PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Refinement by shifting secondary structure elements improves sequence alignments

Jing Tong,[1,2] Jimin Pei,[3] Zbyszek Otwinowski,[1,2] and Nick V. Grishin[1,2,3]*

[1] Department of Biophysics, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390

[2] Department of Biochemistry, University of Texas Southwestern Medical Center at Dallas, Dallas, Texas 75390

[3] Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390

ABSTRACT

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. The web server is available at http://prodata.swmed.edu/sfesa.

Proteins 2015; 83:411–427.

Key words: pairwise alignment; alignment refinement; alignment improvement; contact energy; local secondary structure shifting.

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.[1] With the accumulation of experimentally determined protein structures in the PDB database,[2] homology modeling (also known as template-based modeling) is the most reliable approach to protein structure prediction.[1,3] The 3D structure for a given query sequence can be modeled by aligning the query to one or several protein templates with known structures.[4,5] The model quality relies heavily on the quality of the pairwise or multiple sequence alignment (MSA) between the query and the templates.[6–8] 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.[9–13] 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.[14] For example, the latest significant advance, CNFpred,[15] only has Q-score of 52.4 for the MUSTER benchmark[16] (13.0% average sequence identity by MUSTER's own reference). A number of approaches have been developed for the

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: National Institutes of Health; Grant number: GM094575; Grant sponsor: Welch Foundation; Grant number: I-1505.
*Correspondence to: Nick V. Grishin, Department of Biophysics, University of Texas Southwestern Medical Center at Dallas, Dallas, TX 75390.
E-mail: grishin@chop.swmed.edu
Received 29 August 2014; Revised 25 November 2014; Accepted 10 December 2014
Published online 24 December 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24746

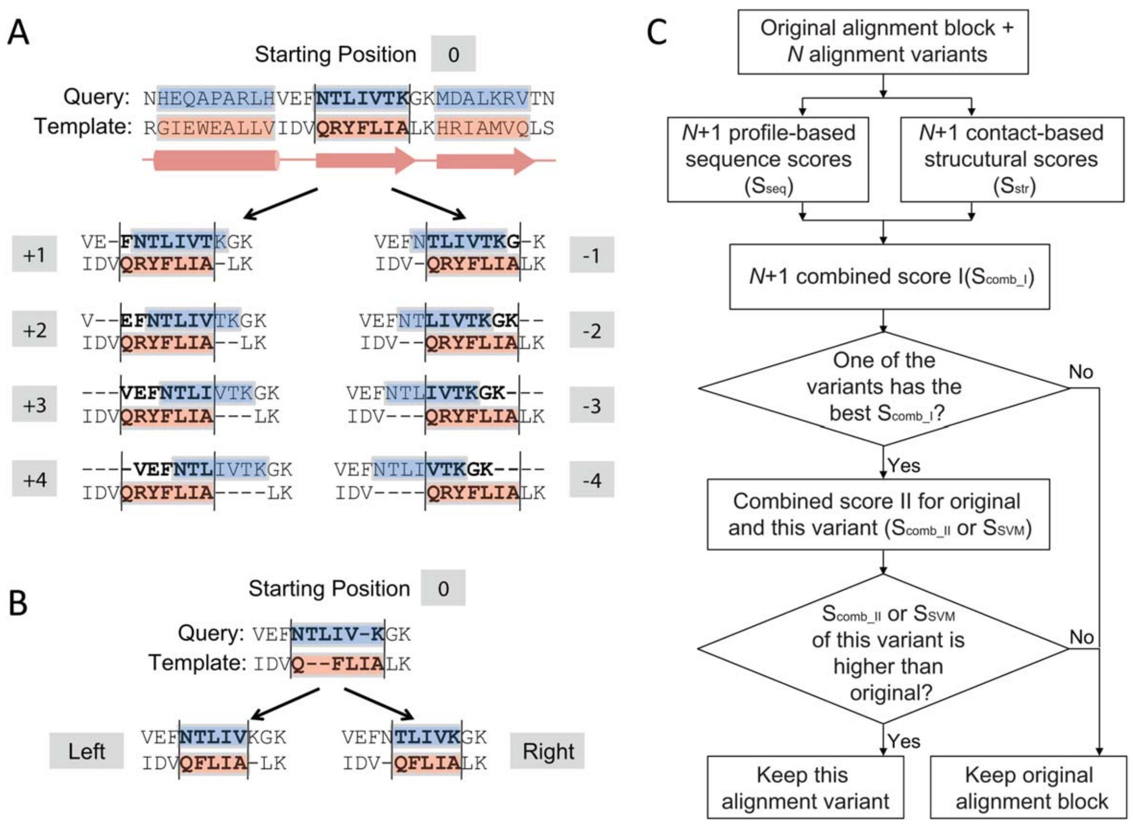
A
Starting Position 0
Query: NHEQAPARLHVEFNTLIVTKGKMDALKRVTN
Template: RGIEWEALLVIDVQRYFLIALKHRIAMVQLS
+1
VE-FNTLIVTKGK
IDVQRYFLIA-LK
-1
VEFNTLIVTKG-K
IDV-QRYFLIALK
+2
V--EFNTLIVTKGK
IDVQRYFLIA--LK
-2
VEFNTLIVTKGK--
IDV--QRYFLIALK
+3
---VEFNTLIVTKGK
IDVQRYFLIA---LK
-3
VEFNTLIVTKGK---
IDV---QRYFLIALK
+4
----VEFNTLIVTKGK
IDVQRYFLIA----LK
-4
VEFNTLIVTKGK----
IDV----QRYFLIALK
B
Starting Position 0
Query: VEFNTLIV-KGK
Template: IDVQ--FLIALK
Left
VEFNTLIVKGK
IDVQFLIA-LK
Right
VEFNTLIVKGK
IDV-QFLIALK
C
Original alignment block + N alignment variants
N+1 profile-based sequence scores (Sseq)
N+1 contact-based strucutural scores (Sstr)
N+1 combined score I(Scomb_I)
One of the variants has the best Scomb_I?
No
Yes
Combined score II for original and this variant (Scomb_II or SSVM)
Scomb_II or SSVM of this variant is higher than original?
No
Yes
Keep this alignment variant
Keep original alignment block

