IMPROVING PROFILE SIMILARITY SEARCH AND ALIGNMENT OF

PROTEIN SEQUENCES

APPROVED BY SUPERVISORY COMMITTEE

_____

Nick Grishin, Ph.D; Advisor

_____

Rama Ranganathan, Ph.D; Committee Chair

_____

Zbyszek Otwinowski, Ph.D

_____

Dominika Borek, Ph.D

IMPROVING PROFILE SIMILARITY SEARCH AND ALIGNMENT OF PROTEIN SEQUENCES

by

JING TONG

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center

Dallas, Texas

December, 2015

IMPROVING PROFILE SIMILARITY SEARCH AND ALIGNMENT OF PROTEIN SEQUENCES

JING TONG

The University of Texas Southwestern Medical Center at Dallas, 2015

Supervising Professor: NICK V. GRISHIN Ph.D

Protein function prediction is one of the most important problems in the field of computational biology. The most reliable method to predict protein function is to detect homologs. Homologous proteins tend to possess conserved sequence motifs, the same structure folds, and similar functional sites. Current sequence-based homology search methods are still unable to detect many similarities evident from protein spatial structures. We present a new method, COMPADRE, to assess the relationship between the query sequence and a hit in the database by considering the similarity between the query and hit's known homologs. This method markedly boosts the homology detection precision rate.

Successful homology-based protein function prediction is also determined by accurate alignment between a protein sequence and its homolog. Alignment errors are the main bottleneck for homology modeling when the query is distantly related to the template. Alignment methods often misalign secondary structural elements by a few residues. We present a refinement method, SFESA, to improve pairwise sequence alignments by evaluating alignment variants generated by local shifts of template-defined secondary structures.

The potential values of these methods for structure/function predictions are illustrated by the detection of homology between evolutionary distant yet structurally similar protein domains.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

Tong, J., Pei, J., Sadreyev, R., and Grishin, N. COMPADRE server for protein homology detection. *Bioinformatics.* 2015 (Submitted)

Tong, J., Sadreyev, R., Pei, J., Kinch, L. and Grishin, N. Using homology relations within a database markedly boosts protein sequence similarity search. *Proceedings of the National Academy of Sciences of the United States of America.* 2015;112(22):7003-8

Tong, J., Pei, J., Otwinowski, Z. and Grishin, N. Refinement by shifting secondary structure elements improves sequence alignments. *Proteins: Structure, Function, and Bioinformatics.* 2015; 83(3):411-27

Tong, J., Pei, J. and Grishin, N. SFESA web server for pairwise alignment refinement by evaluation of secondary structure shifts. *BMC Bioinformatics.* 2015; 16:282

Shi, S., Pei, J., Sadreyev, R, Kinch, L., Majumdar, I., Tong, J., Cheng, H., Kim, B. and Grishin, N. Analysis of CASP8 targets, predictions and assessment methods. *Database (Oxford).* 2009;bap003

List of Figures

List of Tables

I. INTRODUCTION

Knowing protein functions is the key to understanding the nature of the protein universe and in essence, biology (Bork, Dandekar et al. 1998). In the field of computational biology, prediction of protein function is one of the most essential problems (Baker and Sali 2001). Being different than expensive and time-consuming experimental methods to solve structures, e.g., crystallography (Ealick 2000) or NMR spectroscopy (Tyszka, Fraser et al. 2005), computational structure/function prediction methods can economically and efficiently provide biologists with functional hypotheses about their proteins of interest. In recent years, protein function detection is becoming more and more important because of the massive amounts of protein sequence data accumulated by advanced genome sequencing technology (Barnhart 1989) as well experimentally determined protein structures in the PDB database (Berman, Battistuz et al. 2002).

Currently, the most reliable method to predict protein function is to detect homologous proteins (Daga, Patel et al. 2010). Homologous proteins are proteins with a common ancestor that usually possess conserved sequence motifs, the same structure folds, and similar functional sites (Doolittle 1981). High sequence similarity alone or combined sequence-structure similarity is often used to establish homologous relationships between proteins. Distinctive structure features, similar structural folds, conserved sequence motifs and functional sites are often used to further support or verify the inference of homology (Floudas 2007).

When a confident homolog is found below a significant cutoff, alignment construction between homologous proteins can further help functional site prediction or structure modeling (Ginalski 2006). From our recent observation of protein sequence alignments, some errors are prone to occur for short secondary structure alignments. They may have a relatively high similarity score but should not be aligned together according to their structure alignment reference. Thus, alignment errors remain as one of the main bottlenecks in homology modeling and function prediction.

Generally speaking, accurate homology-based protein function prediction is determined by two factors (Marti-Renom, Stuart et al. 2000). The first factor is whether the correct homolog can be identified (the accuracy level of the homology detection method). The second factor is whether the correct alignment can be constructed between the protein sequence and its homolog (the accuracy level of the alignment method).

We will discuss these two topics in the following sections.

HOMOLOGY DETECTION AND SIMILARITY SEARCH

Detecting protein sequence homology to known folds offers the most successful and practically useful strategy to predict protein function (Baker and Sali 2001). As we discussed above, homologous proteins usually possess conserved sequence motifs and share similar structures and functions, but may only have subtle overall sequence similarities. Even very distant homologs can provide reasonable templates for modeling protein sequence targets without closely related structures (Kryshtafovych, Moult et al.

2014). Although significant progress has been attained in this direction, remote sequence similarity is still far from satisfaction (Huang, Mao et al. 2014).

In the history of homology detection method development, BLAST (Basic Local Alignment Search Tool) (Altschul, Gish et al. 1990) is the first method proposed to compare protein sequence similarity by using protein sequence alignment. Also, it is still one of the most widely utilized similarity search programs, which performs very well for proteins with high sequence identity. After BLAST, PSI-BLAST (Position Specific Iteration BLAST) (Altschul, Madden et al. 1997) was developed. For a protein sequence (query), PSI-BLAST can find homologous sequences in the search database using not only the query sequence information but also its homologous proteins. The position specific matrix, or profile, used in PSI-BLAST represents the query and its homologs by a similar set of sequence patterns that includes insertions and deletions. Such work was later followed by methods for profile-profile comparison (Rychlewski, Jaroszewski et al. 2000, Sadreyev and Grishin 2003, Soding 2005, Madera 2008, Margelevicius and Venclovas 2010, Remmert, Biegert et al. 2012), aimed at detecting similarities between distant families. As examples, COMPASS (Sadreyev and Grishin 2003) and HHSearch (Soding 2005) represent the state-of-the-art for profile-profile comparisons methods. In order to more accurately search for sequence similarity, PROCAIN (Wang, Sadreyev et al. 2009) was proposed to incorporate similarity in secondary structure, positional conservation, and sequence motifs into profile–profile scoring.

Besides using profile-based sequence information, a novel direction to further improve sequence-based homology search is to incorporate non-sequence information. In a typical homology search aimed at protein structure and function prediction, a sequence

or family of interest is compared to a database of proteins with known structures. This knowledge allows confident establishment of evolutionary links within the database. In computer science, networks of relationships between database subjects have been successfully used to improve the quality of search methods, most notably web searchers (Brin and Page 1988). In Chapter II and III of this thesis, we show that knowledge of the protein database homology network dramatically increases the accuracy of sequence-based search.

PROREIN SEQUECNE ALIGNMENT

For a protein sequence (query), a homolog (template) can be found by using profile similarity search. In order to predict functional sites or model the three-dimensional structure of a protein sequence, the next important step is to construct an accurate alignment between the query and template (Marti-Renom, Stuart et al. 2000, Ginalski 2006).

Earlier work focused on dynamic programming recursion in the construction of a global or local alignment. Heuristic methods such as FASTA and CLUSTALW (Lipman and Pearson 1985, Thompson, Higgins et al. 1994) were developed to significantly increase the speed of alignment. Subsequently, sequence profiles (Krogh, Brown et al. 1994) were introduced to construct more accurate alignments by using sequence-profile and profile-profile comparison. These comparisons improved pairwise alignments by scoring the similarity between sequence positions in protein families. In addition to pure sequence methods, 3D structural information is valuable for alignment construction

because protein structures tend to evolve more slowly than protein sequences (Chothia and Lesk 1986, Illergard, Ardell et al. 2009).

Although much work has been done in this field, alignments are still not sufficiently accurate for sequences with low similarity (Kryshtafovych, Moult et al. 2014). Automatic aligners such as PROMALS (Pei and Grishin 2007) frequently misalign alignment blocks by a few residues. Better alignment solutions can frequently be found among a limited set of local shifts of alignment blocks (moving residues in the query relative to the template). This observation motivated us to develop a pairwise alignment refinement method, SFESA, which generates candidate alignment variants for each alignment block by shifting the query region. In addition, residue contact based information can complement sequence information to better distinguish among shifted alignments. The details of alignment improvement will be discussed in Chapter IV and V.

II. USING HOMOLOGY RELATIONS WITHIN A DATABASE MARKEDLY BOOSTS PROTEIN SEQUENCE SIMILARITY SEARCH

Inference of homology from protein sequences provides an essential tool for analyzing protein structure, function, and evolution. Current sequence-based homology search methods are still unable to detect many similarities evident from protein spatial structures. In computer science a search engine can be improved by considering networks of known relationships within the search database. Here, we apply this idea to protein sequence-based homology search and show that it dramatically enhances the search accuracy. Our new method, COMPADRE, assesses the relationship between the query sequence and a hit in the database by considering the similarity between the query and hit's known homologs. This approach increases detection quality, boosting the precision rate from 18% to 83% at half-coverage of all database homologs. The increased precision rate allows detection of a large fraction of new protein structural relationships, thus providing structure and function predictions for previously uncharacterized proteins. Our results suggest that this general approach is applicable to a wide variety of methods for detection of biological similarities.

In the field of protein structure prediction, identifying homology to known folds offers the most successful and practically useful strategy to provide protein spatial structure models. For protein sequence targets without closely related structures, even very distant homologs can provide reasonable templates for modeling. Despite significant progress in the field, remote sequence similarity search is far from perfection and fresh

ideas are needed to extend detection limits. In computer science, the concept of utilizing internal relations within a database to improve similarity search was key to the success of search engines such as Google. Here, we show that similar consideration of the homology network within a protein database of structure templates can dramatically improve the accuracy of homology search.

INTRODUCTION

Prediction of protein structure and function by sequence homology is among the most important problems in computational biology of proteins, perhaps next after the grand problem of *de novo* protein folding. The existing gap between the number of known protein sequences and the number of experimentally determined 3D structures is bound to grow with more genomes sequenced by high-throughput technologies(Koboldt, Steinberg et al. 2013, Mardis 2013). Currently, the most reliable and effective way to predict the structure of an uncharacterized protein is to find a sequence homolog with available structural information(Gribskov, McLachlan et al. 1987, Huang, Mao et al. 2014). The chance of finding such a template for a given protein sequence is increasing as sequence space is becoming more extensively covered by 3D structures(Zhang, Hubner et al. 2006). However, there is, and will be for a long time, a significant fraction of proteins for which finding experimentally characterized sequence homologs is challenging or impossible. The structures of many such proteins, when solved, reveal their remote homology to previously known structures that are undetectable by current sequence-based homology search methods(Kryshtafovych, Moult et al. 2014). Therefore, the quality of sequence-based homology search remains key for accurate structure prediction, as consistently confirmed by multiple rounds of the Critical Assessment of protein Structure Prediction (CASP)(Kryshtafovych, Fidelis et al. 2014).

In the last several years, methods for sequence similarity search have been greatly improved by the analysis of sequence patterns reflecting evolutionary, structural, and functional constraints in protein families. Introduction of numerical profiles(Gribskov, McLachlan et al. 1987) and hidden Markov models (HMM)(Eddy 1998) has allowed

comparing a sequence to a multiple sequence alignment (MSA)(Altschul, Madden et al. 1997, Eddy 1998, Karplus, Barrett et al. 1999). Such work was later followed by methods for profile-profile(Rychlewski, Jaroszewski et al. 2000, Sadreyev and Grishin 2003, Margelevicius and Venclovas 2010) and HMM-HMM(Soding 2005, Madera 2008, Remmert, Biegert et al. 2012) comparison, aimed at detecting similarities between distant families. In addition to the residue substitution preferences at sequence positions, MSA can reveal highly informative patterns of inter-dependence between amino acid content at different positions, in the form of MSA motifs and secondary structure predictions(Ginalski, Pas et al. 2003, Soding 2005, Wang, Sadreyev et al. 2009).

Is it possible to further improve sequence-based homology search by considering non-sequence information? In a typical homology search aimed at 3D structure prediction, a sequence or family of interest is compared to a database of proteins with known structures. This knowledge allows confident establishment of evolutionary links within the database. In computer science, networks of relationships between database subjects have been successfully used to improve the quality of search methods, most notably web searchers (Brin and Page 1988). Here, we show that knowledge of the protein database homology network dramatically increases the accuracy of sequence-based search.

To capitalize on this idea, we modified PROCAIN(Wang, Sadreyev et al. 2009), our sensitive method for sequence profile similarity search, by considering the template's homologs within the database (Figure II-1). We designed the new similarity measure of COMPADRE as a linear combination of the original score for the given template with the scores for a set of its homologs identified by structure, function, and sequence. Consistent

similarity of these homologs to the query elevates the original score, which can increase the significance of a marginal sequence-based similarity to a level above detection threshold. On the other hand, a favorable score for a spurious hit becomes less significant if the set of its homologs is consistently dissimilar from the query. Therefore, the new measure improves both sensitivity and specificity of homology detection.

RESULTS

Defining close and remote template homology networks

To establish the network of homologous relationships in the structure database, we defined homologous and non-homologous pairs in a set of 5116 representative structural domains (also known as templates) from the Structural Classification of Proteins (SCOP) database(Andreeva, Howorth et al. 2008). Because a good homology search method should rank homologs according to their distance from the query as well as discriminate homologs from non-homologs, we considered homologous relationships between protein pairs at two different levels. The first level represents "close homologs", and was defined as domain pairs assigned to the same superfamily in SCOP. These templates are typically closely related to each other by 3D structure, tend to share similarities in sequence and function, and should be ranked higher than more remote homologs. The second level represents "all homologs", and includes more distantly homologous protein pairs. To define "all homologs", the SCOP classification was supplemented with a Support Vector Machine (SVM) classifier(Qi, Sadreyev et al. 2007) that uses a number of sequence and structure similarity scores to establish homology (see Methods in Supporting Information for details). This SVM classifier finds similarities between domain pairs that are most likely evolutionarily related and would be meaningful templates for 3D structure modeling.

Establishing a scoring scheme by using homology networks in the search database

The original PROCAIN E-values reflecting the sequence-based similarity between the query and templates were first log-transformed into similarity scores s_0 (see Methods for details). The closest query homolog can often be identified as the top hit by this direct score. To further improve PROCAIN scoring, similarity scores on a particular template (s_0^T) were boosted by the similarity scores of the template's homologs (s_0^H) according to the following equation:

$$S_1 = w_T S_0^T + (1 - w_T) \sum S_0^H \qquad (1)$$

where s_0^T and s_0^H are the similarity scores s_0 for the given template and for a set *H* of its structure-based homologs within a certain evolutionary distance from the template, respectively. The s_0^H similarity score can be calculated using either the close homolog level (H_{close}) or the all homolog level (H_{all}) described above. w_T is a weight optimized for the performance (w_T = 0.8).

Additional information about the query's top hit may help detecting the query's homologs in the database. Indeed, we find (Figure II-2, Table II-1 and II-2) that performance can be additionally improved by transforming the measure in Eq (1) to boost scores for the templates that are homologous to the *top hit* and to reduce scores for non-homologous templates:

$$S_2 = S_1 + h_{\text{top}}(\alpha * S_1 + \beta) \qquad (2)$$

where s_1 is the measure defined by Eq (1), α and β are optimized parameters, and h_{top} depends on the homology of the template to the top hit that has the highest PROCAIN

score: h_{top} = 1 if the template is homologous to the top hit, h_{top} = -1 if it is confidently non-homologous, and h_{top} = 0 if the homology is unclear (see Methods in Supporting Information for details).

The choice of homology networks lead to different homology detection performances

The choice of homolog set *H* in Eq (1) and Eq (2) has a dramatic influence on the method's behavior. Including scores for all template's homologs (H_{all}) results in a wider sampling of protein space around the template and thus should provide more representative information for homolog/non-homolog discrimination. Such a wide sample, however, may lead to a scrambled ranking among detected homologs, with the closest ones being placed below the more distant. For a query's close homologs, the strong direct similarity signal from the first scoring term of Eq (1) may be diluted by the contribution of a diverse set of template's homologs from the second scoring term of Eq (1). Restricting the set's diversity to close homologs (H_{close}) should improve the ranking of close homologs, but may limit the sensitivity of detection at remote homology levels. Therefore, the size of homolog set *H* may require adjustment for different evolutionary distances between query and template.

Indeed, applying different sets *H* (H_{all} or H_{close}) to generate s_1 and s_2 results in a very different performance of the new scoring scheme. We used receiver operator characteristic (ROC) curves to evaluate the homology detection performance of Eq (2) for all query homologs designated as true positives (Figure II-3a) and for only close homologs designated as true positives (Figure II-3b). A ROC curve plots true positive

numbers against false positive numbers at various E-value thresholds. Both plots (Figure II-3a and Figure II-3b) include the curves for several published sequence-based searchers: the popular PSI-BLAST method (Schaffer, Aravind et al. 2001), PROCAIN(Wang, Sadreyev et al. 2009), and a comparable state-of-the-art method HHSearch(Soding 2005) (see Methods for details). These plots are shown together with those produced by our new scoring scheme using two definitions of *H*. When contribution from all template homologs (H_{all}) is allowed in determining $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$, the quality of homolog/non-homolog discrimination is dramatically higher than in other methods (Figure II-3a, Table II-1 and II-3). However, when the contribution is limited to close homologs (H_{close}), the discrimination between homologs and non-homologs becomes worse, especially for more distant homologs, in the area further from the plot's origin (Figure II-3a, Table II-1 and II-3). The situation is opposite when only close homology detection is evaluated (Figure II-3b, Table II-3 and II-4). Inclusion of all template homologs (H_{all}) in determining $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ of Eq (2) results in extremely poor identification of close homologs, suggesting that more distant relationships are often erroneously assigned higher significance. Limiting the set *H* to the close homology level (H_{close}) leads to an accurate close homology detection far surpassing the original PROCAIN performance (Figure II-3b, Table II-3 and II-4).

Improvement of homology detection is consistent among protein classes

The effects observed performance improvement resulting from eq. (1-2) and its strong dependence on the range of template homologs entering eq. (1) might potentially

be confined to one or a few specific query types that strongly affect overall ROC curve. To assess the universality of these effects, we evaluate the performance separately for subsets of queries from individual major domain classes: all-alpha, all-beta, alpha/beta, and alpha+beta, according to SCOP. These effects are consistent among all major protein secondary structure classes (Figure II-4, II-5, II-6 and II-7).

Improvement of COMPADRE scoring scheme by the choice of homology network

The results shown in Figure II-3 suggest that in order to improve both the detection of remote homology and the ranking by evolutionary distance to the query, we can adjust the contribution from more distant homologs in Eqs (1-2) according to the template's distance from the query. Both goals may be achieved if set *H* is kept relatively narrow for close query-template relationships (the left part of orange curve in Figure II-3a, b), and the input from remote template homologs is added only for templates more distant from the query (the right part of brown curve in Figure II-3a, b). We construct a combined scoring function for such an adjustment:

$$\boldsymbol{S}_3 = \boldsymbol{w_c}(\boldsymbol{S}_2^{\boldsymbol{c}}) * \boldsymbol{S}_2^{\boldsymbol{c}} + (1 - \boldsymbol{w_c}(\boldsymbol{S}_2^{\boldsymbol{c}})) * \boldsymbol{S}_2^{\boldsymbol{r}} \qquad (3)$$

where s_2^c and s_2^r are determined by Eq (2) with different definitions of set *H*: only close homologs (H_{close}) for s_2^c and all database homologs (H_{all}) for s_2^r. The weight w_c is a variable depending on score s_2^c as a measure of closeness of template to query. For high s_2^c values (closely similar to the query, when s_2^c is above an upper boundary $s_2^{c(2)}$) this

weight is set to 1, so that the final score includes only the score of close homologs of a template ($s_3 = s_2^c$), whereas for low s_2^c values (distantly similar to the query, when s_2^c is less than a lower boundary $s_2^{c(1)}$) the weight is set to 0 so that the score includes only the score for all template homologs ($s_3 = s_2^r$). For the intermediate values of s_2^c ($s_2^{c(1)} < s_2^c < s_2^{c(2)}$), the weight monotonically grows from 0 to 1, to gradually mix s_2^c with s_2^r. After testing several functions, we find that exponential dependency of w_c on s_2^c (Figure II-8) provides the best performance (see Methods for details).

While consideration of a template's homologs in Eqs (1-3) can boost scores of marginally detectable homologs, it can also reduce the significance of original PROCAIN E-values for highly confident homologs. Thus, we construct a second combined scoring function:

$$\boldsymbol{S_4 = w_3(\ln E_p) * S_3 + (1 - w_3(\ln E_p)) * S_p} \qquad (4)$$

where s_3 is determined by Eq (3) and s_p is the score obtained from the original PROCAIN E-value E_p using the Gumbel extreme value distribution (EVD)(Gumbel 1935), which approximates a distribution of sequence similarity scores of random comparisons(Altschul, Gish et al. 1990, Karlin and Altschul 1990, Dembo 1994, Altschul, Madden et al. 1997). Since it was introduced in sequence analysis by BLAST(Altschul, Gish et al. 1990), this distribution has been widely used to estimate statistical significance of sequence and profile similarity scores (i.e., to compute E-value from a score) in many applications(Altschul, Madden et al. 1997, Sadreyev, Tang et al. 2007, Wang, Sadreyev et al. 2009). Karlin and co-authors(Karlin and Altschul 1990,

Dembo 1994) estimated parameters of EVD and suggested a formula to transform a sequence score to E-value. E-values can be back-transformed to scores using this approximation. The weight w_3 is a function of lnE_p, the logarithm of the PROCAIN E-value. For low lnE_p values (highly confident PROCAIN hits, when lnE_p is less than a lower boundary $lnE_p^{(1)}$) this weight is set to 0, so that the final score is only determined by the original PROCAIN score ($s_4 = s_p$), whereas for high lnE_p values (marginal PROCAIN hits, when lnE_p is above an upper boundary $lnE_p^{(2)}$) this weight is set to 1 so that the final score s_4 is equal to the new score s_3. For the intermediate values of lnE_p ($lnE_p^{(1)} \leq lnE_p \leq lnE_p^{(2)}$), the weight monotonically increases from 0 to 1, to gradually mix s_3 with s_p. Testing several functions, we find that exponential dependency of w_3 on lnE_p (Figure II-9) gives the best performance (see Methods for details). Based on the score s_4 for a given template, statistical significance of the detected similarity is provided in the form of E-value estimated by transforming the score using the EVD approximation.

The final scoring function s_4 offers best performance both in remote homology detection and in ranking by evolutionary distance to a query. Performance of the resulting measure is compared to several methods in Figure II-10a, b and II-11. The inclusion of all templates' homologs (H_{all}) in the set H leads to highly sensitive and accurate retrieval of homology relationships (Figure II-10a, Table II-1 and II-3). At the same time, using a restricted set H for shorter ranges of query-template distance (only close homologs (H_{close})) leads to the correct placement of the close query homologs above others (Figure II-10b, Table II-2 and II-4). One of the most important characteristics of this scoring scheme is precision rate, i.e., the expected proportion of true positives among top hits. As

shown in Table II-3, the scoring function s_4 achieves the precision rate of 83% at half-coverage of all homologs, more than quadruple that of the original PROCAIN rate of 18%. Thus, the combined measure s_4 by far exceeds the current state-of-the-art performance levels in both capturing remote protein relationships and ranking homologs consistently with evolutionary distance. We refer to the resulting detection method as COMPADRE, for COmparison of Multiple Protein sequence Alignments using Database RElationships.

Comparison to structure similarity score

A more detailed analysis of the COMPADRE results suggests that it accurately captures a large fraction of structural similarities that are only weakly reflected in sequence, and at the same time highlights the similarity of local functional motifs that may be missed by an automatic structure comparison method. As an illustration, Figure II-12 shows the comparison of protein groupings based on COMPADRE E-value and on the structural similarity measured by DALI(Holm and Sander 1993) Z-score. We use 1313 representative protein domains from the α/β class in our database to perform hierarchical clustering by all-to-all COMPADRE scores (logarithm of COMPADRE E-values) (Figure II-12a) and by DALI Z-scores (Figure II-12b). The resulting matrices of scores for domain pairs are represented as colored maps, with the scores used for grouping shown above diagonal and the corresponding scores by the other method shown below diagonal.

These groupings provide several notable observations. First, there is a strong general correlation in clustering by both scores, including the identity of major protein groups such as TIM-barrels, Rossmann fold, and SAM-dependent methyltransferases (Figure II-12a, b). Second, as expected, a fraction of structure-based relationships still remains undetected by sequence. These relationships include both similarities outside major clusters (off-diagonal area of the matrices, Figure II-12b) and links within clusters. For example, TIM-barrels have uniformly high DALI Z-scores, whereas the coverage by COMPADRE scores is more fragmented (Figure II-12a, b). Third, COMPADRE produces several clusters of pronounced similarity that stand out from the background (red in Figure II-12a) but are not produced by DALI. These clusters correspond to local functional sequence motifs whose presence is less obvious from structure comparison alone. The most notable example is P-loop nucleoside triphosphatases that are accurately placed together by COMPADRE but split apart by structure similarity (Figure II-12a). As another example, COMPADRE grouping within the TIM-barrel fold highlights a superfamily of (trans)glycosidases that share similar phosphate-binding sites, which is challenging for DALI-based clustering (Figure II-12a).

Detecting more homologs at the SCOP superfamily level

Compared to the original PROCAIN scoring, COMPADRE detects more homologs at the SCOP superfamily level. As an example, in bacterial lysozyme (PDB ID 1jfxA, SCOP family 1, 4-beta-N-acetylmuraminidase, Figure II-13a), the last α/β unit of 1jfx is atypical for TIM barrel folds: it has an antiparallel hairpin replacing the typical

parallel α/β units. A typical TIM barrel structure (PDB ID 1bqcA) is shown in Figure II-13b. Due to this alteration, it is difficult for PROCAIN to detect homologs of 1jfx and the PROCAIN E-values for 1jfx vs. other TIM barrel examples are high. The exception is an N-terminal domain of endolysin (PDB ID 2j8gA, domain 2) and an uncharacterized bacterial protein (PDB ID 1sfsA), which are in the same SCOP family 1, 4-beta-N-acetylmuraminidase. A scatter plot of E-value vs. Dali Z-score shows COMPADRE E-values shifted lower to significant E-values (e-values = 0.005 line displayed in Figure II-13c), while keeping roughly the same ranking as PROCAIN. SCOP classifies all of these structures (dots in Figure II-13c) in the same superfamily: (Trans) Glycosidases and they catalyze reactions with similar chemistry; suggesting they should be homologous. PROCAIN only detects two closest sequences at the SCOP family level (2j8gA and 1sfsA) with a significant E-value, while COMPADRE detects all of the most distant structures with significant E-values.

Detecting homology relationship for a newly resolved structure

For the inference of structure, function, and evolution of a given uncharacterized protein, COMPADRE improves the opportunity for detection of experimentally characterized homologs. As another example, the Zinc-finger antiviral protein (ZAP) is a host factor that specifically inhibits the replication of certain viruses, such as HIV-1(Zhu, Chen et al. 2011). The N-terminal part of ZAP is the major functional region that binds target RNA and recruits the mRNA degradation machinery. The structure of the N-terminal region of ZAP was determined recently(Chen, Xu et al. 2012) (NZAP225, PDB

ID 3u9gA), consisting of an uncharacterized top “cockpit” layer (1-65aa) and four connected zinc-fingers (66-225aa). Figure II-14 shows how the confidence of an originally marginal sequence similarity can be verified for the uncharacterized top “cockpit” domain by using our COMPADRE method. For the query sequence of first 65 residues from NZAP225, no confident homology to proteins with known 3D structure can be detected by direct query-template sequence comparisons. PSI-BLAST search in NCBI nr database converges after three iterations with no hits of known structures, whereas all HHpred and PROCAIN hits in structural databases are outside the significance threshold (best probability of 82.4% for HHpred and best E-value of 76.1 for PROCAIN). COMPADRE, however, is able to assign significant E-values to the similarities between the N-terminal domain of the query and multiple DNA-binding helix-turn-helix (HTH) domains, with the top E-value as low as 2e-3 (Figure II-14). Detected similarity to HTH (Figure II-14b) suggests that the first 65 residue domain of ZAP may also play an important role in recognizing target viral RNA. Sequence alignment to the template (Figure II-14a) points to specific residues that may be involved in RNA recognition and binding, providing potential targets for mutagenesis studies.

DISCUSSION

Comparison with threading methods RaptorX and MUSTER

The RaptorX (Peng and Xu 2011) and the MUSTER (Wu and Zhang 2008) packages are downloaded from their respective websites. In the RaptorX suite, CNFsearch (Ma, Peng et al. 2012) is used to detect homologs for queries. As suggested by the authors of these packages, to attain the best performance, we run them on databases provided by the authors. To speed up the experiments, 100 domains are randomly selected from 16083 ECOD (Cheng, Schaeffer et al. 2014) domain representatives (20% sequence identity cutoff). These domains are used as queries for RaptorX and MUSTER. Top 100 templates are generated for each domain by RaptorX and MUSTER. Since the templates databases used by RaptorX, MUSTER and COMPADRE are different, for each template found by RaptorX and MUSTER, we find a close homolog in ECOD database (all domains, not just representatives) by BLAST (E-value cutoff = 1e-03). All domains in ECOD database are clustered by CD-hit (Li and Godzik 2006), and one representative domain is selected for each cluster. This representative is used instead of any domain in its cluster. Any two domains in the same Homology group of ECOD are defined as “close homologs”. “All homologs” are defined by an SVM classifier based on structure and sequence similarity scores. For RaptorX, there are 5426 query and template pairs in which the template can be mapped to the ECOD database. We sort them by the E-values given by RaptorX and by COMPADRE scores to produce ROC curves (Figure II-15). For MUSTER, there are 5228 pairs in which the template can be mapped to the ECOD database. We sort them by the Z-scores given by MUSTER and by COMPADRE scores to plot the ROC curves (Figure II-16).

COMPADRE outperforms RaptorX and MUSTER on close homologs and on all homologs.

Performance of COMPADRE depends on the balance of weighting scores of close and all homologs of the template

In equation (3) to calculate s_3, the weight $w_c(s_2^c)$ for close homolog score is defined as follows (Figure II-8): $w_c = 0$ if $s_2^c < s_2^{c(1)}$, $w_c = 1$ if $s_2^c > s_2^{c(2)}$, and $w_c = a + be^{\gamma Sc}$ if $s_2^{c(1)} \leq s_c \leq s_2^{c(2)}$, where $s_2^{c(1)} = 33.95$, $s_2^{c(2)} = 111.5$, and γ=0.08; *a, b* can be derived from boundary conditions at $s_2^{c(1)}$ and $s_2^{c(2)}$. Here, we illustrate how the parameters influence the performance (Figure II-17). γ determines the shape of the w_c weighting function. Close homolog score and all homolog score are combined in range [$s_2^{c(1)}$, $s_2^{c(2)}$]. $s_2^{c(2)}$ is the boundary that only close homolog score is used if $s_2^c > s_2^{c(2)}$. Therefore, $s_2^{c(2)}$ is important for ordering high-scoring top hits exclusively based on close homolog score. Since there are two criteria (discrimination of homologs from non-homologs and assignment of closest sequence relationships as top ranks), there is a tradeoff to choose parameters to optimize both (the red curve is plotted by the parameters used in COMPADRE).

In equation (4) to calculate s_4, the weight of s_3 is defined as follows: $w_3 = 0$ if $lnE_p < lnE_p^{(1)}$, $w_3 = 1$ if $lnE_p > lnE_p^{(2)}$, and $w_3 = a + be^{\gamma lnEp}$ if $lnE_p^{(1)} \leq lnE_p \leq lnE_p^{(2)}$, where $lnE_p^{(1)} = -50$, $lnE_p^{(2)} = -15$, and $\gamma = -0.5$; *a, b* are derived from boundary conditions at $lnE_p^{(1)}$ and $lnE_p^{(2)}$. We illustrate how the parameters influence the performance (Figure

II-18). γ determines the shape of the w_3 weighting function. s_3 and the PROCAIN score s_p are combined in the range [$lnE_p^{(1)}$, $lnE_p^{(2)}$]. $lnE_p^{(1)}$ is the boundary that only PROCAIN score is used if $lnE_p < lnE_p^{(1)}$. Therefore, $lnE_p^{(1)}$ is important for ordering the top hits (with low e-values) exclusively based on PROCAIN score. $lnE_p^{(1)}$ and $lnE_p^{(2)}$, and γ all have an impact on the overall behavior of ROC curves (the red curve is based on default settings of COMPADRE) (Figure II-18).

CONCLUSION

Our findings show that defined homology relationships within the search database contain essential information for the improvement of sequence-based homology search. Although this concept is not new to computer science in general, to our knowledge, it has not been successfully applied to protein sequence searches before. Our method, COMPADRE, shows a dramatic increase in the performance of sequence search compared to current methods that are based on traditional query-template similarity measures. The new approach detects a large fraction of structural protein relationships and allows for new predictions of structure and function in previously uncharacterized proteins. Furthermore, our results suggest that this approach may be applicable to a wide variety of methods for the detection of biological similarities.

MATERIALS AND METHODS

Protein databases

As a search database, we use the set of 5116 PSI-BLAST MSAs of homologs for version 1.75 SCOP domain representatives with less than 20% sequence identity that was constructed and extensively used as a part of a previously described benchmarking system. For each protein, multiple sequence alignment of homologs detected in NCBI nr database with default settings is generated by PSI-BLAST. These alignments are used for comparisons by PROCAIN and other profile-based methods. PROCAIN is a sequence profile search method that, in addition to sequences, incorporates similarity in secondary structure, positional conservation, and sequence motifs into profile–profile scoring. When combined with an improved estimation of statistical significance of hits, this scoring results in better performance compared to other methods(Brin and Page 1988).

Definition of H_{all} homolgy network

True and false positive definition combines expert assignments of superfamilies by SCOP and our automated SVM classifier based on multiple scores for sequence and structure similarity of the two proteins, applied as previously described. These homology assignments reduce the number domain relationships classified in SCOP as unclear when the two domains share a SCOP fold but belong to different superfamilies. In brief, we constructed a classifier that combines multiple sequence- and structure-based similarity scores and trained it on a SCOP subset of 1000 pairs that belong to different SCOP

classes and are labeled as negative for SVM training, and 1000 pairs that belong to the same SCOP superfamilies and are labeled positive. The five features of the resulting classifier are as follows: DALI Z-score, FAST score, coverage of FAST alignment, GDT_TS of TM alignment, and BLOSUM score of DALI alignment.

The resulting SVM makes a binary classification of domain pairs into the categories of homologous and non-homologous. However, there is a number of domain pairs that share short regions of similarity but are poor global structural templates for each other (for example, Rossmann-type folds *vs.* TIM barrels). Forcing such cases to either of two categories might bias the evaluation protocol. Therefore, following others (Soding 2005), we use the third category of 'unclear' relations and establish the corresponding lower and higher thresholds of SVM score to define the three areas: non-homologous, unclear and homologous, with unclear pairs comprising ~10% of all pairs.

Two proteins are classified as homologous if they share a SCOP superfamily or have a high SVM score. They are classified non-homologous if do not share a superfamily and have a low SVM score. Relationships between domains from different superfamilies with intermediate SVM score are classified as unclear. We successfully tested and applied the resulting classification to the evaluation of various methods for remote homology detection.

Assessment

ROC (receiver operating characteristic) curve was used to measure performance in both settings: the domain pairs defined as "all homology" are designated as true positives for assessing the power of distinguishing the homologs from non-homologs; the "close homology" pairs are designated as true positives in order to assess the power of placing the homologs according to evolutionary distance. The ROC curve plots the true positive numbers against the false positive numbers for different cutoff points of E-values.

The following statistics were used to compare performance with other methods. We defined the ROC value as $\mathrm{ROC} = \frac{1}{\mathrm{T}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\mathrm{ti}}{\mathrm{n}}$, where t_i is the number of true positives corresponding to the *i*-th false positive found, up to *n* that is the specified number of top false positives. *T* is the overall number of true positives in the database. ROC values and their error estimates are calculated as previously described.

Another essential characteristic for a user is the degree of contamination with false positives (FP), or a proportion of true positives (TP) that is expected in a given list of top hits. Tables 3 and 4 include precision rates (fractions of true positives among top hits: *precision rate* = *TP*/(*TP*+*FP*)) for different levels of sensitivity (detected fraction of all homologs: sensitivity = *TP*/(*TP*+*FN*), *FN* is the number of false negatives). At the sensitivity level of 50% (half of all dataset homologs detected), the precision rate of the new COMPADRE scoring more than triples the original rate of PROCAIN (Table II-3).

Scoring functions

Scores s_0^T and s_0^H in eq (1) are derived as $s_0 = -(\log E / E_c)$ where E is the original PROCAIN E-value and $\log E_c = 6.0$. In equation (3), in order to mix score contributions from the closer template homologs (s_2^c) and all homologs (s_2^r), s_2^r is rescaled so that its values are comparable to that of s_2^c in the area of mixing:

$$(s_2^r)' = s_2^{c(1)} + \frac{(s_2^r - s_2^{r(1)})(s_2^{c(2)} - s_2^{c(1)})}{(s_2^{r(2)} - s_2^{r(1)})}$$

where $s_2^{c(1)}$ = 33.95 and $s_2^{c(2)}$ = 111.5 are lower and upper bounds of s_2^c in the mixing area, $s_2^{r(1)}$ = 333.9 and $s_2^{r(2)}$ = 1998 are lower and upper bounds of s_2^r in the mixing area. The weight of s_2^c is defined as follows: $w_c = 0$ if $s_2^c < s_2^{c(1)}$, $w_c = 1$ if $s_2^c > s_2^{c(2)}$, and $w_c = a + be^{\gamma Sc}$ if $s_2^{c(1)} \le s_2^c \le s_2^{c(2)}$ (Figure II-8). In equation (4), in order to mix original ProCAIn results and our boosting score, $lnE_p^{(1)}$ = -50 and $lnE_p^{(2)}$ = -15 are lower and upper bounds of in the mixing area. The weight of S_c is defined as follows: $w_n = 0$ if $lnE_p < lnE_p^{(1)}$, $w_n = 1$ if $lnE_p > lnE_p^{(2)}$, and $w_n = a + be^{\gamma lnEp}$ if $lnE_p^{(1)} \le lnE_p \le lnE_p^{(2)}$ (Figure II-15). E-values were calculated from the resulting scores based on EVD approximation of total score distributions.

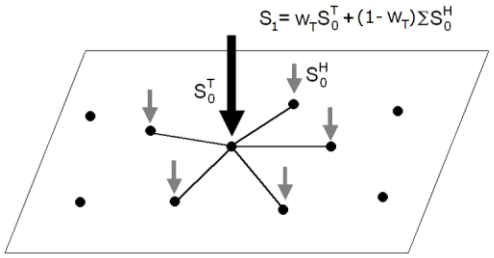


Figure II-1. The context of template's relationships within the database is used to modify the original score (s_0^T) for sequence-based similarity between query and template. The modified measure is a linear combination of s_0^T and scores ($\sum S_0^H$) for the similarity between query and template's homologs.

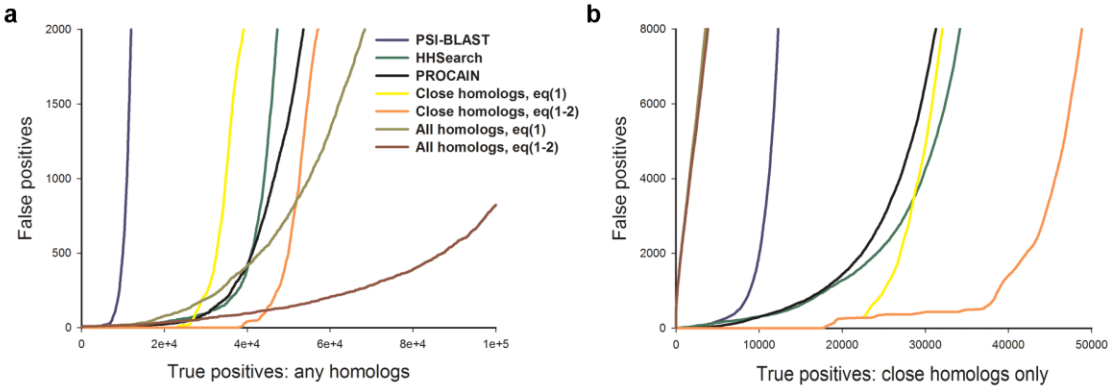


Figure II-2. Performance of similarity measures that include query similarity to template's homologs (eq. (1)) alone and combined with the reward for template's homology to the original top query's hit (eq. (1-2)). ROC plots are shown for two schemes, with equation (1) based on the sets of closer template's homologs and on all pre-determined homologs in the database. These schemes are compared to PSI-BLAST, HHSearch, and PROCAIN. (a) Evaluation of homolog/non-homolog discrimination (any pair of pre-determined homologs is considered a true positive). (b) Evaluation of ranking close homologs (only pairs sharing the same superfamily are considered as true positives).

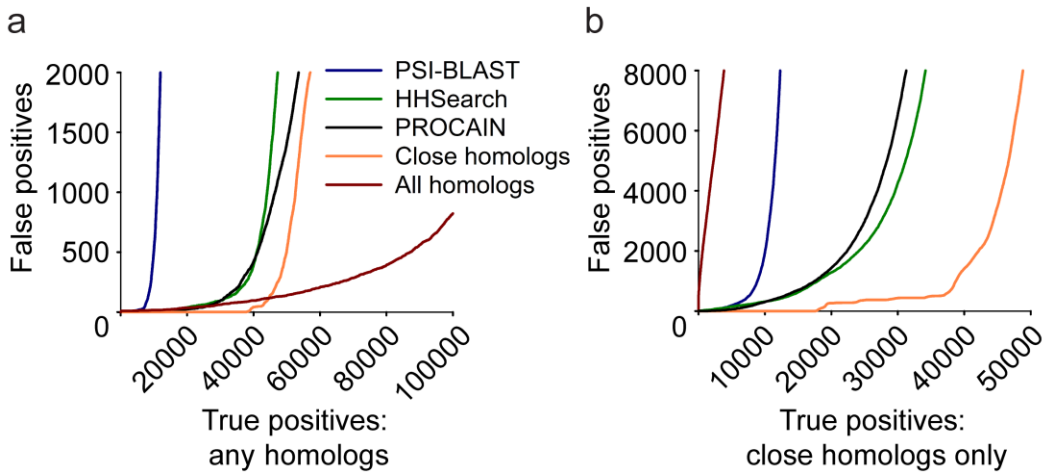


Figure II-3. Detection quality depends on the range of template's homologs included in the similarity scoring. ROC plots for (a) discrimination between homologs and non-homologs and (b) retrieval of closer homologs (relationships outside SCOP superfamily are considered false positives). The scoring based on the inclusion of only closer template homologs H_{close} (orange) is compared to the unrestricted inclusion of all homologs H_{all} (brown).

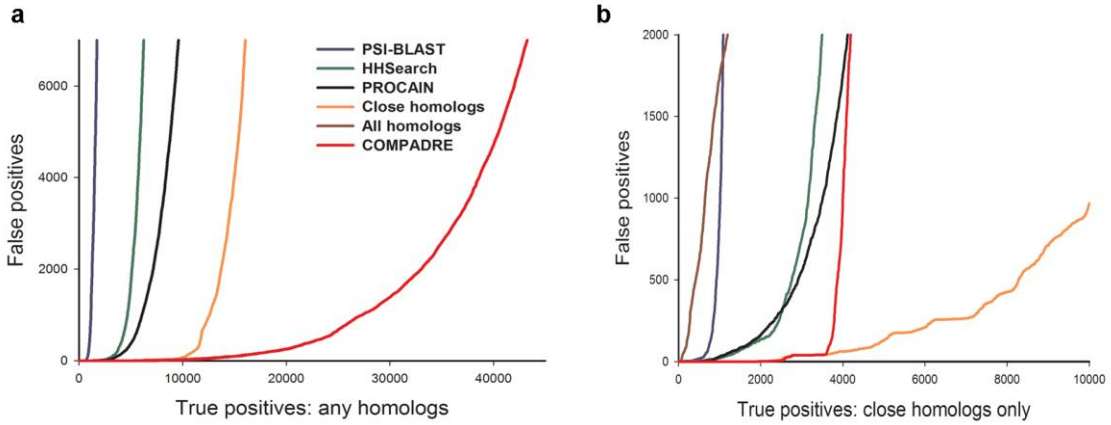


Figure II-4. Detection quality for queries from all-alpha class. (a) Discrimination between homologs and non-homologs. (b) Retrieval of closer homologs.

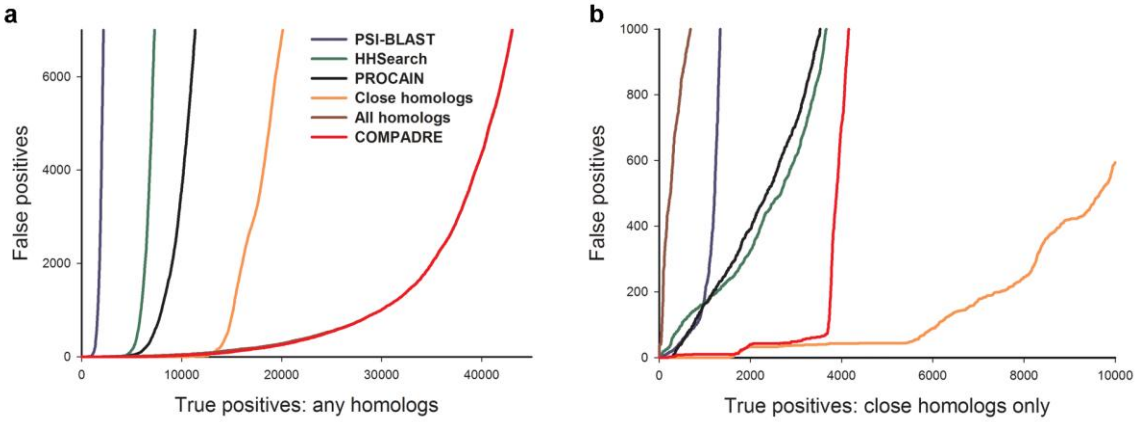


Figure II-5. Detection quality for queries from all-beta class. (a) Discrimination between homologs and non-homologs. (b) Retrieval of closer homologs.

