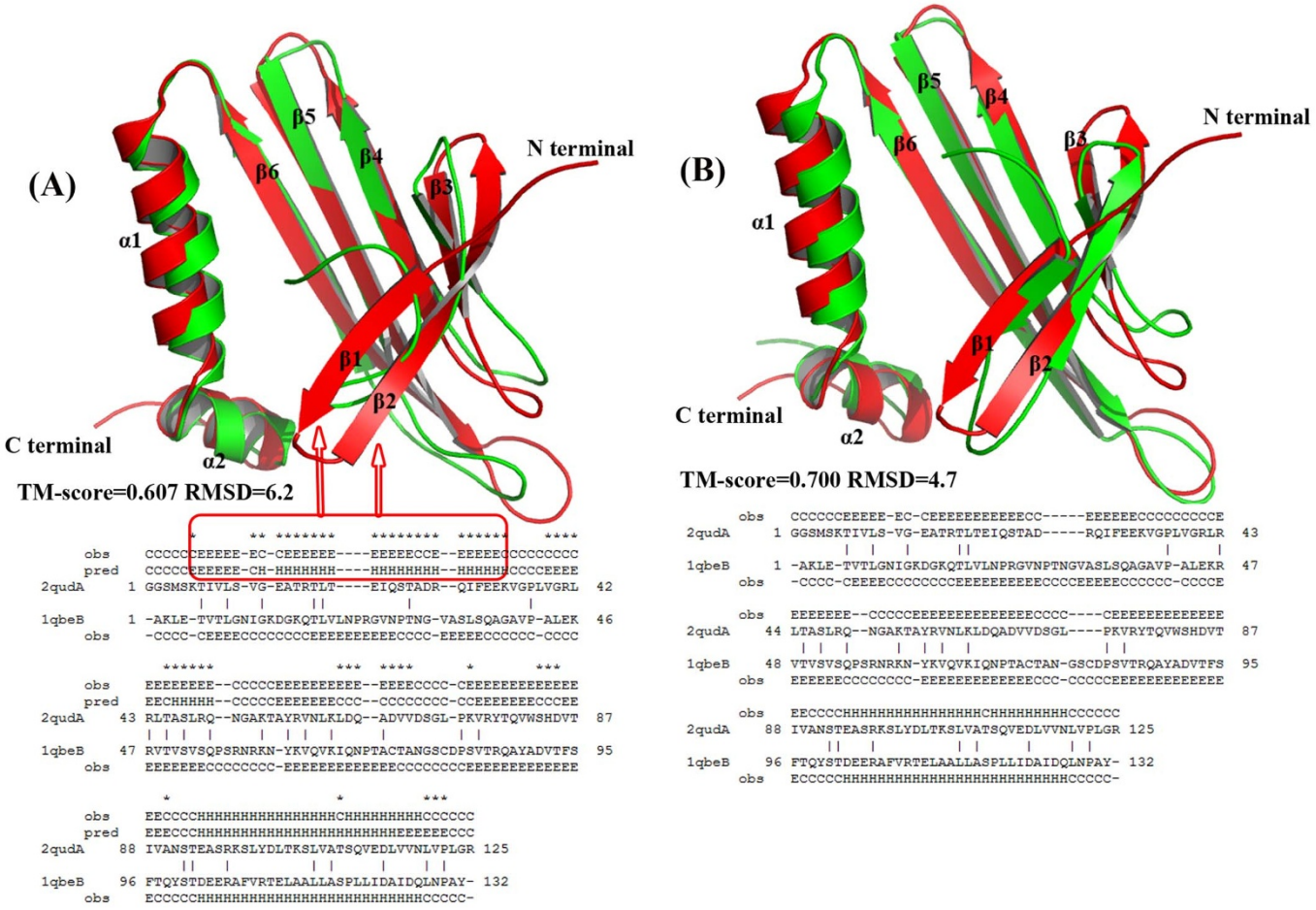

Figure 3 | The illustration of template identifications for 2qudA. (A) MUSTER with predicted secondary structure; (B) $MUSTER^{SS}$ with native secondary structure. The experimental structure and MUSTER models are shown in red and green cartoons, respectively, and the first two beta-strands in (A) are in yellow on the template. The secondary structures are labeled as 'C' for coil, 'E' for strand and 'H' for helix, where 'Pred' and 'Obs' denotes the PSI-pred prediction and the native, respectively. '*' in (A) marks the residues with mis-predicted secondary structure.

generates structural models with an average TM-score 26.5% higher than sequence-profile alignments, and 49.8% higher than single sequence-sequence alignments. The superiority of profile-based alignments over sequence-based alignments was also observed in previous benchmark studies[35,51].