Figure 1

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 eight variants. (**C**). Flowchart of the SFESA method. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

task. Earlier work focused on dynamic programming recursion in construction of a global or local alignment.[17,18] Heuristic methods such as FASTA and BLAST[19,20] were developed to significantly increase the speed of alignment. Subsequently, sequence profiles and hidden Markov models (HMMs)[21] were introduced for comparison of a single sequence and an MSA. Furthermore, profile–profile[22–25] and HMM-HMM[11,12,26] 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.[27,28] The 3D-COFFEE[10] as well as PROMALS3D[12] 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.[15,16,29–34] Information extracted from structure-based alignments of homologous proteins was used to derive amino acid substitution matrices[32,35,36] or position-specific scoring matrices (PSSMs).[37,38] The 3D profile is a position-dependent 20xn scoring matrix derived from protein structures. Such profiles were used to improve sequence-structure alignment.[37,38] Moreover, a 400 × 400 contact-mutation matrix was proposed to improve sequence alignment by using the contacts in template.[39,40] However, how to efficiently and effectively use structural (especially energy-based) information to improve pairwise alignment remains an open question in the field.[41]

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.[42,43] 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.[44–48] 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, that is, α-helices and β-strands)[49,50] alternating with loops (such as turns and coils) connecting these SSEs. SSEs are typically more conserved[51] 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[11] 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,[11] HHpred,[26] and CNFpred[15] on several benchmarks that include both reference-dependent and reference-independent assessment.

MATERIAL AND METHODS

Generation of alignment variants

We partition a pairwise alignment into alignment blocks according to template SSEs defined by the program PALSSE.[52] Short secondary structures (α-helices <8 residues and β-strands <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 [Fig. 1(A)]. 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 Fig. 1(A)]. Similarly, negative shift numbers refer to shifting the query toward 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 [Fig. 1(A)], while the residues neighboring but belonging to a different SSE [such as residue "H" in Fig. 1(A)] are not allowed to shift.

When there are no gaps in the original alignment block, SFESA can generate eight 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 $\pm K$ (up to 4) residues, resulting in eight alignment variants. This strategy is used in SFESA (O) mode (described below).

The second gap treatment strategy is to preprocess gaps before shifting $\pm K$ (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, Fig. 1(B)]. 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[53] iterations. Score for the similarity of residue content in MSA columns is measured by the formula originally implemented in the COMPASS method.[24]

$$S_{\mathrm{seq}} = c_1 \sum_{\mathrm{i}} n_{\mathrm{i}}^1 \ln \frac{Q_{\mathrm{i}}^2}{p_{\mathrm{i}}} + c_2 \sum_{\mathrm{i}} n_{\mathrm{i}}^2 \ln \frac{Q_{\mathrm{i}}^1}{p_{\mathrm{i}}} \tag{1}$$

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:

$$c_1 = \frac{\sum_{\mathrm{i}} n_{\mathrm{i}}^2 - 1}{\sum_{\mathrm{i}} n_{\mathrm{i}}^1 + \sum_{\mathrm{i}} n_{\mathrm{i}}^2 - 2} \tag{2}$$

$$c_2 = \frac{\sum_{\mathrm{i}} n_{\mathrm{i}}^1 - 1}{\sum_{\mathrm{i}} n_{\mathrm{i}}^1 + \sum_{\mathrm{i}} n_{\mathrm{i}}^2 - 2} \tag{3}$$

SFESA further incorporates secondary structure (SS) information into the sequence score. For query, SS is

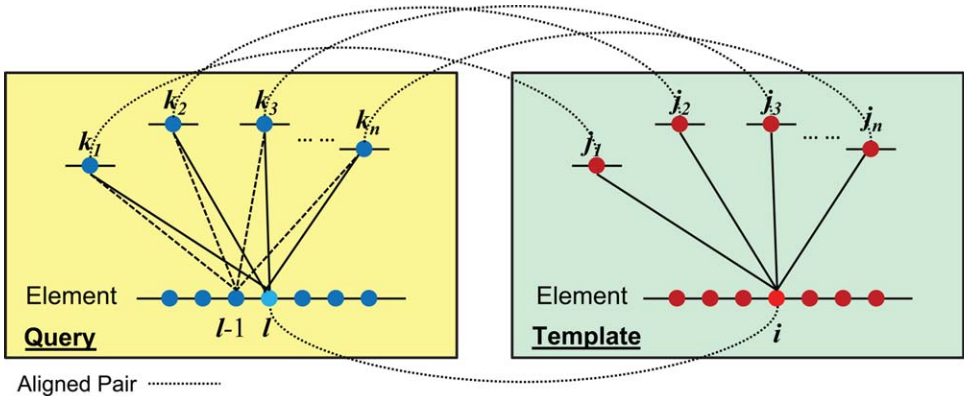

Figure 2

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,. . ., j_n based on template structure. Residue *l* in the query is aligned with *i* and is inferred to be in contact with residues k_1, k_2, k_3,. . ., k_n that are aligned to j_1, j_2, j_3,. . ., 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,. . ., k_n (shown as dashed lines). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

predicted by PSIPRED[54]; for template, SS information in DSSP[50] is used. A three-by-three secondary structure substitution matrix is derived from the structural alignment FAST[55] (considered as query aligned to template) of protein domains from ASTRAL compendium[56] based on SCOP 1.75[57] (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:

$$S'_{seq} = S_{seq} + w_{ss} S_{ss} \tag{4}$$

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, for example, the one derived by Miyazawa and Jernigan,[58] 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 (Fig. 2, 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 2, *i* is one residue in an SSE of template, and j_1, j_2, . . ., 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, . . ., k_n in the query are aligned to j_1, j_2, . . ., j_n, respectively. Thus, the contact residues of *l* are inferred to be k_1, k_2, . . ., k_n. The inferred structure score for residue *l* based on the alignment and the template contact definitions is:

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

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.