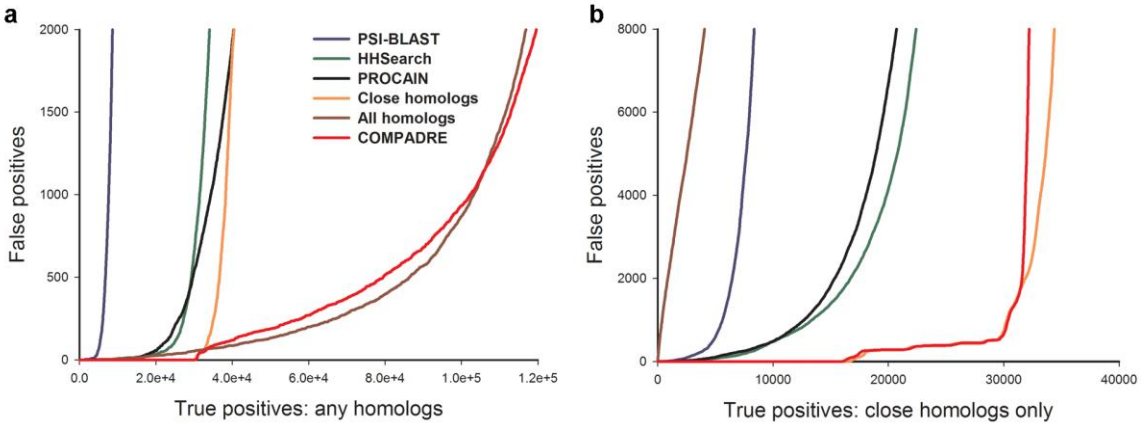


Figure II-6. Detection quality for queries from alpha/beta class. (a) Discrimination between homologs and non-homologs. (b) Retrieval of closer homologs.

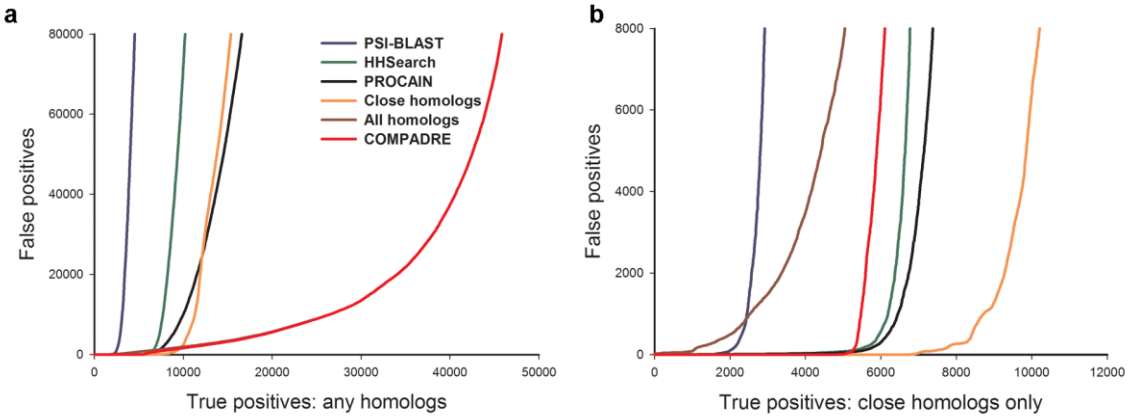


Figure II-7. Detection quality for queries from alpha+beta class. (a) Discrimination between homologs and non-homologs. (b) Retrieval of closer homologs.

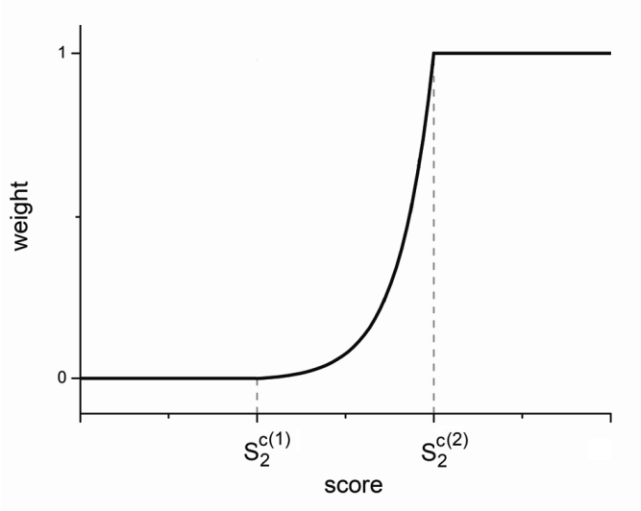


Figure II-8. Dependency of weight for the contribution of closer template homologs s_2^c. For closer query-template distances (high s_2^c), the weight is set to unit. For remote templates, the weight is set to zero. In the intermediate range, the weight grows exponentially with s_2^c (see also main text).

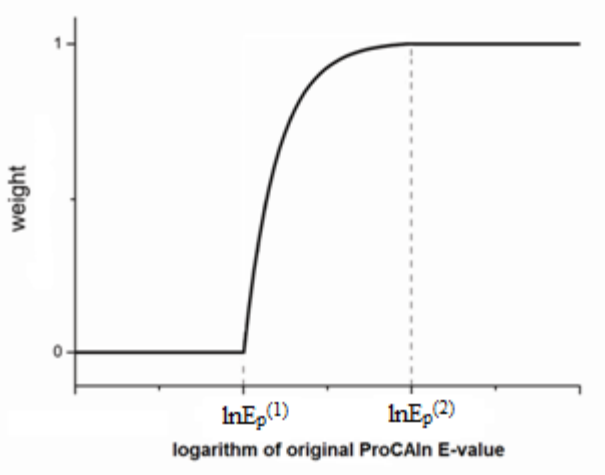


Figure II-9. Dependency of weight for the boosting score lnE_p. For high confident ProCAIn results (low lnE_p), the weight is set to zero. For low confident ProCAIn results (high lnE_p), the weight is set to unit. In the intermediate range, the weight grows exponentially with lnE_p (see also main text).

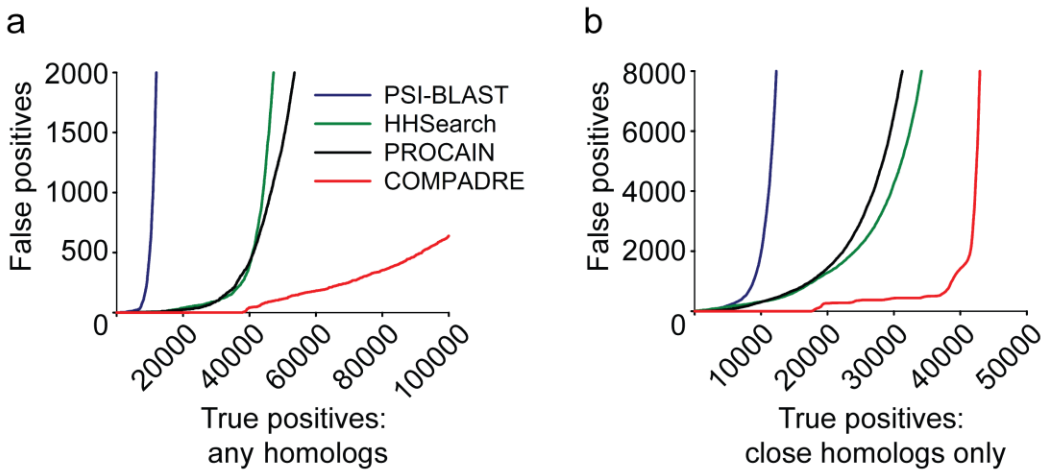


Figure II-10. Performance of combined similarity measure implemented in COMPADRE method. As illustrated by the ROC plots (red), the score both accurately discriminates homologs from non-homologs (a) and assigns top ranks to closest sequence relationships (b).

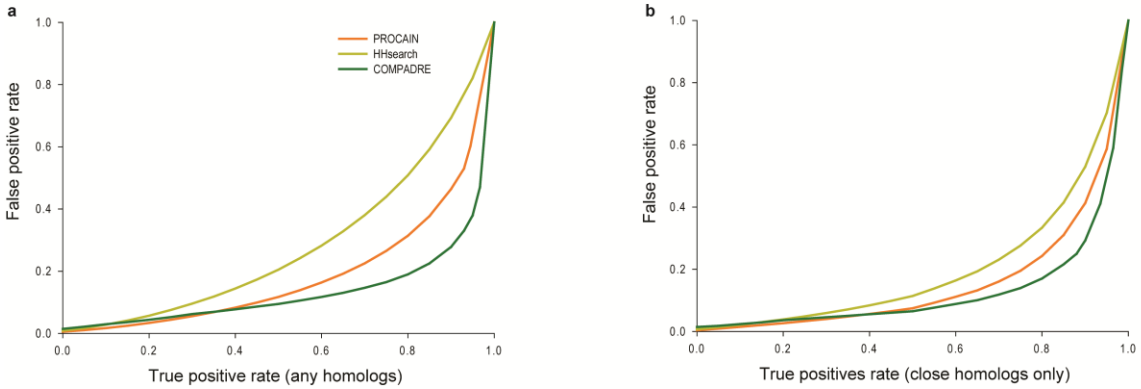


Figure II-11. Performance of COMPADRE, HHsearch and PROCAIN measured by average ROC curves. As illustrated by the curves, COMPADRE score both discriminates homologs from non-homologs (a) and assigns top ranks to closest sequence relationships (b) better than the other two methods.

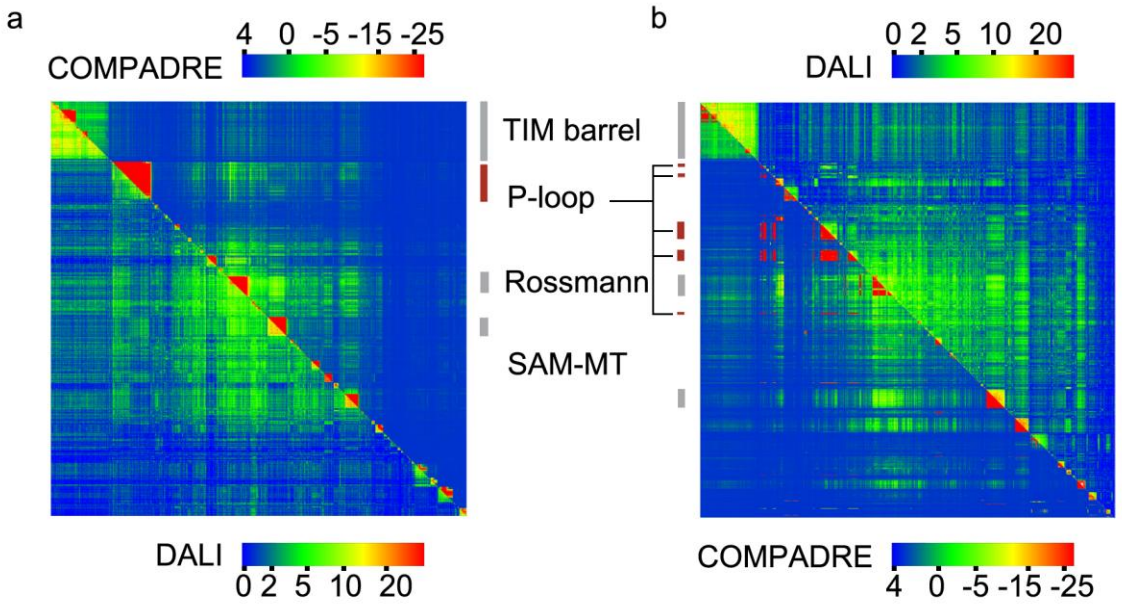


Figure II-12. Protein groupings according to sequence-based similarity scoring by COMPADRE compared to structure-based scoring by DALI. Scores for pairs of 1000 representative domains of α/β class are shown in color. Each panel compares sequence- and structure-based scores, with the score used for clustering shown above diagonal. Major protein groups are labeled on the side. Scale bars show color coding for DALI Z-score and decimal logarithm of COMPADRE E-value. (a) Grouping by COMPADRE score. (b) Grouping by DALI Z-score. Similarity between the groupings suggests that COMPADRE is able to accurately retrieve the majority of structural relationships, although a number of remote similarities still remain a challenge. In some cases, sequence comparison is able to better highlight important local motifs resulting in functionally relevant grouping, with P-loop hydrolases as the most notable example.

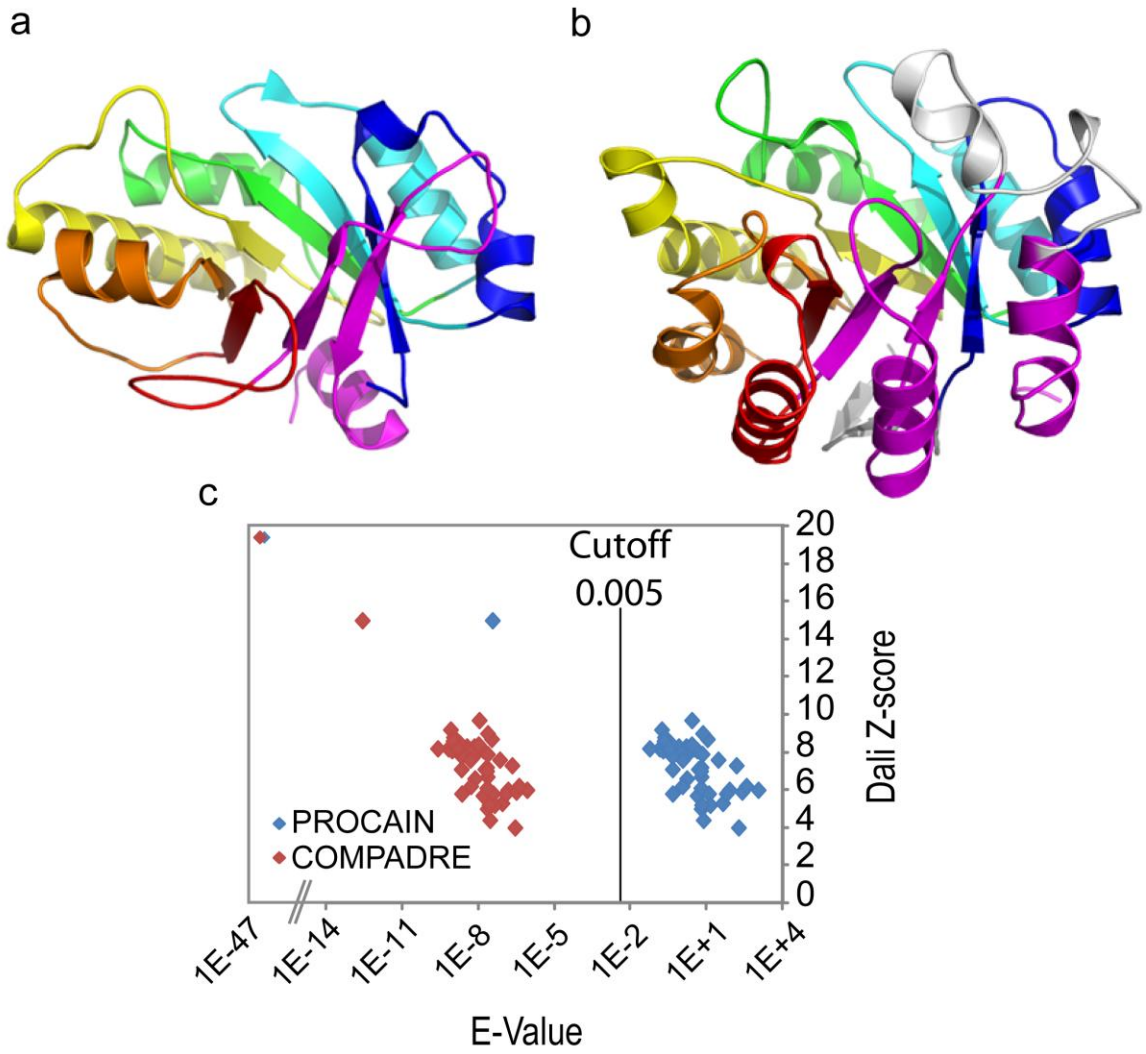


Figure II-13. Detection of more homologs at the SCOP superfamily level. PROCAIN only detects the closest sequence at the SCOP family level (PDB ID 1sfsA) with a significant E-value, while COMPADRE detects all but 8 of the most distant structures with significant E-values. (a) Structure of an atypical TIM barrel fold (PDB ID 1jfxA), which has an antiparallel hairpin replacing the typical parallel α/β unit (magenta), with aligned portion colored. (b) Structure of a typical TIM barrel fold (PDB ID 1bqcA), with aligned portion colored. (c) The scatter plot shows Dali Z-score vs. E-values of COMPADRE (red dots) and PROCAIN (blue dots) for 1jfx and all other same SCOP superfamily structures (e-values = 0.005 line displayed).

a

```
Subject = d1tc3c_ a.4.1.2 DNA/RNA-binding 3-helical bundle
Smith-Waterman score = 27   Original E-value = 7.61e+01
Hom-based score=3.57e+01   Hom-based E-value=2.29e-03
SS_pred1        9       HHHHHHCCCCHHHHHHHHHHCCCHHHHHHHHH 40
QUERY           9       TKILCAHGGRMTLEELLGEIRLPEAQLYELLE 40
CONSENSUS_1     9       TKILCAHGGRMALQELLQHIGLSEAQLREVLQ 40
                        ++++  +   +++ ++  +++ +   ++  ++
CONSENSUS_2     12      AQIDVLKQLGLSIREIARKLKRSRNTIRNYLK 43
d1tc3c_         12      AQLDVMKLLNVSLHEMSRKISRSRHCIRVYLK 43
SS_pred2        12      HHHHHHHHCCCCHHHHHHHHHHHHHHHHHHC 43
```

b

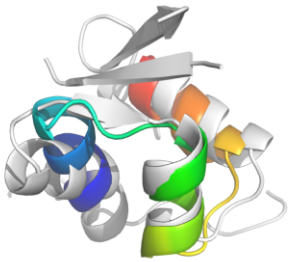


Figure II-14. Confident detection of homology relationship for a newly resolved structure. COMPADRE finds a significant similarity between the top “cockpit” layer of N-terminal ZAP protein and a ‘DNA-binding helix-turn-helix (HTH) domain, suggesting structural fold and specific mode of RNA binding. (a) COMPADRE result, including the alignment of sequences and predicted secondary structures for the query (top) and the hit (bottom). The actual secondary structure of the template is shown as colored arrows below the alignment. (b) Hit structure (PDB ID 1t3cC) superimposed to the target structure (PDB ID 3u9gA, 1-65aa), with aligned portion colored.

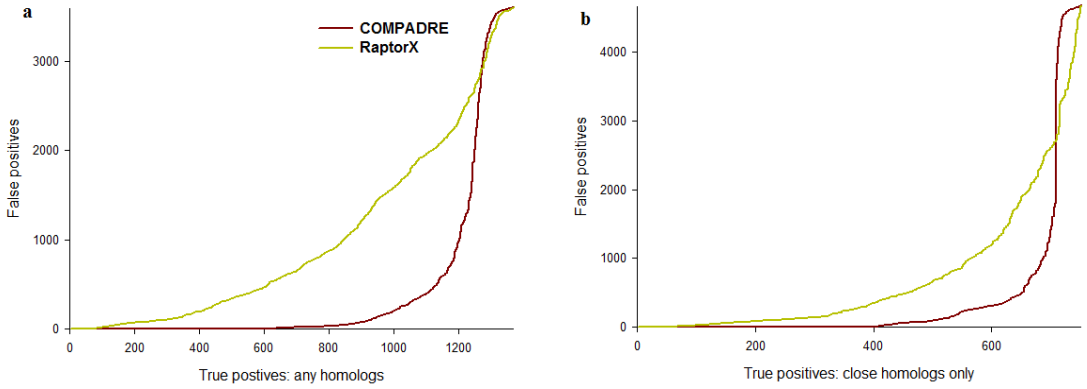


Figure II-15. Comparison of COMPADRE and RaptorX by ROC curves. For the top templates of randomly selected 100 domains, COMPADRE can both better discriminate homologs from non-homologs (a) and assign top ranks to closest sequence relationships (b) than RaptorX.

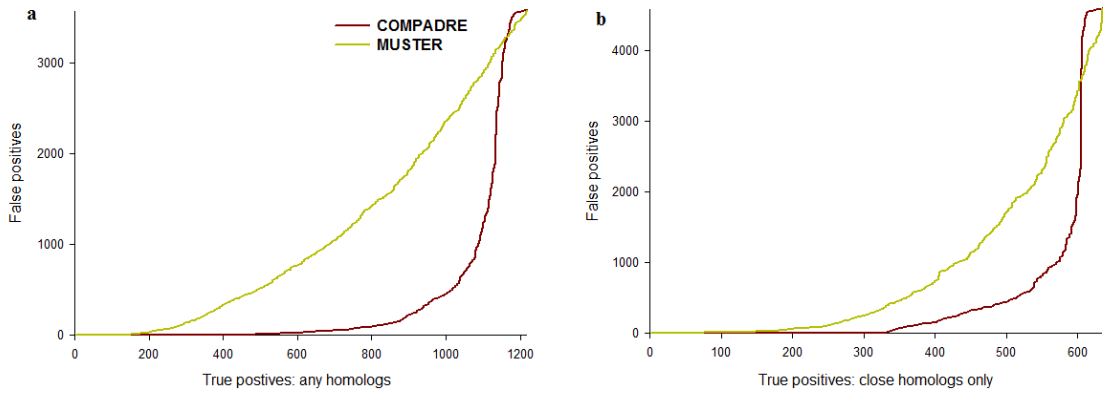


Figure II-16. Comparison of COMPADRE and MUSTER by ROC curves. For the top templates of the randomly selected 100 domains, COMPADRE can both better discriminate homologs from non-homologs (a) and assign top ranks to closest sequence relationships (b) than MUSTER.

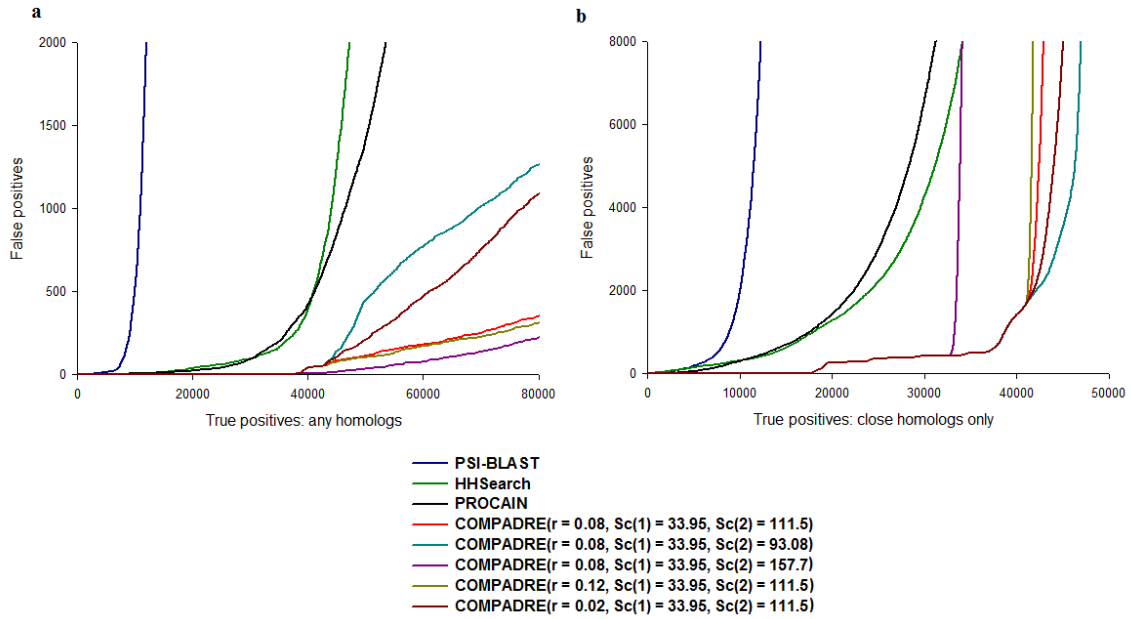


Figure II-17. Performance of COMPADRE evaluated by ROC curves under different parameter settings in equation (3) (defined in the main text). As illustrated by the ROC plots, it is a tradeoff to select parameters to meet the two criteria: discriminating homologs from non-homologs (a) and assigning top ranks to closest sequence relationships (b).

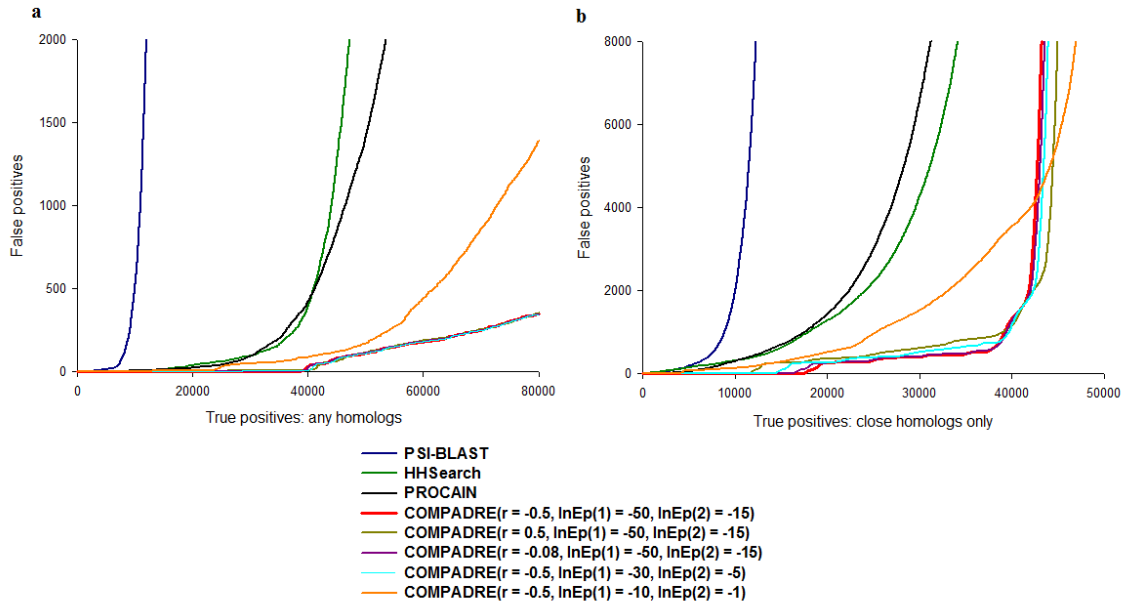


Figure II-18. Performance of COMPADRE evaluated by ROC curves under different parameter settings in equation (4) (defined in the main text). Two criteria are used: discriminating homologs from non-homologs (a) and assigning top ranks to closest sequence relationships (b).

Method	**ROC(25580)**	**ROC(255800)**	**ROC(1015772)**
PSI_BLAST	0.0334 +/- 5e-05	0.0552 +/- 4e-05	0.0906 +/- 5e-05
HHSearch	0.1215 +/- 1e-04	0.1686 +/- 7e-05	0.2303 +/- 4e-05
PROCAIN	0.1689 +/- 2e-04	0.2949 +/- 7e-05	0.4069 +/- 3e-05
Close homologs, eq (1)	0.1290 +/- 3e-04	0.2658 +/- 6e-05	0.4132 +/- 3e-05
Close homologs, eq (1-2)	0.1676 +/- 3e-04	0.2860 +/- 7e-05	0.4210 +/- 3e-05
All homologs, eq (1)	0.2348 +/- 2e-04	0.4073 +/- 6e-05	0.5915 +/- 3e-05
All homologs, eq (1-2)	0.3542 +/- 2e-04	0.6083 +/- 6e-05	0.7538 +/- 3e-05
COMPADRE	0.3702 +/- 2e-04	0.6082 +/- 6e-05	0.7532 +/- 3e-05

Table II-1. ROC values for the discrimination between homologs and non-homologs. Receiver operating characteristics (ROC) for different search methods, calculated for three numbers of top false positives: the mean of 5 top false positives per query (25580 false positives total); the mean of 50 top false positives per query (255800 false positives total), and the point where PROCAIN retrieves half of all true positives in the dataset (1015772 false positives total). The total number of true positives in the set is T= 461222. Methods are denoted the same way as in Figure II-1, II-2 and II-10.

Method	**ROC(25580)**	**ROC(63995)**	**ROC(255800)**
PSI_BLAST	0.1499 +/- 2e-04	0.1711 +/- 1e-04	0.2102 +/- 8e-05
HHSearch	0.4062 +/- 6e-04	0.4561 +/- 3e-04	0.5141 +/- 1e-04
PROCAIN	0.3784 +/- 6e-04	0.4375 +/- 4e-04	0.5278 +/- 2e-04
Close homologs, eq (1)	0.4039 +/- 5e-04	0.4720 +/- 4e-04	0.6332 +/- 4e-04
Close homologs, eq (1-2)	0.6009+/- 6e-04	0.6731 +/- 5e-04	0.8051 +/- 3e-04
All homologs, eq (1)	0.0629 +/- 4e-04	0.1300 +/- 4e-04	0.2630 +/- 3e-04
All homologs, eq (1-2)	0.0723 +/- 5e-04	0.1547 +/- 5e-04	0.3351 +/- 4e-04
COMPADRE	0.5104 +/- 3e-04	0.5267 +/- 1e-04	0.5653 +/- 1e-04

Table II-2. ROC values for the detection of closer homologs (SCOP superfamily level). Receiver operating characteristics (ROC) for different search methods, calculated for three numbers of top false positives: the mean of 5 top false positives per query (25580 false positives total); the mean of 50 top false positives per query (255800 false positives total), and the point where PROCAIN retrieves half of all true positives (proteins sharing the same superfamily) in the dataset (63995 false positives total). The total number of true positives in the set is T= 84286. Methods are denoted the same way as in Figure II-1, II-2 and II-10.

Method \ Sensitivity(%)	**1**	**10**	**25**	**50**	**75**
PSI_BLAST	99.74	7.22	3.48	2.57	2.19
HHSearch	100.00	96.64	15.94	5.03	2.95
PROCAIN	99.96	97.94	68.41	18.50	5.80
Close homologs, eq (1)	100.00	91.72	55.24	22.72	7.66
Close homologs, eq (1-2)	100.00	99.61	65.25	23.00	7.69
All homologs, eq (1)	99.74	98.72	89.80	49.82	23.16
All homologs, eq (1-2)	99.76	99.73	98.67	83.30	53.58
COMPADRE	100.00	99.81	99.14	83.33	53.24

Table II-3. Precision rates for the discrimination between homologs and non-homologs. Precision rates (%) are calculated for several levels of sensitivity from 1% to 75%. The total number of true positives in the set is T= TP+FN = 461222. Methods are denoted the same way as in Figure II-1, II-2 and II-10.

Method \ Sensitivity(%)	**1**	**10**	**25**	**50**	**75**
PSI_BLAST	98.94	91.55	6.46	0.81	0.47
HHSearch	97.12	97.10	93.66	42.51	1.21
PROCAIN	100.00	97.64	92.68	39.71	3.91
Close homologs, eq (1)	100.00	100.00	98.72	52.07	22.36
Close homologs, eq (1-2)	100.00	100.00	98.72	95.49	51.76
All homologs, eq (1)	25.19	29.06	19.48	8.23	5.47
All homologs, eq (1-2)	25.65	32.55	26.41	14.81	10.31
COMPADRE	100.00	100.00	98.69	94.18	10.42

Table II-4. Precision rates for the detection of closer homologs (SCOP superfamily level). Precision rates (%) are calculated for several levels of sensitivity from 1% to 75%. The total number of true positives in the set is T= TP+FN = 84286. Methods are denoted the same way as in Figure II-1, II-2 and II-10.

III. COMAPDRE WEB SERVER FOR PROTEIN SEQUENCE SIMILARITY SEARCH

Accurate detection of biologically meaningful similarities between protein sequences facilitates biomedical research by suggesting hypotheses for experimentation. We recently developed COMPADRE, a sequence profile-based method that uses known homology networks within a protein sequence database to enhance homology inference. Here, we present the COMPADRE web server that searches against a weekly updated database of the evolutionary classification of protein domains (ECOD) using a query sequence or multiple sequence alignment (MSA). Thus, users gain access to proteins with newly released spatial structures. By using homology relationships within the ECOD-based database, COMPADRE improves the performance of sequence search compared with traditional query-template similarity approaches. The output is similar to PSI-BLAST: a list of homologs ranked by E-value and followed by query-template alignments. In addition, 3D structures for several best hits are shown to interactively visualize their similarity. The COMPADRE web server is available at http://prodata.swmed.edu/compadre

INTRODUCTION

Advances in next-generation sequencing over the last decade have resulted in a vast growth of available protein sequences. Efficient and accurate interpretation of the sequence data is a significant challenge, and assignment of structural and functional annotations to proteins is far from being straightforward (Koboldt, Steinberg et al. 2013, Mardis 2013). Experimental methods, such as X-ray crystallography and NMR, are time-consuming and expensive. An efficient shortcut to inferring the spatial structure and molecular function of an uncharacterized protein is to computationally recognize its homologs using a sequence profile search (Gribskov, McLachlan et al. 1987, Baker and Sali 2001). As protein structures accumulate in the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000), the chance of finding a homolog for an uncharacterized protein is also increasing (Zhang, Hubner et al. 2006). However, it is still difficult to confidently detect biologically meaningful similarity between sequences with less than ~20% identity (Soding and Remmert 2011). Hence, the key to success is to increase the sensitivity of homology detection methods, because many proteins have only distant relatives with experimentally determined 3D structures (Huang, Mao et al. 2014).

Much effort has been spent to improve sequence-based homology detection. Earlier work focused on dynamic programming-based alignment methods (Needleman and Wunsch 1970, Smith and Waterman 1981) and related heuristic algorithms (Altschul, Gish et al. 1990, Pearson 1990). Subsequently, introduction of numeric sequence profiles and hidden Markov models (HMMs) (Eddy 1998) allowed comparison of a single sequence against multiple sequence alignments (MSAs) (Altschul, Madden et al. 1997, Eddy 1998, Karplus, Barrett et al. 1999). Furthermore, profile-profile (Sadreyev and

Grishin 2003, Margelevicius and Venclovas 2010) and HMM-HMM (Soding 2005, Madera 2008) comparisons improve homology search by finding the similarities at sequence positions in protein families. Recently, we developed a homology detection tool called COMPADRE (Comparison of Multiple Protein sequence Alignments using Database RElationships) (Tong, Sadreyev et al. 2015) that uses known homology relationships within a database. COMPADRE can dramatically improve protein sequence similarity search and outperform most state-of-art methods, such as PSI-BLAST (Altschul, Madden et al. 1997), PROCAIN (Wang, Sadreyev et al. 2009) and HHSearch (Soding 2005).

Besides algorithm development, the coverage of protein structure space in the search database has a profound impact on homology detection accuracy. A frequently updated search database will increase the probability of finding homologs for a query that is more confidently related to recently released protein structures. Currently, the HHpred server provides a PDB search database with synchronized updates. However, most other current methods (Sadreyev and Grishin 2003, Wang, Sadreyev et al. 2009, Ma, Wang et al. 2014) rely on infrequently updated databases. For example, the structure search databases used in COMPASS (Sadreyev and Grishin 2003) and PROCAIN (Wang, Sadreyev et al. 2009) are extracted from SCOP (Andreeva, Howorth et al. 2008), which is updated irregularly. Therefore, it is important to have a homology detection server based on an advanced algorithm and with a continually updated search database.

Here, we introduce the new version of the COMPADRE web server for homology detection with a continually and automatically updated search database. Compared with the previous version of COMPADRE that relied on the SCOP (Andreeva, Howorth et al.

2008) ASTRAL (Chandonia, Hon et al. 2004) 1.75 database, the new version adds the Evolutionary Classification of protein Domains (Cheng, Schaeffer et al. 2014) (ECOD) database that updates weekly. ECOD is a database that primarily groups domains by evolutionary relationships and consistently classifies weekly releases of PDB structures. The new version of COMPADRE with the ECOD database increases detection accuracy when compared to the current state-of-art methods. Moreover, such improvement allows discovery of new remote homologs that may have been missed in the ECOD database, thus helping to improve it. Therefore, COMPADRE with continual updates is a tool to complement the homology relationships defined in the ECOD database. Given a query sequence, the COMPADRE web server searches for homologs against the up-to-date ECOD database making use of relationships defined in ECOD and relationships predicted by an SVM classifier (Qi, Sadreyev et al. 2007). The web server is available at http://prodata.swmed.edu/compadre.

RESULTS AND DISCUSSION

Improvement of COMPADRE with ECOD search database

We tested COMPADRE on ECOD database version 56 (09/05/2014) that classifies 366,159 domains with available 3D structures. 20% sequence identity filtering of these domains resulted in 16,083 representatives for the COMPADRE search database. Two homology levels were recognized within this search database: "close homology" was defined as domain pairs classified in the same H-group (homology level) in ECOD and the level of "all homology" was defined by adding more distant homologs determined by the SVM classifier (see Materials and methods).

COMPADRE scoring schemes using two homology network definitions (all template homologs and only close template homologs) were compared to COMPADRE with combined scoring from both networks and two other methods PSI-BLAST and PROCAIN (Figure III-1). The inclusion of all template homologs leads to the most sensitive retrieval of all homologous relationships, but closer homologs may not be ranked above distant homologs. Limiting the template's homology to only close homologs can accurately place more closely related homologs above remote homologs in the ranked list; however, more distant homologs are not detected. To achieve optimal performance on both criteria, close homolog scores (Sc) and all homolog scores (Sr) were linearly combined with a weight depending on Sc (see Material and Methods). The final mixed COMPADRE score is best both in remote homology detection and in ranking by evolutionary distance to the query (Figure III-1).

COMPADRE outperforms other methods in terms of ROC value (Tables III-1 and III-2) and precision rate (Tables III-3 and III-4). ROC values reflect the performance of different scoring systems at various cutoffs of false positives. The COMPADRE scoring largely surpasses the other two scoring methods at all false positive thresholds. At the sensitivity level of 50% (half of all dataset homologs detected), the precision rate increases from 16% of PROCAIN ranking to 73% of COMPADRE ranking when all homologs are considered as true positives (Table III-3). A similar trend (the precision rate increases from 7% of PROCAIN to 97% of COMPADRE) is observed when only close homologs are assigned as true positives (Table III-4).

The pipeline for continual update

Followed by the weekly update of the ECOD database, COMPADRE will update its search database and find four boundaries for mixing close and all homolog scores accordingly (Figure III-2). In brief, after selecting domain representatives with sequence identity <20% from the ECOD database, the original all-to-all similarity scores are calculated by PROCAIN. Then, close homologs and all homologs are found in the updated search database and the close homolog (Sc) and all homolog scores (Sr) are computed (see Materials and Methods). Then, the ranks of boundary scores are estimated and four boundaries are found in the sorted lists of Sc and Sr. This pipeline is run after each weekly update of ECOD database.

Because PROCAIN similarity scores need to be calculated for all domain pairs in the updated search database, it is the rate-limiting step of the update pipeline. However, when compared with the previous version of the ECOD database, usually only a very

small number of domains are added to the search database, on the order of 1-3% of the total domain count. Therefore, we conduct two types of updates: a time-consuming complete update to recalculate all PROCAIN scores that will be run every one or two months, and a weekly simplified version, which calculates PROCAIN scores for the newly added domains, while using the PROCAIN scores of other domains in the last complete update. This two-tier update procedure balances the precision of homology detection and the speed of continual updates.

The COMPADRE web server

COMPADRE is a tool to detect homology for a given query sequence or alignment. Since its search database is updated weekly, following the ECOD database update, the COMPADRE web server allows recognizing homology to the most recently released spatial structures.

Users can input or upload a protein sequence or alignment as query and choose a protein database to search: the SCOP or the ECOD databases. The SCOP database consists of 5116 protein domains selected from SCOP ASTRAL 1.75, and the ECOD database is composed of domain representatives from the latest ECOD version. Input searching options include parameters for running PSI-BLAST, such as iteration number and E-value threshold and parameters for further processing of the resulting alignments of detected homologs. Output formatting options include the E-value cutoff to truncate the hit list (expected E-value cutoff), significant E-value threshold and the maximal number of alignments that the user wants to be shown in the result webpage.

The user can access the results via the web browser window from which the input is submitted, or choose to receive an html link to the results by email after the search is finished. The results webpage is composed of two parts. The first part is a list of hits split into two sections: one shows the hits with E-value better than the significant threshold and another contains the hits with E-value worse than the significant threshold but better than the expected E-value cutoff. For each hit, its rank, domain ID in the SCOP or ECOD database, molecule name in the PDB file, SCOP or ECOD classification, COMPADRE score and corresponding E-value are given. The second part of the results gives alignments between the query and each hit. The sequence alignments between the query and each template are generated by PROMALS (Pei and Grishin 2007) that produces alignments superior to those of PROCAIN. For the top scoring hits, besides the alignment and the links to relevant databases (e.g., ECOD, PDB etc.), a JSmol panel is included to interactively display the C-alpha trace of the template to facilitate visualization of structural similarities.

In addition, a search box is added to the main webpage of ECOD database to direct ECOD users to COMPDARE search. The sequence similarity found by COMPADRE can be used to compliment the ECOD classification and detect possible homologs missed in ECOD.

COMPADRE uniquely detects remote sequence similarities

By exploiting the homology network within the search database, COMPADRE increases the probability of finding remote homologs of the query. BWI-2c is a trypsin inhibitor isolated from buckwheat seeds that consists of a helix-loop-helix motif (Oparin,

Mineev et al. 2012). It is characterized by a pattern of cystine residues $C^1X_3C^2X_nC^3X_3C^4$, and its two helices are linked via two disulfide bonds C^1-C^4 and C^2-C^3. When the sequence of BWI-2c was submitted to the COMPADRE server searching against the ECOD database, crambin (PDB: 3NIR) can be detected with an E-value of 1e-5, while the original PROCAIN E-value to crambin was above one (Figure III-3a). Crambin is a member of thionins, a family of antimicrobial factors in plant (Florack and Stiekema 1994). Thionins typically have four disulfide bonds, while one of them is missing in crambins. The thionin helical hairpin can be aligned well with BWI-2c both in sequence and structure (Figure III-3), matching their cysteine patterns $C^1X_3C^2X_nC^3X_3C^4$. Other sequence search methods, for example PSI-BLAST (Altschul, Madden et al. 1997) and HHpred (Soding 2005), failed to detect this remote relationship with either BWI-2c or thionin as query. BWI-2c has another homolog with available structure, Luffin P1, which is a small ribosome-inactivating protein (Ng, Yang et al. 2011) obtained from gourd seeds. Additionally, there are several other plant peptides that are structurally similar to BWI-2c and Luffin P1 (Oparin, Mineev et al. 2012), such as trypsin inhibitor VhTI (Conners, Konarev et al. 2007) and antimicrobial peptide EcAMP1 (Nolde, Vassilevski et al. 2011). The similarity found by COMPADRE suggests a unified superfamily missed in the current version of ECOD and a common origin of these plant defense peptides. Thus, COMPADRE is capable of detecting biologically interesting similarities not found by other sequence-based methods and missed in ECOD.

CONCLUSION

Recently, we developed COMPADRE, a method that uses known homology network within a database to enhance protein homology detection. Here, a continually updatable web server based on this method (http://prodata.swmed.edu/compadre) is developed to detect homologs for a query sequence or alignment provided by users. The new search database composed of ECOD domain representatives is updated weekly following ECOD updates. Such an up-to-date search database allows detecting homology to proteins with recently released spatial structures.

MATERIALS AND METHODS

Overview of the COMPADRE homology detection method

Recently, we developed the COMPADRE method, which detects homology based on a similarity measure that combines the original PROCAIN score for the given template and the PROCAIN scores for a set of its homologs (Tong, Sadreyev et al. 2015). PROCAIN (Wang, Sadreyev et al. 2009) is a sequence profile search method that, in addition to sequences, incorporates similarity in secondary structure, positional conservation, and sequence motifs into profile–profile scoring. Consistent similarities between a template's homologs and the query will boost the original score. As a result, the significance of a marginal sequence similarity score can be raised over the detection threshold. On the other hand, false positives with low scores for their homologs will become less significant.

The details of incorporating the original PROCAIN similarity scores and the scores between template and template's homologs, as well as combining two different homology-based scores have been published previously (Tong, Sadreyev et al. 2015). Here, only specific modifications related to database updates are given. The major challenge of the update is to generate a new search database weekly following the ECOD update and to choose the appropriate cutoffs for mixing various homology based scores automatically.

Database representative selection and homology network definition

The new search database consists of domain representatives with sequence identity <20% from the ECOD database. For domains in the same ECOD F-group (Family level), we generated pairwise TM-align (Zhang and Skolnick 2005) structure alignments of all domain pairs and kept domains with sequence identity <20% based on TM-align alignments.

Distinct from other sequence-based homology search methods, COMPADRE utilizes the homology networks within the search database to improve the discriminating power of the similarity score. Thus, to offer optimal performance, it is essential to update the homology networks regularly, e.g., to couple it with ECOD updates. Two levels of homology are defined in COMPADRE: close homology and all homology. “Close homology” covers domain pairs classified in the same H-group (homology level) in ECOD (Cheng, Schaeffer et al. 2014). These domain pairs are closely related to each other and tend to have similar 3D structures, functions and sequences. They should appear before relatively remote homologs in the search result if ranked by evolutionary distance. The second level of “all homology” includes more distantly related domain pairs. "All homology" is defined by an SVM classifier (Qi, Sadreyev et al. 2007) constructed to complement ECOD classification. The TM-score quantifying structural similarity (Zhang and Skolnick 2005) and the BLOSUM62 score (Qi, Sadreyev et al. 2007) calculated on TM-align structure alignments to quantify sequence similarity are used as input features. The training dataset consists of an ECOD subset of 1000 pairs labeled as negative (i.e., not-homologous) that belong to different ECOD X-groups, and 1000 pairs labeled as positive (i.e., homologous) that belong to the same ECOD H-group.