The sequence profiles are typically constructed by PSI-BLAST and hidden Markov model (HMM) searches, where the former is usually specified by a position-specific scoring matrix (PSSM) and the latter by a trained chain model of matches and insertions/deletions. Our data analysis showed that there is no obvious difference between PSSM and HMM profiles in terms of overall average TM-score, although the HMM based alignments tend to obtain higher TM-scores in Easy targets and generate alignments of higher accuracy but with lower alignment coverage. This data seems in contradiction to the results by Park et al[41] and Madera and Gough[42] who concluded that the HMM based methods consistently outperform PSI-BLAST. We believe that the major reason for the seeming contradiction is due to the difference in sample preparations. In our testing dataset, we intentionally included more medium and hard targets to keep a balanced category distribution in difficulty and all templates with a sequence identity $>$ 30% to the target were excluded. In the experiments by Park et al and Madera and Gough, however, the authors collected large-scale proteins from SCOP without intention to include more hard proteins. In addition, the authors used a sequence identity cutoff 40% for template filtering, which includes homologous templates with a sequence identity in 30–40% that are easy to detect by most threading methods. Therefore, it is anticipated that most of the test proteins in these two studies should correspond to Easy targets in our categorization and their conclusion on the HMM and PSSM profile comparisons is in fact consistent with our analysis on the Easy proteins (Table S1).

The profile-based sequence alignments can be considerably improved by the combination of structure feature predictions. For example, the program of MUSTER, which combines profile alignments with sequence-based secondary structure, residue depth, torsion angle and solvation predictions, has a 9.6% higher TM-score on average when compared to the profile-profile alignment algorithms. The performance of structure-assisted methods relies on the accuracy of the sequence-based structure feature predictions, which can be further improved by nearly 10.8% (or 13.8% in full-length models) if the native structure features extracted from experimental structures are exploited. Nevertheless, the latter is still far worse from the best possible templates as identified by structural alignment program TM-align, which uses the target structure as a probe to generate the optimal alignments. In the Easy, Medium and Hard categories, the TM-score by TM-align is 7.4%, 35.6%, and 79.2% higher than that of MUSTER$^{SS + BTA + SA}$. While filling such a big gap is one of the most urgent goals in template-based protein structure prediction, it apparently cannot be achieved solely by the improvement of structure feature prediction methods. New algorithms with completely novel scoring functions and alignment search engines are probably needed to attack the central problem of sequence alignment, which is essential to template-based protein structure prediction and function annotations.

Methods

Twenty threading/alignment methods, covering different categories of target-template alignment algorithms and possible to install at local computers, are benchmarked in this article. All algorithms without cited references are newly developed in house and first presented in this study.

1. NW-align. NW-align is a sequence-to-sequence alignment program constructed based on the standard Needleman-Wunsch dynamic programming algorithm[10]. The amino acid mutation matrix is from BLOSUM62[52] with gap opening penalty $= -11$ and gap extension penalty $= -1$.

2. SW-align. SW-align is a sequence-to-sequence alignment program using a similar setting as NW-align but with dynamic programming based on the standard Smith-Waterman algorithm[53]. The major difference from NW-align is that the negative score values are set to zero and the alignment trace-back procedure starts from the highest scoring cell and ends with a cell of zero score in SW-align. This setting allows SW-align to identify subsequence motifs having the highest local sequence similarity. The source codes and the executables of both NW-align and SW-align are available at http://zhanglab.ccmb.med.umich.edu/NW-align.

3. BLAST. BLAST[13] is a local sequence alignment tool based on a heuristic searching method, where high-scoring segment pairs (HSPs, or words) are first identified by gapless comparisons. The final alignments are constructed by extension and connection of the HSP regions. The heuristic algorithm in BLAST is often suboptimal but much faster than the optimal dynamic programming algorithms.