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

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):

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

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 <6.5 Å. 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,[59] 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:

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

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 6.5 Å 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 [Fig. 1(C)]. "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 ($S_{\text{comb\_I}}$) 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 $S_{\text{comb\_I}}$) is again compared to the original alignment block by using a different score: $S_{\text{comb\_II}}$ or S_{svm}. If the selected alignment variant also has a better $S_{\text{comb\_II}}$ or S_{svm} than the original alignment block, this alignment variant will replace the original alignment block. Otherwise, the original alignment block is kept.

$S_{\text{comb\_I}}$ and $S_{\text{comb\_II}}$ are linear combinations of sequence score and structure score with different weights:

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

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

where $S'_{\text{seq}}$is the sequence score combined with secondary structure score [Eq. (4)] and $S'_{\text{str}}$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.[60] In our strategy, if any alignment variant passes the first filter (with the top $S_{\text{comb\_I}}$ 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 $S_{\text{comb\_II}}$, 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[55] structural alignments of SCOP domains. Notably, for query, three categories of neural network-predicted rsa values (with PSI-BLAST PSSMs as input)[51] were used based on three equal-sized bins; for template, the real rsa values calculated by NACCESS[61] 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 <20% from the ASTRAL compendium[56] based on SCOP 1.75.[57] All domain pairs with COMPADRE e-value <1e-30 were used. For all domain pairs, we generated DALI structure alignments. Then, we discarded domain pairs with GDT-TS score[62] <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 16,347 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.[63]

Testing benchmarks

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

1. The MUSTER benchmark.[16] This dataset consists of 110 ProSUP protein pairs[64] and 190 pairs selected by the Zhang group with TM-score[65] > 0.5.
2. The SALIGN benchmark.[66] This dataset has 200 protein pairs with about 20% sequence identity, and these pairs have on average about 65 structurally equivalent residues with RMSD < 3.5 Å. Proteins in each pair have very different lengths.
3. The SABmark benchmark.[67] 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).[68] This dataset contains 1682 alignments, and it provides its own reference that is based on the consensus of FSSP structure alignment[69] and CE alignment.[70]

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[71] and CNFpred.[15] 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 (Supporting Information Table SI).

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 article, 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,[59] TM-align,[65] Matt,[72] and Deepalign.[73] (2) The number of alignment blocks improved and deteriorated in different benchmarks after refinement by different SFESA modes. Aforementioned 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[63] score and TM-score[65] 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[8] based on its alignment to the template, and subsequently the score between the query model and the experimental structure was computed.

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 (Supporting Information Fig. S1).

SFESA generates *N* (up to 18) alignment variants [Fig. 1(C)] 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 ($S_{\mathrm{comb\_I}}$) than the original alignment block. If the original alignment block has the best $S_{\mathrm{comb\_I}}$, SFESA keeps it and move to the next alignment block. Otherwise, the alignment variant with the highest $S_{\mathrm{comb\_I}}$ 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 ($S_{\mathrm{comb\_II}}$) or the SVM score (S_{SVM}) (see Materials and Methods). If the selected alignment variant has a higher $S_{\mathrm{comb\_II}}$ 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 toward the C-terminus.

Here we report results of four modes for SFESA (see Materials and Methods for details): SFESA (O) uses up to eight 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 $S_{\mathrm{comb\_II}}$ 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 $\mathrm{S}_{\mathrm{SVM}}$ in the second filter instead of $S_{\mathrm{comb\_II}}$, 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 (Supporting Information Fig. S2).

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)) (Fig. 3). 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 Fig. 3 class b) compared to using parameters trained on class b (the green column in Fig. 3 class b) or using parameters derived from all data (the red column in Fig. 3 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 Fig. 3 class c) and using the parameters trained on the class c (the purple column in Fig. 3 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 [Fig. 4(A)]. 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 six alignments contain more than two deteriorated alignment blocks.

Tests on the MUSTER benchmark

The MUSTER benchmark consists of 300 protein pairs.[16] 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,[69] TMalign,[65] Matt,[72] MUSTER,[16] and DeepAlign.[73] SFESA (O+G+M) and SFESA (O+G+M+S) applied to PROMALS alignments outperform other methods (Table I), regardless of the reference