The SVM classifier can help recognize subtle evolutionary similarities between domain pairs.

Using these definitions of two homology levels, COMPADRE can generate two new similarity scores by combining the original scores for the template and the scores between the query and the template's homologs. The close homolog score (Sc) is produced by combining the scores of the template's close homologs, and the all homolog score (Sr) is produced by combining the scores of the template's all homologs.

Performance assessment

We used two criteria to assess the performance of homology detection methods (Tong, Sadreyev et al. 2015). One better discriminates between homologs and non-homologs (i.e. placing homologs of the query above non-homologous sequences). The other ranks homologs according to their evolutionary distances from the query, so that closer relatives of the query appear as top hits. A ROC (receiver operating characteristic) curve measured performance in both settings. The domain pairs defined as "all homology" represent true positives for assessing the power of distinguishing the homologs from non-homologs, and the "close homology" pairs represent true positives for assessing the power of placing the homologs according to evolutionary distance. The ROC curve plots the true positive numbers against the false positive numbers for different cutoff points of E-values.

The following statistics were used to compare performance with other methods. We defined the ROC value as $\mathrm{ROC} = \frac{1}{\mathrm{T}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\mathrm{ti}}{\mathrm{n}}$, where t_i is the number of true positives

corresponding to the *i*-th false positive found, up to *n* that is the specified number of top false positives. *T* is the overall number of true positives in the database (Fawcett 2006). An essential characteristic for a user is the degree of contamination with false positives (FP), or a proportion of true positives (TP) that is expected in the list of top hits. Precision rates (fractions of true positives among top hits: precision rate = TP/(TP+FP)) for different levels of sensitivity (fraction of detected homologs: sensitivity = TP/(TP+FN), FN is the number of false negatives) were calculated for comparison.

Automatic choice of homology networks

The choice of homology network has an impact on the performance of COMPADRE. The inclusion of all homologs will sample more broadly in protein universe and result in better discrimination of homologs from non-homologs. However, such a wide sample of the template's homologs will dilute the direct signals from close homologs and could scramble the ranking by placing close homologs further down in the ranked list. On the other hand, using only close homologs can better discriminate close homologs from remote homologs and non-homologs, but may limit the sensitivity of remote homology detection.

Per Eq(3) in Tong et al. (Tong, Sadreyev et al. 2015), the solution is to mix the scores from both homology networks with proportions depending on the similarity between a query and template. The weight w_c in Eq(3) is a variable depending on s_2^c (close homolog score) as a measure of closeness between query and template. Using an upper boundary ($s_2^{c(2)}$) and a lower boundary ($s_2^{c(1)}$), we define the contribution only from

close homology (w_c =1 if $s_2^c > s_2^{c(2)}$) or only from all homology (w_c =0 if $s_2^c < s_2^{c(1)}$). For the intermediate values of s_2^c ($s_2^{c(1)} < s_2^c < s_2^{c(2)}$), we mix s_2^c with s_2^r (all homolog score) using the weight w_c exponentially dependent on s_2^c. To mix the two scores (s_2^c and s_2^r) generated by different homology networks, we need to choose the upper and lower boundaries of both s_2^r and s_2^c, and match the scales of the two scores s_2^c and s_2^r using the following equation:

$$(s_2^r)' = s_2^{c(1)} + \frac{(s_2^r - s_2^{r(1)})(s_2^{c(2)} - s_2^{c(1)})}{(s_2^{r(2)} - s_2^{r(1)})}$$

where $s_2^{c(1)}$ and $s_2^{c(2)}$ are lower and upper bounds of s_2^c in the mixing area, $s_2^{r(1)}$ and $s_2^{r(2)}$ are lower and upper bounds of s_2^r in the mixing area.

In the previous COMPADRE version, the four boundaries ($s_2^{c(1)}$, $s_2^{c(2)}$, $s_2^{r(1)}$ and $s_2^{r(2)}$) were selected manually by trial and error to optimize COMPADRE performance. However, an automatic method is required to find these four boundaries for continual database updates. We found there is a linear relationship between the ranks of hits with scores near the boundaries (e.g., the ranks of hits with the scores closest to $s_2^{c(1)}$ and $s_2^{c(2)}$ in the sorted s_2^c list and the ranks of hits with the scores closest to $s_2^{r(1)}$ and $s_2^{r(2)}$ in the sorted s_2^r list) and squared number of sequences in the search database (Supplementary Figure III-4). Thus, the boundaries for mixing can be found in the sorted score list based on the calculated ranks, i.e., from the number of sequences we compute expected rank of a boundary score and take a score for the hit with this rank as the boundary score.

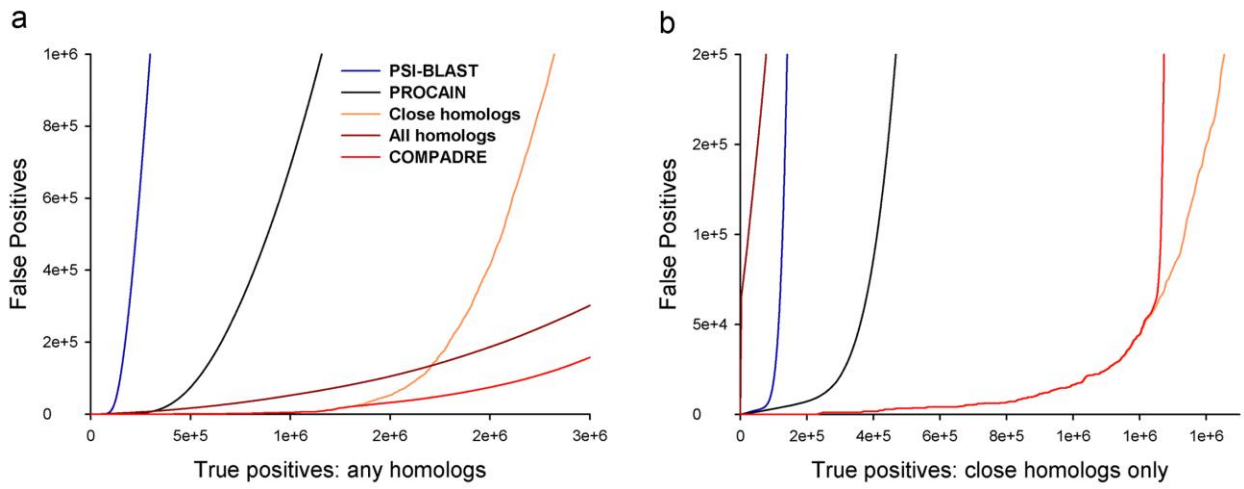


Figure III-1. Detection performance depends on the range of template's homologs included in the similarity scoring. ROC plots for (a) discrimination between homologs and non-homologs and (b) retrieval of closer homologs (relationships outside ECOD homology group are considered false positives) are shown. Performance of the combined similarity measure implemented in COMPADRE method (red) is compared with performance of the scoring based on inclusion of only close template homologs (orange, "Close homologs") and inclusion of all homologs (brown, "All homologs"), as well as PSI-BLAST (blue) and PROCAIN (black) methods.

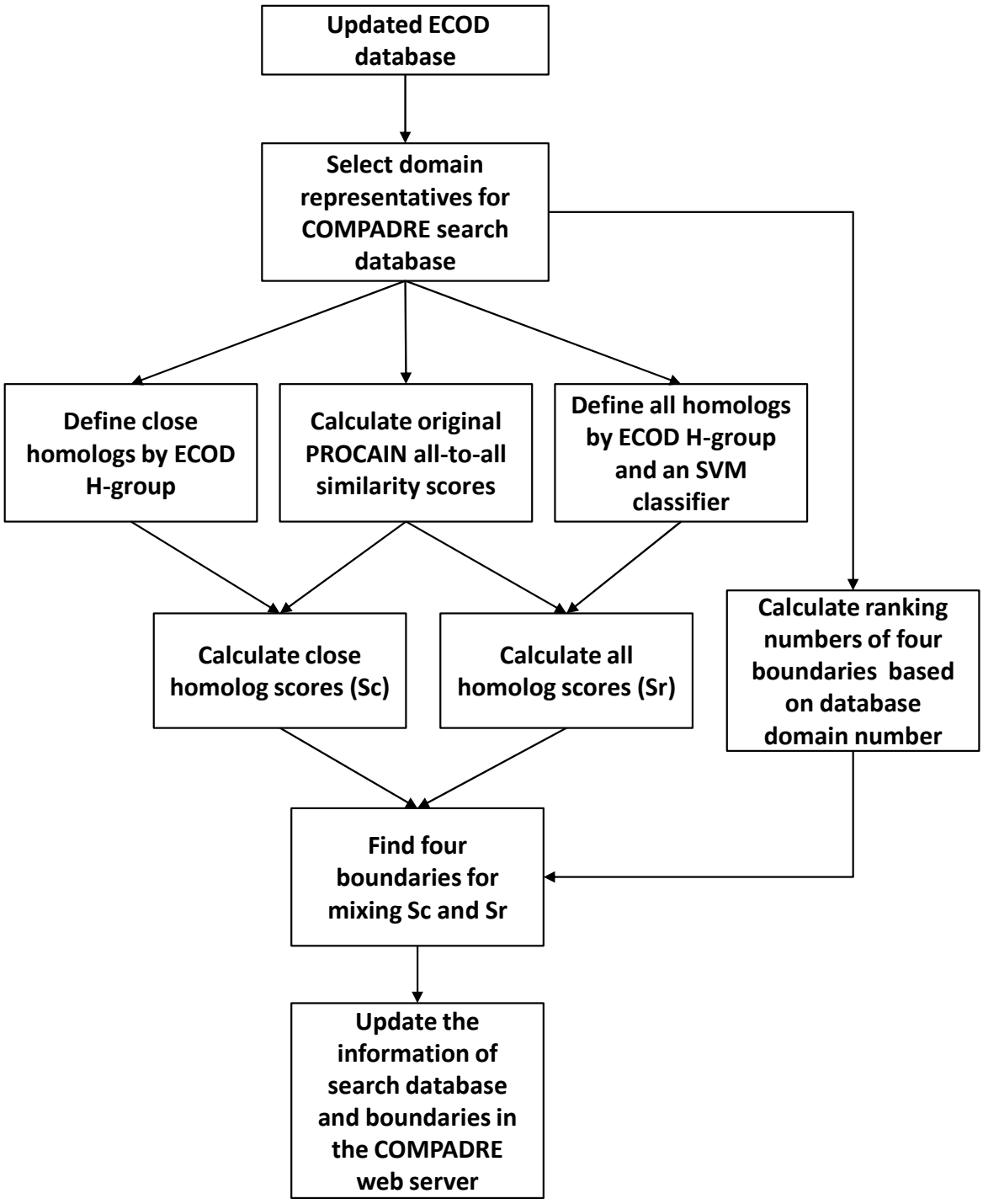


Figure III-2. Pipeline of COMPADRE update following ECOD database update.

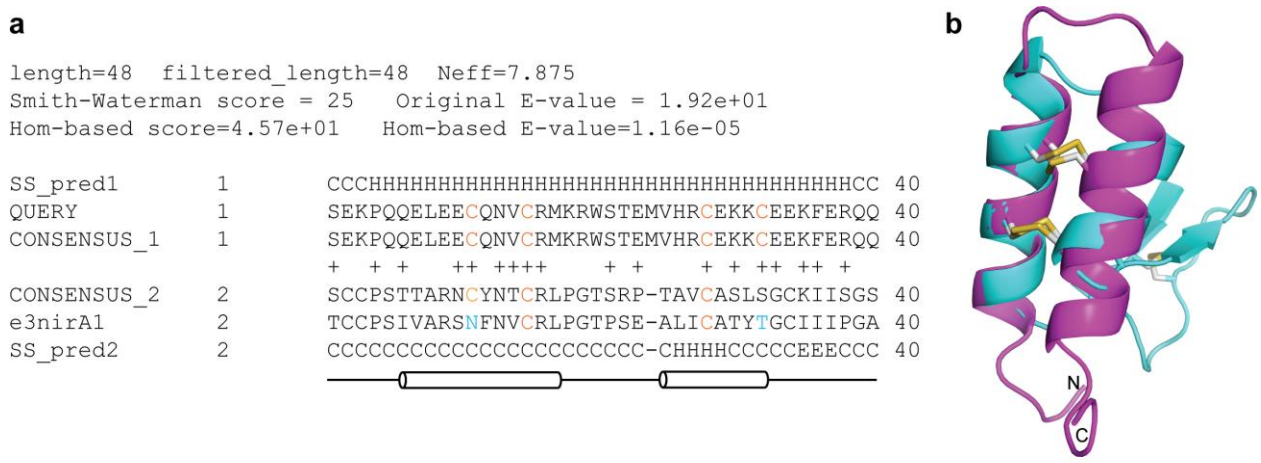


Figure III-3. An example showing remote sequence similarity detected by COMPADRE utilizing homology network. (a) COMPADRE results with BWI-2c sequence as query. Aligned cysteines are colored orange, and cyan positions usually form a disulfe bond in thionins other than crambins (hit: 3NIR). The two aligned helices are indicated by rounded rectangles under the alignment. (b) Structural superposition of BWI-2c (PDB: 2LQX, purple) and beta-hordothionin (PDB: 1WUW, cyan) by TM-align with two pairs of disulfide bond aligned.

.

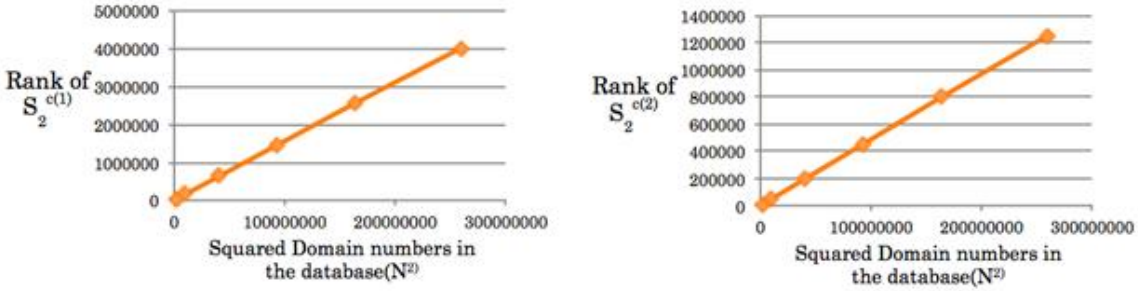


Figure III-4. Relationships between ranking number in the sorted list and database size.

Method	ROC(80,415)	ROC(804,150)	ROC(56,397,442)
PSI-BLAST	0.0103 +/- 5e-06	0.0175 +/- 1e-06	0.1644 +/- 2e-08
PROCAIN	0.0265 +/- 2e-06	0.0504 +/- 4e-07	0.3322 +/- 2e-08
Close homologs	0.0878 +/- 2e-06	0.1231 +/- 4e-07	0.3861 +/- 2e-08
All homologs	0.0418 +/- 1e-06	0.1661 +/- 3e-07	0.6100 +/- 3e-08
COMPADRE	0.1001 +/- 1e-06	0.1983 +/- 3e-07	0.6062 +/- 4e-08

Table III-1. ROC values for the discrimination between homologs and non-homologs. Receiver operating characteristics (ROC) values with standard deviations for different search methods, calculated for three numbers of top false positives: the mean of 5 top false positives per query (80,415 false positives in total); the mean of 50 top false positives per query (804,150 false positives in total), and the point where PROCAIN retrieves half of all true positives in the dataset (56,397,442 false positives in total). The total number of true positives in the set is T= 15,645,053. Methods are denoted the same way as in Figure III-1.

Method	ROC(80,415)	ROC(804,150)	ROC(13,806,119)
PSI-BLAST	0.0470 +/- 2e-05	0.0678 +/- 9e-06	0.1368 +/- 1e-06
PROCAIN	0.1411 +/- 2e-05	0.2265 +/- 5e-06	0.4081 +/- 1e-06
Close homologs	0.4899 +/- 7e-06	0.6637 +/- 5e-06	0.8660 +/- 1e-06
All homologs	0.0013 +/- 2e-05	0.0621 +/- 4e-06	0.5712 +/- 1e-06
COMPADRE	0.4867 +/- 1e-05	0.5613 +/- 8e-06	0.7001 +/- 1e-06

Table III-2. ROC values for the detection of close homologs (ECOD H-group level). Receiver operating characteristics (ROC) values with standard deviations for different search methods, calculated for three numbers of top false positives: the mean of 5 top false positives per query (80,415 false positives in total); the mean of 50 top false positives per query (804,150 false positives in total), and the point where PROCAIN retrieves half of all true positives (domains in the same ECOD H-group) in the dataset (13,806,119 false positives in total). The total number of true positives in the set is T= 2,255,581. Methods are denoted the same way as in Figure III-1.

Method	Sensitivity (%)				
	1	**10**	**25**	**50**	**75**
PSI-BLAST	54.45	9.70	7.86	7.38	6.61
PROCAIN	99.52	43.35	21.32	16.31	9.38
Close homologs	100.00	95.66	32.12	20.38	11.53
All homologs	92.26	92.66	82.04	72.74	25.39
COMPADRE	100.00	97.73	84.33	73.49	25.46

Table III-3. Precision rates for the discrimination between homologs and non-homologs. Precision rates (%) are calculated for several levels of sensitivity from 1% to 75%. The total number of true positives in the set is T= TP+FN = 15,645,053. Methods are denoted the same way as in Figure III-1.

Method	Sensitivity (%)				
	1	**10**	**25**	**50**	**75**
PSI-BLAST	95.44	9.39	2.10	1.28	1.01
PROCAIN	97.28	96.18	51.33	7.55	2.59
Close homologs	100.00	99.99	99.29	97.53	61.16
All homologs	18.96	25.02	24.67	21.77	13.93
COMPADRE	100.00	99.99	99.27	97.51	16.70

Table III-4. Precision rates for the detection of close homologs (ECOD H-group level). Precision rates (%) are calculated for several levels of sensitivity from 1% to 75%. The total number of true positives in the set is T= TP+FN = 2,255,581. Methods are denoted the same way as in Figure III-1.

IV. REFINEMENT BY SHIFTING SECONDARY STRUCTURE ELEMENTS IMPROVES SEQUENCE ALIGNMENTS

Constructing a model of a query protein based on its alignment to a homolog with experimentally determined spatial structure (the template) is still the most reliable approach to structure prediction. Alignment errors are the main bottleneck for homology modeling when the query is distantly related to the template. Alignment methods often misalign secondary structural elements by a few residues. Therefore, better alignment solutions can be found within a limited set of local shifts of secondary structures. We present a refinement method to improve pairwise sequence alignments by evaluating alignment variants generated by local shifts of template-defined secondary structures. Our method SFESA is based on a novel scoring function that combines the profile-based sequence score and the structure score derived from residue contacts in a template. Such a combined score frequently selects a better alignment variant among a set of candidate alignments generated by local shifts and leads to overall increase in alignment accuracy. Evaluation of several benchmarks shows that our refinement method significantly improves alignments made by automatic methods such as PROMALS, HHpred and CNFpred.

INTRODUCTION

Prediction of protein three-dimensional (3D) structures from amino acid sequences is important for biologists to study proteins lacking experimental structures and is one of the key problems in computational biology (Baker and Sali 2001). With the accumulation of experimentally determined protein structures in the PDB database (Berman, Westbrook et al. 2000), homology modeling (also known as template-based modeling) is the most reliable approach to protein structure prediction (Baker and Sali 2001, Zhang 2008). The 3D structure for a given query sequence can be modeled by aligning the query to one or several protein templates with known structures (Schwede, Kopp et al. 2003, Eswar, Webb et al. 2006). The model quality relies heavily on the quality of the pairwise or multiple sequence alignment (MSA) between the query and the templates (Sali, Potterton et al. 1995, Petsko 2006, Peng and Xu 2011). Currently, most MSA methods use a progressive approach that builds up an MSA by aligning the most similar two sequences as a pre-aligned group first and gradually adding more distant sequences or other pre-aligned groups. At each step of progressive alignment, a pairwise alignment method is used to align two sequences, a sequence and a pre-aligned group, or two pre-aligned groups. Thus, pairwise alignment is an integral component in most MSA methods (Notredame, Higgins et al. 2000, O'Sullivan, Suhre et al. 2004, Do, Mahabhashyam et al. 2005, Pei and Grishin 2007, Pei, Kim et al. 2008). An accurate pairwise alignment between the query and the template is essential regardless of whether one or multiple templates are used for homology modeling.

Although pairwise alignment construction has been extensively researched, alignments are still not sufficiently accurate for sequences with low similarity (Rost

1999). For example, the latest significant advance, CNFpred (Ma, Peng et al. 2012), only has Q-score of 52.4 for the MUSTER benchmark (Wu and Zhang 2008) (13.0% average sequence identity by MUSTER's own reference). A number of approaches have been developed for the task. Earlier work focused on dynamic programming recursion in construction of a global or local alignment (Needleman and Wunsch 1970, Smith and Waterman 1981). Heuristic methods such as FASTA and BLAST (Lipman and Pearson 1985, Altschul, Gish et al. 1990) were developed to significantly increase the speed of alignment. Subsequently, sequence profiles and hidden Markov models (HMMs) (Krogh, Brown et al. 1994) were introduced for comparison of a single sequence and an MSA. Furthermore, profile-profile (Yona and Levitt 2002, Mittelman, Sadreyev et al. 2003, Sadreyev and Grishin 2003, Jaroszewski, Rychlewski et al. 2005) and HMM-HMM (Soding 2005, Pei and Grishin 2007, Pei, Kim et al. 2008) comparisons improved pairwise alignments by scoring the similarity between sequence positions in protein families. In addition to pure sequence methods, 3D structural information is valuable for alignment construction because protein structures tend to evolve more slowly than protein sequences (Chothia and Lesk 1986, Illergard, Ardell et al. 2009). 3D-COFFEE (O'Sullivan, Suhre et al. 2004) as well as PROMALS3D (Pei, Kim et al. 2008) use alignment constraints derived from known 3D structures and do not use structure energy-based scoring to explicitly compare a structure to a sequence without 3D structure. Scoring of observed and predicted structural properties, such as secondary structure, solvent accessibility, residue depth, residue contacts and backbone torsion angles, was included in a number of alignment methods (Shi, Blundell et al. 2001, McGuffin and Jones 2003, Zhou and Zhou 2005, Wu and Zhang 2008, Zhang, Liu et al. 2008, Wang,

Sadreyev et al. 2009, Yang, Faraggi et al. 2011, Ma, Peng et al. 2012). Information extracted from structure-based alignments of homologous proteins was used to derive amino acid substitution matrices (Prlic, Domingues et al. 2000, Shi, Blundell et al. 2001, Qiu and Elber 2006) or position-specific scoring matrices (PSSMs) (Luthy, Bowie et al. 1992, Kelley, MacCallum et al. 2000). 3D profile is a position-dependent 20x*n* scoring matrix derived from protein structures. Such profiles were used to improve sequence-structure alignment (Luthy, Bowie et al. 1992, Kelley, MacCallum et al. 2000). Moreover, a 400x400 contact-mutation matrix was proposed to improve sequence alignment by using the contacts in template (Kleinjung, Romein et al. 2004, Dong, Lin et al. 2005). However, how to efficiently and effectively use structural (especially energy-based) information to improve pairwise alignment remains an open question in the field (Pettitt, McGuffin et al. 2005).

Query-template alignment quality is poor when the query is distantly related to the template, and alignment errors remain the main bottleneck in homology modeling (Huang, Mao et al. 2014, Kryshtafovych, Moult et al. 2014). Inevitable shortcomings in each alignment strategy lead to alignment errors. Application of a refinement algorithm to a given alignment can correct such errors. Refinement methods have been used to improve structure-based alignments and progressively constructed MSA (Gotoh 1996, Katoh, Misawa et al. 2002, Thompson, Thierry et al. 2003, Chakrabarti, Lanczycki et al. 2006, Kim, Tai et al. 2009). MSA refinement was often conducted by iteratively dividing an MSA into two sub-alignments and realigning them. However, one obvious drawback of these methods is that no additional information (such as structural information) was added to the iterative refinement.

A template structure can be viewed as regular secondary structure elements (SSEs, i.e., α-helices and β-strands) (Kabsch and Sander 1983, Richards and Kundrot 1988) alternating with loops (such as turns and coils) connecting these SSEs. SSEs are typically more conserved (Huang, Pei et al. 2013) and accurate alignments between SSEs are essential, whereas loops tend to be more evolutionarily plastic and difficult to align. In a given alignment, we define an "alignment block" as the residues in an SSE in the template and their aligned residues in the query. Automatic aligners such as PROMALS (Pei and Grishin 2007) frequently misalign alignment blocks by a few residues . Better alignment solutions can frequently be found among a limited set of local shifts of alignment blocks (moving residues in the query relative to the template). This observation motivated us to develop a pairwise alignment refinement method, SFESA, which generates candidate alignment variants for each alignment block by shifting the query region. We developed a scoring function to judge whether an alignment variant is likely to be more accurate than the original alignment. Our scoring function combines a profile-based sequence score and a novel structural contact-based score derived from residue contacts in template. This combined score was often able to select the best alignment solution among a set of candidates and lead to overall increase in alignment accuracy. Our approach improves alignments generated by a number of methods such as PROMALS (Pei and Grishin 2007), HHpred (Soding 2005) and CNFpred (Ma, Peng et al. 2012) on several benchmarks that include both reference-dependent and reference-independent assessment.

RESULTS

An overview of the SFESA method for pairwise alignment refinement

SFESA is a post-processing tool that can be applied to any pairwise alignment between a query and a template with known spatial structure. It increases alignment quality by locally shifting residues in alignment blocks defined by template SSEs. First, SFESA recognizes alignment blocks in an existing alignment. Each alignment block corresponds to residues in one SSE of the template and their aligned residues in the query. Then, proceeding from N-terminus to C-terminus, SFESA determines if each alignment block should be changed to one of the alignment variants generated by local shifts. Our analysis of PROMALS alignments revealed that SSEs are often misaligned by several residues. Thus, a better alignment solution can be found within a limited set of local shifts of SSEs (Figure IV-1).

SFESA generates *N* (up to 18) alignment variants (Figure IV-2C) by shifting query residues in alignment blocks locally (see MATERIALS and METHODS). Then, both profile-based sequence score (including scoring of secondary structure similarity) and contact-based structure score of aligned residue pairs of the original alignment block and all alignment variants are calculated. We found that a two-filter strategy offers the best performance. The first filter detects alignment variants with a higher combined score I (S_{comb_I}) than the original alignment block. If the original alignment block has the best S_{comb_I}, SFESA keeps it and move to the next alignment block. Otherwise, the alignment variant with the highest S_{comb_I} is selected and passed to the second filter. The second filter compares the selected alignment variant and the original alignment block by using a

different combined score. This combined score is either combined score II (S_{comb_II}) or the SVM score (S_{SVM}) (see MATERIALS AND METHODS). If the selected alignment variant has a higher S_{comb_II} or S_{SVM}, SFESA accepts the alignment variant. Otherwise, SFESA keeps the original alignment block. This refinement procedure is performed for each block in the alignment, starting from the N-terminal block and moving towards the C-terminus.

Here we report results of four modes for SFESA (see MATERIALS AND METHODS for details): SFESA (O) uses up to 8 variants generated by ±4 shifts that keep the gap patterns in the original alignment block and the Miyazawa-Jernigan (MJ) contact matrix for structure score calculation; SFESA (O+G) uses up to 18 variants by considering gap shifts and the MJ contact matrix; SFESA (O+G+M) uses our newly derived contact matrix in addition to gap processing; SFESA (O+G+M+S) differs from SFESA (O+G+M) in that the S_{SVM} instead of S_{comb_II} is used in second filter.

Parameter optimization

Using an in-house dataset of 1675 remote homologous domain pairs (see MATERIALS AND METHODS), we optimized the parameters of four SFESA modes: SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S). Best parameters were found for each mode separately. The Q-score and GDT-TS of the original PROMALS are 62.3 and 0.464, respectively. Each of these SFESA modes improve PROMALS alignments in both reference-dependent (Q-score of DALI) and reference-independent (GDT-TS) assessments. Even the basic mode SFESA (O) that

locally shifts up to ±4 residue positions can increase the average DALI Q-score by 2.0 (from 62.3 to 64.3) and the GDT-TS score by 0.008 (from 0.464 to 0.472). By shifting gaps in the original alignment blocks, the mode that considers 18 alignment variants, SFESA (O+G), can increase Q-score by 3.0 (from 62.3 to 65.3) and GDT-TS by 0.012 (from 0.464 to 0.476). Our new contact matrix, used in SFESA (O+G+M), further increases the alignment quality compared to the MJ matrix. The Q-score and GDT-TS improvement over the original PROMALS are 3.6 (from 62.3 to 65.9) and 0.014 (from 0.464 to 0.478), respectively. Finally, SFESA (O+G+M+S), using S_{SVM} in the second filter instead of S_{comb_II}, increases 3.7 in Q-score (from 62.3 to 66.0) and 0.014 (from 0.464 to 0.478) in GDT-TS. The comparison of numbers of improved and deteriorated alignments also shows that all SFESA modes can improve the original alignments generated by PROMALS (Figure IV-3).

The above results are based on the entire training dataset. To address the possibility of overfitting in parameter training, we divided the training dataset into four subsets based on the four SCOP classes: class a (all α proteins), class b (all β proteins), class c (α and β proteins (a/b)) and class d (α and β proteins (a+b)) (Figure IV-4). We trained the SFESA (O+G+M) parameters (including our new contact energy matrix) on the SCOP class b alignments and tested these parameters on the four subsets separately. Similarly, we trained the parameters on the SCOP class c alignments. There was no significant drop in Q-score on the class b when using parameters trained on class c (the purple column in Figure IV-4 class b) compared to using parameters trained on class b (the green column in Figure IV-4 class b) or using parameters derived from all data (the red column in Figure IV-4 class b). The average Q-scores of class b using parameters

trained on class b, class c and all data are 59.5, 59.3 and 59.4, respectively. Similar results were observed on class c, with no significant Q-score difference between using the parameters trained on the class b (the green column in Figure IV-4 class c) and using the parameters trained on the class c (the purple column in Figure IV-4 class c). Overtraining was not a major issue even in our very stringent, class-specific cross-validation scheme.

Furthermore, we analyzed the distribution of improved and deteriorated alignment block numbers in one alignment (SFESA (O+G+M+S) vs. PROMALS; DALI as a reference) for our training dataset. We found that SFESA sometimes improved several alignment blocks in one alignment, while mostly deteriorating none or only one alignment block (Figure IV-5A). Among 1675 alignments in our training dataset, there are 562 alignments with one improved alignment block while 292 alignments contain only one deteriorated alignment block. There are 268 alignments with two improved alignment blocks while 55 alignments contain two deteriorated alignment blocks. The total number of alignments with more than two improved alignment blocks is 121. In contrast, only 6 alignments contain more than two deteriorated alignment blocks.

Tests on the MUSTER benchmark

The MUSTER benchmark consists of 300 protein pairs (Wu and Zhang 2008). It is a more challenging benchmark with an average DALI Q-score of 51.6 for PROMALS alignments compared to 62.3 of our inhouse dataset. We used a number of structure-based alignment methods as a reference: DALI (Holm and Sander 1998), TMalign (Zhang and Skolnick 2005), Matt (Menke, Berger et al. 2008), MUSTER (Wu and Zhang

2008) and DeepAlign (Wang, Ma et al. 2013). SFESA (O+G+M) and SFESA (O+G+M+S) applied to PROMALS alignments outperform other methods (Table IV-1), regardless of the reference alignments used. SFESA can at most increase 2.6, 2.1, 2.6, 2.5 and 2.2 in Q-score compared with original PROMALS method when either Dali, TMalign, Matt, Muster or DeepAlign is used as a reference. In terms of Q-score based on DALI reference, all SFESA modes are statistically better than the original PROMALS based on the Wilcoxon signed-rank test (p-value less than 0.005, Table IV-2). When applied to alignments generated by HHpred and CNFpred, SFESA also shows improved alignment quality in terms of DALI Q-score, although the improvement is smaller (Table IV-1). Based on the alignment block level and aligned position level comparisons (Table IV-3, IV-4, IV-5, IV-6, IV-7 and IV-8), all SFESA modes generate more improved alignment blocks as well as aligned residue pairs than the deteriorated ones when compared with the original PROMALS. The similar trends can be observed in most SFESA modes applied on HHpred and CNFpred (Table IV-3, IV-4, IV-5, IV-6, IV-7 and IV-8).

SFESA (O+G+M+S) significantly improves PROMALS alignments on the MUSTER benchmark, making 138 alignments better and degrading 49 alignments (Figure IV-6A). In a more detailed comparison, we counted the number of the alignments SFESA improves over PROMALS and the number of alignments PROMALS is descended by SFESA at different Q-score difference cutoffs (Figure IV-6B). SFESA (O+G+M+S) improves PROMALS by at least 5 Q-score on 90 alignments and degrades by this margin on 23 alignments (Figure IV-6B). SFESA (O+G+M+S) also improved CNFpred alignments on this benchmark (Figure IV-6C) with 111 better alignments and

49 worse alignments. SFESA (O+G+M+S) improves CNFpred by at least 5 in Q-score on 48 alignments and failed by this margin on 26 alignments (Figure IV-6D).

Besides the above reference-dependent assessments, reference-independent average TM-score of query models built by MODELLER (Sali, Potterton et al. 1995) also shows that SFESA can improve PROMALS, HHpred and CNFpred alignments (Table IV-1, last column). SFESA (O+G+M+S) applied to PROMALS (Table IV-1, last column) offers the best performance.

Based on the analysis of the improved and deteriorated alignment block numbers in one alignment (SFESA (O+G+M+S) vs. PROMALS; DALI as a reference) for this dataset (Figure IV-5B), we also found that SFESA sometimes improved several alignment blocks in one alignment and mostly deteriorated none or only one alignment block. Among 300 alignments in the MUSTER benchmark, there are 83 alignments with one improved alignment block while 59 alignments contain only one deteriorated alignment block. There are 39 alignments with two improved alignment blocks while 11 alignments contain two deteriorated alignment blocks. The total number of the alignments with more than two improved alignment blocks is 26. In contrast, only 3 alignments contain more than two deteriorated alignment blocks.

Tests on the SALIGN benchmark

The SALIGN benchmark consists of 200 protein pairs. Although it has a similar DALI Q-score of 61.4 on PROMALS alignments compared to 62.3 of our inhouse

dataset, this benchmark is very challenging because proteins in each pair have very different lengths. SFESA applied to PROMALS shows maximal improvement compared to that applied to HHpred and CNFpred (Table IV-9). SFESA improves PROMALS Q-scores by 2.5, 1.9, 2.1 and 2.5 when using either DALI, TMalign, Matt or DeepAlign as a reference. For these references, SFESA shows 0.7, 0.5, 0.5 and 0.5 increases for HHpred and 0.6, 0.5, 0.4 and 0.2 for CNFpred. The reference-independent evaluation (Table IV-9, the last column) shows a similar trend that SFESA has the maximal improvement on PROMALS. The improvement on the SALIGN benchmark is less than that on the MUSTER benchmark, especially for the CNFpred with DALI as a reference (improvement of 1.2 Q-score in MUSTER and 0.6 Q-score in SALIGN). Nevertheless, alignments refined in all SFESA modes are statistically better than PROMALS based on the Wilcoxon signed-rank test (p-value less than 0.005, Table IV-10) in terms of Q-score based on DALI reference. SFESA modes except SFESA (O) are statistically better than HHpred, despite of an increase of less than 1.0 Q-score on HHpred (Table IV-10). For CNFpred, the Wilcoxon signed-rank test shows statistically significant improvement in SFESA (O), SFESA (O+G) and SFESA (O+G+M+S) (Table IV-10) in terms of Q-score. The alignment block level and aligned position level comparisons show that the number of improved alignment blocks or aligned residue pairs is larger than the number of deteriorated ones for all SFESA modes applied on PROMALS and most SFESA modes applied on HHpred and CNFpred (Table IV-11, IV-12, IV-13, IV-14 and IV-15).

Among 200 alignments in the SALIGN benchmark (SFESA(O+G+M+S) vs. PROMALS; DALI as a reference), there are 68 alignments with one improved alignment block while 55 alignments contain only one deteriorated alignment block (Figure IV-5C).

35 alignments contain two improved alignment blocks while 10 alignments contain two deteriorated alignment blocks (Figure IV-5C). In addition, the total number of the alignments with more than two improved alignment blocks is 24 while only 2 alignments contain more than two deteriorated alignment blocks (Figure IV-5C). Thus SFESA refinement improves many alignment blocks without introducing many incorrectly aligned blocks.

Tests on the SABmark benchmark

We separately tested on SABmark's two datasets (Van Walle, Lasters et al. 2005): the "superfamilies" set and the "twilight zone" set. The "superfamilies" set has an average Q-score of 71.1 for PROMALS alignments. On the "superfamilies" set SFESA improved 1.0, 0.7 and 1.3 for PROMALS, HHpred and CNFpred, respectively (Table IV-16). The "twilight zone" set is a more difficult benchmark than the "superfamilies" set with an average Q-score of only 46.2 for the PROMALS alignments. On the "twilight zone" set SFESA improved Q-score for PROMALS, HHpred and CNFpred by 1.9, 0.7 and 0.9, respectively (Table IV-16). Reference-independent average TM-scores of query models built by MODELLER displayed similar trends (Table IV-16). In terms of Q-score based on SABmark's own reference, all SFESA modes in "twilight zone" set as well as SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S) in "superfamilies" set are statistically better than the original PROMALS based on the Wilcoxon signed-rank test (p-value less than 0.005, Table IV-17). Based on the alignment block and aligned position level comparisons, all SFESA modes surpassed the original PROMALS.

In terms of Q-score, TM-score and alignment block/aligned position level, most SFESA modes improved HHpred and CNFpred, but not as much as PROMALS (Table IV-16, IV-18, IV-19, IV-20 and IV-21).

Based on the analysis of the improved and deteriorated alignment block numbers in one alignment (SFESA (O+G+M+S) vs. PROMALS; compared with SABmark's own reference) for "twilight zone" set (Figure IV-7, 209 alignments totally), there are 45 alignments with one improved alignment block while 21 alignments contain one deteriorated alignment block. And there are 8 alignments containing more than one improved alignment blocks in contrast to 5 alignments containing more than one deteriorated alignment blocks. The same analysis on "superfamilies" set (Figure IV-8, 425 alignments totally) shows a similar trend. There are 71 alignments with one improved alignment block while 37 alignments contain one deteriorated alignment block. And there are 16 alignments containing more than one improved alignment blocks while 6 alignments have more than one deteriorated alignment blocks in each alignment. Thus SFESA improves quality of several alignment blocks while avoiding deterioration of many alignment blocks in one alignment.

Tests on the PREFAB benchmark

The PREFAB benchmark (Edgar 2004) contains 1682 protein pairs and is less difficult compared to the previous three benchmarks, with a PROMALS Q-score of 80.3. PREFAB reference alignments (Edgar 2004) are based on a consensus of FSSP (Holm and Sander 1998) structural alignment and CE alignment (Stoyanova, Nicholls et al.

2004). SFESA (O+G+M+S) can increase the Q-score of PROMALS (80.3) and CNFpred (80.5) to 81.3 (Table IV-22, the first column).

In addition, we divided the PREFAB alignments into four equal-sized subsets by sequence identity (Table IV-22). The average sequence identities of the four subsets are 6.8%, 14.9%, 23.1% and 48.4%. In "set1" and "set2" subsets with the lower sequence identity (6.8% and 14.9%), we observed the most prominent improvement of more than 1.0 Q-score unit over PROMALS, HHpred and CNFpred. In the other two less difficult subsets ("set3" and "set4", Table IV-22), SFESA improvement is less dramatic. According to the Wilcoxon signed-rank test, there are statistically significant improvement in "set1" and "set2" (Table IV-23) in terms of Q-score based on PREFAB's own reference. We also observed more improvement in "set1" and "set2" compared with "set3" and "set4" based on alignment block and aligned position comparisons (Table IV-24, IV-25, IV-26, IV-27 and IV-28).

Based on the analysis of the improved and deteriorated alignment block numbers in one alignment (SFESA (O+G+M+S) vs. PROMALS; PREFAB's own reference) for this dataset (Figure IV-9, IV-10, IV-11, IV-12 and IV-13), we also found that SFESA sometimes improved several alignment blocks in an alignment and mostly deteriorated none or only one alignment block. Among 1682 alignments in the PREFAB benchmark, there are 390 alignments with one improved alignment block, while 249 alignments contain only one deteriorated alignment block. And there are 90 alignments containing one improved alignment blocks while 36 alignments contain two deteriorated alignment blocks in each alignment. In addition, 21 alignments are found to have at least two

improved alignment blocks, and 17 alignments contain more than two deteriorated alignment blocks.

Examples of alignments improved by SFESA

Here, we discuss four examples of alignments improved using SFESA. In the first, and very challenging, example (Figure IV-14A), SFESA refined the PROMALS alignment of two SCOP domains from the same superfamily (d.129.3): d2ffsa1 (query) and d2qpva1 (template). The PROMALS alignment of these domains consists of eight alignment blocks. All eight blocks are misaligned by PROMALS, and the Q-score (with the DALI alignment as reference) is only 3.2 (4 out of 125 aligned positions correctly aligned). SFESA changed the alignment in five blocks (S1, S5, S6, S7 and H1) and improved three of them (S1, S7 and H1), resulting in a Q-score of 39.2 (49 out of 125 aligned positions correctly aligned). We observed that both sequence score and structure score contribute to the selection of a better alignment variant. For example, the S1 alignment block in the original PROMALS alignment has a SFESA sequence score of -1.8 and a structure score of 6.0, which increased to 1.8 and 11.4, respectively, in the SFESA alignment.

The second example shows a case with the Q-score increase (1.7) close to the average Q-score difference (the DALI alignment as reference) (Figure IV-14B). The two SCOP domains d1c7qa_ (query) and d1iata_ (template) are from the same SCOP superfamily (c.80.1). Both of them are phosphoglucose isomerases but are from different organisms. The PROMALS alignment of these domains consists of 22 alignment blocks

(13 helices and 9 strands). The original PROMALS alignment has a Q-score of 72.4 when compared with the DALI alignment (296 out of 409 aligned positions correctly aligned). SFESA changed the alignment in one block (S5) and improved this strand, resulting in a Q-score increase of 1.7 (7 out of 409 aligned positions were corrected by SFESA).

Besides these improved alignments, there are a few alignments with accuracy decrease. The third example shows an alignment with accuracy dropping more than 20 Q-score units (Figure IV-14C). These two SCOP domains d1j8yf1 (query) and d1vmaa1 (template) are from the same SCOP superfamily (a.24.13, Domain of the SRP/SRP receptor G-proteins). SFESA incorrectly refined one of the four blocks corresponding to alpha-helices (H4) and led to a decrease of Q-score from 78.7 (48 out of 61 aligned positions correctly aligned) to 52.5 (32 out of 61 aligned positions correctly aligned) when compared with the DALI alignment. We observed a large increase of sequence score (from -1.32 to 4.13) after shifting +3 residues. On the other hand, the original alignment block and the +3 alignment variant have the similar structure score (original: 0.96, +3 variant: 0.94). Thus, +3 alignment variant has the highest combined score 1 among the original alignment block and 8 alignment variants, and this variant has a higher combined score 2 when compared with the original alignment block. As a result, SFESA incorrectly refined the alignment block by +3 shifting. The procedure of +3 shifting in SFESA introduced additional gaps to the right side of the template element (Figure IV-14C). However, no gap penalty is used in SFESA, as our scoring is restricted to the alignment block. From the structure similarity perspective, the C-terminal helix (H4) has a relatively large RMSD (2.81Å) based on DALI alignment compared with

other three helices (H1: 2.21Å, H2: 1.47Å and H3: 1.97Å), suggesting that elements showing large structural deviations between target and template are prone to mistakes by SFESA.

Prediction of active site residues is one of the key goals in alignment construction. Misaligned active site residues can lead to faulty experimental design. The last example shows (Figure IV-15) that SFESA can correct a misaligned active site residue in the alignment of two SCOP domains d1h97a_ (query) and d1tu9a_ (template). Both protein domains are from the SCOP globin family: d1h97a_ is a trematode hemoglobin (Pesce, Dewilde et al. 2001), and d1tu9a_ is a globin-like hypothetical bacterial protein (unpublished). HIS76 in the template and HIS98 in the query are the active site residues (heme-binding) and they occupy structurally equivalent positions according to the DALI alignment. However, in PROMALS alignment, HIS76 in the template is misaligned to LYS96 in the query (Figure IV-15). All SFESA modes succeed in recognizing the misaligned alignment block and correcting it by a shift of -2 (Figure IV-15).

DISCUSSION AND CONCLUSION

SFESA can refine and improve existing alignments

For divergent sequences, alignments generated by automatic methods are error-prone despite significant research efforts. Alignment errors are still the major reason for poor quality of homology models. Alignment refinement is a promising addition to existing alignment methods. Alignment methods often misalign secondary structures by a few residues, and more accurate solutions can be found within a limited set of local shifts of SSEs (Figure IV-1). SFESA aims to refine pairwise alignments by locally shifting alignment blocks defined by template SSEs to correct misaligned blocks.

Alignment errors are frequently caused by incorrect placement of gaps. The simplest SFESA (O) mode keeps original gap patterns while shifting SSEs. This approach generates up to 8 alignment variants for an alignment block. Considering that gaps rarely occur within SSEs, we implemented the SFESA (O+G) mode in which all gaps in an alignment block are moved to one side of the block. This gap shifting approach allows generation of up to 18 alignment variants. Our results show that SFESA (O+G) improves alignments more than SFESA (O).