4. PSI-BLAST. PSI-BLAST[14] is a sequence-to-profile alignment program extended from BLAST which aims to increase the alignment sensitivity of distant homologous proteins by iterative MSA search. It first collects a list of close homologous sequences from a reference database (e.g. NCBI non-redundant sequence database, NR) by BLAST. A PSSM is then derived from the MSA of the sequence homologies, which is used to search against the reference database again to identify a newer set of homologous sequences. The procedure can be repeated a number of times until the PSSM profiles converge. In our test, PSI-BLAST was searched against NR database for 3 iterations using an E-value cutoff, which assesses the significance of the HSP score, below 0.001.

5. PSA. PSA is sequence-to-profile alignment algorithm based on the Needleman-Wunsch dynamic programming. The scoring function of the ith position in the query (q) aligned with the jth position in the template (t) is

$$Score_{\mathrm{PSA}}(i,j) = \sum_{k=1}^{20} F_q(i,k) * B_t(k,j) + shift \quad (2)$$

where $F_q(i, k)$ represents the frequency profile of kth amino acid at ith position of the query. $B_t(k, j)$ denotes a BLOSUM mutation score between the amino acid k and jth residue of the template. The *shift* parameter is introduced to avoid the alignment of unrelated residues in the local regions. Parameters of *shift* (-0.01), gap opening (*go*, -8.6) and gap extension (*ge*, -0.9) penalties were optimized on the ProSup dataset[54].

6. PPA. PPA is an in-house profile-profile alignment method on the Needleman-Wunsch algorithm. The scoring function is defined by

$$Score_{PPA}(i,j) = \sum_{k=1}^{20} F_q(i,k) * L_t(k,j) + shift \quad (3)$$

where $F_q(i, k)$ and $L_t(j, k)$ stand for the sequence frequency profile of query and the log-odds profile of template, respectively. To build the sequence profiles, the sequences are searched against the NR by 3 PSI-BLAST iterations, at an E-value cutoff 0.001. The Henikoff weighting scheme[55] is then used to generate frequency or log-odds profiles. Similarly, the parameters of *shift* (-0.94), *go* (-6.8), and *ge* (-0.52), are optimized by trial and error using the ProSup dataset.

7. PPAS. PPAS is an in-house profile-profile alignment method that combines profile log-odds score and secondary structure comparison. The scoring function is defined by

$$Score_{PPAS}(i,j) = Score_{PPA}(i,j) + C_1\delta\big(S_q(i),S_t(j)\big) \quad (4)$$

where $Score_{PPA}(i, j)$ is defined in Eq. 3, $\delta(S_q(i), S_t(j))$ is the Kronecker delta function to assess the secondary structure match between target and template. $S_q(i)$ is the secondary structure of the ith residue on the target predicted by PSSpred (http://zhanglab.ccmb.med.umich.edu/PSSpred), and $S_t(j)$ is the secondary structure of the jth residue on the template structure assigned by DSSP. A position-specific gap penalty scheme is used in the alignment search, i.e. no gap is allowed inside the secondary structure regions, *go* and *ge* penalties apply to other regions, and the ending gap-penalty is neglected. The four parameters C_1 (0.65), *shift* (-0.96), *go* (-7.0), and *ge* (-0.54), are optimized for PPAS in a similar way as PPA.

8. dPPAS. dPPAS is an in-house profile-profile alignment program extended from PPAS. The only difference from PPAS is that a structure fragment depth profile is added in dPPAS to enhance the alignments, i.e.

$$Score_{dPPAS}(i,j) = \frac{\sum_{k=1}^{20} F_q(i,k)\big(L_{str}(j,k) + L_t(j,k)\big)}{2} + C_1\delta\big(S_q(i),S_t(j)\big) + shift \quad (5)$$

where $F_q(i, k)$ and $L_t(j, k)$ are defined in Eq. 3. $L_{str}(j, k)$ is a frequency depth profile derived from a set of structural fragments that have similar depth as the fragment at jth position of the template[17,19]. Similarly, the parameters ($C_1 = 6.5$, *shift* $= -0.96$, *go* $= -7.0$ and *ge* $= -0.54$) are optimized on the ProSup dataset.