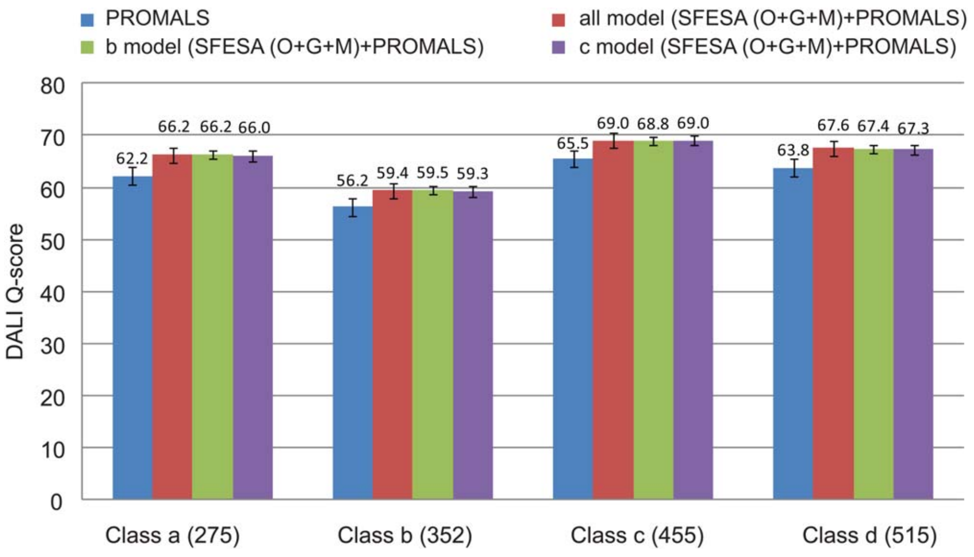

Figure 3

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. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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* values <0.005, Supporting Information Table SII). 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 I). Based on the alignment block level and aligned position level comparisons (Supporting Information Tables SIII–SVIII), 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 (Supporting Information Tables SIII–SVIII).

SFESA (O+G+M+S) significantly improves PROMALS alignments on the MUSTER benchmark, making 138 alignments better and degrading 49 alignments [Fig. 5(A)]. 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 [Fig. 5(B)]. SFESA (O+G+M+S) improves PROMALS by at least 5 *Q*-score on 90 alignments and degrades by this margin on 23 alignments [Fig. 5(B)]. SFESA (O+G+M+S) also improved CNFpred alignments on this benchmark [Fig. 5(C)] 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 [Fig. 5(D)].

Besides the above reference-dependent assessments, reference-independent average TM-score of query models built by MODELLER[8] also shows that SFESA can improve PROMALS, HHpred and CNFpred alignments (Table I, last column). SFESA (O+G+M+S) applied to PROMALS (Table I, 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 [Fig. 4(B)], 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 three alignments contain more than two deteriorated alignment blocks.

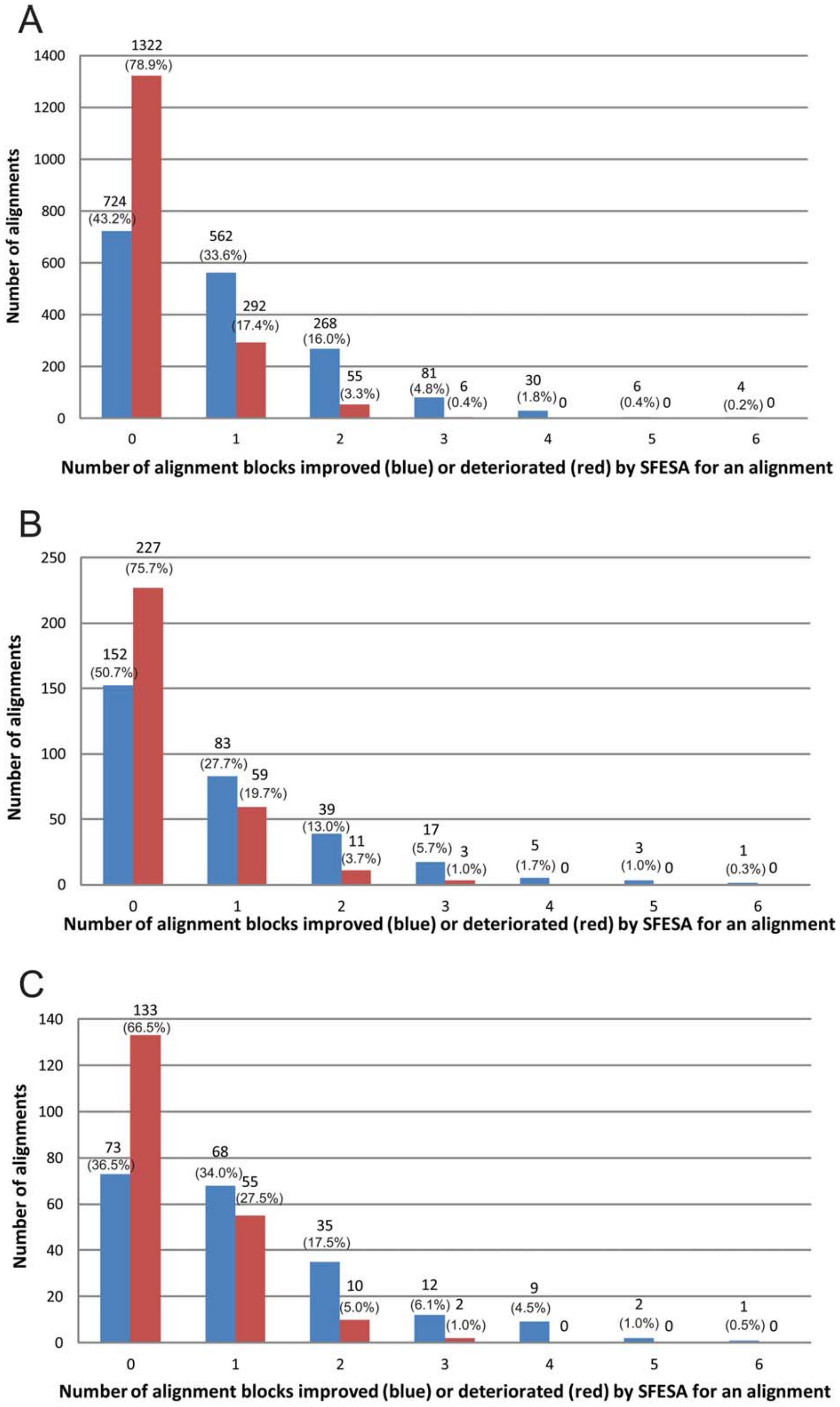

Figure 4

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. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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 II). 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 II, 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* values <0.005, Supporting Information Table SIX) in terms of Q-score based on DALI reference. SFESA modes except SFESA (O) are statistically better than HHpred, despite of an increase of <1.0 *Q*-score on HHpred (Supporting Information Table SIX). For CNFpred, the Wilcoxon signed-rank test shows statistically significant improvement in SFESA (O), SFESA (O+G) and SFESA (O+G+M+S) (Supporting Information Table SIX) 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 (Supporting Information Tables SX–SXIV).