A limitation of the SFESA approach is that shifting involves only residues within an alignment block and its adjacent loops. Residues in other alignment blocks are not allowed to move into the current alignment block. Therefore, SFESA will not correct blocks with residues misaligned to non-equivalent SSEs. However, such errors frequently occur in alignments of proteins with very different lengths, e.g., those in the SALIGN benchmark. Thus, SFESA shows less improvement on SALIGN alignments compared to

the other benchmarks. Alternative methods need to be developed to deal with non-local alignment errors.

The advantages of contact energy matrix and two-filter strategy used in SFESA

In addition to the profile-based sequence score, we included a contact-based structure score. A residue-residue contact is defined as a residue pair within a distance cutoff. In the template, a residue's contacts contribute to its structural environment. The correctly aligned equivalent residues in the query should pack more favorably in such a structural environment than incorrectly aligned residues. Thus, the estimated contact energy is an essential source of structural information and could be used as a scoring function for alignment evaluation. Unlike position-specific profile scores used in programs such as PSI-BLAST and HHpred, pairwise contact scores, in the form of two-body interactions, are difficult to incorporate into a polynomial-time algorithm (e.g. dynamic programming) to find the optimal alignment, since the interaction partners for a position are not known before the alignment is obtained. Thus, heuristic methods are needed to deal with this NP-hard problem (Lathrop 1994), such as linear programming (Xu, Li et al. 2003, Ma, Wang et al. 2013), branch and bound (Horton 1996, Horton 2001) and dead end elimination (Lukashin and Rosa 1999). However, our task is to refine an existing alignment. Using the existing alignment, contacts for a position in the query can be deduced from those contacts defined in its aligned position in the template and the query-template alignment. The resulting contact score is a positional score like the

profile-based sequence score. If the initial alignment is generally accurate, with only a few blocks misaligned, such a deduction works well.

We tested a number of contact energy matrices to derive the contact-based structure score. Firstly, we used Miyazawa and Jernigan (MJ) (Miyazawa and Jernigan 1999) contact energy matrix in SFESA (O) and SFESA (O+G). This matrix was designed for threading improved alignments. Secondly, we designed secondary structure-dependent contact energy matrices (data not shown), but they did not lead to additional improvement. Thirdly, we tested four body contact potentials (Feng, Kloczkowski et al. 2007), and they also did not give promising results. These more complex matrices were not designed for the alignment refinement task. Since our task is to select the most accurate alignment among a set of alignment variants generated by local shifts, we finally computed a new contact energy matrix specific for this task by log-odds scoring that compares contacts deduced from the correctly aligned positions to those deduced from the incorrectly aligned positions. Using the new contact energy matrix, SFESA (O+G+M) outperformed SFESA (O+G) using the MJ matrix. Another direction to improve contact energy is to explore the definition of contacts. MJ contacts are limited to one fixed distance (6.5Å) between centers of side chains. We tested several definitions of contacts to deduce our contact energy matrix. The best definition was a fixed distance (6.5Å) between any side chain atoms of two residues. A number of distance-based potentials such as DFIRE (Zhang, Liu et al. 2004), DOPE (Shen and Sali 2006) and EPAD energy (Zhao and Xu 2012) have been proposed, and some of them consider side chain orientation-dependent terms (Yang and Zhou 2008, Zhou and Skolnick 2011). Many of these potentials are all-atom based, and their application to alignment refinement would

require constructing structure models at the atomic level. A simple coarse-grained residue-contact energy matrix we used may be more appropriate for alignment scoring than atomic-level energy potentials, because atomic details of contacts may differ greatly between distant homologs, while residues could still be in similar environments and the residue-residue contacts for homologous positions are largely preserved in the structures of the template and the query.

SFESA uses a combination of profile-based sequence score and contact-based structure score to maximize the chance that the correct alignment variant is selected. First, we tested a one-filter strategy by choosing the variant with the best combination score after weight optimization. However, this strategy resulted in many false positives, i.e., the alignment variant with the best score has, on average, fewer correctly aligned positions. In practice, we found that a two-filter strategy performs better. The first filter is to inspect if there are any alignment variants with a higher combined score I. If not, the original alignment block is kept. Otherwise the alignment variant with the highest combination score is selected and passed into the second filter. If this variant has a higher combination score II than the original alignment block, the alignment variant is accepted. Otherwise the original alignment block is kept. The optimal weights for sequence vs. structure score are different in the two filtering steps. More weight is placed on the sequence score in the first filtering step, but the opposite is true for the second step.

SFESA performance is influenced by residue contacts

We observed that contact-based structural information can improve alignment, but it has limitations. We found that this structure scoring works well when there are sufficient contacts in the template as well as sufficient corresponding aligned residues in the query. However, if an SSE is involved in too few contacts (e.g. exposed edge β-strands) the remaining contacts are insufficient to define a complete structural environment and SFESA is less effective. To probe the effects of contact number and secondary structure type, we divided alignment blocks in our inhouse dataset into three categories: helix, edge strand (with hydrogen bonds on only one side) and non-edge strand (with hydrogen bonds on two sides) (Table IV-29. Edge strands have fewer contacts (average contact number is 12.2) than non-edge strands (average contact number is 25.7). Indeed, SFESA is more likely to succeed in correctly shifting non-edge strands (3.0 success/failure rate) than edge strands (1.5 success/failure rate) (Table IV-29). The helices have an average contact number of 23.7 and have a 1.8 success/failure rate. Moreover, success/failure rate positively correlates with the increase of contact number for SSEs in each of the three categories. For example, helices with less than 11 contacts have a 0.8 success/failure rate while helices with more than 36 contacts have a 9.8 success/failure rate. The same general trend is also observed in edge strands and non-edge strands. Thus, SFESA performs better when there are more contacts in an alignment block.

MATERIALS AND METHODS

Generation of alignment variants

We partition a pairwise alignment into alignment blocks according to template SSEs defined by the program PALSSE (Majumdar, Krishna et al. 2005). Short secondary structures (α-helices less than 8 residues and β-strands less than 4 residues) are not considered and are treated as loop regions. Each alignment block is defined as the residues in one template SSE and their aligned residues in the query. Eight additional alignment variants can be generated for one alignment block by shifting the original alignment in the block up to ±4 residues (Figure IV-2A). We use +*K* shift to refer to the alignment variant that shifts the query in the alignment block toward the C-terminus by *K* residues. Residues in the neighboring loop regions can be placed inside an alignment block after the shift (e.g., residue "F" in the query in +1 shift in Figure IV-2A). Similarly, negative shift numbers refer to shifting the query towards the N-terminus. SFESA does not allow residues in neighboring alignment blocks to shift. For example, in the +4 shift, the neighboring residue "V" is the last one shifted into the alignment block (Figure IV-2A), while the residues neighboring but belonging to a different SSE (such as residue "H" in Figure IV-2A) are not allowed to shift.

When there are no gaps in the original alignment block, SFESA can generate 8 alignment variants according to above procedure. If gaps are present in the query and/or template in the alignment block, there are two gap processing strategies. The first one is to keep the gap pattern in the original alignment block when shifting ±*K* (up to 4)

residues, resulting in 8 alignment variants. This strategy is used in SFESA (O) mode (described below).

The second gap treatment strategy is to preprocess gaps before shifting ±*K* (up to 4) residues. As gaps rarely occur in the middle of SSEs, we move the gaps to the same side (left or right) without interrupting the SSEs. Residues in an alignment block can be pushed to leftmost or rightmost while all gaps are put to the opposite side, resulting in two alignment variants (left and right, Figure IV-2B). Each of these two alignment variants is then used as a starting point to generate 8 additional alignment variants by ±4 shifting while keeping the modified gap patterns. Therefore, if gaps exist in the original alignment, SFESA with gap shifting can generate up to 18 (1+8+1+8) alignment variants. This strategy is used in SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S) (described below).

Profile-based sequence score

Profiles are generated from multiple sequence alignments (MSAs) generated from three PSI-BLAST (Altschul, Madden et al. 1997) iterations. Score for the similarity of residue content in MSA columns is measured by the formula originally implemented in the COMPASS method (Sadreyev and Grishin 2003).

$$S_{seq} = c_1 \sum_i n_i^1 \ln \frac{Q_i^2}{p_i} + c_2 \sum_i n_i^2 \ln \frac{Q_i^1}{p_i} \qquad (1)$$

where n_i^1 and n_i^2 are effective numbers of residue type i in the query column 1 and template column 2, Q_i^1 and Q_i^2 are estimated residue frequencies of the two compared columns, and p_i is the background residue frequency. Parameters c_1 and c_2 are calculated as:

$$c_1 = \frac{\sum_i n_i^2 - 1}{\sum_i n_i^1 + \sum_i n_i^2 - 2} \quad (2)$$

$$c_2 = \frac{\sum_i n_i^1 - 1}{\sum_i n_i^1 + \sum_i n_i^2 - 2} \quad (3)$$

SFESA further incorporates secondary structure (SS) information into the sequence score. For query, SS is predicted by PSIPRED (Jones 1999); for template, SS information in DSSP (Kabsch and Sander 1983) is used. A three-by-three secondary structure substitution matrix is derived from the structural alignment FAST (Zhu and Weng 2005) (considered as query aligned to template) of protein domains from ASTRAL compendium (Chandonia, Hon et al. 2004) based on SCOP 1.75 (Murzin, Brenner et al. 1995) (see Training dataset below). For each residues pair, the SS score S_{ss} and S_{seq} are combined to get the new sequence score S'_{seq} as:

$$S'_{seq} = S_{seq} + w_{ss} S_{ss} \quad (4)$$

where w_{ss} is the constant weight for the secondary structure score term and is set to 0.06 in our study.

Contact-based structure score

A residue contact is defined as a residue pair within a distance cutoff. In the template of one alignment, the residue contacts can be identified using the known structure of the template. Correctly aligned equivalent residues in the query should have similar structural environment as in the template. Based on a residue-residue contact energy matrix, e.g., the one derived by Miyazawa and Jernigan (Miyazawa and Jernigan 1999), the total contact energies of query residues in an alignment block can be inferred from the query-template alignment and the contacts defined by the template structure (Figure IV-16, explained below). Our contact-based structure score, corresponding to the negative of the inferred contact energies of the query, should reflect the fitness of the query residues in the structural environment defined by the template structure. Our hypothesis is that an alignment variant with a higher inferred contact energy score is likely to be more accurate.

In Figure IV-16, i is one residue in an SSE of template, and $j_1, j_2, \ldots, j_n$ are the contact residues of i based on the template structure. According to the alignment, l is the residue in the query aligned to i, and $k_1, k_2, \ldots, k_n$ in the query are aligned to $j_1, j_2, \ldots, j_n$, respectively. Thus, the contact residues of l are inferred to be $k_1, k_2, \ldots, k_n$. The inferred structure score for residue l based on the alignment and the template contact definitions is:

$$S_{contact}(l) = -\sum_{m=1}^{n} e_{l,k_m} \quad (5)$$

where e_{l,k_m} is the pairwise contact energy for residues l and k_m in the query.

The total structure score of the alignment block is the sum of the contact energy scores for all the query residues.

$$S_{str} = \sum_{l} S_{contact}(l) \quad (6)$$

When a profile is used instead of a sequence, the structure score S'_{str} is the average of all contact energies of equivalent residues in homologs of query (including the query itself):

$$S'_{str} = \frac{1}{N} \sum_{a=1}^{N} S_{str}(a) \quad (7)$$

where N is the total number of homologs of the query in the PSI-BLAST multiple sequence alignment, and $S_{str}(a)$ is the structure score calculated by Eq(6) for the homologous sequence a.

Two contact energy matrices are explored. One is the Miyazawa-Jernigan contact energy matrix with contacts defined as residue pairs with side chain centers less than 6.5Å. The other matrix is a new contact energy matrix trained on PROMALS alignments. For each alignment, each alignment block is allowed to shift ±4 residues (eight variants), and then the best-scoring variant (showing the best agreement with DALI reference (Holm and Sander 1996), that is, the variant having the most number of common aligned

positions with DALI alignment) among the original alignment and the eight variants (nine total) is selected based on and the other eight variants (alignment variants or the original alignment) are considered as background. The contact energy is formulated as follows:

$$e_{ij} = -\ln(\frac{b_{ij}}{d_{ij}/8}) + C \qquad (8)$$

where b_{ij} is the occurrence of aligned residue pair of i and j in the best-scoring alignment; d_{ij} is the occurrence of aligned residue pair of i and j in background alignments; C is a constant (C = 0.25). The cutoff for contact definition is 6.5Å between any side chain atoms of two residues.

Evaluation of the original alignment and alignment variants for an alignment block

Two filtering steps to evaluate the combined (sequence and structure) score are used to determine whether the original alignment block is kept or replaced by one of the alignment variants (Figure IV-2C). "Original alignment block" refers to blocks prior to refinement by SFESA. In the first filtering step, if the best-scoring alignment variant has a higher score (S_{comb_I}) than the original alignment block, this variant will be selected and passed to the second filter. Otherwise, SFESA keeps the original alignment block. In the second filtering step, the selected alignment variant (with the best S_{comb_I}) is again compared to the original alignment block by using a different score: S_{comb_II} or S_{svm}. If the selected alignment variant also has a better S_{comb_II} or S_{svm} than the original alignment

block, this alignment variant will replace the original alignment block. Otherwise, the original alignment block is kept.

S_{comb_I} and S_{comb_II} are linear combinations of sequence score and structure score with different weights:

$$S_{comb\_I} = w_1 S'_{seq} + (1 - w_1) S'_{str} \quad (9)$$

$$S_{comb\_II} = w_2 S'_{seq} + (1 - w_2) S'_{str} \quad (10)$$

where S'_{seq} is the sequence score combined with secondary structure score (Eq(4)) and S'_{str} is the contact-based structure score with consideration of query homologs (Eq(7)). SFESA has four modes (described below). w_1 and w_2 are optimized to be 0.8 and 0.1 in SFESA (O), 0.4 and 0.1 in SFESA (O+G), and 0.12 and 0.02 in SFESA (O+G+M). In SFESA (O+G+M+S), w_1 is optimized to be 0.12. S_{svm} used in second filter of SFESA (O+G+M+S) is a score generated by a SVM classifier described below.

The SVM score

A support vector machine (SVM), implicitly mapping its inputs into high-dimensional feature space, is widely used in binary classification (Cortes 1995). In our strategy, if any alignment variant passes the first filter (with the top S_{comb_I} compared to all other variants and the original alignment block), the second filter is a two-category classification problem – either accepting this selected alignment variant or keeping the

original alignment block. Besides S_{comb_II}, an SVM was trained in the second filter to aid this decision. Ten features were used in such an SVM binary classifier. Two features are binary representatives for secondary structure type: helix as (1, 0) and strand as (0, 1). Four features represent the scores of the original alignment block: S_{seq}, S_{ss}, $S_{str}^{'}$ and S_{rsa}, and another four corresponding features are used for the selected alignment variant. Similarly to S_{ss}, S_{rsa} is based on a three-by-three relative solvent accessibility (rsa) substitution matrix derived from FAST (Zhu and Weng 2005) structural alignments of SCOP domains. Notably, for query, three categories of neural network-predicted rsa values (with PSI-BLAST PSSMs as input) (Huang, Pei et al. 2013) were used based on three equal-sized bins; for template, the real rsa values calculated by NACCESS (Hubbard and Thornton 1993) are used to generated three categories based on three equal-sized bins. Two-fold cross validation was used in our SVM training. The linear, polynomial and radial basis functions were tried as kernels. The linear model was found to be optimal. The criterion to accept the alignment variant is set to SVM score above -0.6 for optimal performance of alignment accuracy.

Training dataset

The training dataset consists of protein domain pairs with sequence identity less than 20% from the ASTRAL compendium (Chandonia, Hon et al. 2004) based on SCOP 1.75 (Murzin, Brenner et al. 1995). All domain pairs with COMPADRE e-value less than 1e-30 were used. For all domain pairs, we generated DALI structure alignments. Then, we discarded domain pairs with GDT-TS score (Zemla 2003) less than 0.5 in DALI

alignment. We also included at most 10 domains from any individual SCOP superfamily, and ensured that each domain is present no more than twice in domain pairs. The final training dataset consists of 1675 domain pairs with 2305 protein domains. We generated PROMALS alignments for these domain pairs and deduced 16347 alignment blocks in these alignments. There are 3061 incorrectly aligned alignment blocks (at least one residue misaligned compared to DALI reference alignment) in PROMALS alignments. The parameters of w_{ss} in Eq (4), C in Eq (8), w_1, w_2 in Eq (9) and Eq (10) and all SVM parameters were trained on this dataset. The assessment of alignment quality is Q-score (alignment quality score, described below) compared with reference DALI alignment and reference-independent GDT-TS score (Zemla, Venclovas et al. 1999).

Testing benchmarks

We used the following four public datasets to test the method:

1. The MUSTER benchmark (Wu and Zhang 2008). This dataset consists of 110 ProSUP protein pairs (Lackner, Koppensteiner et al. 2000) and 190 pairs selected by the Zhang group with TM-score (Zhang and Skolnick 2005) > 0.5.

2. The SALIGN benchmark (Marti-Renom, Madhusudhan et al. 2004). This dataset has 200 protein pairs with about 20% sequence identity, and these pairs have on average about 65 structurally equivalent residues with RMSD < 3.5Å. Proteins in each pair have very different lengths.

3. The SABmark benchmark (Van Walle, Lasters et al. 2005). This benchmark is designed for testing multiple sequence alignments (MSAs). SABmark dataset (version 1.65) has two benchmark sets: the "twilight zone" set has 209 groups of SCOP fold-level domains with very low similarity, whereas the "superfamilies" set has 425 groups of same-superfamily domains. We randomly selected one domain pair from each group to test our pairwise alignment refinement method.

4. The PREFAB benchmark (version 4.0) (Edgar 2004). This dataset contains 1682 alignments, and it provides its own reference that is based on the consensus of FSSP structure alignment (Holm and Sander 1998) and CE alignment (Stoyanova, Nicholls et al. 2004).

For these four benchmarks, we applied our refinement method to PROMALS alignments, as well as alignments generated by two profile-based and structure-aided methods: HHpred (Soding, Biegert et al. 2005) and CNFpred (Ma, Peng et al. 2012). Here, HHpred was used in the global alignment mode as its local alignment mode often results in short alignments and shows lower alignment accuracy than global alignments (Table IV-30, IV-31, IV-32 and IV-33).

The evaluation criteria include:

1. Reference-dependent evaluation. (1) Q-score is the fraction of correctly aligned residue pairs in a test alignment among all aligned residue pairs in a reference alignment. In this paper, the range of Q-score is from 0 to 100 (e.g. 100 means 100% agreement with reference). The reference alignments were constructed by several structure alignment methods: DALI (Holm and Sander 1996), TM-align (Zhang and

Skolnick 2005), Matt (Menke, Berger et al. 2008) and Deepalign (Wang, Ma et al. 2013). (2) The number of alignment blocks improved and deteriorated in different benchmarks after refinement by different SFESA modes. Above-mentioned structural alignment methods are used as references. One alignment block is considered as an improved block when the correctly aligned residue pair number (compared with reference) in the block is increased. One alignment block with less correctly aligned residue pairs (compared with reference) is treated as a deteriorated block. (3) The number of aligned positions improved and deteriorated in different benchmarks after refinement by different SFESA modes. Abovementioned structural alignment methods are used as references. This number provides the residue position-level alignment accuracy comparison.

2. Reference-independent score. Alignment-based GDT-TS (Zemla, Venclovas et al. 1999) score and TM-score (Zhang and Skolnick 2005) were used in our study to evaluate alignment quality. GDT-TS score is based on the number of structurally equivalent pairs of C-alpha atoms that are within specified distance cutoffs (1Å, 2Å, 4Å and 8Å) based on the sequence-independent superpositions of two protein structure. TM-score is a simpler template modeling score, which evaluates the similarity of two protein structures in a single superposition by weighting the close atom pairs stronger than the distant matches. For TM-score, a 3D model was built for the query protein by MODELLER (Sali, Potterton et al. 1995) based on its alignment to the template, and subsequently the score between the query model and the experimental structure was computed.

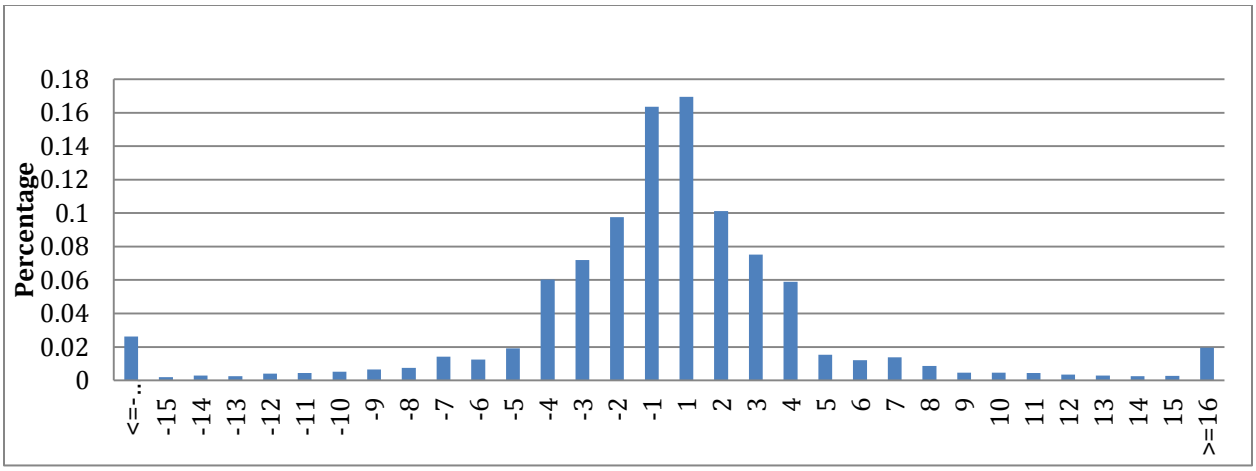


Figure IV-1. Distribution of shifting numbers that resulted in improved alignment quality in alignment blocks. Those alignment blocks for which correct aligned positions (Dali as a reference) cannot be increased by shifts are not included. The shifting number *n* means that the query block in the alignment block is shifted by *n* residue positions. Positive numbers mean that the shifting is towards C-terminus, while negative numbers mean that the shifting is towards N-terminus.

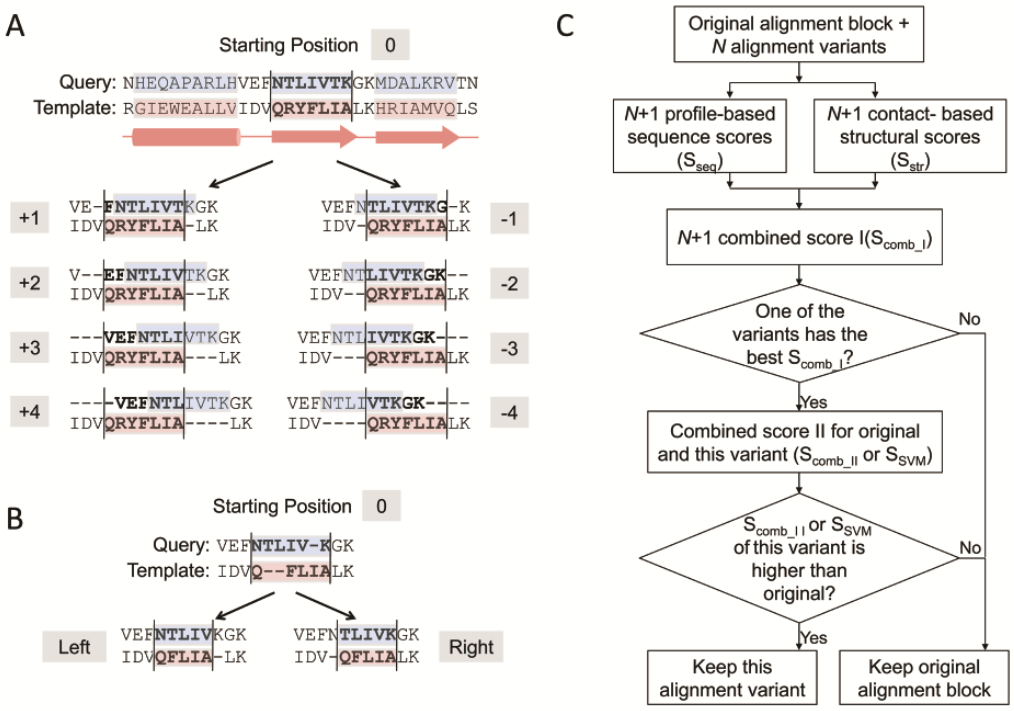


Figure IV-2. An overview of the SFESA method. (A) For each alignment block, SFESA generates up to ±4 variants by shifting (marked as -1, -2, -3, -4, +1, +2, +3, and +4). The pink boxes show the SSEs recognized from template structure and the blue boxes are corresponding regions in the query aligned to such SSEs. Residues and gaps in one corresponding blue and pink boxes compose an alignment block. The corresponding black lines provide the boundaries between which sequence and structure scores are calculated for each aligned residue pairs. (B) If gap shifting is considered, two variants (left and right) are generated by putting gaps on the same side (left or right) before generating the above 8 variants. (C). Flowchart of the SFESA method.

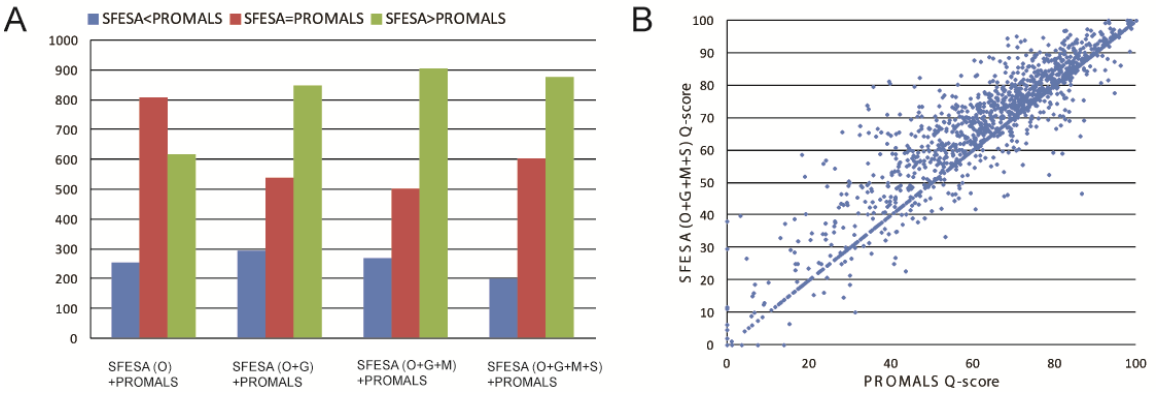


Figure IV-3. Tests on our training dataset. (A) DALI Q-score comparison of PROMALS and SFESA for different SFESA modes. The columns in blue, red and green represent the numbers of alignments that DALI Q-Score of SFESA is less than, the same as, and more than PROMALS, respectively. SFESA (O) improved 617 alignments generated by PROMALS while making 215 alignments worse. SFESA (O+G) improved 847 PROMALS alignments and made 292 alignments worse. The number of improved SFESA (O+G+M) alignments is 904 while PROMALS is better than SFESA (O+G+M) on 268 alignments. SFESA (O+G+M+S) produced 875 higher quality alignments than PROMALS and 199 worse alignments. SFESA (O+G+M+S) performance is quite similar to SFESA (O+G+M) but is more likely to have an equivalent Q-score with the original PROMALS alignment (601 alignments in SFESA (O+G+M+S) and 503 alignments in SFESA (O+G+M)), thus being more conservative compared with SFESA (O+G+M). (B) Scatter plot of SFESA (O+G+M+S) Q-score vs. PROMALS Q-score for 1675 alignments in the training dataset. Each point is one domain pair.

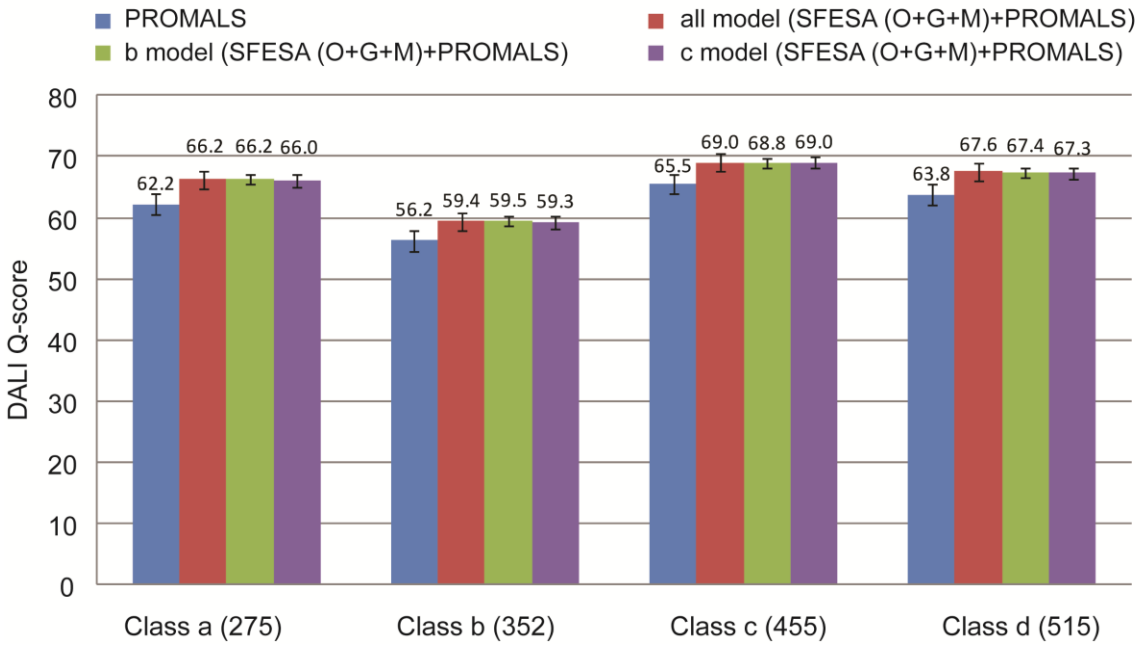
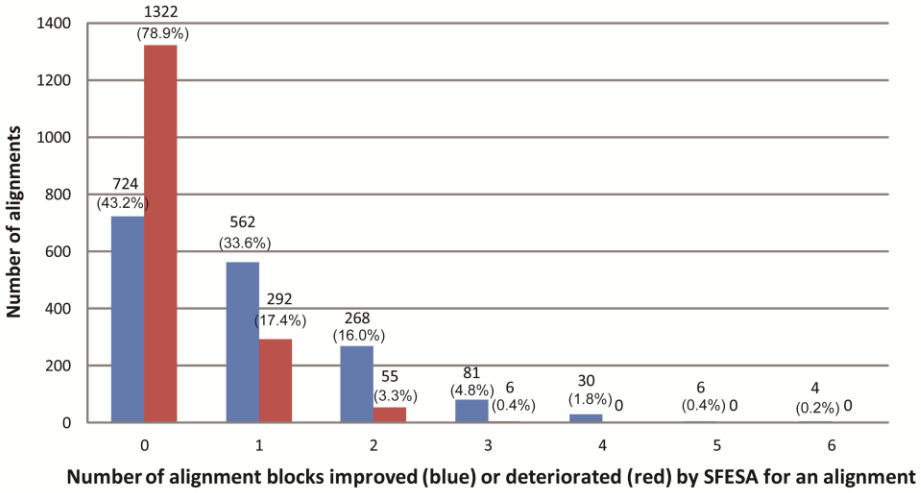
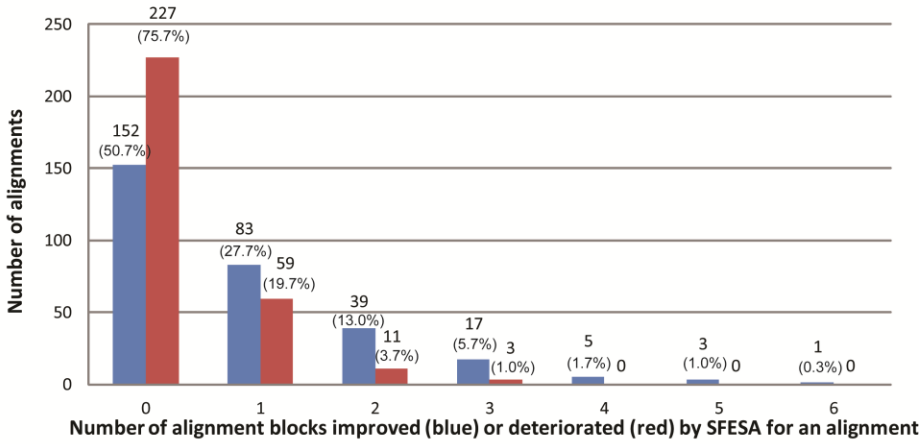


Figure IV-4. Tests on our training subsets divided by four SCOP classes. DALI Q-score is compared in different subsets: 275 class a alignments (all α proteins), 352 class b alignments (all β proteins), 455 class c alignments (α and β proteins (α/β)) and 515 class d alignments (α and β proteins (α+β)). The blue column represents the performance of PROMALS alignments. The red column shows the SFESA (O+G+M) results with parameters derived from all data (1675 alignments). The green and purple columns are the SFESA (O+G+M) results trained on class b and class c, respectively. The error bars (standard error of the mean) are showed.

A


1322
(78.9%)
724
(43.2%)
562
(33.6%)
292
(17.4%)
268
(16.0%)
55
(3.3%)
81
(4.8%)
6
(0.4%)
30
(1.8%)
0
6
(0.4%)
0
4
(0.2%)
0
1400
1200
1000
800
600
400
200
0
Number of alignments
0
1
2
3
4
5
6
Number of alignment blocks improved (blue) or deteriorated (red) by SFESA for an alignment

B


227
(75.7%)
152
(50.7%)
83
(27.7%)
59
(19.7%)
39
(13.0%)
11
(3.7%)
17
(5.7%)
3
(1.0%)
5
(1.7%)
0
3
(1.0%)
0
1
(0.3%)
0
250
200
150
100
50
0
Number of alignments
0
1
2
3
4
5
6
Number of alignment blocks improved (blue) or deteriorated (red) by SFESA for an alignment

C

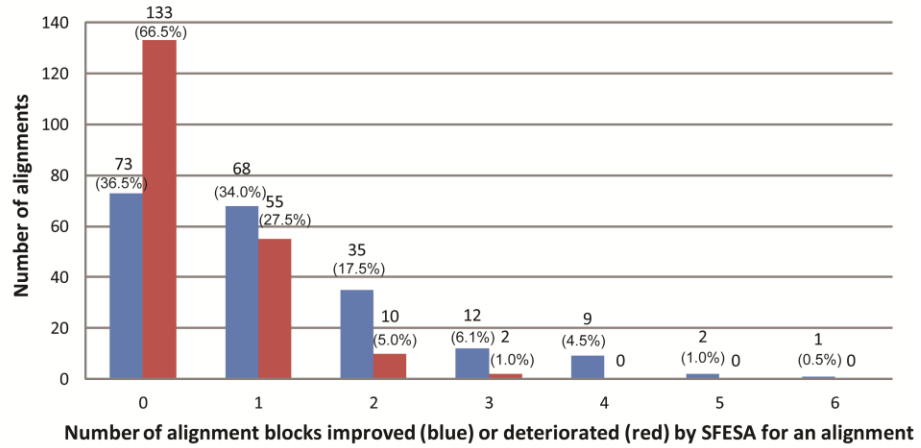

133
(66.5%)
73
(36.5%)
68
(34.0%)
55
(27.5%)
35
(17.5%)
10
(5.0%)
12
(6.1%)
2
(1.0%)
9
(4.5%)
0
2
(1.0%)
0
1
(0.5%)
0
140
120
100
80
60
40
20
0
Number of alignments
0
1
2
3
4
5
6
Number of alignment blocks improved (blue) or deteriorated (red) by SFESA for an alignment

Figure IV-5. Alignment block-level evaluation of SFESA performance on different datasets. (A) Evaluation on our training dataset (1675 alignments). (B) Evaluation on the MUSTER benchmark (300 alignments). (C) Evaluation on the SALIGN benchmark (200 alignments). SFESA (O+G+M+S) is used to refine alignments generated by PROMALS and Dali structure alignment is used as the reference. The blue column represents the number of alignments in which a certain number of aligned blocks were improved by SFESA. The red column represents the number of alignments in which a certain number of aligned blocks were deteriorated by SFESA. Columns of the "0" in the x-axis show the number of alignments where none of the alignment blocks were improved (blue) by SFESA and the number of alignments where none of the alignment blocks were deteriorated (red) by SFESA. The number of alignment cases in each category and the percentage is shown above each column.

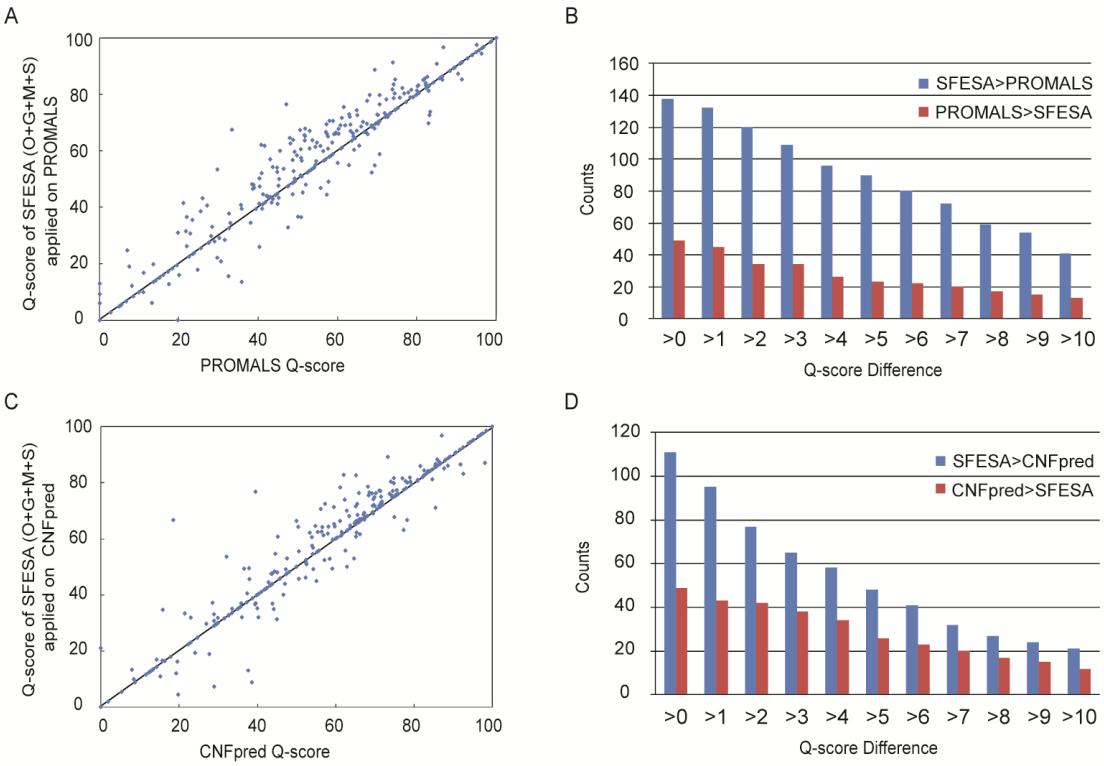


Figure IV-6. DALI Q-score for the MUSTER benchmark. (A) Scatter plot of SFESA (O+G+M+S) Q-score (applied to PROMALS) vs. PROMALS Q-score. Each point represents one domain pair. (B) The number of alignments that SFESA is better than PROMALS in Q-score and the number of alignments that PROMALS is better than SFESA at different Q-score difference cutoffs. (C) Scatter plot of SFESA (O+G+M+S) Q-score (applied to CNFpred) vs. CNFpred Q-score. (D) The number of the alignments that SFESA is better than CNFpred in Q-score and the number of the alignments that CNFpred is better than SFESA at different Q-score difference cutoffs.

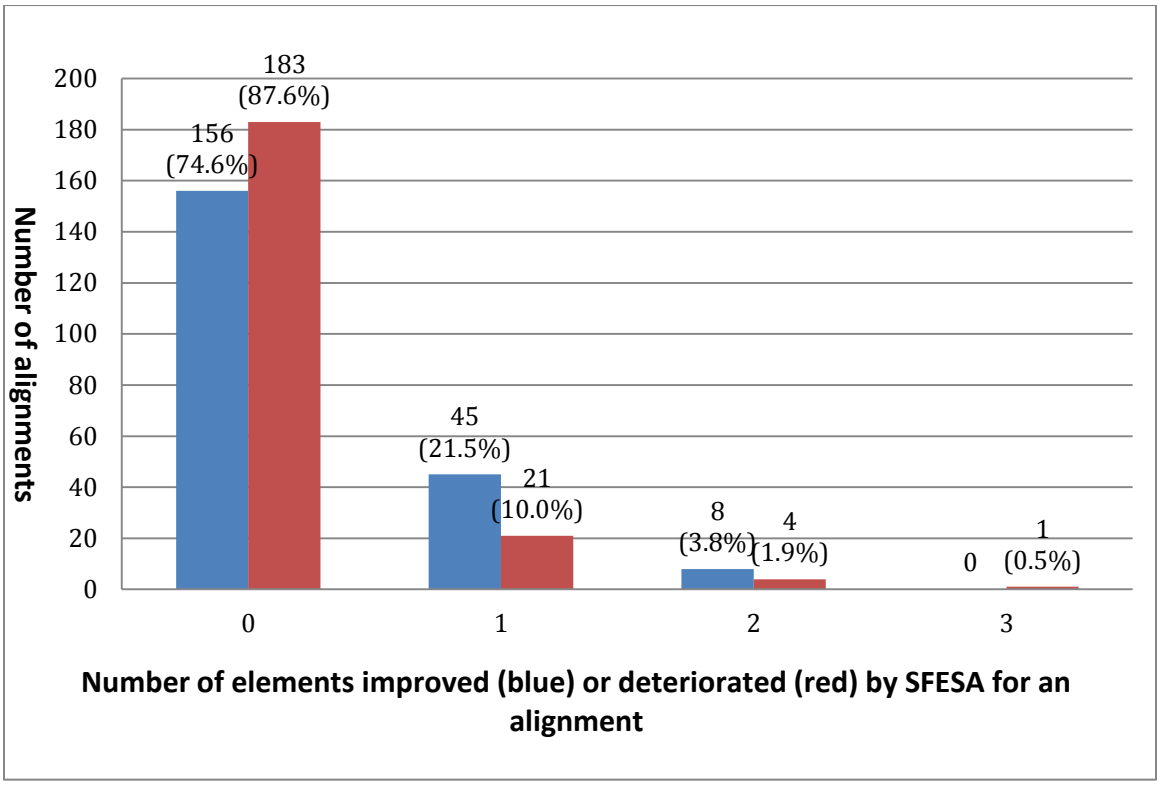


Figure IV-7. Element-level evaluation of SFESA performance on the SABMARK twilight dataset. SFESA (O+G+M+S) is used and SABmark's own reference is used. The blue column represents the number of alignments in which a certain number of aligned blocks were improved by SFESA. The red column represents the number of alignments in which a certain number of aligned blocks were deteriorated by SFESA. Columns of the "0" in the x-axis show the number of alignments where none of the alignment blocks were improved (blue) by SFESA and the number of alignments where none of the alignment blocks were deteriorated (red) by SFESA. The number of alignment cases in each category and the percentage is shown above each column.

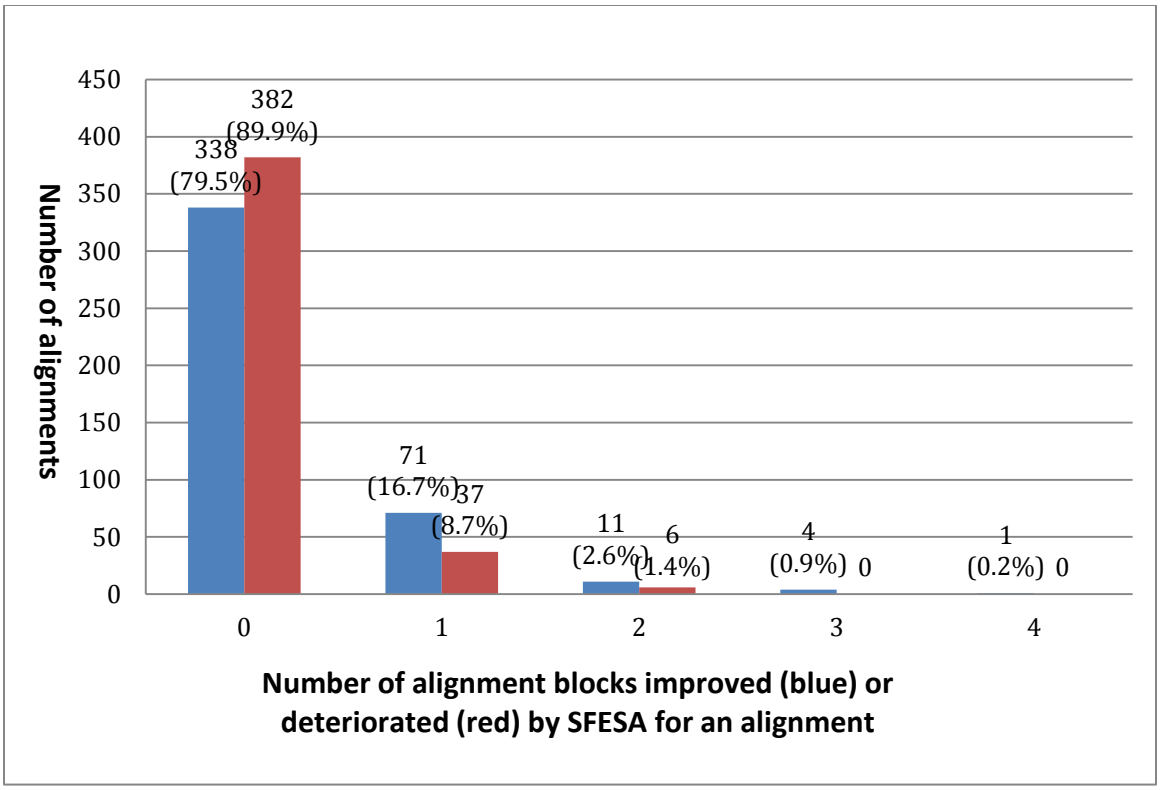


Figure IV-8. Alignment block-level evaluation of SFESA performance on the SABMARK superfamily dataset. SFESA (O+G+M+S) is used and SABmark's own reference is used. The blue column represents the number of alignments in which a certain number of aligned blocks were improved by SFESA. The red column represents the number of alignments in which a certain number of aligned blocks were deteriorated by SFESA. Columns of the "0" in the x-axis show the number of alignments where none of the alignment blocks were improved (blue) by SFESA and the number of alignments where none of the alignment blocks were deteriorated (red) by SFESA. The number of alignment cases in each category and the percentage is shown above each column.