9. MUSTER. MUSTER[17] is a profile-profile based threading program which combines multiple sequence and structure matching information. In addition to the sequence profiles obtained by PSI-BLAST searches, the scoring function contains secondary structure match (SS), fragment depth profiles, solvent accessibility (SA), backbone torsion angles (BTA), and hydrophobic scoring matrix. The optimal

alignment is generated by Needleman-Wunsch dynamic programming. Compared to dPPAS, MUSTER contains additional terms from SA, BTA, and hydrophobic scoring matrix matches, whereby the weighting parameters are re-trained by a new dataset.

To further examine the potential of structure-assisted threading algorithms, we developed three variants of MUSTER programs, MUSTER^{SS}, MUSTER$^{SS + BTA}$ and MUSTER$^{SS + BTA + SA}$, which exploit the SS, BTA, and SA features extracted from the experimental structures of the target. Similar to MUSTER, all parameters in these algorithms are optimized in a separate training set of 100 non-redundant proteins by maximizing the TM-score.

10. SAM. SAM[56] is a hidden Markov model (HMM) based protein fold-recognition method. Starting from the PSI-BLAST search, SAM constructs a HMM profile based on the iterative MSA searches. The HMM profile is then used to search through the PDB library to identify structural templates. SAM can conduct both local and global alignment searches and we use the local alignment mode in this work.

11. PRC. PRC[57] is a program for scoring and aligning profile HMMs. To run PRC, we first construct HMMs of both target and template sequences by SAM[56]. The HMM–HMM based alignments are then computed and ranked by PRC which is designed to find the Viterbi path that maximizes the sum of forward–backward odds scores.

12. HHsearch-I and HHsearch-II. HHsearch[24] is a HMM-HMM based alignment program which combines the profile log-odds score and the secondary structure prediction in the Viterbi dynamic programming. We run two versions of HHsearch: HHsearch-I uses PSI-BLAST to start the MSA search for building the profile HMMs for target and template sequences, while HHsearch-II uses HHblits to construct the profile HMM for target sequences. HHblits uses a discretized-profile prefilter that can generate HMM profiles faster than PSI-BLAST[58]. The final query-template alignments are constructed by the same HHsearch program. Both HHsearch-I and HHsearch-II are in the local alignment mode.

13. PROSPECT. PROSPECT[21] is a sequence profile-profile alignment algorithm assisted with a residue-level contact potential and SS predictions. A global optimization of target-template alignment is generated by the divide-and-conquer searching method.

14. SPARKS and SP3. Both SPARKS[18] and SP3[19] were developed in Zhou Lab. In SPARKS, the authors exploit a sequence profile–profile alignment combined with a single-body statistical potential; in SP3, they use a residue depth dependent structure profile to replace the single-body potential used in SPARKS.

15. FFAS. FFAS[59] is a sequence profile-profile based alignment program. It calculates the sequence profile by PSI-BLAST searching against the NR85s database with 5 iterations. A dot-product scoring function is then used to align two sequence profiles. The alignment score is finally translated into a statistical measure by comparing it with the distribution of scores obtained for pairs of unrelated proteins.

1. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
2. Zhang, Y. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* **18**, 342–348 (2008).
3. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
4. Zhang, Y. & Skolnick, J. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. USA* **102**, 1029–1034 (2005).
5. Skolnick, J., Zhou, H. Y. & Brylinski, M. Further Evidence for the Likely Completeness of the Library of Solved Single Domain Protein Structures. *Journal of Physical Chemistry B* **116**, 6654–6664 (2012).
6. Sanchez, R. & Sali, A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A* **95**, 13597–13602 (1998).
7. Malmstrom, L. *et al.* Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *Plos Biol* **5**, e76 (2007).
8. Zhang, Y. & Skolnick, J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA* **101**, 7594–7599 (2004).
9. Xu, D. & Zhang, Y. Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. *Sci Rep* **3**, 1895 (2013).
10. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **48**, 443–453 (1970).
11. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
12. Lipman, D. J. & Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441 (1985).
13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
14. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
15. Bowie, J. U., Luthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170 (1991).
16. Gribskov, M., McLachlan, A. D. & Eisenberg, D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A* **84**, 4355–4358 (1987).
17. Wu, S. & Zhang, Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* **72**, 547–556 (2008).
18. Zhou, H. & Zhou, Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **55**, 1005–1013 (2004).
19. Zhou, H. & Zhou, Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* **58**, 321–328 (2005).
20. Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* **9**, 232–241 (2000).
21. Xu, Y. & Xu, D. Protein threading using PROSPECT: design and evaluation. *Proteins* **40**, 343–354 (2000).
22. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**, 1501–1531 (1994).
23. Karplus, K., Barrett, C. & Hughey, R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856 (1998).
24. Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
25. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
26. Wu, S. & Zhang, Y. ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. *PLoS ONE* **3**, e3400 (2008).
27. Chen, H. & Zhou, H. X. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res* **33**, 3193–3199 (2005).
28. Yang, Y., Faraggi, E., Zhao, H. & Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **27**, 2076–2082 (2011).
29. Skolnick, J., Kihara, D. & Zhang, Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Protein* **56**, 502–518 (2004).
30. Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. & Tramontano, A. Critical assessment of methods of protein structure prediction - Round VIII. *Proteins* **77 Suppl 9**, 1–4 (2009).
31. Battey, J. N. *et al.* Automated server predictions in CASP7. *Proteins* **69**, 68–82 (2007).
32. Fischer, D., Rychlewski, L., Dunbrack, R. L., Jr., Ortiz, A. R. & Elofsson, A. CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins* **53 Suppl 6**, 503–516 (2003).
33. Rychlewski, L. & Fischer, D. LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci* **14**, 240–245 (2005).
34. Kinch, L. N. *et al.* CASP9 target classification. *Proteins* **79 Suppl 10**, 21–36 (2011).
35. Sauder, J. M., Arthur, J. W. & Dunbrack, R. L., Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* **40**, 6–22 (2000).
36. Qi, Y., Sadreyev, R. I., Wang, Y., Kim, B. H. & Grishin, N. V. A comprehensive system for evaluation of remote sequence similarity detection. *BMC Bioinformatics* **8**, 314 (2007).
37. Wang, G. & Dunbrack, R. L., Jr. Scoring profile-to-profile sequence alignments. *Protein Sci* **13**, 1612–1626 (2004).
38. Raghava, G. P., Searle, S. M., Audley, P. C., Barber, J. D. & Barton, G. J. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics* **4**, 47 (2003).
39. Van Walle, I., Lasters, I. & Wyns, L. SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics* **21**, 1267–1268 (2005).
40. Elofsson, A. A study on protein sequence alignment quality. *Proteins* **46**, 330–339 (2002).
41. Park, J. *et al.* Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* **284**, 1201–1210 (1998).
42. Madera, M. & Gough, J. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res* **30**, 4321–4328 (2002).
43. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
44. Wu, S. T. & Zhang, Y. LOMETS: A local meta-threading-server for protein structure prediction. *Nucl. Acids. Res.* **35**, 3375–3382 (2007).
45. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
46. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
47. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic. Acids Res.* **33**, 2302–2309 (2005).
48. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
49. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

50. Ginalski, K., Elofsson, A., Fischer, D. & Rychlewski, L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**, 1015–1018 (2003).
51. Edgar, R. C. & Sjolander, K. A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics* **20**, 1301–1308 (2004).
52. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915–10919 (1992).
53. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *Journal of molecular biology* **147**, 195–197 (1981).
54. Domingues, F. S., Lackner, P., Andreeva, A. & Sippl, M. J. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *Journal of molecular biology* **297**, 1003–1013 (2000).
55. Henikoff, S. & Henikoff, J. G. Position-based sequence weights. *J Mol Biol* **243**, 574–578 (1994).
56. Karplus, K. *et al.* Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* **53 Suppl 6**, 491–496 (2003).
57. Madera, M. Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics (Oxford, England)* **24**, 2630–2631 (2008).
58. Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173–175 (2012).
59. Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. FFAS03: a server for profile–profile sequence alignments. *Nucleic acids research* **33**, W284–288 (2005).

Acknowledgements

We are grateful to Dr. Jeffrey Brender for reading the manuscript. The project is supported in part by the NSF Career Award (DBI 1027394), the National Institute of General Medical Sciences (GM083107, GM084222), and the NSFC (31128004).

Author contributions

R.Y. and Y.Z. conceived the project; R.Y. conducted the calculations and analyzed the data; D.X., J.Y. and S.W. participated in discussions; R.Y. and Y.Z. wrote the manuscript. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Yan, R., Xu, D., Yang, J., Walker, S. & Zhang, Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci. Rep.* **3**, 2619; DOI:10.1038/srep02619 (2013).