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 [Fig. 4(C)]. The 35 alignments contain two improved alignment blocks while 10 alignments contain two deteriorated alignment blocks [Fig. 4(C)]. 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 [Fig. 4(C)]. 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[67]: 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 III). 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 III). Reference-independent average TM-scores of query models built by MODELLER displayed similar trends (Table III). 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)

Table I
Test on the MUSTER Database

	Reference-dependent (*Q*-score)					Reference-independent (TM-score)
Methods	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**

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.

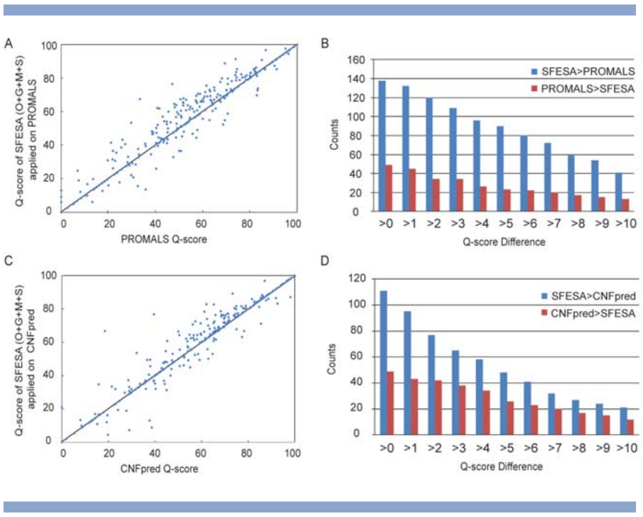

Figure 5

DALI *Q*-score for the MUSTER benchmark. (**A**) Scatter plot of SFESA (O+G+M+S) Q-score (applied to PROMALS) versus 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) versus 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. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

in "superfamilies" set are statistically better than the original PROMALS based on the Wilcoxon signed-rank test (*P* values <0.005, Supporting Information Table SXV). 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 III, Supporting Information Table SXVI–SXIX).

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 (Supporting Information Fig. S3, 209 alignments totally), there are 45 alignments with one improved alignment block while 21 alignments contain one deteriorated alignment block. And there are eight alignments containing more than one improved alignment blocks in contrast to five alignments containing more than one deteriorated alignment blocks. The same analysis on "superfamilies" set (Supporting Information Fig. S4, 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 six 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[68] 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[68] are based on a consensus of FSSP[69] structural alignment and CE alignment.[70] SFESA (O+G+M+S) can increase the Q-score of PROMALS (80.3) and CNFpred (80.5) to 81.3 (Table IV, the first column).

Table II

Test on the SALIGN Database

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**

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.

Table III
Test on the SABmark Database

	Reference-dependent (*Q*-score)		Reference-independent (TM-score)	
Methods on subsets	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**

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.

In addition, we divided the PREFAB alignments into four equal-sized subsets by sequence identity (Table IV). 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), SFESA improvement is less dramatic. According to the Wilcoxon signed-rank test, there are statistically significant improvement in "set1" and "set2" (Supporting Information Table SXX) 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 (Supporting Information Tables SXXI–SXXV).

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 (Supporting Information Figs. S5–S9), 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

Table IV
Test on the PREFAB Database

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

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

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 [Fig. 6(A)], 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) [Fig. 6(B)]. 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