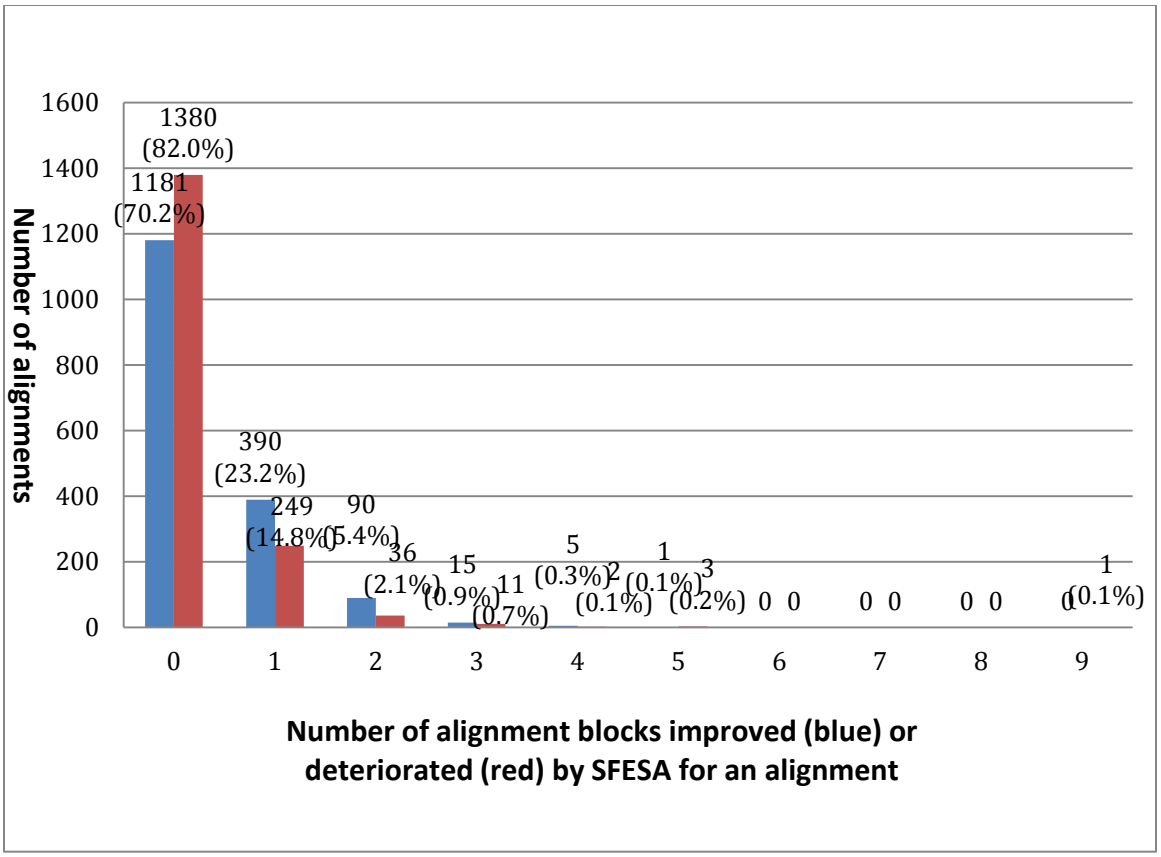


Figure IV-9. Alignment block-level evaluation of SFESA performance on the PREFAB dataset. SFESA (O+G+M+S) is used and PREFAB's own reference is used. The blue column represents the number of alignments in which a certain number of aligned blocks were improved by SFESA. The red column represents the number of alignments in which a certain number of aligned blocks were deteriorated by SFESA. Columns of the "0" in the x-axis show the number of alignments where none of the alignment blocks were improved (blue) by SFESA and the number of alignments where none of the alignment blocks were deteriorated (red) by SFESA. The number of alignment cases in each category and the percentage is shown above each column.

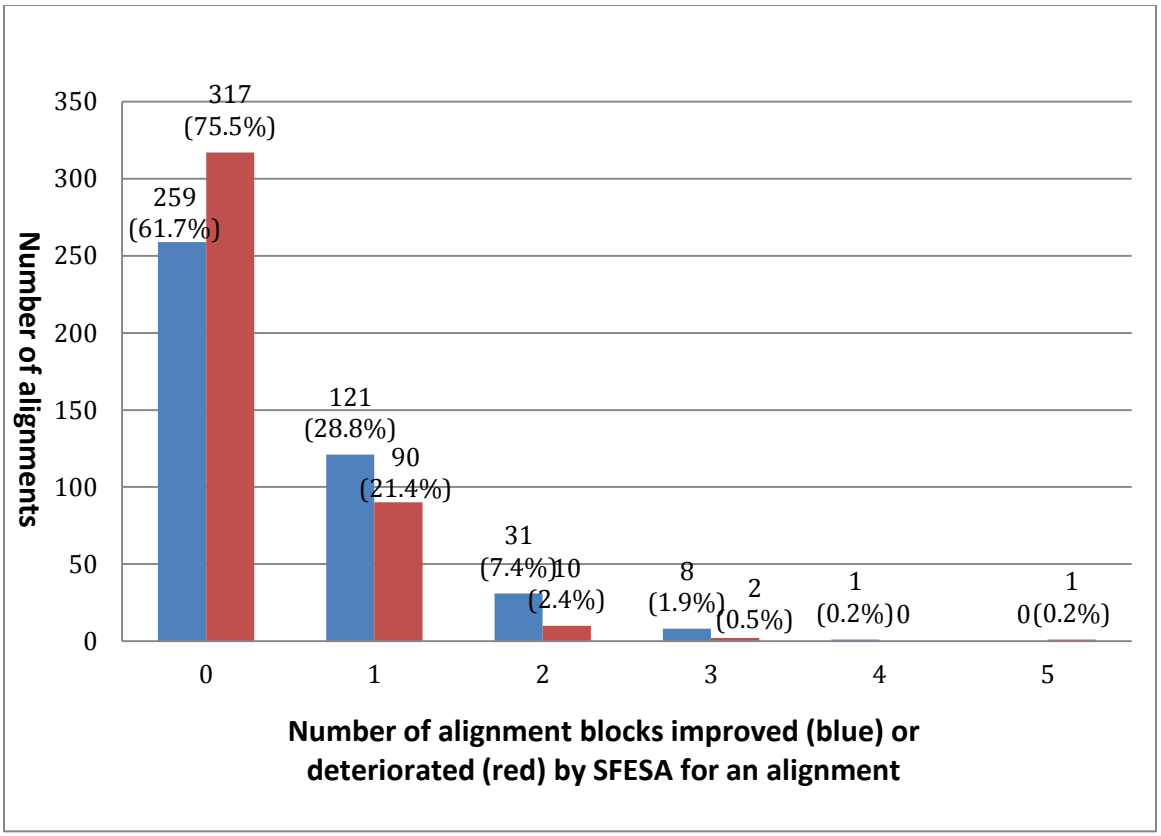


Figure IV-10. Alignment block-level evaluation of SFESA performance on the PREFAB "set1". SFESA (O+G+M+S) is used and PREFAB's own reference is used. The blue column represents the number of alignments in which a certain number of aligned blocks were improved by SFESA. The red column represents the number of alignments in which a certain number of aligned blocks were deteriorated by SFESA. Columns of the "0" in the x-axis show the number of alignments where none of the alignment blocks were improved (blue) by SFESA and the number of alignments where none of the alignment blocks were deteriorated (red) by SFESA. The number of alignment cases in each category and the percentage is shown above each column.

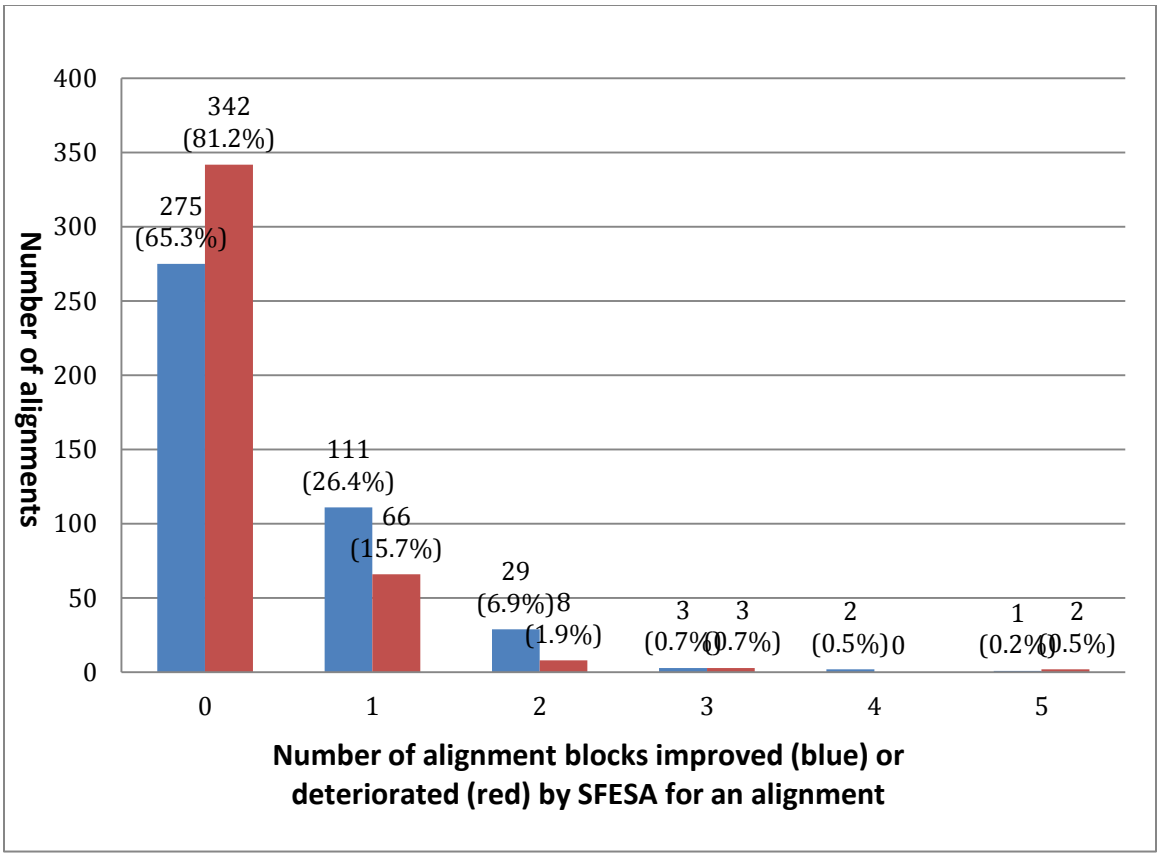


Figure IV-11. Alignment block-level evaluation of SFESA performance on the PREFAB "set2". SFESA (O+G+M+S) is used and PREFAB's own reference is used. The blue column represents the number of alignments in which a certain number of aligned blocks were improved by SFESA. The red column represents the number of alignments in which a certain number of aligned blocks were deteriorated by SFESA. Columns of the "0" in the x-axis show the number of alignments where none of the alignment blocks were improved (blue) by SFESA and the number of alignments where none of the alignment blocks were deteriorated (red) by SFESA. The number of alignment cases in each category and the percentage is shown above each column.

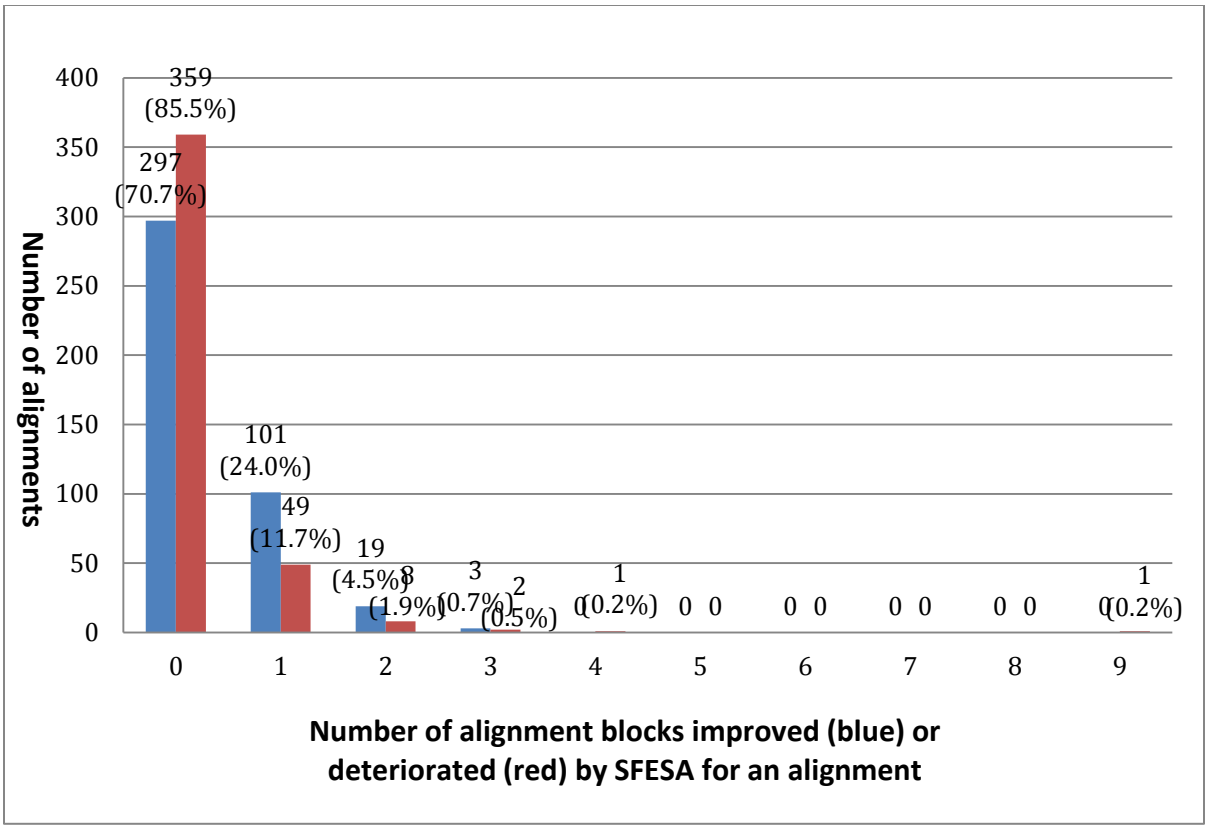


Figure IV-12. Alignment block-level evaluation of SFESA performance on the PREFAB "set3". SFESA (O+G+M+S) is used and PREFAB's own reference is used. The blue column represents the number of alignments in which a certain number of aligned blocks were improved by SFESA. The red column represents the number of alignments in which a certain number of aligned blocks were deteriorated by SFESA. Columns of the "0" in the x-axis show the number of alignments where none of the alignment blocks were improved (blue) by SFESA and the number of alignments where none of the alignment blocks were deteriorated (red) by SFESA. The number of alignment cases in each category and the percentage is shown above each column.

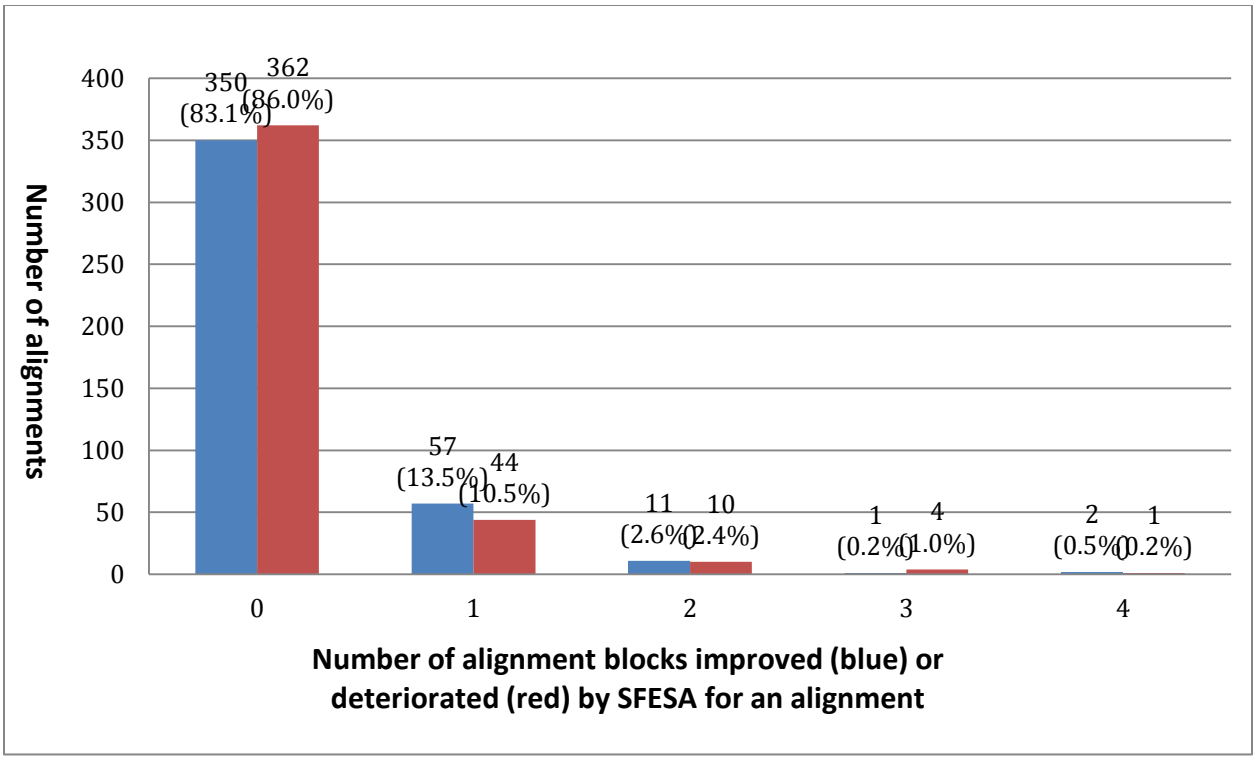


Figure IV-13. Alignment block-level evaluation of SFESA performance on the PREFAB "set4". SFESA (O+G+M+S) is used and PREFAB's own reference is used. The blue column represents the number of alignments in which a certain number of aligned blocks were improved by SFESA. The red column represents the number of alignments in which a certain number of aligned blocks were deteriorated by SFESA. Columns of the "0" in the x-axis show the number of alignments where none of the alignment blocks were improved (blue) by SFESA and the number of alignments where none of the alignment blocks were deteriorated (red) by SFESA. The number of alignment cases in each category and the percentage is shown above each column.

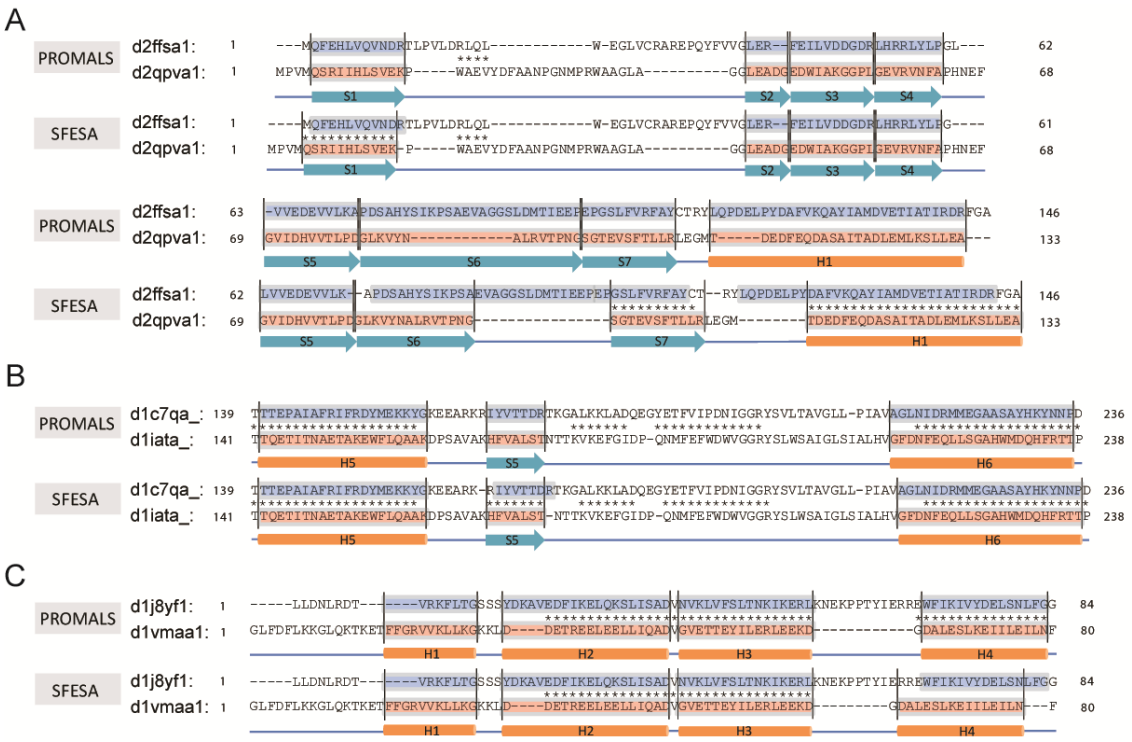


Figure IV-14. Three examples of SFESA refinement. (A) The alignments between d2ffsa1 (query) and d2qpva1 (template) generated by PROMALS and SFESA (O+G) + PROMALS. (B) The partial alignments between d1c7qa_ (query) and d1iata_ (template) generated by PROMALS and SFESA (O) + PROMALS. (C) The alignments between d1j8yf1 (query) and d1vmaa1 (template) generated by PROMALS and SFESA (O) + PROMALS. The pink boxes show the SSEs recognized from template and the blue boxes are those regions in the query aligned to such SSEs. Each corresponding blue and pink regions is an alignment block. The asterisk between two aligned residues indicates this aligned residue pair is in agreement with DALI alignment (reference).

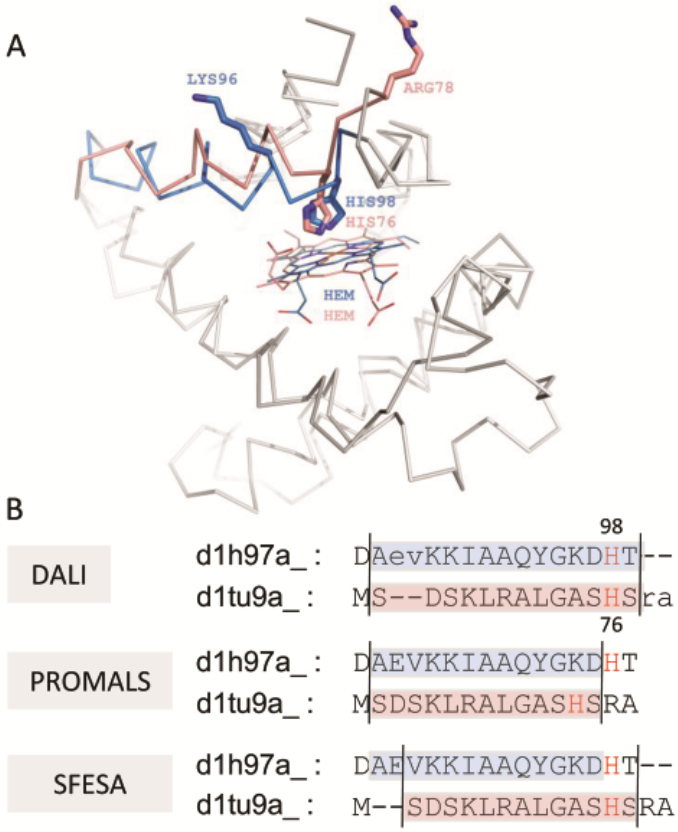


Figure IV-15. An example of SFESA correction of a misaligned active site residue. (A). Superposition of d1h97a_ (query) and d1tu9a_ (template) based on the DALI structure alignment (reference). The blue (query) and pink (template) α-helical regions indicate the alignment block. The histidine residues are the active site residues in contact with hemes (shown in lines). LYS96 and HIS98 in the query are incorrectly aligned to HIS76 and ARG78in the template in the PROMALS alignment, respectively. The sidechains of these residues are shown in sticks. (B). Alignments of DALI (reference), PROMALS and SFESA in this region. All SFESA modes can generate such alignment refinement.

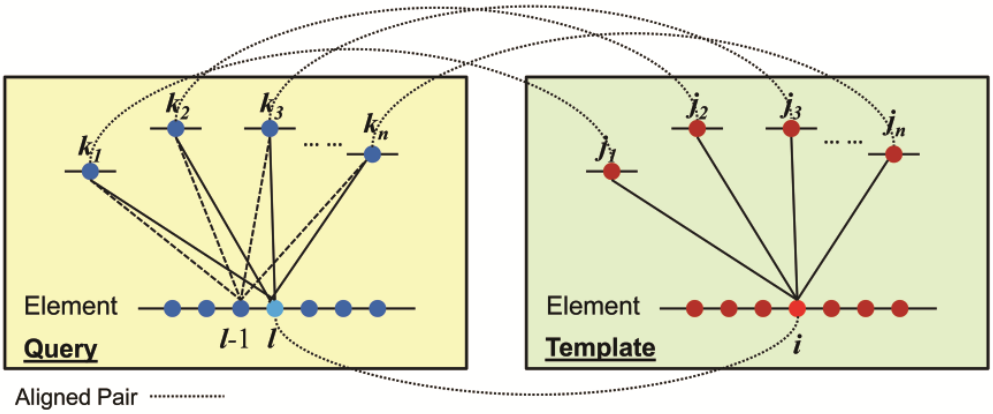


Figure IV-16. The template contact residue pairs are transferred to the query by original alignment to calculate structure score for the original alignment block and alignment variants. The blue and red filled circles represent residues in query and template, respectively. The dashed lines connect aligned residue pairs in the original alignment. Residue *i* is in contact with residues j_1, j_2, $j_3 \ldots j_n$ based on template structure. Residue *I* in the query is aligned with *i* and is inferred to be in contact with residues k_1, k_2, $k_3 \ldots k_n$ that are aligned to j_1, j_2, $j_3 \ldots j_n$. The contact-based score for residue *l* is calculated by Eq (5). In the case of +1 shift, residue *l*-1 is aligned to residue *i*, and the inferred contacts are between residue *l*-1 and k_1, k_2, $k_3 \ldots k_n$ (shown as dashed lines).

Methods	Reference-dependent (Q-score)					Reference-independent (TM-score)
	Dali	**TMalign**	**Matt**	**MUSTER**	**Deep Align**	
PROMALS	51.6	48.1	49.5	51.5	53.5	0.515
SFESA (O)+PROMALS	53.4	49.6	51.5	53.2	55.0	0.521
SFESA (O+G)+PROMALS	53.6	49.6	51.6	53.4	55.1	0.522
SFESA (O+G+M)+PROMALS	**54.2**	**50.2**	**52.1**	**54.0**	55.3	0.523
SFESA (O+G+M+S)+PROMALS	**54.2**	50.0	52.0	53.8	**55.7**	**0.525**
HHpred	49.3	**45.3**	46.7	49.0	49.7	0.490
SFESA (O)+HHpred	49.2	45.2	46.8	48.8	49.8	0.490
SFESA (O+G)+HHpred	49.4	45.2	47.0	**49.1**	**50.0**	**0.491**
SFESA (O+G+M)+HHpred	49.4	45.1	47.2	49.0	49.7	0.490
SFESA (O+G+M+S)+HHpred	**49.6**	**45.3**	**47.3**	**49.1**	49.9	**0.491**
CNFpred	51.5	48.2	49.2	52.4	53.7	0.511
SFESA (O)+CNFpred	52.0	48.3	49.9	52.5	54.0	0.511
SFESA (O+G)+CNFpred	52.2	48.4	50.1	52.5	54.1	0.512
SFESA (O+G+M)+CNFpred	52.6	48.7	50.4	52.9	54.0	0.512
SFESA (O+G+M+S)+CNFpred	**52.7**	**49.0**	**50.7**	**53.3**	**54.8**	**0.515**

Table IV-1. Test on MUSTER database. Columns 2-6 indicate five different structure alignment methods to generate reference alignments (Reference-dependent evaluation). Column 7 indicates the average of query model's TM-score built by Modeller (Reference-independent evaluation). Bold indicates the best performance in the subsection. Bold with underscore indicates the overall best performance in one column. Average Q-score and TM-score are reported.

Method/(p-value)	PROMALS
SFESA(O)+PROMALS	3.90E-08
SFESA(O+G)+PROMALS	2.50E-06
SFESA(O+G+M)+PROMALS	6.70E-09
SFESA(O+G+M+S)+PROMALS	0

Method/(p-value)	HHpred
SFESA(O)+HHpred	0.36
SFESA(O+G)+HHpred	0.094
SFESA(O+G+M)+HHpred	0.1
SFESA(O+G+M+S)+HHpred	0.034

Method/(p-value)	CNFpred
SFESA(O)+CNFpred	0.064
SFESA(O+G)+CNFpred	0.064
SFESA(O+G+M)+CNFpred	0.0034
SFESA(O+G+M+S)+CNFpred	0.00024

Table IV-2. Statistically significant Q-score improvement of SFESA on PROMALS, HHpred and CNFpred (Dali as a reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on MUSTER dataset. P-values are calculated based on the paired Wilcoxon signed-rank test. P-values below 0.05 are marked green and below 0.005 are marked pink.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3159	151	56	164	1246	404	41806
SFESA (O+G)+PROMALS	2678	283	172	397	2015	972	40469
SFESA (O+G+M)+PROMALS	2577	326	201	426	2363	1186	39907
SFESA (O+G+M+S)+PROMALS	2997	253	90	190	1844	636	40976
SFESA (O)+HHpred	2990	17	25	51	119	160	43177
SFESA (O+G)+HHpred	2588	156	103	236	524	459	42473
SFESA (O+G+M)+HHpred	2340	251	184	308	1009	923	41524
SFESA (O+G+M+S)+HHpred	2706	162	91	124	649	520	42287
SFESA (O)+CNFpred	3011	81	52	90	646	394	42416
SFESA (O+G)+CNFpred	2395	255	189	395	1420	1092	40944
SFESA (O+G+M)+CNFpred	2367	291	188	388	1664	1167	40625
SFESA (O+G+M+S)+CNFpred	2821	183	83	147	1103	534	41819

Table IV-3. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (Dali as a reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on MUSTER dataset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3159	139	65	167	1130	442	44221
SFESA (O+G)+PROMALS	2678	269	177	406	1852	1035	42906
SFESA (O+G+M)+PROMALS	2577	316	200	437	2186	1226	42381
SFESA (O+G+M+S)+PROMALS	2997	232	97	204	1633	706	43454
SFESA (O)+HHpred	2990	16	26	51	97	148	45548
SFESA (O+G)+HHpred	2588	137	109	249	455	434	44904
SFESA (O+G+M)+HHpred	2340	231	197	315	906	940	43947
SFESA (O+G+M+S)+HHpred	2706	139	98	140	567	526	44700
SFESA (O)+CNFpred	3011	72	54	97	548	427	44818
SFESA (O+G)+CNFpred	2395	240	194	405	1257	1107	43429
SFESA (O+G+M)+CNFpred	2367	271	206	390	1433	1170	43190
SFESA (O+G+M+S)+CNFpred	2821	167	82	164	976	541	44276

Table IV-4. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (TMalign as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on MUSTER dataset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3159	145	53	173	1176	330	40278
SFESA (O+G)+PROMALS	2678	271	150	431	1816	837	39131
SFESA (O+G+M)+PROMALS	2577	311	178	464	2146	1079	38559
SFESA (O+G+M+S)+PROMALS	2997	239	85	209	1656	577	39551
SFESA (O)+HHpred	2990	20	19	54	130	107	41547
SFESA (O+G)+HHpred	2588	146	95	254	520	394	40870
SFESA (O+G+M)+HHpred	2340	225	171	347	974	819	39991
SFESA (O+G+M+S)+HHpred	2706	155	80	142	684	456	40644
SFESA (O)+CNFpred	3011	76	42	105	587	291	40906
SFESA (O+G)+CNFpred	2395	230	170	439	1331	964	39489
SFESA (O+G+M)+CNFpred	2367	253	174	440	1496	1033	39255
SFESA (O+G+M+S)+CNFpred	2821	168	71	174	1104	446	40234

Table IV-5. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (Matt as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on MUSTER dataset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3159	135	58	178	1098	386	40230
SFESA (O+G)+PROMALS	2678	257	161	434	1776	911	39027
SFESA (O+G+M)+PROMALS	2577	299	181	473	2076	1084	38554
SFESA (O+G+M+S)+PROMALS	2997	231	87	215	1615	644	39455
SFESA (O)+HHpred	2990	16	25	52	96	154	41464
SFESA (O+G)+HHpred	2588	132	90	273	437	399	40878
SFESA (O+G+M)+HHpred	2340	218	174	351	837	849	40028
SFESA (O+G+M+S)+HHpred	2706	136	87	154	540	487	40687
SFESA (O)+CNFpred	3011	68	55	100	511	425	40778
SFESA (O+G)+CNFpred	2395	226	185	428	1203	1093	39418
SFESA (O+G+M)+CNFpred	2367	251	195	421	1363	1127	39224
SFESA (O+G+M+S)+CNFpred	2821	163	76	174	965	515	40234

Table IV-6. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (MUSTER as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on MUSTER dataset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3159	145	68	158	1196	461	42743
SFESA (O+G)+PROMALS	2678	286	188	378	1997	1090	41313
SFESA (O+G+M)+PROMALS	2577	314	233	406	2291	1385	40724
SFESA (O+G+M+S)+PROMALS	2997	244	108	181	1800	729	41871
SFESA (O)+HHpred	2990	21	22	50	144	140	44116
SFESA (O+G)+HHpred	2588	160	110	225	576	444	43380
SFESA (O+G+M)+HHpred	2340	240	214	289	1044	1058	42298
SFESA (O+G+M+S)+HHpred	2706	149	107	121	653	587	43160
SFESA (O)+CNFpred	3011	79	57	87	629	434	43337
SFESA (O+G)+CNFpred	2395	270	196	373	1500	1200	41700
SFESA (O+G+M)+CNFpred	2367	280	221	366	1578	1383	41439
SFESA (O+G+M+S)+CNFpred	2821	182	80	151	1112	544	42744

Table IV-7. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (Deepalign as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on MUSTER dataset.

Domain 1	Domain 2	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
1aba__	1gp1A_	7	0	0	0	0	0	80
1acf__	1pne__	8	2	0	0	3	0	119
1afi__	1aps__	4	0	1	0	1	10	52
1agjA_	1elg__	12	1	0	2	7	0	178
1agqD_	1tgj__	7	0	0	0	0	0	87
1aizB_	1rcy__	9	0	0	0	0	0	102
1apmE_	1erk__	16	0	2	0	0	16	249
1apmE_	1irk__	11	2	2	0	22	13	219
1ash__	1binA_	7	0	0	0	0	0	135
1ash__	1bvd__	6	0	0	0	0	0	139
1ash__	1cpcA_	4	1	1	1	7	11	95
1ax4A_	1cl1A_	18	2	0	1	26	2	266
1bbhB_	1nbbB_	4	0	0	0	0	0	126
1bbpD_	1hbq__	11	2	0	0	14	0	139
1bcpL_	1prtB_	7	0	1	0	0	8	76
1bdiA_	2dri__	20	0	1	0	0	8	252
1bdmB_	6ldh__	18	2	0	1	14	3	288
1bfmA_	1tafB_	3	0	0	0	0	0	66
1binA_	1bvd__	6	0	0	0	0	0	138
1binA_	2hbg__	7	0	0	0	0	0	135
1brz__	1gps__	4	0	0	0	0	0	44
1btn__	1dynB_	7	1	0	0	1	0	89
1btn__	1irsA_	6	1	0	0	1	0	85
1btn__	1mai__	7	1	0	1	6	0	82
1bvd__	2hbg__	7	0	0	0	0	0	138
1cewI_	1molA_	4	0	1	0	0	1	80
1cewI_	1ounA_	4	0	1	0	0	13	64
1cnv__	1nar__	14	2	0	2	25	2	219
1cpcA_	1colA_	5	0	1	1	0	8	106
1cpcA_	1cpcB_	9	0	0	0	0	0	159
1cpcA_	2hbg__	7	0	0	0	0	0	122
1ctj__	1cxc__	7	0	0	0	0	0	75
1ctj__	2mtaC_	3	1	0	0	12	0	73
1dat__	1afrF_	5	1	0	0	10	1	141
1dat__	1ryt__	4	1	0	1	2	0	148
1dat__	1xikA_	5	1	0	1	30	0	117
1den__	1tcp__	2	1	0	0	8	0	38
1dhkA_	1bag__	25	0	0	0	0	0	394
1dynB_	1irsA_	6	1	0	1	7	0	86

1dynB_	1mai__	8	1	0	0	6	0	78
1eaf__	3cla__	13	0	1	1	0	3	173
1eceA_	1edg__	19	1	0	0	18	1	261
1ecmB_	1csmB_	3	0	0	0	0	0	92
1ecpA_	1ula__	16	1	0	0	6	0	217
1elg__	1havA_	17	1	0	2	4	1	169
1elg__	2alp__	12	3	0	1	9	0	147
1erk__	1irk__	15	1	0	0	7	0	241
1esl__	1lit__	9	1	0	0	4	0	110
1eur__	2sim__	23	4	0	2	34	3	287
1fecA_	1nhq__	29	2	0	0	10	0	389
1fmb__	1smeB_	10	0	1	0	0	1	100
1fna__	1mspA_	7	0	0	1	0	0	71
1gsa__	1bncA_	19	2	0	3	20	1	219
1gsa__	1iow__	18	3	0	0	28	3	217
1gtqA_	1gtpA_	5	1	1	0	10	1	81
1hce__	1i1b__	9	1	1	1	9	2	104
1hce__	4fgf__	10	2	0	0	12	2	97
1hfc__	1iag__	9	1	0	0	19	0	112
1hrdA_	1lehA_	22	1	0	0	9	3	314
1i1b__	4fgf__	9	2	1	0	14	7	98
1idk__	1air__	41	0	0	1	0	0	268
1iow__	1bncA_	22	2	0	0	7	0	245
1irsA_	1mai__	4	2	3	0	15	29	42
1kpcD_	1hxqB_	7	0	0	0	0	0	99
1lea__	1ruoB_	3	1	0	0	6	0	54
1lpe__	1nbbB_	4	0	0	0	0	0	103
1lpe__	1vltA_	4	0	0	0	0	0	111
1ltiD_	1bcpL_	6	0	1	0	0	5	79
1ltiD_	1prtB_	7	0	2	0	0	18	63
1ltiD_	1tiiD_	7	0	1	0	0	10	85
1ndh__	1fnb__	17	1	0	1	1	0	228
1ndh__	2pia__	13	1	0	0	5	0	208
1nulA_	1hgxA_	12	1	0	1	10	0	116
1ounA_	1std__	4	2	0	1	7	0	113
1pgs__	1phm__	13	0	0	5	0	0	207
1phd__	2hpdB_	21	3	1	0	9	1	358
1plq__	2polA_	17	2	0	0	8	0	213
1pot__	1sbp__	18	4	0	0	32	0	235
1ptvA_	1ytn__	14	0	0	0	0	0	246
1qpaA_	2cyp__	17	0	0	0	0	0	262
1ris__	1spbP_	4	0	2	0	0	9	53
1rmg__	1bhe__	49	3	0	0	11	2	308

1ryt__	1afrF_	8	0	0	1	0	0	137
1ryt__	1xikA_	7	0	0	0	0	0	146
1ryt__	1xsm__	5	0	0	0	0	0	140
1sbp__	2abh__	18	0	0	0	0	0	221
1spbP_	1nueA_	4	0	1	0	0	8	52
1ste__	2tssA_	17	0	0	0	0	0	183
1tafA_	1bfmA_	3	0	0	0	0	0	66
1tafA_	1tafB_	3	0	0	0	0	0	63
1tdj__	2tysB_	21	1	1	0	3	1	305
1tiiD_	3ullA_	6	0	0	1	0	0	75
1ulo__	2ayh__	11	0	1	1	0	5	122
1vltA_	1nbbB_	4	0	0	0	0	0	99
1wba__	1i1b__	7	5	0	1	41	0	79
1wba__	4fgf__	9	1	0	2	7	0	104
1xikA_	1afrF_	14	0	0	1	0	0	224
1xikA_	1xsm__	10	0	0	0	0	0	274
1xsm__	1afrF_	10	0	0	3	0	0	237
2alp__	1havA_	16	1	0	0	3	0	149
2bltB_	3pte__	20	0	0	0	0	0	305
2gmfB_	1rcb__	5	0	0	0	0	0	91
2hfh__	1hstA_	5	0	0	0	0	0	67
2hhmA_	1spiD_	16	3	0	1	17	0	202
2lefA_	1hma__	4	0	0	0	0	0	66
2pia__	1fnb__	15	1	2	0	2	5	208
2pii__	1aps__	4	0	0	1	0	0	70
3nll__	1qrdB_	9	0	0	0	0	0	131
3ullA_	1bcpL_	5	0	0	0	0	0	80
193l__	153l__	7	0	0	0	0	0	95
1p53A3	1fna__	3	1	0	2	6	0	59
1tig__	1pavA_	4	0	0	0	0	0	58
1vkfA_	1i0dA_	11	0	0	5	0	0	141
1rwhA1	1r76A_	11	0	0	1	0	0	159
1fseA_	1rr7A_	3	0	0	0	0	0	41
1k5jA_	2bbvA_	4	0	0	2	0	0	75
1ul7A_	1t17A_	5	0	1	0	0	3	72
1te4A_	1hz4A_	4	0	0	3	0	0	100
1nznA_	1b89A_	4	0	0	1	0	0	85
1rliA_	1g2iA_	11	1	0	0	1	0	91
1m3uA_	1nqkA_	13	3	0	0	19	0	150
1xvhA1	1r8iA_	3	0	0	0	0	0	60
1xmbA1	1rxyA_	10	1	0	2	8	1	155
1dqgA_	1avaC_	7	0	0	3	0	0	108
1egjA_	1ollA1	8	0	1	0	0	10	66

1jigA_	1nogA_	4	0	0	0	0	0	100
1h1nA_	1r3sA_	11	0	3	1	0	16	219
2tpsA_	1a0cA_	9	0	0	5	0	0	186
1wmhA_	1t0yA_	6	1	0	0	7	0	62
1qopA_	1tv8A_	15	0	1	0	0	1	167
1y6dA_	1jogA_	6	0	0	0	0	0	82
2ila__	1avaC_	7	2	0	1	14	0	107
1t9fA_	1md6A_	11	1	0	1	7	1	118
1p6rA_	1gxqA_	4	2	0	0	12	0	50
1vhwA_	1vheA2	11	0	0	3	0	0	157
2plc__	1odzA_	14	0	0	1	0	0	192
1gzgA_	1eokA_	11	0	0	3	0	0	184
1ukuA_	1nu4A_	4	1	0	2	5	2	67
1ep3A_	1uuqA_	19	1	1	1	4	7	199
1p1xA_	1k77A_	18	1	1	0	6	5	181
1pii_2	1fcqA_	8	0	1	0	0	7	142
1dcfA_	1fyeA_	7	3	0	0	18	0	71
1u8sA2	1s99A_	4	0	0	1	0	0	69
1lucA_	1ujpA_	9	0	0	6	0	0	194
1qfoA_	1rowA_	7	1	0	1	5	0	78
1vhnA_	1qnrA_	20	0	1	2	0	2	189
1oeyJ_	1v5oA_	4	1	0	0	2	0	68
1pfbA_	1vie__	5	0	0	1	0	0	43
1m5wA_	1ur4A_	11	3	1	0	25	4	182
1dcfA_	1l9xA_	10	0	1	1	0	9	95
1pii_2	1f8mA_	14	0	1	0	0	6	159
2pth__	1vheA2	6	0	2	2	0	24	98
1ukuA_	1utaA_	3	2	0	1	20	0	48
1ub3A_	1qopA_	13	3	1	0	23	8	142
1vk8A_	1y0hA_	4	1	0	1	10	0	67
1k66A_	1n57A_	7	1	1	2	6	8	99
1uwdA_	1josA_	4	0	0	1	0	0	73
1u8sA2	1ulrA_	5	1	0	0	7	0	64
1j6oA_	1jfxA_	10	0	0	3	0	0	156
1urrA_	1tr0A_	3	0	0	3	0	0	71
1jfxA_	1rhcA_	10	2	1	0	12	6	144
1u8sA2	1itpA_	6	0	0	0	0	0	69
1izcA_	1r30A_	17	2	0	1	22	7	156
1ix2A_	1owwA_	6	0	0	1	0	0	80
1bbzA_	1jb0E_	2	0	0	0	0	0	48
1nbcA_	1ayoA_	9	2	0	0	18	0	83
1roaA_	1oh0A_	3	0	1	1	0	11	66
1wmhB_	1xd3B_	4	0	1	0	0	8	60

1h05A_	1ydgA_	8	1	0	0	13	0	94
1f0zA_	1v2yA_	6	0	0	0	0	0	60
1j27A_	1q8bA_	6	0	1	1	0	7	72
1u1jA2	2ebn__	15	0	1	2	0	2	195
1eceA_	1d8wA_	12	0	1	2	0	17	209
1sj1A_	1u8sA1	5	0	1	0	0	2	53
1od6A_	1mjhA_	6	1	0	1	15	0	92
1rhcA_	1eokA_	15	0	1	0	0	2	197
1knmA_	1j0sA_	8	3	0	0	15	0	97
1ukuA_	1vmbA_	6	0	0	1	0	0	81
1ujpA_	1i1wA_	13	2	1	1	12	10	168
1j27A_	1s99A_	4	0	1	1	0	7	65
1jixA_	1a8i__	22	0	0	1	0	0	296
1pjaA_	1lzlA_	11	0	0	1	0	0	183
1bkb_2	1qzgA_	5	0	0	0	0	0	53
1ntvA_	1ntyA2	7	1	0	1	16	0	80
1ptmA_	1u7nA_	13	0	0	1	0	0	233
1mi8A_	1at0__	13	0	0	0	0	0	131
1pjqA1	1qycA_	8	1	0	0	6	0	95
1h70A_	1g61A_	8	5	2	2	30	17	132
1efdN_	1pszA_	13	2	1	1	13	6	178
1xtpA_	1o54A_	15	1	0	0	1	0	148
1t4hA_	1nw1A_	15	1	0	1	4	3	151
1sacA_	1y4wA1	16	0	0	1	0	0	126
1gv9A_	2sli_1	15	1	0	2	7	0	128
1j97A_	1rlmA_	14	0	0	1	0	0	135
1bjaA_	1jgsA_	7	0	0	0	0	0	84
1gklA_	1brt__	11	1	1	1	10	3	165
1fjjA_	1qouA_	9	1	0	0	5	0	111
1it2A_	1allA_	6	2	0	0	28	0	90
1xpjA_	1qyiA_	8	0	0	1	0	0	99
1a78A_	1s2bA_	15	0	0	1	0	0	113
1k2eA_	1vk6A2	7	1	0	1	12	0	101
1oeyA_	1c9fA_	4	1	0	1	1	0	56
1qx2A_	1s6jA_	4	0	0	0	0	0	56
1rqjA_	1di1A_	8	1	2	1	2	34	183
1s1qA_	1jatA_	6	1	0	1	9	0	94
1wfzA_	1mzgA_	5	1	0	0	13	0	83
1c4rA_	1oq1A_	10	0	1	1	0	4	126
1e5kA_	1i52A_	10	3	0	1	18	0	157
1mek__	1lu4A_	8	0	0	1	0	0	92
1tjoA_	1mtyD_	6	0	0	0	0	0	145
1f3yA_	1vk6A2	9	0	0	0	0	0	120

1m2dA_	1wouA_	4	1	0	1	8	0	63
1xbbA_	1nd4A_	13	2	0	0	16	0	148
1m6sA_	1eg5A_	17	4	1	0	25	4	275
1npsA_	1f53A_	7	0	0	0	0	0	63
1ll2A_	1h7eA_	13	0	0	1	0	0	163
1h70A_	1vkpA_	13	3	3	2	20	11	201
1xg8A_	2trxA_	7	1	0	0	7	0	77
1dlwA_	1h97A_	5	0	0	0	0	0	102
1wdkA2	1txgA1	9	3	0	1	18	0	135
1eg5A_	2dkb__	21	2	0	0	10	0	301
1dlwA_	1it2A_	6	1	0	0	3	0	95
1grj_2	1fd9A_	5	1	0	0	6	0	51
1rjdA_	1ri5A_	16	1	0	1	7	0	197
1ogmX2	1oflA_	38	0	0	2	0	0	270
1mdoA_	1m6sA_	18	2	2	0	15	20	258
1d2fA_	1s0aA_	17	6	1	0	48	5	255
1lsuA_	1rkxA_	10	1	0	0	8	0	113
1js3A_	1fg7A_	16	1	0	1	7	0	300
1kyhA_	1ub0A_	17	1	0	0	5	0	188
1wf6A_	1cdzA_	7	0	0	0	0	0	93
1wmgA_	3ygsP_	6	0	0	0	0	0	72
1gcwB_	1jboA_	6	0	1	0	0	15	102
1qx2A_	1y1xA_	4	0	0	0	0	0	66
1vj5A2	1tca__	12	0	0	1	0	0	192
1y0uA_	1stzA1	4	1	0	0	4	0	53
1keaA_	1ornA_	12	0	0	0	0	0	207
1h6yA_	1jhjA_	9	2	1	0	20	5	89
1nkiA_	1u6lA_	10	2	0	2	8	0	98
1ajsA_	1ohwA_	15	2	1	3	12	7	309
1e87A_	1o7bT_	6	0	1	0	0	3	67
1q6wA_	1mkaA_	4	1	1	1	2	6	98
1hekA_	1ecwA_	4	1	0	0	16	0	79
1txdA2	1wguA_	8	1	0	0	9	0	89
1eh2__	1uhnA_	4	0	1	0	0	3	71
1a3k__	1umzA_	10	0	0	1	0	0	125
1v70A_	1eybA_	7	2	0	0	11	0	83
1cpy__	1vj5A2	16	0	2	0	0	20	233
1or4A_	1xg0C_	8	0	0	0	0	0	126
1omzA_	1qg8A_	15	3	0	0	21	1	166
1qz9A_	1ax4A_	18	5	0	2	44	4	279
1tvgA_	1umhA_	5	2	1	0	13	5	83
1xc2A1	1txgA1	8	4	0	0	33	3	104
1p9aG_	1jl5A_	29	0	0	2	0	0	189

1is3A_	1kqrA_	7	0	0	3	0	0	101
1snyA_	1vjpA1	7	2	0	4	24	0	140
1wi9A_	1j75A_	4	0	1	0	0	1	53
1qlwA_	1b6g__	13	2	0	0	13	0	180
2cpgA_	1bazA_	2	0	0	0	0	0	43
1cvl__	1mnaA_	11	1	0	0	10	0	181
1xtpA_	1i9gA_	14	1	0	1	1	0	151
1nkgA2	1k42A_	8	1	0	1	9	0	104
1a78A_	1gv9A_	10	1	0	2	6	0	121
1rljA_	1oboA_	8	1	0	0	7	0	110
1imjA_	1din__	12	1	0	0	6	0	183
1rrqA2	1q33A_	7	1	0	0	16	0	94
1jigA_	1t0qB_	6	0	0	0	0	0	140
1rrqA2	1k2eA_	9	2	0	0	18	0	94
1sjwA_	1q40A_	5	0	0	1	0	0	105
1h8bA_	1uhnA_	4	0	0	0	0	0	62
1fg7A_	1ax4A_	21	4	0	0	39	6	261
1iqzA_	7fd1A_	4	1	0	0	5	1	51
1jg1A_	1i9gA_	12	0	0	2	0	0	145
1uliB_	1oh0A_	8	0	0	3	0	0	116
1s0aA_	1vp4A_	17	3	1	0	19	1	302
1nvmB1	1xg5A_	7	1	1	1	14	4	109
1esc__	1vjgA_	10	0	0	0	0	0	183
1dgnA_	1wh4A_	6	1	0	0	8	0	77
1guiA_	1h6yA_	10	2	0	0	18	2	106
1a3k__	1c4rA_	8	2	1	1	6	8	106
1ne9A2	1p0hA_	8	2	0	1	20	0	123
1mkp__	1wchA_	9	1	0	0	3	1	131
1h80A_	1k5cA_	32	3	1	0	19	4	239
1sjwA_	1hkxA_	9	0	0	0	0	0	109
1k75A_	1a4sA_	24	1	1	2	7	9	261
1s0aA_	1m6sA_	21	1	0	0	14	0	295
1rh6A_	1j9iA_	5	0	0	0	0	0	47
1ca1_2	1lox_2	9	0	0	0	0	0	105
1s3jA_	1bjaA_	6	0	0	0	0	0	82
1js3A_	1c7nA_	25	0	1	0	0	2	337
1mnaA_	1vj5A2	12	3	0	0	33	2	170
1m8zA_	1jdhA_	22	0	0	2	0	0	292
1vk4A_	1vi9A_	14	0	0	0	0	0	216
1senA_	1a8l_1	7	0	0	1	0	0	84
1p5tA_	1wi1A_	2	1	0	1	14	0	67
1rttA_	1sqsA_	10	2	0	0	15	0	141
1od5A2	1x7nA_	10	2	0	1	11	2	104

1ohwA_	1svvA_	15	1	2	2	7	4	282
1ttzA_	1t1vA_	6	0	0	0	0	0	71
1n1jB_	1tzyB_	3	0	0	0	0	0	77

Table IV-8. The alignment block-level and aligned position-level comparison for each alignment of SFESA (O+G+M+S) applied on PROMALS on the MUSTER dataset (Dali as a reference).

Methods	Reference-dependent (Q-score)				Reference-independent (TM-score)
	DALI	TMalign	Matt	DeepAlign	
PROMALS	61.4	59.5	60.2	62.6	0.582
SFESA (O)+PROMALS	62.7	60.5	61.2	63.9	0.585
SFESA (O+G)+PROMALS	63.4	61.0	61.9	64.6	0.588
SFESA (O+G+M)+PROMALS	63.7	61.1	62.2	64.8	**0.589**
SFESA (O+G+M+S)+PROMALS	**63.9**	**61.4**	**62.3**	**65.1**	**0.589**
HHpred	63.0	60.6	62.7	64.4	0.589
SFESA (O)+HHpred	63.1	60.6	62.7	64.4	0.590
SFESA (O+G)+HHpred	63.1	60.6	62.7	64.5	0.590
SFESA (O+G+M)+HHpred	**63.7**	**61.1**	**63.2**	**64.9**	**0.592**
SFESA (O+G+M+S)+HHpred	63.5	**61.1**	**63.2**	64.8	**0.592**
CNFpred	64.7	62.2	62.6	66.3	0.595
SFESA (O)+CNFpred	65.1	62.5	62.6	66.4	0.596
SFESA (O+G)+CNFpred	**65.3**	**62.7**	62.5	66.4	**0.598**
SFESA (O+G+M)+CNFpred	64.8	62.2	62.6	66.0	0.595
SFESA (O+G+M+S)+CNFpred	65.2	**62.7**	**63.0**	**66.5**	**0.598**

Table IV-9. Test on SALIGN database. Columns 2-5 indicate five different structure alignment methods to generate reference alignments (Reference-dependent evaluation). Column 6 indicates the average of query model's TM-score built by Modeller (Reference-independent evaluation). Bold indicates the best performance in the subsection. Bold with underscore indicates the overall best performance in one column. Average Q-score and TM-score are reported.

method/(p-value)	PROMALS
SFESA(O)+PROMALS	3.40E-08
SFESA(O+G)+PROMALS	0
SFESA(O+G+M)+PROMALS	0
SFESA(O+G+M+S)+PROMALS	0

method/(p-value)	HHpred
SFESA(O)+HHpred	0.19
SFESA(O+G)+HHpred	0.035
SFESA(O+G+M)+HHpred	3.60E-04
SFESA(O+G+M+S)+HHpred	0.0026

method/(p-value)	CNFpred
SFESA(O)+CNFpred	0.031
SFESA(O+G)+CNFpred	0.018
SFESA(O+G+M)+CNFpred	0.086
SFESA(O+G+M+S)+CNFpred	0.0026

Table IV-10. Statistically significant Q-score improvement of SFESA on PROMALS, HHpred and CNFpred (Dali as a reference). All SFESA options by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on SALIGN dataset. P-values are calculated based on the paired Wilcoxon signed-rank test. P-values below 0.05 are marked green and below 0.005 are marked pink.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3275	122	47	104	985	352	45502
SFESA (O+G)+PROMALS	2900	242	133	273	1652	701	44486
SFESA (O+G+M)+PROMALS	2801	278	159	310	1957	848	44034
SFESA (O+G+M+S)+PROMALS	3113	226	81	128	1622	453	44764
SFESA (O)+HHpred	3323	30	16	27	205	131	46503
SFESA (O+G)+HHpred	2932	162	132	170	573	460	45806
SFESA (O+G+M)+HHpred	2726	247	172	251	1053	667	45119
SFESA (O+G+M+S)+HHpred	3020	166	103	107	733	436	45670
SFESA (O)+CNFpred	3277	63	44	74	518	348	45973
SFESA (O+G)+CNFpred	2794	222	165	277	1219	1012	44608
SFESA (O+G+M)+CNFpred	2691	235	201	331	1197	1150	44492
SFESA (O+G+M+S)+CNFpred	3038	169	102	149	915	630	45294

Table IV-11. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (Dali as a reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the SALIGN dataset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3275	118	54	101	978	449	47348
SFESA (O+G)+PROMALS	2900	229	144	275	1625	819	46331
SFESA (O+G+M)+PROMALS	2801	261	185	301	1881	1003	45891
SFESA (O+G+M+S)+PROMALS	3113	215	94	126	1571	556	46648
SFESA (O)+HHpred	3323	24	25	24	165	182	48428
SFESA (O+G)+HHpred	2932	148	136	180	548	492	47735
SFESA (O+G+M)+HHpred	2726	231	174	265	976	711	47088
SFESA (O+G+M+S)+HHpred	3020	160	95	121	697	449	47629
SFESA (O)+CNFpred	3277	67	46	68	531	349	47895
SFESA (O+G)+CNFpred	2794	227	177	260	1253	1027	46495
SFESA (O+G+M)+CNFpred	2691	230	218	319	1206	1184	46385
SFESA (O+G+M+S)+CNFpred	3038	170	108	142	945	641	47189

Table IV-12. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (TMalign as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the SALIGN dataset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3275	113	50	110	902	411	44340
SFESA (O+G)+PROMALS	2900	233	128	287	1524	730	43399
SFESA (O+G+M)+PROMALS	2801	262	155	330	1774	836	43043
SFESA (O+G+M+S)+PROMALS	3113	210	87	138	1458	468	43727
SFESA (O)+HHpred	3323	25	22	26	173	168	45312
SFESA (O+G)+HHpred	2932	153	125	186	514	474	44665
SFESA (O+G+M)+HHpred	2726	227	162	281	919	653	44081
SFESA (O+G+M+S)+HHpred	3020	160	92	124	674	411	44568
SFESA (O)+CNFpred	3277	60	45	76	448	425	44780
SFESA (O+G)+CNFpred	2794	204	158	302	986	999	43668
SFESA (O+G+M)+CNFpred	2691	215	184	368	1054	1095	43504
SFESA (O+G+M+S)+CNFpred	3038	157	97	166	825	625	44203

Table IV-13. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (Matt as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the SALIGN dataset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3275	128	54	91	1062	432	46381
SFESA (O+G)+PROMALS	2900	260	132	256	1804	825	45246
SFESA (O+G+M)+PROMALS	2801	286	171	290	2059	1006	44810
SFESA (O+G+M+S)+PROMALS	3113	230	88	117	1720	536	45619
SFESA (O)+HHpred	3323	28	20	25	189	172	47514
SFESA (O+G)+HHpred	2932	165	124	175	596	525	46754
SFESA (O+G+M)+HHpred	2726	247	177	246	1060	797	46018
SFESA (O+G+M+S)+HHpred	3020	163	105	108	747	513	46615
SFESA (O)+CNFpred	3277	68	49	64	521	438	46916
SFESA (O+G)+CNFpred	2794	227	175	262	1258	1199	45418
SFESA (O+G+M)+CNFpred	2691	232	212	323	1207	1341	45327
SFESA (O+G+M+S)+CNFpred	3038	167	110	143	919	746	46210

Table IV-14. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred (Deepalign as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the SALIGN dataset.

Domain 1	Domain 2	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
1qfeA_	1rlcL_	14	0	0	2	0	0	206
3adkA_	1nstA_	6	2	0	2	25	3	130
1barA_	1xyfA_	11	1	0	2	1	0	110
1f6yA_	1reqA_	12	3	0	0	11	0	202
1dioA_	1qfeA_	14	0	0	0	0	0	208
1ez3A_	1fewA_	2	0	1	0	0	1	120
1dioA_	1d9eA_	18	0	0	3	0	0	209
1qfeA_	1dxeA_	14	1	0	0	16	0	164
1qfeA_	1aldA_	9	0	0	4	0	0	199
1nsjA_	1reqA_	11	0	0	1	0	0	188
1qfeA_	1nsjA_	11	2	2	0	15	19	142
1dxeA_	1f61A_	14	1	1	0	5	5	189
2asrA_	1occC_	4	0	0	0	0	0	126
1fioA_	3c98B_	6	0	0	0	0	0	177
1cb8A_	1qazA_	9	1	0	1	17	0	228
1qfeA_	1f6yA_	13	1	1	2	6	1	187
1nal1_	1qtwA_	13	0	1	2	0	7	190
1rpxA_	1dioA_	14	1	0	0	6	0	178
1dtyA_	1b9hA_	20	2	2	0	14	9	269
1czfA_	1tspA_	44	0	1	4	0	6	274
1bs2A_	1a8hA_	16	2	0	0	38	0	324
1dorB_	1nal1_	11	4	1	1	25	1	170
1piiA_	1dxeA_	11	1	0	2	12	1	176
1aw1B_	1dosA_	12	1	2	0	4	4	205
1fq0A_	1qrqA_	17	2	1	0	10	1	176
1nsjA_	2reqB_	13	0	0	1	0	0	185
5rubB_	1thfD_	12	2	0	1	10	0	171
1cqqA_	1cu1A_	12	0	0	4	0	0	135
1dxeA_	1rpxA_	12	1	1	0	6	1	181
1wkdA_	1qfeA_	14	2	1	0	19	11	175
2mnrA_	1dorB_	18	0	0	3	0	0	189
1ak1A_	1qgoA_	13	1	1	0	5	4	225
1de5A_	1a0cA_	19	2	0	3	13	1	306
1fq0A_	1pscA_	15	0	0	3	0	0	183
1daeA_	1nipA_	9	2	1	1	10	3	165
1a80A_	1fq0A_	14	0	1	0	0	5	177
1nal1_	1fq0A_	12	2	1	0	13	7	158
1qhtA_	1noyA_	22	0	0	0	0	0	282
1nal1_	1ezwA_	16	1	0	2	9	0	183

1dxeA_	1fq0A_	12	3	1	0	20	4	152
2ercA_	1eizA_	11	1	0	0	7	6	131
1nal1_	1aw1B_	8	2	1	3	26	9	150
1mpfA_	1a0tP_	20	0	0	0	0	0	282
2thiA_	1anfA_	25	1	0	0	6	0	308
1aw1B_	1f6yA_	10	1	2	3	6	5	174
1cu1A_	7lprA_	16	0	0	3	0	0	137
1fohB_	1qq2A_	9	0	0	1	0	0	140
1dfoA_	1b9hA_	19	5	1	0	40	7	268
1vjsA_	1bvzA_	25	4	0	0	33	1	339
1fq0A_	2tpsA_	13	1	0	0	8	0	183
1igsA_	1ad1B_	10	3	3	0	30	19	125
1udrB_	1qo0D_	7	1	0	0	9	4	102
1b9hA_	1qgnA_	17	4	0	3	23	0	274
4kbpB_	1qhwA_	14	0	3	1	0	17	230
2tpsA_	1nal1_	12	2	1	1	17	7	157
1nksF_	1nipA_	11	1	0	0	12	0	136
1ad1B_	1qr7B_	8	5	1	2	40	10	156
1noyA_	1tgoA_	22	2	0	1	9	0	173
1qgxA_	1fpkA_	14	3	0	2	16	0	223
1cj0A_	1b9hA_	18	4	0	1	26	1	286
1ftsA_	1daeA_	8	4	0	1	39	4	126
1eokA_	1d2kA_	14	1	2	0	9	6	219
1dfoA_	1bs0A_	21	4	1	1	28	3	298
1dfoA_	1elqA_	15	6	1	0	46	8	268
1qgnA_	1dfoA_	22	1	1	0	7	3	276
1dqyA_	1broA_	13	1	0	2	11	0	189
1fofA_	1skfA_	18	0	0	0	0	0	196
1d2fB_	1cj0A_	22	3	0	0	26	0	293
1nzyA_	1tyfB_	12	1	1	0	1	1	146
1qlwA_	1auoA_	11	2	0	1	9	6	171
1aw1B_	1fq0A_	13	0	2	0	0	7	169
1edtA_	1ctnA_	15	0	1	1	0	1	234
1chmA_	1xgmA_	17	0	0	0	0	0	209
1b5lA_	1evsA_	4	0	0	0	0	0	127
1c3qA_	1bx4A_	14	2	0	1	9	0	198
1aw1B_	1aldA_	6	1	1	3	6	17	182
1c0aA_	12asA_	16	1	0	0	13	0	251
1ad1B_	1dioA_	14	0	0	2	0	0	211
1b9hA_	1bs0A_	22	2	0	0	15	0	289
1b3uA_	1ibrB_	16	0	0	2	0	0	351
1judA_	1fezA_	13	0	0	2	0	0	138
1a9nA_	1d0bA_	7	0	0	0	0	0	123

1d9eA_	1fq0A_	13	2	0	1	20	2	162
1igsA_	1fq0A_	11	3	1	0	19	9	157
1amoA_	4nllA_	8	1	0	1	9	1	126
1plqA_	1dmlA_	13	4	1	2	32	11	187
1ad1B_	2tpsA_	13	0	1	0	0	7	168
1n5wC_	1fiqB_	19	1	0	0	7	0	267
1galA_	3coxA_	27	2	0	2	14	1	385
1cj0A_	1qgnA_	20	2	1	1	20	2	263
2hrvA_	1cu1A_	8	0	0	3	0	0	115
1b54A_	1qu4D_	15	1	0	0	12	0	184
1ahuA_	1f0xA_	30	0	0	0	0	0	419
1eutA_	1czvA_	14	1	0	0	6	0	128
1dubA_	1tyfB_	13	1	0	0	1	0	135
1rkdA_	1c3qA_	17	1	0	0	5	2	198
1ez0C_	1ad3A_	31	1	0	1	11	1	404
1vptA_	1eizA_	8	1	1	2	7	2	146
2tmdA_	1f6mA_	24	0	0	1	0	0	257
1ohvA_	2gsaA_	22	1	0	0	12	0	370
1cqqA_	2hrvA_	12	1	1	1	6	10	114
1d9eA_	1de5A_	11	0	0	3	0	0	206
1ad1B_	1qfeA_	12	2	0	1	21	0	174
1qu4D_	1sftB_	18	4	2	0	20	5	256
1aj6A_	1b62A_	10	2	0	1	16	0	146
1qrrA_	1a4uA_	11	2	1	0	20	3	182
1itgA_	1bcoA_	9	1	0	1	5	1	130
3adkA_	1gkyA_	12	3	0	0	22	3	127
1havA_	2hrvA_	11	2	0	1	5	0	128
1qu0A_	1c4rE_	13	1	1	0	5	2	155
3minA_	1mioB_	27	0	0	0	0	0	406
1bciA_	2isdA_	5	2	1	0	18	12	86
1taqA_	1a76A_	17	1	0	0	5	0	228
1ldcA_	1dorB_	17	2	0	1	17	4	213
1eizA_	2admA_	9	2	1	0	17	3	130
2pueA_	1bykA_	18	1	0	0	1	0	252
1chmA_	1matA_	20	0	0	0	0	0	220
1ciuA_	1vjsA_	31	3	1	0	28	2	345
1dcnB_	1c3cA_	19	0	0	0	0	0	380
1rptA_	1ihpA_	16	1	0	0	12	1	280
12asA_	1b8aA_	16	2	0	1	15	1	259
1smpA_	1kuhA_	7	0	0	0	0	0	122
1xgmA_	1a16A_	12	1	1	0	6	2	197
1d9eA_	1nsjA_	13	1	0	1	12	0	167
1dm0L_	1tcsA_	15	1	0	0	1	0	214

1qr7B_	1d9eA_	19	0	1	0	0	2	241
1gnwA_	1eemA_	10	0	1	0	0	1	190
1dpeA_	1rkmA_	32	1	0	0	9	0	460
1ar1B_	1fftB_	12	1	0	0	1	0	211
1ttpB_	1oasA_	18	0	1	0	0	2	291
1nal1_	1thfD_	13	1	1	1	1	8	186
1ag8A_	1ez0C_	31	3	0	0	10	5	425
1qgiA_	1chkA_	12	0	0	1	0	0	215
1cqxA_	1qfjA_	16	0	1	0	0	1	129
1dilA_	1eutA_	20	4	0	1	35	1	292
1ad1B_	1d9eA_	12	2	2	2	6	9	196
1d0bA_	1yrgA_	23	0	0	0	0	0	153
1c3cA_	1fuoA_	16	0	1	0	0	8	366
1dfjI_	1yrgA_	28	0	0	3	0	0	313
1tdjA_	1oasA_	19	0	0	0	0	0	294
1eemA_	1gsdB_	10	0	1	1	0	13	183
1havA_	1cqqA_	16	0	0	1	0	0	170
1uokA_	1ciuA_	29	1	0	1	9	0	394
1lrvA_	1b3uA_	16	0	0	0	0	0	214
1dgdA_	1ohvA_	22	2	0	0	15	1	382
1bw9A_	3mw9A_	23	1	0	0	8	0	318
1occC_	1fftC_	5	0	0	0	0	0	177
1xelA_	1db3A_	15	1	0	1	4	0	286
1bqgA_	1fhuA_	19	1	1	0	1	1	269
1matA_	1bn5A_	17	2	0	1	9	0	227
1ahnA_	1amoA_	10	0	0	0	0	0	145
1fblA_	1kuhA_	5	2	0	0	6	0	114
1dfjI_	1d0bA_	9	0	0	0	0	0	182
1daaA_	1et0A_	20	2	0	0	15	0	234
1imaA_	1qgxA_	17	2	0	2	10	0	241
1ciuA_	2aaaA_	23	3	1	0	19	11	409
1db3A_	1bxkA_	17	2	1	0	8	1	282
1whsB_	1ivyA_	8	0	0	0	0	0	147
1ovaA_	1sekA_	22	3	0	0	21	1	337
1eurA_	1sllA_	22	1	2	0	6	6	327
7ahlA_	1pvlA_	18	0	0	0	0	0	233
1gpmA_	1qdlB_	16	1	1	0	8	2	176
1whsB_	1cpyA_	9	0	0	0	0	0	142
1ipsB_	1dcsA_	14	1	0	0	1	0	243
1b5fA_	1fknA_	17	1	0	0	10	0	217
1darA_	1dpfA_	11	0	0	2	0	0	146
2rebA_	1cr2A_	10	3	2	0	13	23	147
1tlfC_	2pueA_	19	0	1	0	0	13	259

1dxyA_	1psdA_	22	0	0	0	0	0	311
1psdA_	1gdhA_	20	1	0	0	7	0	301
4tf4B_	1nbcA_	12	0	0	0	0	0	132
1lyaD_	1fknA_	19	0	1	0	0	1	228
1xffB_	1ct9A_	16	1	0	0	4	0	178
1fiqC_	1vlbA_	36	1	0	0	10	0	695
1ac5A_	1cpyA_	24	1	0	1	7	0	393
2shpA_	1yptA_	12	1	0	1	17	2	223
1iovA_	1ehiA_	19	1	0	0	12	0	289
4mhtA_	1dctA_	18	0	0	0	0	0	276
1reqA_	1cb7A_	11	0	0	0	0	0	129
1cpyA_	1ivyA_	24	2	1	0	21	1	356
2plcA_	2isdA_	15	0	1	0	0	4	207
1d2kA_	1ctnA_	24	1	0	0	2	0	353
1xgmA_	1matA_	18	0	1	0	0	1	229
1ag8A_	1euhA_	35	0	0	0	0	0	466
1whsA_	1cpyA_	14	1	0	1	13	0	224
1ahsB_	1bvp4_	11	0	0	0	0	0	125
1au1B_	1b5lA_	4	1	0	0	1	0	143
1agrE_	1emuA_	8	0	0	0	0	0	124
1atiA_	1qf6A_	25	1	1	0	6	1	349
1ahuA_	1diiA_	30	0	0	0	0	0	511
1hrdA_	1gtmA_	26	1	0	0	11	0	399
1ciyA_	1dlcA_	31	1	1	0	1	1	553
1kobB_	1a06A_	14	1	0	0	2	1	260
1froA_	1fa6B_	10	0	0	0	0	0	124
1etuA_	1darA_	13	0	1	0	0	8	144
1tcsA_	3rtjA_	19	0	0	0	0	0	240
1larA_	1ptyA_	16	1	0	1	4	0	272
1vlbA_	1n5wA_	11	0	0	1	1	1	155
1vlbA_	1fiqA_	9	1	0	0	1	0	149
1eepA_	1dorB_	19	1	0	0	13	0	223

Table IV-15. The alignment block-level and aligned position-level comparison for each alignment of SFESA (O+G+M+S) applied on PROMALS on the SALIGN dataset (Dali as a reference).

Methods On subsets	Reference-dependent (Q-score)		Reference-independent (TM-score)	
	SABmark_TWI	**SABmark_SUP**	**SABmark_TWI**	**SABmark_SUP**
PROMALS	46.2	71.1	0.413	0.583
SFESA (O)+PROMALS	47.3	71.3	**0.416**	0.585
SFESA (O+G)+PROMALS	48.0	71.8	**0.416**	0.585
SFESA (O+G+M)+PROMALS	47.9	71.9	**0.416**	0.586
SFESA (O+G+M+S)+PROMALS	**48.1**	**72.1**	**0.416**	**0.587**
HHpred	40.7	68.9	0.371	0.570
SFESA (O)+HHpred	40.6	69.0	0.371	0.570
SFESA (O+G)+HHpred	41.3	69.1	0.372	**0.571**
SFESA (O+G+M)+HHpred	**41.4**	**69.6**	0.372	**0.571**
SFESA (O+G+M+S)+HHpred	41.3	69.4	**0.373**	**0.571**
CNFpred	41.5	66.1	0.368	0.543
SFESA (O)+CNFpred	41.6	66.4	0.367	**0.545**
SFESA (O+G)+CNFpred	42.3	67.0	0.370	**0.545**
SFESA (O+G+M)+CNFpred	**42.4**	**67.4**	**0.371**	**0.545**
SFESA (O+G+M+S)+CNFpred	42.2	66.9	0.370	**0.545**

Table IV-16. Test on SABmark database. Columns 2-3 indicate the alignment Q-score based on their reference on two subsets of the SABmark benchmark: "twilight zone" and "superfamilies" respectively (Reference-dependent evaluation). Columns 4-5 indicate the average of query model's TM-score built by Modeller on two subsets of the SABmark benchmark: "twilight zone" and "superfamilies" respectively (Reference-independent evaluation). Bold indicates the best performance in the subsection. Bold with underscore indicates the overall best performance in one column.

Method/(p-value)	SABmark_twi	SABmark_sup
SFESA(O)+PROMALS	4.39e-04	0.15
SFESA(O+G)+PROMALS	3.07e-03	1.28e-03
SFESA(O+G+M)+PROMALS	2.35e-03	3.33e-04
SFESA(O+G+M+S)+PROMALS	2.30e-04	1.43e-06

Method/(p-value)	HHpred	HHpred
SFESA(O)+HHpred	0.18	0.48
SFESA(O+G)+HHpred	3.79e-02	0.099
SFESA(O+G+M)+HHpred	1.37e-02	1.41e-04
SFESA(O+G+M+S)+HHpred	1.11e-02	3.74e-05

Method/(p-value)	CNFpred	CNFpred
SFESA(O)+CNFpred	0.43	3.29e-02
SFESA(O+G)+CNFpred	0.068	4.25e-04
SFESA(O+G+M)+CNFpred	3.67e-02	2.38e-06
SFESA(O+G+M+S)+CNFpred	2.74e-02	1.57e-05

Table IV-17. Statistically significant Q-Score improvement of SFESA on PROMALS, HHpred and CNFpred (SABMARK as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on SABMARK dataset. P-values are calculated based on the paired Wilcoxon signed-rank. P-values below 0.05 are marked green and below 0.005 are marked pink.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	1629	31	16	68	242	74	13199
SFESA (O+G)+PROMALS	1367	75	53	249	558	329	12628
SFESA (O+G+M)+PROMALS	1344	79	57	264	562	321	12632
SFESA (O+G+M+S)+PROMALS	1512	61	32	139	432	180	12903
SFESA (O)+HHpred	1293	2	7	18	11	25	13479
SFESA (O+G)+HHpred	1137	36	23	124	150	79	13286
SFESA (O+G+M)+HHpred	1089	55	25	151	202	113	13200
SFESA (O+G+M+S)+HHpred	1202	35	15	68	154	60	13301
SFESA (O)+CNFpred	1152	15	17	39	131	100	13284
SFESA (O+G)+CNFpred	924	71	41	187	391	281	12843
SFESA (O+G+M)+CNFpred	903	73	39	208	370	238	12907
SFESA (O+G+M+S)+CNFpred	1035	49	23	116	283	176	13056

Table IV-18. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred. All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the SABMARK “twilight zone” subset.

Domain 1	Domain 2	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
1h97a_	1jboa_	6	0	1	0	0	3	103
1lrza1	1seta1	2	0	0	0	0	0	49
1c52__	1e29a_	6	0	0	0	0	0	54
1bjaa_	1jhga_	4	1	0	0	14	0	38
1nvma1	1oaia_	3	0	0	0	0	0	39
1g4da_	1nd9a_	4	1	0	0	13	0	13
1hcia4	1jnra1	3	0	0	0	0	0	61
1gjta_	1k1xa1	4	0	0	0	0	0	39
1erd__	2erl__	3	0	0	0	0	0	24
1ail__	1fyja_	2	0	0	0	0	0	42
1k99a_	1qrva_	4	0	0	0	0	0	54
1kx5d_	1n1jb_	3	0	1	0	0	1	61
1eexg_	1om2a_	3	0	0	1	0	0	0
1he1a_	2liga_	3	0	0	1	0	0	65
1mxra_	1noga_	5	0	0	0	0	0	73
1hzia_	1lki__	3	0	0	1	0	0	70
1f4la1	1ivsa2	7	0	0	0	0	0	86
1dv5a_	1l0ia_	4	0	0	0	0	0	60
1is2a1	1jqia1	6	0	0	0	0	0	84
1joya_	1nkd__	1	0	0	1	0	0	0
1ic8a2	1vpwa1	4	0	0	0	0	0	41
1a0aa_	1mdya_	2	0	0	0	0	0	40
1dqea_	2cbla1	5	0	0	0	0	0	54
1bkra_	1h67a_	5	0	0	0	0	0	71
1baza_	2cpga_	2	0	0	0	0	0	29
1k0ma1	1k3ya1	7	0	0	0	0	0	90
1bmta1	1khda1	4	0	0	0	0	0	53
1hs7a_	1lvfa_	3	0	0	0	0	0	76
1de4c1	2cbla2	3	0	1	0	0	17	40
1bea__	1fk5a_	4	0	0	1	0	0	47
1b4fa_	1cuk_2	4	1	0	0	9	0	46
1ed1a_	1mn8a_	4	0	0	1	0	0	56
1l9la_	1n69a_	4	0	0	0	0	0	30
1g7da_	1m2vb1	3	0	0	1	0	0	50
1h4ld_	1jkw_2	4	0	0	2	0	0	59
1d2zb_	1n3ka_	5	0	0	0	0	0	67
1jr3a1	1jr3d1	6	0	0	0	0	0	75
1ko9a1	1ngna_	6	0	0	0	0	0	87
1f0ya1	1n1ea1	4	0	0	1	0	0	43

1lf6a1	1qaza_	7	1	0	1	17	0	103
1pbwa_	1wer__	8	0	0	1	0	0	48
1h6ka1	1ld8a_	10	0	0	2	0	0	64
1lwba_	1poc__	5	0	0	0	0	0	28
1e79i_	1tbge_	1	0	0	0	0	0	0
1ft5a_	3caoa_	5	0	0	0	0	0	19
1ji1a1	1jmxa4	9	0	0	3	0	0	47
1e5ba_	1h6fa_	6	0	0	2	0	0	57
1f86a_	1qhoa2	4	0	0	4	0	0	58
1jzga_	1kzqa1	9	0	0	0	0	0	72
1dcea2	1qpxa2	5	2	1	0	17	2	46
1bhu__	1c01a_	4	0	0	1	0	0	45
1jz8a3	1xnaa_	10	0	1	0	0	1	83
1ahsa_	1flca1	6	0	0	3	0	0	51
1dmza_	1g6ga_	11	0	0	0	0	0	89
1n1ta1	2nlra_	12	0	0	4	0	0	90
1jz8a4	1n7oa3	11	1	1	2	5	9	117
1fqta_	1o7na1	12	0	0	0	0	0	82
1i1ja_	1vie__	6	0	0	0	0	0	41
1auua_	1p3ha_	3	0	0	3	0	0	30
1kwaa_	1mfga_	7	1	0	0	8	0	64
1d3ba_	1mgqa_	9	0	0	0	0	0	62
1i40a_	1o7ia_	6	1	0	2	7	0	48
1bfg__	1hcd__	10	2	0	0	14	1	78
1a8p_1	1n08a_	7	1	0	0	2	0	64
1flma_	1i0ra_	10	0	1	0	0	1	64
1befa_	1hava_	15	1	0	0	8	0	101
1eu1a1	1g8ka1	10	0	0	1	0	0	87
1k5db_	1mixa2	6	0	1	0	0	2	64
1ifc__	1qfta_	6	1	1	3	6	8	59
1ei5a2	1jmxa5	6	0	0	1	0	0	65
1e8ua_	1k32a2	14	1	0	8	8	0	36
1fwxa2	1k32a3	25	0	0	2	0	0	135
1g5aa1	1gjwa1	5	1	1	0	3	8	42
1goia1	1pina1	2	0	0	1	0	0	11
1i5pa2	1vmoa_	11	0	0	2	0	0	89
1daba_	1k4za_	18	0	0	2	0	0	29
1qrea_	3tdt__	22	0	0	3	0	0	100
1gy9a_	1ig0a1	9	0	0	1	0	0	65
1bdo__	1k8ma_	7	1	0	0	5	0	57
1euwa_	4ubpb_	5	0	0	0	0	0	27
1itua_	1kbla1	17	0	0	2	0	0	71
1hxha_	1ooea_	12	1	0	0	6	0	174

1gtea3	1m6ia2	9	0	0	1	0	0	83
1h16a_	1hk8a_	24	0	0	1	0	0	0
1kid__	1l5ja2	16	0	0	2	0	0	65
1a4ya_	1igra1	14	0	0	1	0	0	61
1on3a1	1tyfa_	9	1	1	0	1	9	96
1cdza_	1l0ba1	8	0	0	0	0	0	43
1oe4a_	3euga_	8	0	2	0	0	26	94
1fjgb_	1m2fa_	6	0	3	0	0	13	75
1ep3b2	1fdr_2	9	0	0	0	0	0	98
1k92a1	1n2ea_	9	1	1	0	8	6	95
1b6ra2	1iow_1	6	0	1	0	0	6	62
1poxa1	1pvda1	12	0	0	0	0	0	110
1oi2a1	1tuba1	8	0	0	2	0	0	86
1keka1	1qgda1	11	1	1	0	11	6	105
1e6ca_	1khta_	11	1	0	0	6	0	127
1eaf__	1l5aa1	9	0	0	0	0	0	94
1iiba_	1jf8a_	4	2	0	0	14	0	61
1d5ra2	1fpza_	14	0	0	0	0	0	126
1knga_	1qmha1	4	2	0	0	17	0	47
1keka3	1qgda3	9	1	0	0	4	0	87
1gyta1	1hjza_	9	1	0	0	5	0	107
1crza2	1ex2a_	7	1	0	1	4	0	57
1dmua_	1fiua_	14	0	0	0	0	0	66
1dt9a1	1j54a_	6	0	0	1	0	0	80
1a2za_	1lam_2	10	0	0	1	0	0	122
1c1da2	1lu9a2	5	0	0	1	0	0	68
1fsga_	1g2qa_	14	0	0	1	0	0	89
1atza_	1m1xb2	11	0	1	1	0	8	115
1af7_2	1nw3a_	12	0	0	0	0	0	86
1c4ka2	1elua_	21	1	0	1	21	0	180
1hm9a2	1i52a_	13	2	0	1	14	0	162
1ju3a2	3tgl__	14	0	1	1	0	1	52
1duvg1	1jfla2	7	1	0	0	7	0	63
1j6na_	1qopb_	18	2	0	0	9	0	223
1k2yx2	1kfia2	6	0	1	0	0	1	55
1n2za_	1qgoa_	12	1	0	2	6	0	70
2liv__	8abp__	16	1	2	5	4	6	98
1a8e__	3thia_	17	0	0	0	0	0	59
1hnja2	1ox0a2	9	0	0	0	0	0	98
1aln_1	1aln_2	9	0	0	0	0	0	59
1i0va_	1lnia_	4	0	1	0	0	6	40
153l__	1qgia_	5	0	0	2	0	0	85
1gx3a_	2cb5a_	13	0	0	1	0	0	0

1fr2b_	1ql0a_	6	0	0	2	0	0	29
1d9na_	1gcca_	3	0	0	1	0	0	21
1guqa2	1l9va1	8	0	1	0	0	6	58
1pkp_1	1ueka1	6	0	0	1	0	0	53
1f52a1	1l5pa_	6	0	0	0	0	0	28
1c0pa2	1ju2a2	6	0	0	0	0	0	68
1gy7a_	1oaca2	4	0	0	1	0	0	51
1hdmb2	1jfma_	6	1	0	0	2	0	57
1jnda2	1pina2	5	0	0	1	0	0	16
1kw3b1	1lqpa_	10	0	0	0	0	0	40
1lo7a_	1mkaa_	5	1	0	1	1	0	92
1brwa3	1n62b1	7	0	0	1	0	0	42
1buoa_	1t1da_	6	1	0	1	6	0	43
1kn0a_	1qu6a2	2	1	1	0	9	5	47
1dtja_	1k1ga_	6	0	0	0	0	0	53
1egaa2	1fjgc1	5	0	0	0	0	0	68
1h0hb_	1nxia_	5	1	0	0	16	0	14
1fjgd_	1h3fa2	6	1	0	1	5	0	49
1f7ua3	1iq0a3	8	0	0	0	0	0	58
1hc7a3	1nj8a2	5	0	0	0	0	0	49
1n6za_	1xxaa_	6	0	0	0	0	0	62
1jj2f_	1tuba2	5	0	0	0	0	0	56
1gyxa_	1otfa_	4	0	0	0	0	0	47
1cf2o2	1p1ja2	5	1	0	0	10	0	47
1mo9a3	3grs_3	8	1	0	0	6	0	86
1l2ma_	1m55a_	7	0	0	1	0	0	71
1ast__	1jaka2	9	0	2	0	0	14	54
1lkka_	2plda_	5	1	0	1	1	0	68
1dq3a3	1dq3a4	5	1	0	0	3	0	65
1b66a_	1uox_1	5	0	0	0	0	0	77
12asa_	1seta2	12	0	1	2	0	2	132
1iyka2	1qsma_	8	1	0	0	11	0	112
1d4xg_	1m4ja_	7	0	0	3	0	0	88
1h8ma_	1n9la_	3	2	0	0	19	0	38
1hzta_	1k2ea_	13	0	0	0	0	0	99
1byqa_	1ei1a2	11	1	0	0	9	0	108
1g61a_	1h70a_	10	0	0	6	0	0	91
1f46a_	1ytba1	6	1	0	1	4	0	66
1b77a2	1iz5a2	8	2	0	0	15	0	70
1rl6a1	1rl6a2	8	0	0	0	0	0	59
1ckma2	1kbla3	11	0	0	1	0	0	61
1f0xa2	1n62c2	10	1	0	1	6	0	107
1f7la_	1qr0a1	7	1	0	0	1	0	53

1gdoa_	1ofda3	16	0	0	1	0	0	68
1e5da2	1k07a_	16	1	0	0	9	0	164
1ii7a_	1nnwa_	15	0	0	2	0	0	141
1b8pa2	1ceqa2	11	1	0	0	4	1	130
1a26_2	1ikpa2	10	0	0	1	0	0	80
1chua3	1kssa3	8	0	0	0	0	0	72
1g1ta1	1prtb2	5	1	0	1	9	0	42
1hhsa_	1l3sa2	17	0	0	4	0	0	0
1feza_	1k1ea_	11	0	0	0	0	0	108
1ikpa3	1k3ka_	7	0	0	1	0	0	0
1p4ta_	2mpra_	9	0	0	1	0	0	15
1i2ua_	1lpba2	2	0	0	1	0	0	6
1kbaa_	3ebx__	5	0	0	0	0	0	49
1tocr1	1tocr2	2	1	0	0	8	0	17
1ewsa_	1ijva_	3	0	0	0	0	0	23
1hkya_	1i8na_	5	0	2	1	0	11	42
1i71a_	1l6ja4	4	0	0	0	0	0	9
1iw4a_	4sgbi_	3	0	0	0	0	0	10
1aoca_	1hcnb_	4	0	0	1	0	0	30
1ly2a1	1quba1	5	0	0	0	0	0	48
1exta1	1oqdk_	3	0	0	0	0	0	9
1o9aa2	1tpg_2	5	0	0	0	0	0	40
1iuaa_	2hipa_	5	0	0	0	0	0	37
1fu9a_	2glia2	1	0	0	0	0	0	19
1d66a1	1zmec1	2	0	0	0	0	0	27
1fjgn_	1k3xa3	3	0	0	1	0	0	0
1dsva_	1eska_	2	0	0	0	0	0	17
1yua_2	1zaka2	1	0	0	0	0	0	0
1fbva4	1n87a_	5	0	0	0	0	0	42
1jjda_	4mt2__	1	0	0	0	0	0	0
1e53a_	1kbea_	3	0	0	0	0	0	36
1ezva2	1ezvb1	11	0	0	1	0	0	131
1gjja1	1h1js_	2	0	0	0	0	0	35
1jyoa_	1l2wi_	2	0	0	2	0	0	0
1k82a1	1mu5a1	4	0	0	0	0	0	69
1jaja_	1knya2	6	0	0	1	0	0	54
1c0va_	1fftb2	1	0	0	2	0	0	0
1kqfc_	1nekd_	4	0	0	0	0	0	72
1ocrk_	1ocrm_	2	0	0	0	0	0	16
1dzla_	1hx6a1	9	0	0	2	0	0	76
1go3e2	1owta_	4	1	0	0	5	0	36

Table IV-19. The alignment block-level and aligned position-level comparison for each alignment of SFESA (O+G+M+S) applied on PROMALS on the SABMARK twilight dataset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	3959	38	35	98	287	202	40063
SFESA (O+G)+PROMALS	3555	131	95	349	789	491	39272
SFESA (O+G+M)+PROMALS	3510	144	109	367	879	519	39154
SFESA (O+G+M+S)+PROMALS	3781	109	49	191	755	275	39522
SFESA (O)+HHpred	3748	10	11	43	79	79	40394
SFESA (O+G)+HHpred	3467	73	48	224	256	167	40129
SFESA (O+G+M)+HHpred	3335	134	67	276	519	228	39805
SFESA (O+G+M+S)+HHpred	3591	78	27	116	323	94	40135
SFESA (O)+CNFpred	3385	35	29	71	294	182	40076
SFESA (O+G)+CNFpred	2880	185	97	358	834	497	39221
SFESA (O+G+M)+CNFpred	2857	195	97	371	863	420	39269
SFESA (O+G+M+S)+CNFpred	3146	137	48	189	676	297	39579

Table IV-20. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred. All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the SABMARK “superfamily” subset.