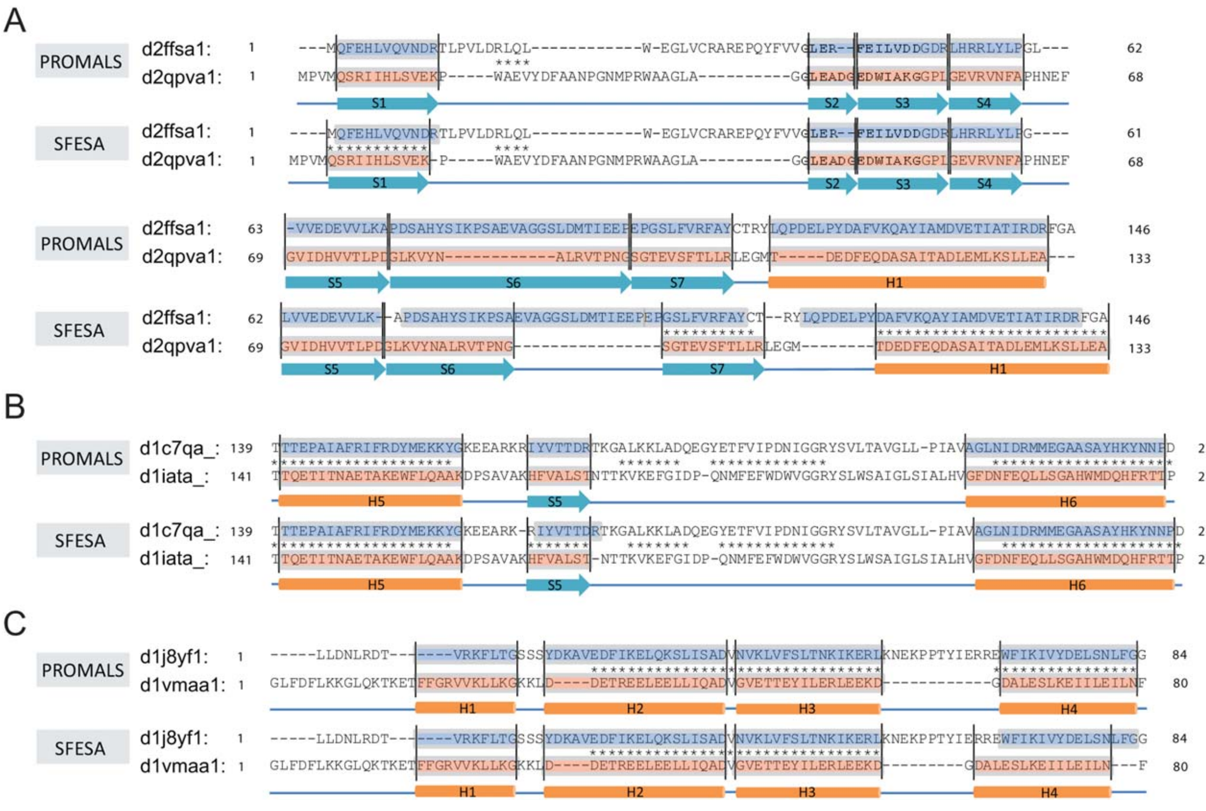
A
PROMALS
d2ffsa1: 1 ---MQFEHLVQVNDRTLPVLDRLQL-----------W-EGLVCRAREPQYFVVGLER--FEILVDDGDRLHRRLYLPGL--- 62
d2qpva1: 1 MPVMQSRIIHLSVEKP-----WAEVYDFAANPGNMPRWAAGLA---------GGLEADGEDWIAKGGPIGEVRVNFAPHNEF 68
S1 S2 S3 S4
SFESA
d2ffsa1: 1 ----MQFEHLVQVNDRTLPVLDRLQL-----------W-EGLVCRAREPQYFVVGLER--FEILVDDGDRLHRRLYLPG---- 61
d2qpva1: 1 MPVMQSRIIHLSVEK-P-----WAEVYDFAANPGNMPRWAAGLA---------GGLEADGEDWIAKGGPIGEVRVNFAPHNEF 68
S1 S2 S3 S4
PROMALS
d2ffsa1: 63 -VVEDEVVLKAPDSAHYSIKPSAEVAGGSLDMTIEEPEPGSLFVRFAYCTRYLQPDELPYDAFVKQAYIAMDVETIATIRDRFGA 146
d2qpva1: 69 GVIDHVVTLPDGLKVYN-----------ALRVTPNGSGTEVSFTLLRLEGMT-----DEDFEQDASAITADLEMLKSLLEA--- 133
S5 S6 S7 H1
SFESA
d2ffsa1: 62 LVVEDEVVLK--APDSAHYSIKPSAEVAGGSLDMTIEEPEPGSLFVRFAYCT--RYLQPDELPYDAFVKQAYIAMDVETIATIRDRFGA 146
d2qpva1: 69 GVIDHVVTLPDGLKVYNALRVTPNG---------------SGTEVSFTLLRLEGM--------TDEDFEQDASAITADLEMLKSLLEA 133
S5 S6 S7 H1
B
PROMALS
d1c7qa_: 139 TTTEPAIAFRIFRDYMEKKYGKEEARKRIYVTTDRTKGALKKLADQEGYETFVIPDNIGGRYSVLTAVGLL-PIAVAGLNIDRMMEGAASAYHKYNNPD 236
d1iata_: 141 TTQETITNAETAKEWFLQAAKDPSAVAKHFVALSTNTTKVKEFGIDP-QNMFEFWDWVGGRYSLWSAIGLSIALHVGFDNFEQLLSGAHWMDQHFRTTP 238
H5 S5 H6
SFESA
d1c7qa_: 139 TTTEPAIAFRIFRDYMEKKYGKEEARK-RIYVTTDRTKGALKKLADQEGYETFVIPDNIGGRYSVLTAVGLL-PIAVAGLNIDRMMEGAASAYHKYNNPD 236
d1iata_: 141 TTQETITNAETAKEWFLQAAKDPSAVAKHFVALST-NTTKVKEFGIDP-QNMFEFWDWVGGRYSLWSAIGLSIALHVGFDNFEQLLSGAHWMDQHFRTTP 238
H5 S5 H6
C
PROMALS
d1j8yf1: 1 -----LLDNLRDT-------VRKFLTGSSSYDKAVEDFIKELQKSLISADVNVKLVFSLTNKIKERLKNEKPPTYIERREWFIKIVYDELSNLFGG 84
d1vmaa1: 1 GLFDFLKKGLQKTKETFFGRVVKLLKGKKLD----DETREELEELLIQADVGVETTEYILERLEEKD-----------GDALESLKEIILEILNF 80
H1 H2 H3 H4
SFESA
d1j8yf1: 1 -----LLDNLRDT-------VRKFLTGSSSYDKAVEDFIKELQKSLISADVNVKLVFSLTNKIKERLKNEKPPTYIERREWFIKIVYDELSNLFGG 84
d1vmaa1: 1 GLFDFLKKGLQKTKETFFGRVVKLLKGKKLD----DETREELEELLIQADVGVETTEYILERLEEKD---------GDALESLKEIILEILN---F 80
H1 H2 H3 H4

Figure 6

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). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

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 >20 *Q*-score units [Fig. 6(C)]. 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 eight 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 [Fig. 6(C)]. 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 (2.81 Å) based on DALI alignment compared with other three helices (H1: 2.21 Å, H2: 1.47 Å and H3: 1.97 Å), 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 (Fig. 7) 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,[74] 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 (Fig. 7). All SFESA modes succeed in recognizing the misaligned alignment block and correcting it by a shift of −2 (Fig. 7).