Domain 1	Domain 2	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
1b8da_	1la6a_	6	0	1	0	0	2	114
1gtea1	1nekb1	5	1	0	0	11	0	57
1gh6a_	1xbl__	3	0	0	0	0	0	42
1eiya1	1ivsa1	3	0	0	0	0	0	53
1ap6a1	1bsma1	2	0	0	0	0	0	65
1ql3a_	1ycc__	6	0	0	0	0	0	91
1e3oc1	1ig7a_	3	0	0	0	0	0	45
1mgta1	1sfe_1	4	0	0	0	0	0	75
1c20a_	1ig6a_	6	0	0	0	0	0	59
1jgsa_	1p4xa2	7	0	0	0	0	0	103
1a04a1	1fsea_	4	0	0	0	0	0	58
1f4ia_	1otra_	3	0	0	0	0	0	39
1fjgm_	1k3xa1	3	0	0	0	0	0	53
1d4ua1	1g4da_	4	0	0	1	0	0	50
1cuna1	2spca_	2	0	0	0	0	0	51
1kf6a1	1qlaa1	8	0	0	0	0	0	107
1gab__	1gjta_	3	0	0	0	0	0	35
1erd__	1hd6a_	3	0	0	0	0	0	28
1h4ra1	1mixa1	4	0	0	0	0	0	74
1a32__	1fyja_	2	0	0	0	0	0	38
1i11a_	1k99a_	4	0	0	0	0	0	52
1kx5c_	1n1jb_	4	0	0	0	0	0	64
1cpq__	1jafa_	4	0	0	0	0	0	110
1cgme_	1ei7a_	9	0	0	0	0	0	84
1is2a2	1jqia1	5	0	0	2	0	0	77
1o9ra_	1qgha_	5	0	0	0	0	0	129
1i1rb_	1m4ra_	4	1	0	0	19	0	48
1ffya1	1li5a1	4	0	0	0	0	0	43
1af8__	1dnya_	4	0	0	0	0	0	50
1d1da1	1qrjb1	4	0	0	0	0	0	61
1fts_1	1ls1a1	4	0	0	0	0	0	59
1eqfa1	1eqfa2	6	0	0	0	0	0	105
1lmb3_	1ner__	4	0	0	0	0	0	55
1an4a_	1mdya_	2	0	0	0	0	0	28
1m31a_	1wdcc_	6	0	0	0	0	0	60
1aoa_2	1mb8a2	5	0	0	0	0	0	87
1cmba_	1irqa_	1	0	1	0	0	10	31
1duga1	1f2ea1	5	0	0	0	0	0	109
1o17a1	2tpt_1	4	0	0	0	0	0	64

1ez3a_	1lvfa_	2	0	0	1	0	0	77
1bea__	1l6ha_	4	0	0	1	0	0	43
1owfa_	1owfb_	8	0	0	0	0	0	62
1fo4a1	1hlra1	4	0	0	0	0	0	61
1b4fa_	1dxsa_	5	0	0	0	0	0	45
1cuk_2	1dgsa1	4	0	0	0	0	0	62
1a77_1	1bgxt1	4	0	0	0	0	0	27
1a6s__	1mn8a_	4	0	0	0	0	0	49
1l9la_	1m12a_	4	0	0	1	0	0	60
1axn__	1n00a_	18	0	0	0	0	0	292
1e79a1	1fx0a1	6	0	0	0	0	0	75
1em9a_	2eiaa2	5	0	0	0	0	0	102
1aisb2	1vola1	5	0	0	0	0	0	87
1d2za_	1dgna_	4	1	0	1	10	2	61
1jr3a1	1jr3d1	6	0	0	0	0	0	75
1f5xa_	1ki1b1	10	0	0	0	0	0	94
1cmza_	1omwa1	7	0	0	0	0	0	106
1mn2__	1mwva1	18	0	0	0	0	0	160
1mpga1	2abk__	9	0	0	0	0	0	132
1np7a1	1qnf_1	10	0	0	0	0	0	227
1mv8a1	1pgja1	3	1	0	0	10	0	64
1fp3a_	1g9ga_	21	0	0	3	0	0	141
1j0ma1	1n7oa1	18	0	0	0	0	0	294
1c3d__	1ld8b_	15	0	0	0	0	0	197
1a59__	1ioma_	19	0	0	0	0	0	299
1e9xa_	1n6ba_	24	1	0	0	16	0	213
1qhba_	1vns__	17	0	0	0	0	0	0
1pbwa_	1wer__	8	0	0	1	0	0	48
1h6ka1	1n8va_	5	0	0	0	0	0	43
1awcb_	1ycsb1	8	0	0	0	0	0	106
1a17__	1iyga_	7	0	0	0	0	0	94
1dvpa1	1elka_	8	0	0	0	0	0	82
1n83a_	2prga_	11	1	0	0	5	0	202
1ak0__	1ca1_1	9	1	0	0	17	0	126
1kxpd2	1n5ua3	9	0	0	0	0	0	163
1hy0a_	1jswa_	18	0	0	0	0	0	271
1g4ia_	1lfja_	5	0	0	0	0	0	110
1dxrc_	3cyr__	4	0	0	0	0	0	0
1b88a_	1jmaa_	5	1	0	2	11	0	73
1cfb_2	1n6va2	4	2	0	1	16	0	51
1jz8a1	1jz8a2	6	0	0	1	0	0	64
1f13a2	1f13a3	7	1	0	0	3	0	74
1l3wa3	1l3wa4	7	0	0	0	0	0	91

1akp__	1j48a_	9	1	0	0	1	0	103
1do5a_	1ej8a_	9	0	0	0	0	0	89
1gyva_	1kyfa1	9	0	0	0	0	0	90
1cwva2	1cwva4	6	0	0	2	0	0	44
1e5ba_	1g1ka_	6	0	0	1	0	0	60
1amx__	1n67a2	8	3	1	0	20	2	69
1bvoa_	1h6fa_	9	1	0	2	4	0	81
1pama2	1qhoa2	6	1	0	0	1	0	93
1eo9a_	1eo9b_	11	0	0	0	0	0	120
1aoza2	1kv7a2	11	1	0	0	9	0	97
1dsya_	1k5wa_	7	1	0	1	1	0	109
1k2fa_	1lb6a_	5	0	0	4	0	0	92
1bhu__	1h4ax1	5	0	0	1	0	0	49
1pgs_1	1phm_2	7	0	0	1	0	0	72
1hx6a2	1ruxa2	9	0	0	1	0	0	92
1gmea_	1shsa_	8	0	0	0	0	0	61
1kgya_	1of4a_	12	0	0	0	0	0	89
1ahsa_	1qhda2	8	1	0	1	10	0	56
1h7za_	1kkea1	4	0	0	0	0	0	10
1kxga_	2tnfa_	10	1	0	0	12	0	108
1f1sa3	1j0ma2	6	0	1	2	0	1	54
1g6ga_	1gxca_	11	0	0	0	0	0	91
1kit_2	3btaa1	10	0	0	2	0	0	111
1fqta_	1rfs__	11	1	0	2	8	1	71
1bia_2	1igqa_	3	0	1	0	0	3	29
1jo8a_	1neb__	6	0	0	0	0	0	48
1vie__	2ahjb_	5	1	0	1	2	0	43
1jj2s_	1m1ga2	6	0	0	0	0	0	43
1aono_	1g31a_	4	0	0	0	0	0	58
1ihja_	1ntea_	6	0	0	0	0	0	76
1d3ba_	1h641_	9	0	0	0	0	0	65
1enfa1	3chbd_	5	1	0	1	6	0	53
1br9__	1jb3a_	9	0	1	1	0	4	73
1fgua2	1gm5a2	8	0	0	0	0	0	70
1e9ga_	2prd__	12	0	0	0	0	0	141
1guta_	1h9ma2	6	0	0	0	0	0	59
1nuna_	1qqla_	12	0	0	0	0	0	121
1ggpb1	1m2tb2	12	1	0	0	5	0	100
1epwa2	1eyla_	12	0	0	0	0	0	128
1dfca2	1hcd__	11	0	0	0	0	0	83
1f60a1	1n0ua1	8	1	0	0	2	0	72
1exma2	1f60a2	8	0	0	0	0	0	55
1flma_	1i0ra_	10	0	1	0	0	1	64

1ekbb_	1rfna_	20	0	0	0	0	0	210
1e79a2	1fx0b2	5	0	0	1	0	0	50
1idaa_	1lf2a_	8	2	0	1	8	0	66
1ffya2	1h3na2	12	0	0	2	0	0	117
1cz4a1	1h0ha1	9	1	0	0	10	0	54
1h4ra2	1mixa2	7	1	0	0	1	0	81
1a49a1	1e0ta1	10	0	0	0	0	0	71
1ggla_	1koia_	5	1	0	3	10	0	70
1a33__	1lopa_	19	0	0	0	0	0	124
1itva_	1pex__	16	1	0	0	9	0	158
1eur__	2bat__	13	0	2	6	0	11	104
1nr0a2	1p22a2	23	2	2	2	15	3	220
1e43a1	1ji2a2	4	3	0	1	15	1	59
1o6wa2	1pina1	3	0	0	0	0	0	26
1jd0a_	1kopa_	14	0	1	0	0	2	178
1ciy_2	1i5pa2	6	0	1	3	0	10	99
1jpc__	1kj1d_	12	0	0	0	0	0	77
1k5ca_	1qcxa_	30	2	0	4	7	0	174
1lxa__	1qrea_	25	0	0	2	0	0	0
1ep0a_	2phla1	14	1	1	0	2	10	68
1gp6a_	1odma_	16	0	0	1	0	0	125
1i5za2	1rgs_2	11	0	0	0	0	0	99
1dd2a_	1htp__	6	0	0	2	3	3	51
1dv1a1	1e2wa2	2	0	0	3	0	0	34
1gpr__	2gpr__	15	0	0	0	0	0	132
1e9ya1	1ejxb_	7	0	0	0	0	0	84
1euwa_	1ogha_	13	0	0	0	0	0	94
1lyxa_	1n55a_	14	0	0	1	0	0	200
1a53__	1eixa_	10	2	1	1	15	15	145
1ep3a_	1oyb__	19	0	0	0	0	0	160
1j96a_	1lqaa_	20	1	0	0	5	0	237
1iexa1	7taa_2	15	0	0	1	0	0	139
1bf6a_	1k6wa2	20	0	0	2	0	0	206
1n8fa_	1o0ya_	12	1	0	3	11	0	122
1jpma1	2mnr_1	17	0	0	0	0	0	184
1dxea_	1f8ma_	17	0	0	0	0	0	157
1a0ca_	1qtwa_	13	1	0	1	10	6	227
1ezwa_	1lucb_	22	0	0	1	0	0	164
2plc__	2ptd__	15	0	0	0	0	0	101
1ccwb_	7reqa1	16	0	1	1	0	5	269
1kewa_	1qg6a_	10	3	1	0	11	2	134
1feca2	1lvl_2	9	0	0	1	0	0	103
1dysa_	1tml__	18	1	0	0	12	2	176

1h16a_	1hk8a_	24	0	0	1	0	0	0
1fqva2	1io0a_	10	0	0	1	0	0	130
1a9na_	1p9ag_	14	0	0	0	0	0	92
1nzya_	1on3a1	11	1	1	0	5	12	121
1dgtb3	1l0ba1	7	0	0	0	0	0	39
1c2ya_	1ejba_	9	0	0	0	0	0	131
1laue_	1oe4a_	13	0	2	0	0	22	111
1m2fa_	1tmy__	10	0	0	0	0	0	84
1e5da1	2fcr__	9	1	0	0	7	0	118
1bmta2	7reqb2	10	0	0	0	0	0	97
1esc__	1k7ca_	15	1	0	0	1	0	88
1gqoa_	1j2ya_	9	0	0	0	0	0	132
1pe0a_	1qdlb_	12	1	0	0	8	0	103
1que_2	2cnd_2	11	0	0	0	0	0	102
1coza_	1f7ua2	8	0	1	0	0	1	97
1j20a1	1jmva_	7	0	1	1	0	5	88
1dnpa2	1qnf_2	10	0	0	0	0	0	131
1gsoa2	2hgsa1	6	1	0	2	4	0	68
1d4oa_	1m2ka_	11	0	0	1	0	0	118
1ofua1	1tuba1	12	1	0	0	8	0	153
1im5a_	1yaca_	12	0	0	0	0	0	135
1keka2	1ovma2	12	0	0	0	0	0	112
1f60a3	1jj7a_	8	0	0	2	0	0	46
1p8ja2	1thm__	12	2	2	0	8	9	223
1c3pa_	1d3va_	12	1	1	0	4	13	144
1eaf__	1nocb_	13	0	0	1	0	0	155
1jf8a_	1phr__	7	0	0	0	0	0	119
1d5ra2	1lara2	11	0	0	0	0	0	126
1e0ca1	1rhs_1	10	0	0	0	0	0	66
1a8l_2	1erv__	8	0	0	0	0	0	83
1dtwb2	1itza3	10	0	0	0	0	0	107
1a3wa3	1a49a3	9	0	0	0	0	0	86
1gyta1	1hjza_	9	1	0	0	5	0	107
1hc7a1	1nj8a1	7	0	0	1	0	0	123
1f1za2	1fiua_	7	0	0	1	0	0	55
1ekja_	1g5ca_	10	0	0	0	0	0	100
1bu6o1	1czan2	7	0	0	3	0	0	44
1j54a_	1jl1a_	8	0	0	1	0	0	33
1fjgk_	1ilya_	4	1	1	0	7	3	37
1b8oa_	1k9sa_	14	1	0	0	15	0	155
1loka_	1m4la_	11	0	0	1	0	0	137
1c1da2	1nyta2	7	0	0	2	0	0	70
1h2ea_	3pgm__	11	0	0	1	0	0	161

1i5ea_	1qb7a_	14	0	1	0	0	3	124
1ijba_	1m1xb2	13	0	0	0	0	0	147
1jg1a_	1jqea_	10	0	0	0	0	0	109
1jf9a_	1tpla_	21	2	0	1	13	0	201
1gx4a_	1hv9a2	12	1	2	4	9	10	115
1ju3a2	1ku0a_	13	0	0	1	0	0	116
1df7a_	1vdra_	12	0	0	1	0	0	133
1o14a_	1rkd__	17	3	0	0	19	0	137
1e19a_	1gs5a_	21	0	0	1	0	0	147
1e4bp_	1k0wa_	14	0	1	0	0	1	158
1ed8a_	1k7ha_	23	0	1	1	0	3	288
1lwda_	1xaa__	19	1	0	1	1	0	205
1a1s_2	1ml4a2	10	0	0	0	0	0	143
1b74a1	1b74a2	6	1	0	1	3	2	55
1j6na_	1tdj_1	19	0	0	0	0	0	266
1c7qa_	1iata_	24	0	0	2	0	0	236
1g8ka2	1tmo_2	29	0	1	1	0	4	201
1a4sa_	1ez0a_	32	1	1	0	19	2	340
1k2yx1	3pmga3	9	0	0	0	0	0	56
16pk__	1hdia_	31	0	0	1	0	0	350
1f0ka_	1jixa_	17	0	0	4	0	0	0
1nnsa_	4pgaa_	23	0	0	0	0	0	289
1hrka_	1qgoa_	14	0	1	0	0	2	195
1n2za_	1psza_	16	0	0	0	0	0	137
1jyea_	2liv__	18	1	0	1	10	0	93
1amf__	1wdna_	18	0	0	2	0	0	105
1hzpa1	1mzja2	11	0	0	0	0	0	91
1aln_2	1uaqa_	5	1	1	0	4	6	57
1a2pa_	1i0va_	6	0	0	1	0	0	47
1gd6a_	1qgia_	5	0	0	0	0	0	77
1qmya_	2cb5a_	9	0	1	1	0	6	71
1e7la2	1fr2b_	5	0	0	0	0	0	21
1agi__	1rnfa_	10	0	0	0	0	0	88
1j9oa_	1qg7a_	4	0	0	0	0	0	37
1e0ba_	1knaa_	4	1	0	0	2	0	44
1kjka_	1qk9a_	3	1	0	0	4	0	30
1jj2l_	1n88a_	3	0	0	2	0	0	51
1guqa2	1kpf__	7	0	0	0	0	0	67
1ei1a1	1kkha1	10	0	0	1	0	0	80
1l7ya_	1mg8a_	4	1	0	0	6	0	60
1d4ba_	1f2ri_	7	0	0	0	0	0	46
1f0za_	1fm0d_	5	0	0	1	0	0	44
1hlra2	1krha3	5	0	0	1	0	0	67

1bmlc3	1qqra_	7	0	0	0	0	0	43
1an8_2	3tss_2	8	0	0	0	0	0	96
1c0pa2	1mxta2	3	0	1	1	0	17	45
1cewi_	1mola_	5	0	0	0	0	0	69
1ksia2	1oaca2	6	0	0	0	0	0	69
1gy6a_	1ocva_	7	2	1	1	31	7	63
1jfma_	3frua2	12	0	0	0	0	0	131
1c4zd_	1j7da_	7	1	0	0	7	0	93
1dzoa_	1oqva_	5	0	0	0	0	0	33
1j6ya_	1jnsa_	8	0	0	0	0	0	76
1goia3	1hjxa2	7	0	0	0	0	0	60
1ecsa_	1lqpa_	9	0	0	0	0	0	100
1c8ua2	1lo7a_	6	1	0	0	9	0	74
1csei_	1lw6i_	7	0	0	0	0	0	57
1fo4a3	1hlra3	9	0	0	0	0	0	96
1qapa2	1qpoa2	8	0	0	0	0	0	101
1t1da_	3kvt__	8	0	0	0	0	0	74
1efub2	1efub4	5	0	0	0	0	0	56
1bsma2	1ma1a2	8	0	0	0	0	0	111
1di2a_	1kn0a_	3	1	0	0	10	0	46
1k1ga_	2fmr__	5	0	0	0	0	0	51
1hh2p2	1k0ra2	5	0	0	0	0	0	74
1onea2	2mnr_2	7	0	0	2	0	0	93
1jnrb_	7fd1a_	8	0	0	0	0	0	41
1aye_2	1jqga2	7	0	0	0	0	0	77
1i1ga2	1lq9a_	3	1	1	1	7	1	42
1nzaa_	1p1la_	9	0	0	0	0	0	77
1ehwa_	1nhkl_	9	0	0	0	0	0	131
1owxa_	1qm9a2	6	0	0	0	0	0	46
1dar_4	1n0ua4	4	0	0	2	0	0	35
1aw0__	1fe0a_	6	0	0	0	0	0	63
1phza1	1psda3	4	2	0	0	21	0	41
1h72c2	1kvka2	8	0	1	0	0	1	107
1e6yb2	1hbnb2	12	0	0	0	0	0	49
1dj0a1	1k8wa4	8	0	0	0	0	0	55
1dm9a_	1jh3a_	5	0	0	1	0	0	57
1eh1a_	1is1a_	10	0	0	0	0	0	182
1bwvs_	1gk8i_	6	0	0	0	0	0	66
1dzfa2	1eika_	4	0	0	0	0	0	14
1qd9a_	1ufya_	4	0	1	1	0	8	61
1ofua2	1tubb2	6	1	0	0	9	0	70
1ck9a_	1e7ka_	8	0	0	0	0	0	83
1hfoa_	1otga_	6	0	0	0	0	0	96

1dih_2	1mb4a2	5	0	0	1	0	0	62
1dxla3	3lada3	8	0	0	0	0	0	123
1fo4a4	1jroa3	8	0	0	0	0	0	81
1l2ma_	1tbd__	5	1	0	0	1	0	78
1nox__	1vfra_	12	0	0	0	0	0	166
1c7ka_	1eb6a_	5	0	0	1	0	0	66
1jaka2	1qba_4	10	0	0	0	0	0	69
1a81a2	1mil__	9	1	0	0	5	1	79
1k1ca_	1pch__	7	0	0	0	0	0	80
1dq3a3	1dq3a4	5	1	0	0	3	0	65
1b66a_	1b9la_	6	0	0	0	0	0	71
12asa_	1jjca_	14	1	0	0	5	0	152
1iyka2	1ufha_	8	1	0	0	3	0	121
1jhwa3	1m4ja_	5	0	2	0	0	9	63
1acf__	1ypra_	10	0	0	1	0	0	108
1mc0a2	1mkma2	6	2	1	0	19	5	80
1ew0a_	1ll8a_	8	0	0	0	0	0	87
1a6ja_	1hynp_	12	0	0	0	0	0	103
1k2ea_	1ktga_	9	1	0	1	3	1	106
1bkpa_	1tis__	17	0	0	0	0	0	236
1cxya_	1kbia2	7	0	0	0	0	0	58
1byqa_	1l0oa_	6	0	0	0	0	0	76
1iqqa_	1ucaa_	12	0	0	0	0	0	164
1g61a_	1jdw__	14	4	0	1	26	1	85
1az9_2	1o0xa_	16	0	0	0	0	0	190
1f52a2	1qh4a2	11	0	0	1	0	0	117
1ko9a2	1ytba1	7	1	0	0	6	0	54
1kfia4	3pmga4	10	0	0	0	0	0	71
1icxa_	1jssa_	7	0	0	3	0	0	116
1mxa_3	1qm4a2	3	0	1	1	0	9	43
1iz5a1	2pola3	8	1	1	0	7	2	89
1fo4a5	1n62b2	32	1	0	2	2	0	441
1ckv__	1g10a_	10	0	0	0	0	0	41
1jj2e2	1rl6a1	6	1	0	0	4	0	54
1iow_2	1kbla3	12	0	0	1	0	0	113
1blxa_	1j7la_	15	0	0	0	0	0	126
1diqa2	1f0xa2	14	1	0	0	7	0	153
1f7la_	1qr0a2	7	0	0	1	0	0	65
1gph12	1j2pa_	11	0	0	2	0	0	89
1e5da2	1jjea_	16	1	0	1	11	0	163
1jk7a_	1utea_	14	1	0	1	6	0	117
1b8pa2	1ez4a2	11	0	0	1	3	3	155
1ggpa_	1qi7a_	15	0	0	0	0	0	191

1giqa1	1qs1a2	13	0	0	0	0	0	169
1jnra3	1neka3	6	0	0	0	0	0	97
1jwib_	1jzna_	11	0	0	0	0	0	95
1fzcb1	1fzda_	11	0	0	0	0	0	158
1k9oi_	1lj5a_	22	0	0	1	0	0	192
1ei5a3	1es5a_	15	2	0	0	15	0	129
1cf9a2	1gwea_	25	0	0	0	0	0	440
1jqia2	3mdda2	14	0	0	0	0	0	202
1lbva_	2hhma_	20	0	0	0	0	0	195
1jiha_	1tgoa2	16	0	0	4	0	0	67
1daaa_	1iyea_	19	1	0	0	9	0	230
1qq5a_	1zrn__	14	0	0	0	0	0	182
1oaoa_	1oaoc_	13	0	0	2	0	0	0
1a87__	1cola_	10	0	0	0	0	0	139
1ciy_3	1i5pa3	9	0	0	0	0	0	182
1f16a_	1k3ka_	9	0	0	0	0	0	78
1lgha_	1lghb_	1	0	0	0	0	0	0
1mm4a_	1qj8a_	8	0	0	0	0	0	79
1by5a_	2mpra_	18	0	0	5	0	0	76
1mmc__	9wgaa2	3	0	0	0	0	0	23
1dl0a_	1eit__	2	0	0	0	0	0	27
1jxca_	1npia_	3	0	0	0	0	0	30
1igra3	1ivoa4	1	0	0	1	0	0	0
1imt_1	1lpba1	2	0	0	0	0	0	15
1cvua2	1urk_1	4	0	0	0	0	0	18
1dec__	1skz_2	0	0	0	0	0	0	17
1hc9a_	3ebx__	5	0	0	0	0	0	49
1aapa_	1bik_1	3	0	0	0	0	0	53
1d6ba_	1kj6a_	2	0	1	0	0	2	15
1bhta1	1i8na_	6	0	0	1	0	0	60
1d2ja_	1k7ba_	2	0	0	0	0	0	32
1bhp__	1ejga_	4	0	0	0	0	0	45
1h8pa2	2hpqp_	2	0	0	0	0	0	26
1pce__	1sgpi_	4	0	0	0	0	0	36
1hi7a_	2pspa2	3	0	0	0	0	0	41
1jpya_	1lxia_	5	0	1	0	0	3	38
1g40a2	1gkna1	5	0	0	0	0	0	37
1n4ya_	2ech__	4	0	0	0	0	0	3
1eaic_	1hx2a_	4	0	0	0	0	0	42
1jmab1	1oqek_	2	0	0	0	0	0	11
1o9aa2	1tpg_2	5	0	0	0	0	0	40
1hpi__	2hipa_	5	0	0	0	0	0	49
2glia4	5znf__	3	0	0	0	0	0	23

1d66a1	1zmec1	2	0	0	0	0	0	27
1ibia2	1lata_	3	0	0	0	0	0	26
1a6bb_	1dsva_	2	0	0	0	0	0	17
1aky_2	1zin_2	4	0	0	0	0	0	14
1i50i1	1yua_1	1	0	0	2	0	0	24
1h7va_	1rb9__	7	0	0	0	0	0	27
1jj22_	1nvha_	5	0	0	0	0	0	9
1iyma_	1jm7a_	5	0	0	2	0	0	0
1dmc__	1jjda_	0	0	0	2	0	0	5
1kbea_	1ptq__	6	0	0	0	0	0	42
1dvpa2	1fp0a1	1	0	0	1	0	0	22
1e31a_	1qbha_	9	0	0	0	0	0	39
1a8p_1	2pia_1	7	0	0	0	0	0	79
1ezva1	1l0lb1	13	0	0	0	0	0	189
1gjja1	1jeia_	2	0	0	0	0	0	35
1ifwa_	1jgna_	6	0	0	0	0	0	42
1h3za_	1oi1a2	3	1	0	0	6	0	35
1je3a_	1pava_	4	0	0	0	0	0	50
1h3qa_	1h8ma_	7	0	0	1	0	0	102
1nj1a2	1nj8a2	6	0	0	0	0	0	43
1jeqa1	1kcfa1	2	0	0	0	0	0	19
1ayl_2	1ii2a2	14	0	0	0	0	0	154
1kvna_	1lnga_	5	0	0	0	0	0	74
1jz8a4	1nsza_	24	1	0	5	6	0	96
1clc_2	1edqa1	7	1	0	1	9	0	46
1f5aa4	1fa0a4	9	0	0	0	0	0	82
1e12a_	1h2sa_	9	0	0	0	0	0	92
1k4cc_	1orsc_	4	0	0	1	0	0	43
1kf6c_	1nekd_	3	0	0	1	0	0	34
1fftc_	1ocrc_	5	0	0	0	0	0	129
1m56b2	1ocrb2	2	0	0	0	0	0	83
1dxrm_	1qovl_	9	0	0	0	0	0	217
1ee8a2	1nnja2	11	0	0	0	0	0	103
1m9sa3	1m9sa4	8	0	0	0	0	0	57
1h3ia2	1n3ja_	12	0	0	0	0	0	59
1ihma_	1k5ma_	12	0	1	0	0	5	113
1c8da_	1gff2_	7	0	0	3	0	0	0
1dzla_	1vpsa_	11	0	0	0	0	0	105
1a34a_	1stma_	8	0	0	0	0	0	75
1ku3a_	1l0oc_	3	0	0	0	0	0	39
1g8fa1	1iq8a3	8	0	0	1	0	0	47
1iw7f3	1or7a2	5	0	0	0	0	0	69

Table IV-21. The alignment block-level and aligned position-level comparison for each alignment of SFESA (O+G+M+S) applied on PROMALS on the SABMARK superfamily dataset.

Method	All (1682/23.0)	Set 1 (420/6.8)	Set 2 (421/14.9)	Set 3 (420/23.1)	Set 4 (421/48.4)
PROMALS	80.3	56.7	80.2	90.0	**94.2**
SFESA (O)+PROMALS	80.4	57.4	80.5	89.9	93.8
SFESA (O+G)+PROMALS	80.1	57.6	80.3	89.4	93.2
SFESA (O+G+M)+PROMALS	80.3	57.3	81.1	89.7	93.1
SFESA (O+G+M+S)+PROMALS	**81.3**	**58.7**	**81.7**	**90.5**	94.1
HHpred	78.0	46.6	80.2	90.3	94.7
SFESA (O)+HHpred	78.0	46.7	80.2	90.3	94.7
SFESA (O+G)+HHpred	78.3	47.2	80.6	90.5	**94.9**
SFESA (O+G+M)+HHpred	**78.6**	**47.8**	**81.3**	**90.6**	94.8
SFESA (O+G+M+S)+HHpred	**78.6**	47.7	81.1	**90.6**	**94.9**
CNFpred	80.5	56.3	81.7	89.6	94.5
SFESA (O)+CNFpred	81.0	57.1	82.1	90.2	94.6
SFESA (O+G)+CNFpred	81.2	57.3	82.3	90.3	**94.7**
SFESA (O+G+M)+CNFpred	81.2	57.2	**82.8**	90.3	94.6
SFESA (O+G+M+S)+CNFpred	**81.3**	**57.5**	**82.8**	**90.4**	94.6

Table IV-22. Test on PREFAB database. Average Q-score is reported. The total 1682 PREFAB alignments are divided to four semi-equal-sized sets according to sequence identity of the PROMALS alignment. The number of alignments and the average sequence identity are in parenthesis beneath the set names. Bold indicates the best performance in the subsection. Bold with underscore indicates the overall best performance in one column.

	All (1682/23.0)	Set 1 (420/6.8)	Set 2 (421/14.9)	Set 3 (420/23.1)	Set 4 (421/48.4)
Method/(p-value)	PROMALS	PROMALS	PROMALS	PROMALS	PROMALS
SFESA(O)+PROMALS	0.016	9.39e-03	3.37e-02	0.14	8.31e-03 **(worse)**
SFESA(O+G)+PROMALS	0.13	3.98e-03	0.1	0.44	1.16e-03 **(worse)**
SFESA(O+G+M)+PROMALS	0.058	3.56e-02	5.45e-04	0.29	2.59e-05 **(worse)**
SFESA(O+G+M+S)+PROMALS	0	7.93e-09	0	6.99e-05	0.46

Method/(p-value)	HHpred	HHpred	HHpred	HHpred	HHpred
SFESA(O)+HHpred	0.45	0.41	0.21	0.15	0.34
SFESA(O+G)+HHpred	0	3.20e-04	5.77e-04	1.94e-04	1.14e-02
SFESA(O+G+M)+HHpred	0	7.74e-07	4.00e-09	3.75e-04	0.15
SFESA(O+G+M+S)+HHpred	0	2.61e-09	0	2.03e-06	6.23e-03

Method/(p-value)	CNFpred	CNFpred	CNFpred	CNFpred	CNFpred
SFESA(O)+CNFpred	3.8e-09	1.68e-03	3.15e-03	1.32e-06	2.62e-03
SFESA(O+G)+CNFpred	0	9.34e-04	3.10e-04	0	9.27e-09
SFESA(O+G+M)+CNFpred	0	2.03e-02	1.92e-06	1.36e-08	1.29e-03
SFESA(O+G+M+S)+CNFpred	0	1.98e-04	3.18e-07	0	3.60e-06

Table IV-23. Statistically significant Q-Score improvement of SFESA on PROMALS, HHpred and CNFpred (PREFAB as reference). All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on PREFAB dataset. P-values are calculated based on the paired Wilcoxon signed-rank. P-values below 0.05 are marked green and below 0.005 are marked pink.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	20856	389	403	560	2972	3145	259032
SFESA (O+G)+PROMALS	19905	632	714	957	4189	5148	255812
SFESA (O+G+M)+PROMALS	18962	872	942	1432	5450	6449	253250
SFESA (O+G+M+S)+PROMALS	20496	640	386	686	4109	2427	258613
SFESA (O)+HHpred	20722	47	60	139	300	273	264576
SFESA (O+G)+HHpred	19537	497	305	629	1749	1057	262343
SFESA (O+G+M)+HHpred	18492	858	518	1100	3182	1954	260013
SFESA (O+G+M+S)+HHpred	19688	577	232	471	2177	946	262026
SFESA (O)+CNFpred	20619	272	160	309	2093	1114	261942
SFESA (O+G)+CNFpred	19440	768	346	806	3179	1923	260047
SFESA (O+G+M)+CNFpred	18221	1165	611	1363	4473	3240	257436
SFESA (O+G+M+S)+CNFpred	19549	836	308	667	3367	1862	259920

Table IV-24. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred. All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the PREFAB dataset (all).

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	4596	135	113	390	1019	761	37519
SFESA (O+G)+PROMALS	4155	207	186	686	1436	1108	36755
SFESA (O+G+M)+PROMALS	3655	276	274	1029	1911	1706	35682
SFESA (O+G+M+S)+PROMALS	4403	211	121	499	1475	721	37103
SFESA (O)+HHpred	3961	21	30	99	137	122	39040
SFESA (O+G)+HHpred	3466	146	93	406	580	319	38400
SFESA (O+G+M)+HHpred	3015	233	142	721	965	502	37832
SFESA (O+G+M+S)+HHpred	3598	160	64	289	656	237	38406
SFESA (O)+CNFpred	4150	93	77	220	741	467	38091
SFESA (O+G)+CNFpred	3673	203	133	531	1068	712	37519
SFESA (O+G+M)+CNFpred	3156	289	219	876	1477	1177	36645
SFESA (O+G+M+S)+CNFpred	3796	218	115	411	1167	701	37431

Table IV-25. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred. All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the PREFAB “Set 1” subset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	4467	116	109	122	863	880	54791
SFESA (O+G)+PROMALS	4262	177	193	182	1228	1412	53894
SFESA (O+G+M)+PROMALS	4076	263	229	246	1705	1518	53311
SFESA (O+G+M+S)+PROMALS	4401	191	101	121	1268	567	54699
SFESA (O)+HHpred	4665	12	20	23	71	96	56367
SFESA (O+G)+HHpred	4342	150	93	135	546	360	55628
SFESA (O+G+M)+HHpred	4060	278	146	236	1148	604	54782
SFESA (O+G+M+S)+HHpred	4354	189	71	106	795	318	55421
SFESA (O)+CNFpred	4518	79	40	54	599	301	55634
SFESA (O+G)+CNFpred	4241	207	96	147	933	602	54999
SFESA (O+G+M)+CNFpred	3951	329	154	257	1385	836	54313
SFESA (O+G+M+S)+CNFpred	4255	225	76	135	1031	499	55004

Table IV-26. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred. All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the PREFAB “Set 2” subset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	5648	91	90	33	727	741	79542
SFESA (O+G)+PROMALS	5488	155	162	57	993	1304	78713
SFESA (O+G+M)+PROMALS	5339	213	210	100	1217	1494	78299
SFESA (O+G+M+S)+PROMALS	5586	148	84	44	896	633	79481
SFESA (O)+HHpred	5811	11	7	13	56	34	80920
SFESA (O+G)+HHpred	5620	113	55	54	368	223	80419
SFESA (O+G+M)+HHpred	5428	204	118	92	642	457	79911
SFESA (O+G+M+S)+HHpred	5616	133	51	42	428	214	80368
SFESA (O)+CNFpred	5706	77	33	24	602	280	80128
SFESA (O+G)+CNFpred	5488	208	65	79	818	425	79767
SFESA (O+G+M)+CNFpred	5252	328	130	130	1076	736	79198
SFESA (O+G+M+S)+CNFpred	5480	240	62	58	819	395	79796

Table IV-27. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred. All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the PREFAB “Set 3” subset.

Methods	Unshifted Blocks	Shifted Blocks (SFESA> Original)	Shifted Blocks (SFESA< Original)	Shifted Blocks (SFESA= Original)	Aligned Positions (SFESA> Original)	Aligned Positions (SFESA< Original)	Aligned Positions (SFESA= Original)
SFESA (O)+PROMALS	6145	47	91	15	363	763	87180
SFESA (O+G)+PROMALS	6000	93	173	32	532	1324	86450
SFESA (O+G+M)+PROMALS	5892	120	229	57	617	1731	85958
SFESA (O+G+M+S)+PROMALS	6106	90	80	22	470	506	87330
SFESA (O)+HHpred	6285	3	3	4	36	21	88249
SFESA (O+G)+HHpred	6109	88	64	34	255	155	87896
SFESA (O+G+M)+HHpred	5989	143	112	51	427	391	87488
SFESA (O+G+M+S)+HHpred	6120	95	46	34	298	177	87831
SFESA (O)+CNFpred	6245	23	10	11	151	66	88089
SFESA (O+G)+CNFpred	6038	150	52	49	360	184	87762
SFESA (O+G+M)+CNFpred	5862	219	108	100	535	491	87280
SFESA (O+G+M+S)+CNFpred	6018	153	55	63	350	267	87689

Table IV-28. The alignment block-level and aligned position-level comparison of SFESA on PROMALS, HHpred and CNFpred. All SFESA modes by using different parameters (SFESA (O), SFESA (O+G), SFESA (O+G+M) and SFESA (O+G+M+S)) are compared with three existing alignment methods on the PREFAB “Set 4” subset.

Blocks Categories	Total block number	Succeeded block number (SFESA>PROMALS)	Failed block number (SFESA<PROMALS)	Success/Failure rate
All	16347	1000	495	2.0
Helix	5778	328	181	1.8
Edge Strand	4439	283	185	1.5
Nonedge Strand	6130	389	129	3.0
Helix Category (Average Contact number = 23.7)				
0-11	1013	46	61	0.8
12-17	1105	83	34	2.4
18-21	775	46	28	1.6
22-27	972	59	28	2.1
28-36	1009	55	26	2.1
36-114	904	39	4	9.8
Edge Strand Category (Average Contact number = 12.2)				
0-5	866	29	53	0.5
6-8	641	37	33	1.1
9-11	723	46	33	1.4
12-15	898	74	29	2.6
16-19	624	47	22	2.1
20-46	687	50	15	3.3
Nonedge Strand Category (Average Contact number = 25.7)				
0-16	1053	63	39	1.6
16-21	1026	81	21	3.9
22-25	1032	77	25	3.1
26-30	1145	71	12	5.9
31-35	935	48	19	2.5
36-77	939	49	13	3.8

Table IV-29. SFESA (O+G+M) success/failure rate depends on SSE types and contact numbers on of alignment blocks. Edge strands are those strands that have backbone-to-backbone hydrogen bonding interactions (defined by DSSP) on only one side. For each SSE type, data is divided into equal-size bins based on contact numbers (see last three sub tables).

Methods	Reference-dependent (Q-score)					Reference-independent (TM-score)
	Dali	TMalign	Matt	MUSTER	Deep Align	
HHpred (Local)	44.0	40.5	42.2	44.1	43.9	0.414
HHpred (Global)	49.3	45.3	46.7	49.0	49.7	0.490

Table IV-30. HHpred local and global mode performance on the MUSTER dataset. Columns 2-6 indicate five different structure alignment methods to generate reference alignments (Reference-dependent evaluation). Column 7 indicates the average of query model's TM-score built by Modeller (Reference-independent evaluation).

Methods	Reference-dependent (Q-score)				Reference-independent (TM-score)
	Dali	TMalign	Matt	DeepAlign	
HHpred (Local)	58.9	56.7	59.0	60.2	0.528
HHpred (Global)	63.0	60.6	62.7	64.4	0.589

Table IV-31. HHpred local and global mode performance on the SALIGN dataset. Columns 2-5 indicate five different structure alignment methods to generate reference alignments (Reference-dependent evaluation). Column 6 indicates the average of query model's TM-score built by Modeller (Reference-independent evaluation).

Methods On subsets	Reference-dependent (Q-score)		Reference-independent (TM-score)	
	SABmark_TWI	**SABmark_SUP**	**SABmark_TWI**	**SABmark_SUP**
HHpred (Local)	37.1	65.5	0.316	0.528
HHpred (Global)	40.7	68.9	0.371	0.570

Table IV-32. HHpred local and global mode performance on the SABMARK dataset. Columns 2-3 indicate the alignment Q-score based on their reference on two subsets of the SABmark benchmark: "twilight zone" and "superfamilies" respectively (Reference-dependent evaluation). Columns 4-5 indicate the average of query model's TM-score built by Modeller on two subsets of the SABmark benchmark: "twilight zone" and "superfamilies" respectively (Reference-independent evaluation).

Methods	All (1682/23.0)	Set 1 (420/6.8)	Set 2 (421/14.9)	Set 3 (420/23.1)	Set 4 (421/48.4)
HHpred (Local)	75.8	42.8	76.9	89.2	94.2
HHpred (Global)	78.0	46.6	80.2	90.3	94.7

Table IV-33. HHpred local and global mode performance on the PREFAB dataset. Average Q-score is reported. The total 1682 PREFAB alignments are divided to four semi-equal-sized sets according to sequence identity of the PROMALS alignment. The number of alignments and the average sequence identity are in parenthesis beneath the set names.

V. SFESA WEB SERVER FOR PAIRWISE ALIGNMENT REFINEMENT BY SECONDARY STRUCTURE SHIFTS

Protein sequence alignment is essential for a variety of tasks such as homology modeling and active site prediction. Alignment errors remain the main cause of low-quality structure models. A bioinformatics tool to refine alignments is needed to make protein alignments more accurate. We developed the SFESA web server to refine pairwise protein sequence alignments. Compared to the previous version of SFESA, which required a set of 3D coordinates for a protein, the new server will search a sequence database for the closest homolog with an available 3D structure to be used as a template. For each alignment block defined by secondary structure elements in the template, SFESA evaluates alignment variants generated by local shifts and selects the best-scoring alignment variant. A scoring function that combines the sequence score of profile-profile comparison and the structure score of template-derived contact energy is used for evaluation of alignments. PROMALS pairwise alignments refined by SFESA are more accurate than those produced by current advanced alignment methods such as HHpred and CNFpred. In addition, SFESA also improves alignments generated by other software. SFESA is a web-based tool for alignment refinement, designed for researchers to compute, refine, and evaluate pairwise alignments with a combined sequence and structure scoring of alignment blocks. To our knowledge, the SFESA web server is the only tool that refines alignments by evaluating local shifts of secondary structure elements. The SFESA web server is available at http://prodata.swmed.edu/sfesa.

INTRODUCTION

Homology modeling that constructs a structural model of a "query" protein based on its similarity to a homologous protein with known 3-dimensional structure (the "template") remains the most reliable method of structure prediction. In most homology modeling methods, an essential step requires the input or construction of a pairwise sequence alignment between the query and the template, from which structurally equivalent residue pairs are deduced. Pairwise alignment is also the foundation for most multiple sequence alignment (MSA) methods. For example, the progressive method for MSA construction assembles a multiple sequence alignment by a series of pairwise alignments of sequences or pre-aligned groups (Feng and Doolittle 1987).

Early methods of pairwise protein alignments apply dynamic programming algorithms that rely on general substitution matrices of amino acid residues and pre-defined gap penalties (Needleman and Wunsch 1970, Smith and Waterman 1981). Heuristic pairwise alignment tools such as BLAST (Altschul, Gish et al. 1990) excel in speed and are suitable for sequence database searches. Numerical sequence profiles have been designed to incorporate information of homologous proteins to help aligning divergent sequences. PSI-BLAST (Altschul, Madden et al. 1997) and HMMER (Eddy 1998) are examples of sequence-profile comparison methods that are generally more accurate than methods of sequence-sequence comparison. The subsequent development of profile-profile comparison methods (Rychlewski, Jaroszewski et al. 2000, Yona and Levitt 2002, Sadreyev and Grishin 2003, Gniewek, Kolinski et al. 2012) further enhanced alignment quality and the ability to detect homologous relationships. In addition to amino acid sequence profiles, predicted structural information, e.g., secondary structure and

solvent accessibility, was also included in various alignment methods (Soding 2005, Pei and Grishin 2007, Ma, Peng et al. 2012). Three-dimensional structure information has been used in alignment construction methods in various ways, such as those based on structure-dependent profiles (Bowie, Luthy et al. 1991, Kelley, MacCallum et al. 2000) and a Monte Carlo-based alignment method that samples a set of moves of gapless alignment stretches and scores based on a template contact map (Gniewek, Kolinski et al. 2014).

Despite continuous method development in the alignment field, obtaining high-quality alignments for distantly related proteins remains a challenge. Alignment errors are still the main cause for the low quality of models built by homology. One common type of alignment error is the local misalignment, often by only a few residues, of secondary structure elements (α-helices and β-strands). Such errors often reflect the periodic nature of regular secondary structures. For example, many α-helices can be shifted by three or four residues while still maintaining a similar pattern of hydrophobic residues and polar residues. Therefore, one possible direction for refining an alignment lies in the generation of alignment variants by locally shifting secondary structure elements and evaluating the sequence and structure fitness of these alignment variants to determine which one is more likely to be correct.

Here we describe the SFESA web server, which refines pairwise protein alignments by evaluating alignment variants resulting from locally shifting secondary structure elements. The SFESA web server enables researchers to compute, refine, and evaluate pairwise alignments with a combined sequence and structure scoring of alignment blocks. The previous version of SFESA required the upload of a predefined

template structure. In contrast, the new web server allows for a template to be specified by its PDB and chain identifiers. Furthermore, if no structure is provided, the SFESA server will search the database of sequences with experimentally determined 3D structures for the closest template, and this will then be used in the alignment refinement. The server facilitates further analysis of alignments at the level of secondary structure, providing detailed results of sequence and structure scores for local shifts of secondary structure elements. To our knowledge, the SFESA web server is the only online tool that refines alignments by evaluating local shifts of secondary structure elements.

RESULTS

The SFESA web server

The SFESA web server is a tool for constructing, refining, and evaluating pairwise protein alignments (Figure V-1). The workflow of the server is shown in Figure V-1. Compared to the previously reported version of SFESA (Tong, Pei et al. 2015), in which a user must provide a structure for the template sequence, the updated server will search against our inhouse protein structure database to find the closest (to either sequence) homolog with available 3D structure to improve the alignment.

Users can input or upload sequences for the query and template either as a pairwise alignment or as two unaligned sequences in FASTA format. If two unaligned sequences are provided, the server uses PROMALS (Pei and Grishin 2007) to automatically construct a pairwise alignment. Input of a 3-dimensional structure (in pdb format) with high sequence similarity to the template is optional but recommended. A user can input a PDB identifier and a chain identifier, instead of a coordinate set, to directly use the structure from the RCSB PDB database (Berman, Westbrook et al. 2000). If no structure is provided for the template, the server uses BLAST (Altschul, Madden et al. 1997) to automatically search for homologs for either the query or template in a database of representative spatial structures and selects the best hit as the homologous structure, used as a template for structure score calculation.

Four SFESA alignment refinement modes are available in the web server: SFESA (O) uses up to 8 variants generated by ±4 shifts that keep the gap patterns of the original alignment block and the Miyazawa-Jernigan (MJ) (Miyazawa and Jernigan 1999) contact

matrix for structure score calculation; SFESA (O+G) uses up to 18 variants by considering gap shifts and the MJ contact matrix; and SFESA (O+G+M) (default) uses a newly derived contact matrix in addition to gap processing; SFESA (O+G+M+S) differs from SFESA (O+G+M) in that an SVM-derived score is used in the second filtering step instead of S_{comb_II}.

Several parameters are provided. One parameter is the sequence identity threshold between the template sequence and its homolog with a known 3D structure. SFESA refinement is applied only when the sequence identity between the template and its structure homolog is higher than the threshold (default = 0.5). Another parameter is the maximal number of residue positions to shift (default = 4, i.e., shifts are applied from −4 to +4 positions). Increasing this parameter generates more alignment variants, but also increases the probability that a wrong variant is accepted. The third parameter is the threshold for the fraction of non-gapped residue pairs above which an alignment block is used in the refinement process (default = 0.5). We also provide parameters for running and processing PSI-BLAST (Altschul, Madden et al. 1997) results to generate the sequence profile used for the sequence score calculation, such as the number of iterations, the e-value inclusion cutoff, and a sequence identity cutoff to remove divergent hits.

The output page of the SFESA web server includes the starting alignment (the input alignment or in the case of the input of unaligned sequences, the automatically generated PROMALS alignment), the refined alignment, and the refinement details for each evaluated alignment block. Figure V-2 shows one example of the output page. The first part of the output page (Figure V-2A) contains the starting alignment and the refined alignment with colored alignment blocks. PSIPRED (Jones 1999) predicts secondary

structure elements of the query and secondary structure elements of the template based on PALSSE (Majumdar, Krishna et al. 2005) and DSSP (Kabsch and Sander 1983), and these predicted elements are shown above the query sequence and below the template sequence, respectively. Evaluated alignment blocks are depicted in red and orange for α-helices and blue and dark green for β-strands to distinguish them. In the SFESA-refined alignment, the modified alignment blocks are marked with underscores.

The second part of the output page (Figure V-2B) is a table summarizing the refinement results of the evaluated alignment blocks, numbered from the N-terminus to the C-terminus. Each row in the table provides the element start and end position numbers in the template, the element secondary structure type, the original alignment block, the shift result, and the refined alignment block. The shift result column shows Gap Mode and Shift Number. Gap Mode can be "Left" (gap pattern preprocessed by moving residues all the way to the left), "Right" (gap pattern preprocessed by moving residues all the way to the right), or "Original" (no gap preprocessing). Shift Number (in brackets) is the number of positions the residues in the query are shifted by, relative to the template. The "+" and "-" signs in Shift Number denote that the query residues in the alignment block are shifted towards the C-terminus or the N-terminus, respectively. If no alignment variant was accepted for an alignment block (i.e., the original refinement retained), "No shift" is shown in the shift result column and "-" is shown in the column of Refined Alignment Block. The third part of the output page contains tables with scoring details for the alignment variants. A table is provided for each alignment block evaluated by SFESA and presents each alignment variant and its sequence score, structure score, and combined scores I and II. Figure V-2C provides an example for alignment block

number 4. The residues in the original alignment block are colored blue and pink for the query and the template, respectively. The scoring details for alignment variants may help users manually evaluate and select alternative alignments.

An example of an alignment improved by the SFESA server

In the example shown in Figure V-2, the input consisted of two SCOP domains, d1ja1a3 (query) and d2piaa2 (template), and the 3D structure of the template. SFESA used PROMALS to obtain the starting alignment, which was refined to generate the refined alignment with the default option SFESA (O+G+M). Out of the seven alignment blocks evaluated by SFESA, five alignment blocks were kept without shifts and two alignment blocks were modified according to SFESA refinement scores (Figure V-2B). Both of these modified alignment blocks are in better agreement with the Dali structural alignment (Holm and Sander 1996) of the query and the template compared to the original alignment blocks. We generated structure models for the query based on the starting alignment (Figure V-2D, left panel) and the refined alignment (Figure V-2D, right panel). Both models (in light grey and red ribbons) were superimposed upon the real structure of the query (in dark grey and green ribbons). The GDT-TS scores (Zemla 2003) for models generated from the starting alignment and the refined alignment are 57.7 and 67.1, respectively. The query secondary structure element in the fourth evaluated alignment block is highlighted in both structure superpositions (green for the real structure and red for the model). This element, misaligned by two residues in the starting alignment (Figure V-2D, left panel), has been corrected in the refined alignment

(Figure V-2D, right panel). As a result, the RMSD for this secondary structure element between the model and the real structure improved from 5.3Å for the model generated by the starting alignment to 2.0Å for the model generated by the refined alignment.

DISCUSSION

Despite many significant research efforts, it is still challenging to correctly align weakly similar but homologous protein sequences. Alignment errors remain the main reason for the poor quality of homology models. Refining the alignments generated by automatic methods is a promising approach for increasing alignment quality. We found that secondary structure elements are often misaligned by only a few residues and that more accurate solutions can be identified within a limited set of local shifts of secondary structure elements. Therefore, we developed the SFESA method in order to refine alignments by evaluating the alignment variants generated by local shifts of template-defined secondary structures.

In the SFESA scoring system, both a profile-based sequence score and a novel contact-based structure score of the aligned residue pairs in the original alignment block and the alignment variants are calculated. Thus, an insufficient number of contacts can limit the quality of the alignment refinement. We found that structure scoring works well when there are sufficient contacts in the template as well as sufficient corresponding aligned residues in the query (Tong, Pei et al. 2015). However, if a secondary structure element is involved in too few contacts (e.g. exposed edge β-strands), the remaining contacts are insufficient to define a complete structural environment. SFESA is less effective in these cases. This observation suggests that dedicated efforts on misaligned blocks with insufficient contacts are required to improve alignments further.

CONCLUSION

SFESA is a web-based tool to compute, refine, and evaluate pairwise alignments with a combined sequence and structure scoring of alignment blocks. Taking a pairwise alignment as input, the SFESA web server searches against an in-house database of protein spatial structures to find the closest homolog of either sequence. It then refines the pairwise alignment by combining the sequence profile similarity and residue-residue contact information that were obtained from the homolog with the structure. Finally, it facilitates further analysis of the alignment results at the level of secondary structure, providing details about scoring for all shifts of secondary structure elements.

MATERIALS AND METHODS

Recently we developed SFESA (Tong, Pei et al. 2015), a method that refines pairwise protein sequence alignment by evaluating alignment variants generated from local shifts of secondary structure elements. SFESA first delineates alignment blocks from a starting pairwise protein alignment. Each alignment block corresponds to a regular secondary structural element (α-helix or β-strand as delineated by PALSSE (Majumdar, Krishna et al. 2005)) in the template and the corresponding aligned region in the query. For each alignment block, SFESA generates a set of alignment variants by locally shifting query residues relative to template residues. Then, both a profile-based sequence score and a contact-based structure score of the aligned residue pairs in the original alignment block and the alignment variants are calculated. We have shown that the best-scoring alignment variant has the highest probability of being correct, e.g., showing the best agreement with the structure-based alignment.

SFESA uses two local shifting strategies to generate alignment variants with different treatments of gaps in the original alignment block. In the first strategy, up to 8 alignment variants are generated by shifting query residues up to four positions left or right relative to the template while maintaining the gap pattern in the original alignment block. However, we observed that gaps rarely occur in the middle of secondary structure elements in structure-based alignments. Therefore, in the second strategy, SFESA preprocesses the gap pattern in the original alignment block by eliminating gaps in the middle of the secondary structure elements. To achieve this, residues of an alignment block in both the query and template are shifted all the way to the left or right while all gaps are placed on the opposite side. Two preprocessed alignment blocks are generated:

one by shifting residues to the left and filling the right side with gaps and the other by shifting residues to the right and filling the left side with gaps. Each of these two alignment variants is then used as a starting point to generate 8 additional alignment variants by ±4 shifts while keeping the modified gap patterns. This procedure gives rise to up to 18 (1+8+1+8) unique alignment variants .

For the sequence score, we use the profile-profile COMPASS score (Sadreyev and Grishin 2003). Sequence profiles are generated from PSI-BLAST multiple sequence alignments (Altschul, Madden et al. 1997). For the structure score, we define residue contacts based on the structure of the template. A residue contact is defined as a residue pair within a distance cutoff. In the template of an alignment, the residue contacts can be identified using the known structure of the template. We then evaluate the contact energy of corresponding contact residue pairs in the query that are inferred from query-template alignment. For example, if residue *i* in the template makes contact with residues *j*, *k*, and *m* in the template structure (i.e., contact pairs are (*i*, *j*), (*i*, *k*), and (*i*, *m*)), and the corresponding aligned residues for *i*, *j*, *k*, and *m* in the query are *i'*, *j'*, *k'*, and *m'*, respectively, then the inferred contact pairs in the query are (*i'*, *j'*), (*i'*, *k'*), and (*i'*, *m'*). The structure score for the aligned residue pair *i* and *i'* is CE(*i'*, *j'*) + CE(*i'*, *k'*) + CE(*i'*, *m'*), reflecting the structural fitness of the inferred query contact residue pairs. Here, CE is a matrix of the contact energy for residue pairs. We used two contact energy matrices: one is derived by Miyazawa and Jernigan (Miyazawa and Jernigan 1999) with contacts defined as residue pairs with side chain centers less than 6.5Å, and the other is developed by us to best discriminate correct alignment variants from incorrect alignment variants .

Regarding our derived contact matrix, the cutoff for contact definition is 6.5Å between any side chain atoms of two residues.

In practice, the SFESA method uses a two-filter strategy to compare the scores of the original alignment block and the alignment variants and determines whether the original alignment block should be kept or changed to one of the alignment variants. The first filter checks if there are any alignment variants with a higher combined score I (S_{comb_I}, a linear combination of sequence score and structure score) than the original alignment block. If none of the alignment variants has a S_{comb_I} higher than the original alignment block, SFESA rejects all the alignment variants and keeps the original alignment block. Otherwise, the alignment variant with the highest S_{comb_I} is selected and passed to the second filter. In the second filter, SFESA uses combined score II (S_{comb_II}, a linear combination of sequence score and structure score) or an SVM score (S_{SVM}) to compare the selected alignment variant and the original alignment block. If the selected alignment variant still has a higher S_{comb_II} or S_{SVM}, SFESA will accept this alignment variant. Otherwise, SFESA keeps the original alignment block. The weights of the sequence score and structure score in S_{comb_I} and S_{comb_II} are optimized separately. S_{SVM} is a score reported by a support vector machine (SVM) that was trained to differentiate correct alignment variants from incorrect alignment variants by using a number of features including a COMPASS-based sequence score (Sadreyev and Grishin 2003), a contact-based structure score, a solvent accessibility score and a secondary structure score. The solvent accessibility score is based on a three-by-three relative solvent accessibility substitution matrix derived from FAST (Zhu and Weng 2005) structural alignments of SCOP (Andreeva, Howorth et al. 2008) domains. Similarly, the secondary

structure score is based on a three-by-three secondary structure substitution matrix derived from FAST (Zhu and Weng 2005) structural alignments of SCOP (Andreeva, Howorth et al. 2008) domains. The secondary structure is predicted by PSIPRED (Jones 1999) for the query; the secondary structure information in DSSP (Kabsch and Sander 1983) is used for the template. For each alignment block, starting from the N-terminus and proceeding to the C-terminus, SFESA decides whether to keep the original alignment block or to accept one of the alignment variants.

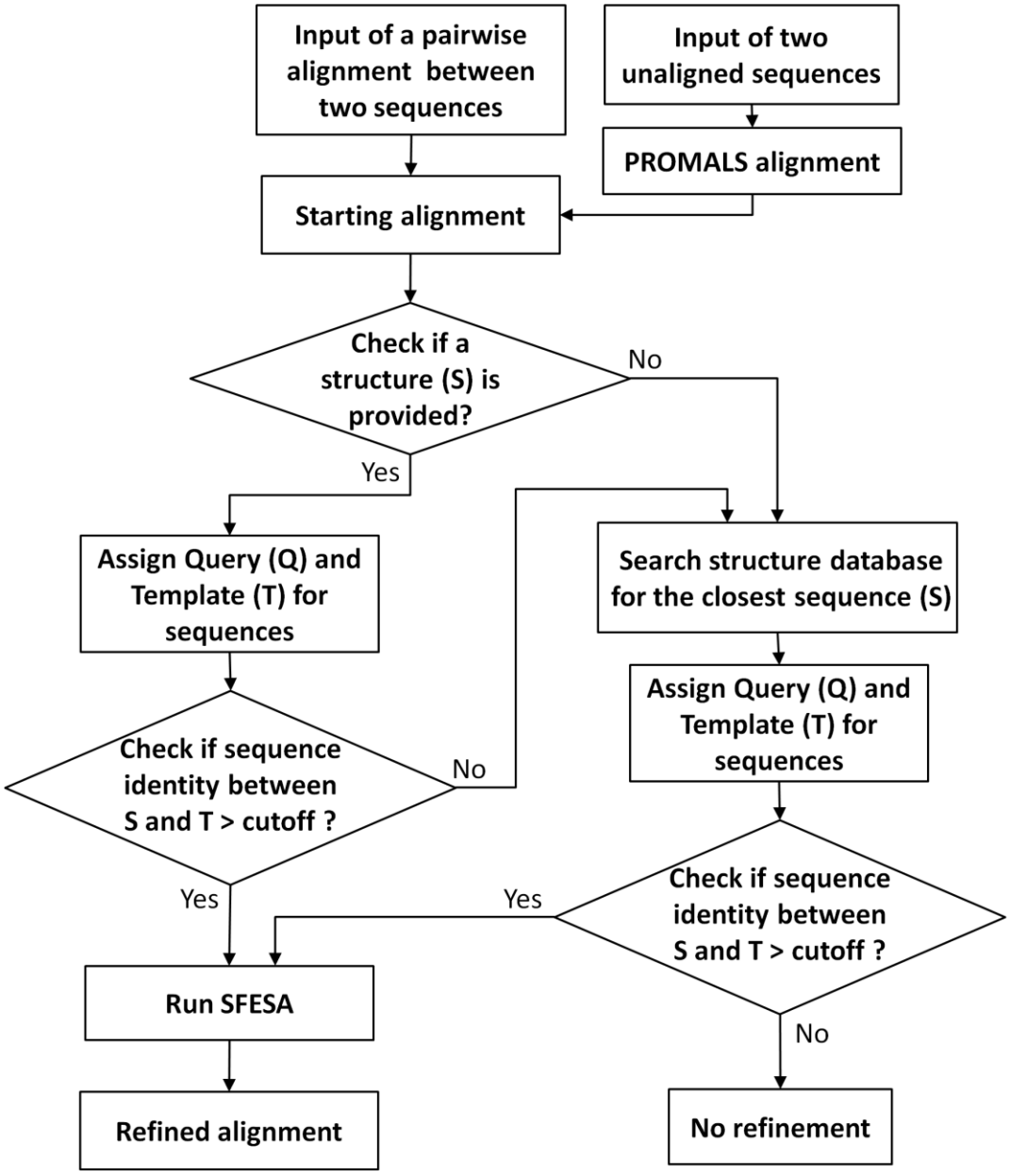


Figure V-1. Flowchart of the SFESA web server. The sequence that is found to be closest to the provided structure or the structure database is assigned as the Template (T). The other sequence is assigned as the Query (Q).

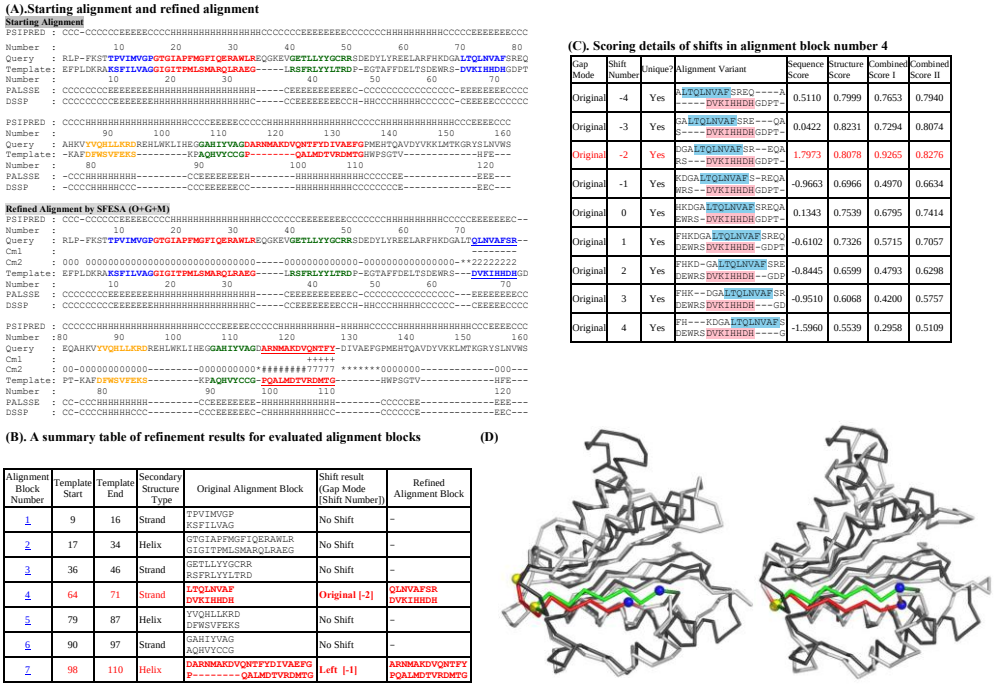


Figure V-2. An example showing the output of the SFESA server and its ability to improve the alignment. (A) Output from the starting alignment and SFESA-refined alignment with secondary structure and colored alignment block. Predicted secondary structures for the query and the real secondary structures for the template are shown ("H"-Helix, "S"-Strand and "C"- Coil). “Number” shows the position number of the residue above the query and below the template, respectively. “Cm1" and "Cm2" represent the positional differences between the refined alignment and starting alignment. "Cm1" shows the sign of the query residue shifting ("+": query residue shifted towards C-terminal; "-": query residue shifted towards N-terminal) while "Cm2" shows the query residue shift number. If the query residue is aligned to a gap in both the starting and refined alignments, "Cm1" is left blank and "Cm2" shows the gap character "-". If the query residue is aligned to one residue in the starting alignment but aligned to a gap in the refined alignment, "Cm1" is left blank and "Cm2" shows "*". If a template residue is aligned to a gap in the starting alignment, both "Cm1" and "Cm2" are left blank. □-helix alignment blocks are shown alternately in red and orange. □-strand alignment blocks are shown alternately in blue and dark green. The refined alignment blocks are marked with underscores. (B) A table summarizing refinement results for the evaluated alignment blocks. The alignment block number is ordered from N-terminus to C-terminus. The sixth column indicates the refinement results of this alignment block. If refined, a format of "Gap mode [shift number]" is shown. Rows of the refined alignment blocks are colored red. (C) One example of the scoring details of shifts for alignment block number 4. This

table contains the original alignment block and all alignment variants. The first column in the table is gap mode. There are three gap modes if there are gaps in this alignment block: Original (no change of the original alignment block), Left (residues in alignment blocks can be aligned all the way to the left while all gaps are put to the opposite side before shifting) and Right (residues in alignment blocks can be aligned all the way to the right while all gaps are put to the opposite side before shifting). The second column is the shift number. The third column indicates if such a variant is a unique one or the same as a variant shown previously. The fourth column shows the alignment variants with extended residues in both ends. The residues in the original alignment block are colored blue (query) and pink (template). The last four columns show the sequence score, structure score, combined score I and combined score II of each alignment variant. The row colored red corresponds to the alignment variant that is the final choice in the refined alignment. (D). Structure superpositions of query structure models (light grey ribbon) and query real structure (dark grey ribbon). Structure models were generated by MODELLER based on the starting alignment (left panel) and the SFESA-refined alignment (right panel). The strand ("QLNYAFSR") in alignment block number 4 is highlighted. This strand is shown in red and green in the structure model and the real structure, respectively. Blue spheres and yellow spheres mark the N-terminal boundary ("Q") and the C-terminal boundary ("R"), respectively.

VI. CONCLUDING REMARKS

Protein function prediction is one of most essential topics in the field of computational biology but still an ongoing quest for computational biologist. Currently, homology detection is still the most reliable method to predict protein function. After a homolog is recognized, the alignment construction between the protein sequence and its homologs is important to predict functional sites and modeling structure. Although many methods are developed within recent many years, it is a very challenging task to search sequence similarity or align protein sequences accurately.

The work described in this dissertation can be viewed on the perspective of finding more accurate homologous proteins and constructing more accurate alignment between protein sequence and its homolog.

In Chapter II, a new homology detection method, COMPADRE, assesses the relationship between the query sequence and a hit in the database by considering the similarity between the query and hit's known homologs. This approach increases detection quality, boosting the precision rate from 18% to 83% at half-coverage of all database homologs. The increased precision rate allows detection of a large fraction of new protein structural relationships, thus providing structure and function predictions for previously uncharacterized proteins. Our results suggest that this general approach is applicable to a wide variety of methods for detection of biological similarities. In Chapter III, a continually updatable web server based on this method (http://prodata.swmed.edu/compadre) was developed to detect homologs for the query sequence or alignment provided by users. Besides protein representatives from SCOP

database, a new search database composed of ECOD domain representatives can be updated weekly following ECOD updates. Such an up-to-date search database allows detecting homology of recently released protein structures.

In Chapter IV, we developed a pairwise alignment refinement method, SFESA, which generates candidate alignment variants for each alignment block by shifting the query region. We also designed a scoring function to judge whether an alignment variant is likely to be more accurate than the original alignment. Our scoring function combines a profile-based sequence score and a novel structural contact-based score derived from residue contacts in template. Our results prove the new contact-based score to be of ability to help protein sequence and homolog alignment. Our approach improves alignments generated by a number of state-of-the-art methods on several benchmarks that include both reference-dependent and reference-independent assessment. In Chapter V, We developed the SFESA web server to refine pairwise protein sequence alignments. It is a web-based tool for alignment refinement, designed for researchers to compute, refine, and evaluate pairwise alignments with a combined sequence and structure scoring of alignment blocks. SFESA can refine the input of pairwise protein alignment or construct alignment for the input of two unaligned protein sequences (one is treat as query, while the another is treated as template). For each alignment block defined by secondary structure elements in the template, SFESA evaluates alignment variants generated by local shifts and selects the best-scoring alignment variant. The SFESA web server is available at http://prodata.swmed.edu/sfesa.

VII. BIBLIOGRAPHY

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Andreeva, A., D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia and A. G. Murzin (2008). "Data growth and its impact on the SCOP database: new developments." Nucleic Acids Res **36**(Database issue): D419-425.

Baker, D. and A. Sali (2001). "Protein structure prediction and structural genomics." Science **294**(5540): 93-96.

Barnhart, B. J. (1989). "The Department of Energy (DOE) Human Genome Initiative." Genomics **5**(3): 657-660.

Berman, H. M., T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook and C. Zardecki (2002). "The Protein Data Bank." Acta Crystallogr D Biol Crystallogr **58**(Pt 6 No 1): 899-907.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Bork, P., T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen and Y. Yuan (1998). "Predicting function: from genes to genomes and back." J Mol Biol **283**(4): 707-725.

Bowie, J. U., R. Luthy and D. Eisenberg (1991). "A method to identify protein sequences that fold into a known three-dimensional structure." Science **253**(5016): 164-170.

Brin, S. and L. Page (1988). The Anatomy of a Large-Scale Hypertextual Web Search Engine. 7th international conference on World Wide Web (WWW), Brisbane, Australia.

Chakrabarti, S., C. J. Lanczycki, A. R. Panchenko, T. M. Przytycka, P. A. Thiessen and S. H. Bryant (2006). "Refining multiple sequence alignments with conserved core regions." Nucleic Acids Res **34**(9): 2598-2606.

Chandonia, J. M., G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt and S. E. Brenner (2004). "The ASTRAL Compendium in 2004." Nucleic Acids Res **32**(Database issue): D189-192.

Chen, S., Y. Xu, K. Zhang, X. Wang, J. Sun, G. Gao and Y. Liu (2012). "Structure of N-terminal domain of ZAP indicates how a zinc-finger protein recognizes complex RNA." Nat Struct Mol Biol **19**(4): 430-435.

Cheng, H., R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B. H. Kim and N. V. Grishin (2014). "ECOD: an evolutionary classification of protein domains." PLoS Comput Biol **10**(12): e1003926.

Chothia, C. and A. M. Lesk (1986). "The relation between the divergence of sequence and structure in proteins." EMBO J **5**(4): 823-826.

Conners, R., A. V. Konarev, J. Forsyth, A. Lovegrove, J. Marsh, T. Joseph-Horne, P. Shewry and R. L. Brady (2007). "An unusual helix-turn-helix protease inhibitory motif in a novel trypsin inhibitor from seeds of Veronica (Veronica hederifolia L.)." J Biol Chem **282**(38): 27760-27768.

Cortes, C., Vapnik, N. (1995). "Support-Vector Networks." Machine Learning **20**.

Daga, P. R., R. Y. Patel and R. J. Doerksen (2010). "Template-based protein modeling: recent methodological advances." Curr Top Med Chem **10**(1): 84-94.

Dembo, A., Karlin, S. & Zeitouni, O. (1994). "Limit distribution of maximal non-aligned two-sequence segmental score." Ann. Prob. **22**: 18.

Do, C. B., M. S. Mahabhashyam, M. Brudno and S. Batzoglou (2005). "ProbCons: Probabilistic consistency-based multiple sequence alignment." Genome Res **15**(2): 330-340.

Dong, Q. W., L. Lin, X. L. Wang and M. H. Li (2005). "Contact-based Simulated Annealing Protein Sequence Alignment Method." Conf Proc IEEE Eng Med Biol Soc **3**: 2798-2801.

Doolittle, R. F. (1981). "Similar amino acid sequences: chance or common ancestry?" Science **214**(4517): 149-159.

Ealick, S. E. (2000). "Advances in multiple wavelength anomalous diffraction crystallography." Curr Opin Chem Biol **4**(5): 495-499.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.

Eddy, S. R. (1998). "Profile hidden Markov models." Bioinformatics **14**(9): 755-763.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Eswar, N., B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper and A. Sali (2006). "Comparative protein structure modeling using Modeller." Curr Protoc Bioinformatics **Chapter 5**: Unit 5 6.

Fawcett, T. (2006). "An introduction to ROC analysis." Pattern Recognition Letters **27**(8): 861-874.

Feng, D. F. and R. F. Doolittle (1987). "Progressive sequence alignment as a prerequisite to correct phylogenetic trees." J Mol Evol **25**(4): 351-360.

Feng, Y., A. Kloczkowski and R. L. Jernigan (2007). "Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys." Proteins **68**(1): 57-66.

Florack, D. E. and W. J. Stiekema (1994). "Thionins: properties, possible biological roles and mechanisms of action." Plant Mol Biol **26**(1): 25-37.

Floudas, C. A. (2007). "Computational methods in protein structure prediction." Biotechnol Bioeng **97**(2): 207-213.

Ginalski, K. (2006). "Comparative modeling for protein structure prediction." Curr Opin Struct Biol **16**(2): 172-177.

Ginalski, K., J. Pas, L. S. Wyrwicz, M. von Grotthuss, J. M. Bujnicki and L. Rychlewski (2003). "ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure." Nucleic Acids Res **31**(13): 3804-3807.

Gniewek, P., A. Kolinski and D. Gront (2012). "Optimization of profile-to-profile alignment parameters for one-dimensional threading." J Comput Biol **19**(7): 879-886.

Gniewek, P., A. Kolinski, A. Kloczkowski and D. Gront (2014). "BioShell-Threading: versatile Monte Carlo package for protein 3D threading." BMC Bioinformatics **15**: 22.

Gotoh, O. (1996). "Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments." J Mol Biol **264**(4): 823-838.

Gribskov, M., A. D. McLachlan and D. Eisenberg (1987). "Profile analysis: detection of distantly related proteins." Proc Natl Acad Sci U S A **84**(13): 4355-4358.

Gumbel, E. J. (1935). "Les valeurs extrêmes des distributions statistiques." Ann. Inst. Henri Poincaré **5**(2): 44.

Holm, L. and C. Sander (1993). "Protein structure comparison by alignment of distance matrices." J Mol Biol **233**(1): 123-138.

Holm, L. and C. Sander (1996). "Mapping the protein universe." Science **273**(5275): 595-603.

Holm, L. and C. Sander (1998). "Touring protein fold space with Dali/FSSP." Nucleic Acids Res **26**(1): 316-319.

Horton, P. (1996). "A branch and bound algorithm for local multiple alignment." Pac Symp Biocomput: 368-383.

Horton, P. (2001). "Tsukuba BB: a branch and bound algorithm for local multiple alignment of DNA and protein sequences." J Comput Biol **8**(3): 283-303.

Huang, I. K., J. Pei and N. V. Grishin (2013). "Defining and predicting structurally conserved regions in protein superfamilies." Bioinformatics **29**(2): 175-181.

Huang, Y. J., B. Mao, J. M. Aramini and G. T. Montelione (2014). "Assessment of template-based protein structure predictions in CASP10." Proteins **82 Suppl 2**: 43-56.

Hubbard, S. and J. Thornton (1993). "Available athttp://www.bioinf.manchester.ac.uk/naccess/(last accessed date July 28, 2008)." Department of Biochemistry and Molecular Biology. University College London.

Illergard, K., D. H. Ardell and A. Elofsson (2009). "Structure is three to ten times more conserved than sequence--a study of structural response in protein cores." Proteins **77**(3): 499-508.

Jaroszewski, L., L. Rychlewski, Z. Li, W. Li and A. Godzik (2005). "FFAS03: a server for profile--profile sequence alignments." Nucleic Acids Res **33**(Web Server issue): W284-288.

Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." J Mol Biol **292**(2): 195-202.

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**(12): 2577-2637.

Karlin, S. and S. F. Altschul (1990). "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." Proc Natl Acad Sci U S A **87**(6): 2264-2268.

Karplus, K., C. Barrett, M. Cline, M. Diekhans, L. Grate and R. Hughey (1999). "Predicting protein structure using only sequence information." Proteins **Suppl 3**: 121-125.

Katoh, K., K. Misawa, K. Kuma and T. Miyata (2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic Acids Res **30**(14): 3059-3066.

Kelley, L. A., R. M. MacCallum and M. J. Sternberg (2000). "Enhanced genome annotation using structural profiles in the program 3D-PSSM." J Mol Biol **299**(2): 499-520.

Kim, C., C. H. Tai and B. Lee (2009). "Iterative refinement of structure-based sequence alignments by Seed Extension." BMC Bioinformatics **10**: 210.

Kleinjung, J., J. Romein, K. Lin and J. Heringa (2004). "Contact-based sequence alignment." Nucleic Acids Res **32**(8): 2464-2473.

Koboldt, D. C., K. M. Steinberg, D. E. Larson, R. K. Wilson and E. R. Mardis (2013). "The next-generation sequencing revolution and its impact on genomics." Cell **155**(1): 27-38.

Krogh, A., M. Brown, I. S. Mian, K. Sjolander and D. Haussler (1994). "Hidden Markov models in computational biology. Applications to protein modeling." J Mol Biol **235**(5): 1501-1531.

Kryshtafovych, A., K. Fidelis and J. Moult (2014). "CASP10 results compared to those of previous CASP experiments." Proteins **82 Suppl 2**: 164-174.

Kryshtafovych, A., J. Moult, P. Bales, J. F. Bazan, M. Biasini, A. Burgin, C. Chen, F. V. Cochran, T. K. Craig, R. Das, D. Fass, C. Garcia-Doval, O. Herzberg, D. Lorimer, H.

Luecke, X. Ma, D. C. Nelson, M. J. van Raaij, F. Rohwer, A. Segall, V. Seguritan, K. Zeth and T. Schwede (2014). "Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10." Proteins **82 Suppl 2**: 26-42.

Lackner, P., W. A. Koppensteiner, M. J. Sippl and F. S. Domingues (2000). "ProSup: a refined tool for protein structure alignment." Protein Eng **13**(11): 745-752.

Lathrop, R. H. (1994). "The protein threading problem with sequence amino acid interaction preferences is NP-complete." Protein Eng **7**(9): 1059-1068.

Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.

Lipman, D. J. and W. R. Pearson (1985). "Rapid and sensitive protein similarity searches." Science **227**(4693): 1435-1441.

Lukashin, A. V. and J. J. Rosa (1999). "Local multiple sequence alignment using dead-end elimination." Bioinformatics **15**(11): 947-953.

Luthy, R., J. U. Bowie and D. Eisenberg (1992). "Assessment of protein models with three-dimensional profiles." Nature **356**(6364): 83-85.

Ma, J., J. Peng, S. Wang and J. Xu (2012). "A conditional neural fields model for protein threading." Bioinformatics **28**(12): i59-66.

Ma, J., S. Wang, Z. Wang and J. Xu (2014). "MRFalign: protein homology detection through alignment of Markov random fields." PLoS Comput Biol **10**(3): e1003500.

Ma, J., S. Wang, F. Zhao and J. Xu (2013). "Protein threading using context-specific alignment potential." Bioinformatics **29**(13): i257-265.

Madera, M. (2008). "Profile Comparer: a program for scoring and aligning profile hidden Markov models." Bioinformatics **24**(22): 2630-2631.

Majumdar, I., S. S. Krishna and N. V. Grishin (2005). "PALSSE: a program to delineate linear secondary structural elements from protein structures." BMC Bioinformatics **6**: 202.

Mardis, E. R. (2013). "Next-generation sequencing platforms." Annu Rev Anal Chem (Palo Alto Calif) **6**: 287-303.

Margelevicius, M. and C. Venclovas (2010). "Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison." BMC Bioinformatics **11**: 89.

Marti-Renom, M. A., M. S. Madhusudhan and A. Sali (2004). "Alignment of protein sequences by their profiles." Protein Sci **13**(4): 1071-1087.

Marti-Renom, M. A., A. C. Stuart, A. Fiser, R. Sanchez, F. Melo and A. Sali (2000). "Comparative protein structure modeling of genes and genomes." Annu Rev Biophys Biomol Struct **29**: 291-325.

McGuffin, L. J. and D. T. Jones (2003). "Improvement of the GenTHREADER method for genomic fold recognition." Bioinformatics **19**(7): 874-881.

Menke, M., B. Berger and L. Cowen (2008). "Matt: local flexibility aids protein multiple structure alignment." PLoS Comput Biol **4**(1): e10.

Mittelman, D., R. Sadreyev and N. Grishin (2003). "Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments." Bioinformatics **19**(12): 1531-1539.

Miyazawa, S. and R. L. Jernigan (1999). "An empirical energy potential with a reference state for protein fold and sequence recognition." Proteins **36**(3): 357-369.

Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." J Mol Biol **247**(4): 536-540.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-453.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-453.

Ng, Y. M., Y. Yang, K. H. Sze, X. Zhang, Y. T. Zheng and P. C. Shaw (2011). "Structural characterization and anti-HIV-1 activities of arginine/glutamate-rich polypeptide Luffin P1 from the seeds of sponge gourd (Luffa cylindrica)." J Struct Biol **174**(1): 164-172.

Nolde, S. B., A. A. Vassilevski, E. A. Rogozhin, N. A. Barinov, T. A. Balashova, O. V. Samsonova, Y. V. Baranov, A. V. Feofanov, T. A. Egorov, A. S. Arseniev and E. V. Grishin (2011). "Disulfide-stabilized helical hairpin structure and activity of a novel antifungal peptide EcAMP1 from seeds of barnyard grass (Echinochloa crus-galli)." J Biol Chem **286**(28): 25145-25153.

Notredame, C., D. G. Higgins and J. Heringa (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-217.

O'Sullivan, O., K. Suhre, C. Abergel, D. G. Higgins and C. Notredame (2004). "3DCoffee: combining protein sequences and structures within multiple sequence alignments." J Mol Biol **340**(2): 385-395.

Oparin, P. B., K. S. Mineev, Y. E. Dunaevsky, A. S. Arseniev, M. A. Belozersky, E. V. Grishin, T. A. Egorov and A. A. Vassilevski (2012). "Buckwheat trypsin inhibitor with helical hairpin structure belongs to a new family of plant defence peptides." Biochem J **446**(1): 69-77.

Pearson, W. R. (1990). "Rapid and sensitive sequence comparison with FASTP and FASTA." Methods Enzymol **183**: 63-98.

Pei, J. and N. V. Grishin (2007). "PROMALS: towards accurate multiple sequence alignments of distantly related proteins." Bioinformatics **23**(7): 802-808.

Pei, J., B. H. Kim and N. V. Grishin (2008). "PROMALS3D: a tool for multiple protein sequence and structure alignments." Nucleic Acids Res **36**(7): 2295-2300.

Peng, J. and J. Xu (2011). "RaptorX: exploiting structure information for protein alignment by statistical inference." Proteins **79 Suppl 10**: 161-171.

Pesce, A., S. Dewilde, L. Kiger, M. Milani, P. Ascenzi, M. C. Marden, M. L. Van Hauwaert, J. Vanfleteren, L. Moens and M. Bolognesi (2001). "Very high resolution structure of a trematode hemoglobin displaying a TyrB10-TyrE7 heme distal residue pair and high oxygen affinity." J Mol Biol **309**(5): 1153-1164.

Petsko, G. A. (2006). "An introduction to modeling structure from sequence." Curr Protoc Bioinformatics **Chapter 5**: Unit 5 1.

Pettitt, C. S., L. J. McGuffin and D. T. Jones (2005). "Improving sequence-based fold recognition by using 3D model quality assessment." Bioinformatics **21**(17): 3509-3515.

Prlic, A., F. S. Domingues and M. J. Sippl (2000). "Structure-derived substitution matrices for alignment of distantly related sequences." Protein Eng **13**(8): 545-550.

Qi, Y., R. I. Sadreyev, Y. Wang, B. H. Kim and N. V. Grishin (2007). "A comprehensive system for evaluation of remote sequence similarity detection." BMC Bioinformatics **8**: 314.

Qiu, J. and R. Elber (2006). "SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs." Proteins **62**(4): 881-891.

Remmert, M., A. Biegert, A. Hauser and J. Soding (2012). "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." Nat Methods **9**(2): 173-175.

Richards, F. M. and C. E. Kundrot (1988). "Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure." Proteins **3**(2): 71-84.

Rost, B. (1999). "Twilight zone of protein sequence alignments." Protein Eng **12**(2): 85-94.

Rychlewski, L., L. Jaroszewski, W. Li and A. Godzik (2000). "Comparison of sequence profiles. Strategies for structural predictions using sequence information." Protein Sci **9**(2): 232-241.

Sadreyev, R. and N. Grishin (2003). "COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance." J Mol Biol **326**(1): 317-336.

Sadreyev, R. I., M. Tang, B. H. Kim and N. V. Grishin (2007). "COMPASS server for remote homology inference." Nucleic Acids Res **35**(Web Server issue): W653-658.

Sali, A., L. Potterton, F. Yuan, H. van Vlijmen and M. Karplus (1995). "Evaluation of comparative protein modeling by MODELLER." Proteins **23**(3): 318-326.

Schaffer, A. A., L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin and S. F. Altschul (2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." Nucleic Acids Res **29**(14): 2994-3005.

Schwede, T., J. Kopp, N. Guex and M. C. Peitsch (2003). "SWISS-MODEL: An automated protein homology-modeling server." Nucleic Acids Res **31**(13): 3381-3385.

Shen, M. Y. and A. Sali (2006). "Statistical potential for assessment and prediction of protein structures." Protein Sci **15**(11): 2507-2524.

Shi, J., T. L. Blundell and K. Mizuguchi (2001). "FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties." J Mol Biol **310**(1): 243-257.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-197.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-197.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." Bioinformatics **21**(7): 951-960.

Soding, J., A. Biegert and A. N. Lupas (2005). "The HHpred interactive server for protein homology detection and structure prediction." Nucleic Acids Res **33**(Web Server issue): W244-248.

Soding, J. and M. Remmert (2011). "Protein sequence comparison and fold recognition: progress and good-practice benchmarking." Curr Opin Struct Biol **21**(3): 404-411.

Stoyanova, R., A. W. Nicholls, J. K. Nicholson, J. C. Lindon and T. R. Brown (2004). "Automatic alignment of individual peaks in large high-resolution spectral data sets." J Magn Reson **170**(2): 329-335.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-4680.

Thompson, J. D., J. C. Thierry and O. Poch (2003). "RASCAL: rapid scanning and correction of multiple sequence alignments." Bioinformatics **19**(9): 1155-1161.

Tong, J., J. Pei, Z. Otwinowski and N. V. Grishin (2015). "Refinement by shifting secondary structure elements improves sequence alignments." Proteins **83**(3): 411-427.

Tong, J., R. I. Sadreyev, J. Pei, L. N. Kinch and N. V. Grishin (2015). "Using homology relations within a database markedly boosts protein sequence similarity search." Proc Natl Acad Sci U S A **112**(22): 7003-7008.

Tyszka, J. M., S. E. Fraser and R. E. Jacobs (2005). "Magnetic resonance microscopy: recent advances and applications." Curr Opin Biotechnol **16**(1): 93-99.

Van Walle, I., I. Lasters and L. Wyns (2005). "SABmark--a benchmark for sequence alignment that covers the entire known fold space." Bioinformatics **21**(7): 1267-1268.

Wang, S., J. Ma, J. Peng and J. Xu (2013). "Protein structure alignment beyond spatial proximity." Sci Rep **3**: 1448.

Wang, Y., R. I. Sadreyev and N. V. Grishin (2009). "PROCAIN: protein profile comparison with assisting information." Nucleic Acids Res **37**(11): 3522-3530.

Wang, Y., R. I. Sadreyev and N. V. Grishin (2009). "PROCAIN: protein profile comparison with assisting information." Nucleic Acids Res.

Wu, S. and Y. Zhang (2008). "MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information." Proteins **72**(2): 547-556.

Xu, J., M. Li, D. Kim and Y. Xu (2003). "RAPTOR: optimal protein threading by linear programming." J Bioinform Comput Biol **1**(1): 95-117.

Yang, Y., E. Faraggi, H. Zhao and Y. Zhou (2011). "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**(15): 2076-2082.

Yang, Y. and Y. Zhou (2008). "Specific interactions for ab initio folding of protein terminal regions with secondary structures." Proteins **72**(2): 793-803.

Yona, G. and M. Levitt (2002). "Within the twilight zone: a sensitive profile-profile comparison tool based on information theory." J Mol Biol **315**(5): 1257-1275.

Yona, G. and M. Levitt (2002). "Within the twilight zone: a sensitive profile-profile comparison tool based on information theory." J Mol Biol **315**(5): 1257-1275.

Zemla, A. (2003). "LGA: A method for finding 3D similarities in protein structures." Nucleic Acids Res **31**(13): 3370-3374.

Zemla, A., C. Venclovas, J. Moult and K. Fidelis (1999). "Processing and analysis of CASP3 protein structure predictions." Proteins **Suppl 3**: 22-29.

Zhang, C., S. Liu and Y. Zhou (2004). "Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential." Protein Sci **13**(2): 391-399.

Zhang, W., S. Liu and Y. Zhou (2008). "SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model." PLoS One **3**(6): e2325.

Zhang, Y. (2008). "Progress and challenges in protein structure prediction." Curr Opin Struct Biol **18**(3): 342-348.

Zhang, Y., I. A. Hubner, A. K. Arakaki, E. Shakhnovich and J. Skolnick (2006). "On the origin and highly likely completeness of single-domain protein structures." Proc Natl Acad Sci U S A **103**(8): 2605-2610.

Zhang, Y. and J. Skolnick (2005). "TM-align: a protein structure alignment algorithm based on the TM-score." Nucleic Acids Res **33**(7): 2302-2309.

Zhao, F. and J. Xu (2012). "A position-specific distance-dependent statistical potential for protein structure and functional study." Structure **20**(6): 1118-1126.

Zhou, H. and J. Skolnick (2011). "GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction." Biophys J **101**(8): 2043-2052.

Zhou, H. and Y. Zhou (2005). "Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments." Proteins **58**(2): 321-328.

Zhu, J. and Z. Weng (2005). "FAST: a novel protein structure alignment algorithm." Proteins **58**(3): 618-627.

Zhu, Y., G. Chen, F. Lv, X. Wang, X. Ji, Y. Xu, J. Sun, L. Wu, Y. T. Zheng and G. Gao (2011). "Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting

multiply spliced viral mRNAs for degradation." Proc Natl Acad Sci U S A **108**(38): 15834-15839.