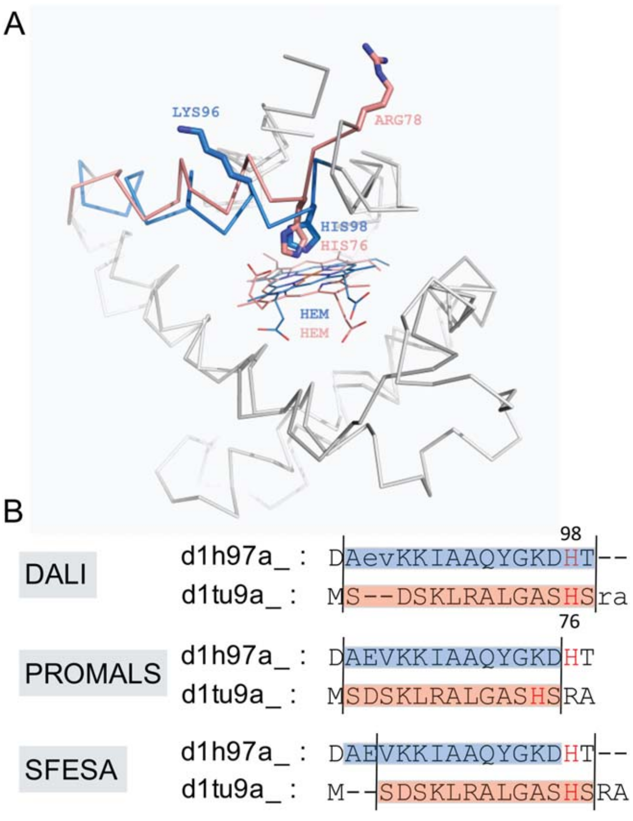
A
LYS96
ARG78
HIS98
HIS76
HEM
HEM
B
98
DALI
d1h97a_ : DAevKKIAAQYGKDHT--
d1tu9a_ : MS--DSKLRALGASHSra
76
PROMALS
d1h97a_ : DAEVKKIAAQYGKDHT
d1tu9a_ : MSDSKLRALGASHSRA
SFESA
d1h97a_ : DAEVKKIAAQYGKDHT--
d1tu9a_ : M--SDSKLRALGASHSRA

Figure 7

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.

DISCUSSION

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 (Supporting Information Fig. S1). SFESA aims to refine pairwise alignments by locally shifting alignment blocks defined by template SSEs to correct misaligned blocks.

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, for example, 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 nonlocal alignment errors.

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 eight 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).

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,[75] such as linear programming,[76,77] branch and bound[78,79] and dead end elimination.[80] 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. First, we used Miyazawa-Jernigan (MJ)[58] contact energy matrix in SFESA (O) and SFESA (O+G). This matrix was designed for threading improved alignments. Second, 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,[81] and they also did not give promising results. These more complex matrices were not designed for the alignment refinement task. Because 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 (6.5 Å) between centers of side chains. We tested several definitions of contacts to deduce our contact energy matrix. The best definition was a fixed distance (6.5 Å) between any side-chain atoms of two residues. A number of distance-based potentials such as DFIRE,[82] DOPE[83] and EPAD energy[84] have been proposed, and some of them consider side-chain orientation-dependent terms.[85,86] 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, that is, 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 versus 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.

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) (Supporting Information

Table SXXVI. 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) (Supporting Information Table SXXVI). 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 <11 contacts have a 0.8 success/failure rate while helices with >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.

ACKNOWLEDGMENT

The authors thank Lisa Kinch and Qian Cong for the helpful discussions and thank Dustin Schaeffer for reading.

REFERENCES

1. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
2. Berman HM, Westbrook J, Feng Z, Gillil G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
3. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol 2008;18:342–348.
4. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 2003; 31:3381–3385.
5. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative protein structure modeling using Modeller. Curr Protoc Bioinform 2006; 15:5.6:5.6.1–5.6.30
6. Peng J, Xu J. RaptorX: exploiting structure information for protein alignment by statistical inference. Proteins 2011;79 (Suppl 10):161–171.
7. Petsko GA. An introduction to modeling structure from sequence. Curr Protoc Bioinform 2006; 15:5.1:5.1.1–5.1.3.
8. Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. Evaluation of comparative protein modeling by MODELLER. Proteins 1995;23: 318–326.
9. Notredame C, Higgins DG, Heringa J. TCoffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302: 205–217.
10. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol 2004;340:385–395.
11. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 2007;23: 802–808.
12. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 2008;36: 2295–2300.
13. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 2005;15:330–340.
14. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
15. Ma J, Peng J, Wang S, Xu J. A conditional neural fields model for protein threading. Bioinformatics 2012;28:i59–i66.
16. Wu S, Zhang Y. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 2008;72:547–556.
17. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
18. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
19. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science 1985;227:1435–1441.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
21. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994;235:1501–1531.
22. Mittelman D, Sadreyev R, Grishin N. Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. Bioinformatics 2003;19:1531–1539.
23. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 2002;315:1257–1275.
24. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 2003;326:317–336.
25. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 2005; 33(Web Server issue):W284–W288.
26. Soding J. Protein homology detection by HMM–HMM comparison. Bioinformatics 2005;21:951–960.
27. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
28. Illergard K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins 2009;77:499–508.
29. Wang Y, Sadreyev RI, Grishin NV. PROCAIN: protein profile comparison with assisting information. Nucleic Acids Res 2009;37:3522–3530.
30. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58:321–328.
31. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 2011;27:2076–2082.
32. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310:243–257.
33. Zhang W, Liu S, Zhou Y. SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. PLoS One 2008;3:e2325.
34. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 2003;19:874–881.
35. Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 2000;13:545–550.
36. Qiu J, Elber R. SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. Proteins 2006;62:881–891.
37. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–85.

38. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
39. Kleinjung J, Romein J, Lin K, Heringa J. Contact-based sequence alignment. Nucleic Acids Res 2004;32:2464–2473.
40. Dong QW, Lin L, Wang XL, Li MH. Contact-based simulated annealing protein sequence alignment method. Conf Proc IEEE Eng Med Biol Soc 2005;3:2798–2801.
41. Pettitt CS, McGuffin LJ, Jones DT. Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 2005;21:3509–3515.
42. Huang YJ, Mao B, Aramini JM, Montelione GT. Assessment of template-based protein structure predictions in CASP10. Proteins 2014;82 (Suppl 2):43–56.
43. Kryshtafovych A, Moult J, Bales P, Bazan JF, Biasini M, Burgin A, Chen C, Cochran FV, Craig TK, Das R, Fass D, Garcia-Doval C, Herzberg O, Lorimer D, Luecke H, Ma X, Nelson DC, van Raaij MJ, Rohwer F, Segall A, Seguritan V, Zeth K, Schwede T. 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 2014;82 (Suppl 2):26–42.
44. Kim C, Tai CH, Lee B. Iterative refinement of structure-based sequence alignments by seed extension. BMC Bioinform 2009;10:210.
45. Thompson JD, Thierry JC, Poch O. RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics 2003;19: 1155–1161.
46. Chakrabarti S, Lanczycki CJ, Panchenko AR, Przytycka TM, Thiessen PA, Bryant SH. Refining multiple sequence alignments with conserved core regions. Nucleic Acids Res 2006;34:2598–2606.
47. Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol 1996;264:823–838.
48. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30:3059–3066.
49. Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 1988;3:71–84.
50. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
51. Huang IK, Pei J, Grishin NV. Defining and predicting structurally conserved regions in protein superfamilies. Bioinformatics 2013;29: 175–181.
52. Majumdar I, Krishna SS, Grishin NV. PALSSE: a program to delineate linear secondary structural elements from protein structures. BMC Bioinform 2005;6:202.
53. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
54. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
55. Zhu J, Weng Z. FAST: a novel protein structure alignment algorithm. Proteins 2005;58:618–627.
56. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. Nucleic Acids Res 2004;32(Database issue):D189–D192.
57. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
58. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins 1999;36:357–369.
59. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
60. Cortes C, Vapnik N. Support-vector networks. Machine Learn 1995;20.
61. Hubbard S, Thornton J. Available at: http://www.bioinf.manchester.ac.uk/naccess/(last accessed date July 28, 2008). Department of Biochemistry and Molecular Biology University College London 1993.
62. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
63. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;(Suppl 3):22–29.
64. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS. ProSup: a refined tool for protein structure alignment. Protein Eng 2000;13: 745–752.
65. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
66. Marti-Renom MA, Madhusudhan MS, Sali A. Alignment of protein sequences by their profiles. Protein Sci 2004;13:1071–1087.
67. Van Walle I, Lasters I, Wyns L. SABmark—a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics 2005;21:1267–1268.
68. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–1797.
69. Holm L, Sander C. Touring protein fold space with dali/FSSP. Nucleic Acids Res 1998;26:316–319.
70. Stoyanova R, Nicholls AW, Nicholson JK, Lindon JC, Brown TR. Automatic alignment of individual peaks in large high-resolution spectral data sets. J Magn Reson 2004;170:329–335.
71. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33(Web Server issue):W244–W248.
72. Menke M, Berger B, Cowen L. Matt: local flexibility aids protein multiple structure alignment. PLoS Comput Biol 2008;4:e10.
73. Wang S, Ma J, Peng J, Xu J. Protein structure alignment beyond spatial proximity. Sci Rep 2013;3:1448.
74. Pesce A, Dewilde S, Kiger L, Milani M, Ascenzi P, Marden MC, Van Hauwaert ML, Vanfleteren J, Moens L, Bolognesi M. Very high resolution structure of a trematode hemoglobin displaying a TyrB10-TyrE7 heme distal residue pair and high oxygen affinity. J Mol Biol 2001;309:1153–1164.
75. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng 1994;7: 1059–1068.
76. Ma J, Wang S, Zhao F, Xu J. Protein threading using context-specific alignment potential. Bioinformatics 2013;29:i257–265.
77. Xu J, Li M, Kim D, Xu Y. RAPTOR: optimal protein threading by linear programming. J Bioinform Comput Biol 2003;1:95–117.
78. Horton P. Tsukuba BB: a branch and bound algorithm for local multiple alignment of DNA and protein sequences. J Comput Biol 2001;8:283–303.
79. Horton P. A branch and bound algorithm for local multiple alignment. Pac Symp Biocomput 1996:368–383.
80. Lukashin AV, Rosa JJ. Local multiple sequence alignment using dead-end elimination. Bioinformatics 1999;15:947–953.
81. Feng Y, Kloczkowski A, Jernigan RL. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. Proteins 2007;68:57–66.
82. Zhang C, Liu S, Zhou Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci 2004;13: 391–399.
83. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
84. Zhao F, Xu J. A position-specific distance-dependent statistical potential for protein structure and functional study. Structure 2012; 20:1118–1126.
85. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 2008;72:793–803.
86. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 2011;101:2043–2052.