

# Effect of using suboptimal alignments in template-based protein structure prediction

Hao Chen<sup>1</sup> and Daisuke Kihara<sup>1,2,3\*</sup>

<sup>1</sup>Department of Biological Sciences, College of Science, Purdue University, West Lafayette, Indiana

<sup>2</sup>Department of Computer Science, College of Science, Purdue University, West Lafayette, Indiana

<sup>3</sup>Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, Indiana

## ABSTRACT

Computational protein structure prediction remains a challenging task in protein bioinformatics. In the recent years, the importance of template-based structure prediction is increasing because of the growing number of protein structures solved by the structural genomics projects. To capitalize the significant efforts and investments paid on the structural genomics projects, it is urgent to establish effective ways to use the solved structures as templates by developing methods for exploiting remotely related proteins that cannot be simply identified by homology. In this work, we examine the effect of using suboptimal alignments in template-based protein structure prediction. We showed that suboptimal alignments are often more accurate than the optimal one, and such accurate suboptimal alignments can occur even at a very low rank of the alignment score. Suboptimal alignments contain a significant number of correct amino acid residue contacts. Moreover, suboptimal alignments can improve template-based models when used as input to Modeller. Finally, we use suboptimal alignments for handling a contact potential in a probabilistic way in a threading program, SUPRB. The probabilistic contacts strategy outperforms the partly thawed approach, which only uses the optimal alignment in defining residue contacts, and also the re-ranking strategy, which uses the contact potential in re-ranking alignments. The comparison with existing methods in the template-recognition test shows that SUPRB is very competitive and outperforms existing methods.

Proteins 2010; 00:000–000.  
© 2010 Wiley-Liss, Inc.

**Key words:** template-based protein structure prediction; threading; fold recognition; suboptimal alignment; knowledge-based residue contact potential.

## INTRODUCTION

Computational protein structure prediction remains a challenging task in bioinformatics and computational biophysics.<sup>1–3</sup> Methods that were developed in the past years can be roughly classified into two categories; ones that use structures of known proteins as a global template (template-based methods)<sup>4–8</sup> and those that use a coarse-grained<sup>9–11</sup> or full atomic protein model to explore the fold space (*ab initio* methods). Identifying appropriate global/local templates in the database is crucially important not only for template-based structure prediction methods but also for *ab initio* methods (with notable exception of methods that attempt to fold a protein model using only the first principles in physics<sup>12,13</sup>) because the strategy to combine known fragment structures has been successful in several *ab initio* methods.<sup>14,15</sup> In recent years, the importance of effective use of template structures is highlighted because more and more protein structures are being solved by the structural genomics projects.<sup>16–19</sup> To capitalize the significant efforts and investments paid on the structural genomics projects, it is urgent to establish effective ways to use solved structures as templates with a special emphasis on developing methods for exploiting remotely related proteins that cannot be simply identified by homology. It is also worthwhile to note the recent interesting discussions on the continuity of the protein structure space that revisit commonality among structures of different overall folds (and thus obviously do not share any ancestral relationship)<sup>20–22</sup> and intriguing attempt to use such structures of different fold for structure modeling.<sup>23</sup> Following these discussions, it might be able to establish a method that uses structure templates for modeling that are not conventionally considered to be similar to a query protein sequence.

Various methods have been proposed in the past for identifying and aligning remotely related templates, which include those that use sequence information extensively,<sup>24,25</sup> those that use structure-related scoring terms,<sup>26,27</sup> ones that use the hidden Markov Mod-

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Institutes of Health; Grant number: GM075004; Grant sponsor: National Science Foundation; Grant numbers: DMS800568, EF0850009, and IIS0915801.

\*Correspondence to: Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN 47907. E-mail: dkihara@purdue.edu

Received 5 March 2010; Revised 24 August 2010; Accepted 27 August 2010

Published online 17 September 2010 in Wiley Online Library (wileyonlinelibrary.com).

DOI: 10.1002/prot.22885

els<sup>28,29</sup> and other machine-learning techniques,<sup>30</sup> and meta-server approaches.<sup>31,32</sup>

In this work, we examine the effect of using suboptimal alignments in template-based protein structure prediction. The algorithms for computing suboptimal alignments themselves have been developed more than a decade ago. Vingron and co-workers<sup>33–35</sup> presented a dynamic programming (DP)-based algorithm that produces suboptimal alignments, each of which is constrained to contain a certain residue pair. Saqi and Sternberg's<sup>36</sup> method produces alternative alignments by penalizing the path of the optimal alignment in the DP matrix. Suboptimal alignments can also be constructed by perturbing alignment parameters.<sup>37</sup> Some other works use thermodynamic partition functions to consider probabilities of suboptimal alignments.<sup>38–41</sup> These probabilities of suboptimal alignments can also be derived from the hidden Markov model.<sup>42,43</sup> Suboptimal alignments can be useful for biological sequence alignments because mathematically optimal alignment of two sequences does not always agree with the biologically correct alignment.<sup>33,34,43,44</sup> It was also shown that consistent regions in optimal and suboptimal alignments correspond to correctly aligned regions in many cases.<sup>34</sup> Previously, we developed a quality assessment score<sup>45</sup> for protein structure models named the SPAD (SubOptimal Alignment Diversity) score, which considers the consistency of the optimal and suboptimal alignments.<sup>46</sup> We showed that the SPAD score has a better correlation to the root mean square deviation (RMSD) of structure models compared with several other scores, which are derived from sequence alignment properties. In our recent subsequent work, we have shown that combination of the SPAD score with several structure-related scores further improves the prediction accuracy of the quality of structure models.<sup>47</sup> Previous works on suboptimal alignments focused on applications to sequence analyses. There was an attempt to use suboptimal alignments in homology modeling,<sup>48</sup> but this work is the first in which suboptimal alignments are thoroughly investigated and implemented in a threading algorithm.

The rest of the article is organized as follows: First, we show that suboptimal alignments frequently are more correct than the optimal alignments. Moreover, we show that they contain a substantial amount of correctly aligned pairs in alignments and correct amino acid residue contacts when tertiary structure models are built based on the alignments. These results indicate the strong potential of suboptimal alignments in template-based protein structure prediction. Next, we show that using suboptimal alignments indeed improves the accuracy of homology models of protein structures constructed by Modeller,<sup>49,50</sup> a popular homology modeling program. Finally, we design a template-based structure prediction method named SUPRB (threading with Suboptimal alignment-based PRoBABilistic residue contact informa-

tion), which uses suboptimal alignment information. The main motivation of using the suboptimal alignments is to explore a better way to handle a two-body amino acid contact potential<sup>51–53</sup> in template-based structure prediction. Amino acid contact potentials have been shown to be effective in protein tertiary structure prediction both in template-based methods<sup>4,53,54</sup> and *ab initio* methods.<sup>11,14,55</sup> Therefore, effective use of contact potentials is one of the keys to assess structural compatibility of the target sequence to a template especially for recognizing very distantly related templates. However, contact potentials have been underused in template-based methods because of the difficulty in obtaining the optimal alignments for the contact potential by DP algorithm. An alternative choice of alignment optimization for contact potentials is by the Monte Carlo approach as proposed by Madej *et al.*<sup>56</sup>; however, DP has an advantage over Monte Carlo in terms of less-expensive computational time to cope with the growing size of a template database. To be able to handle a contact potential by DP, Godzik *et al.*<sup>57</sup> have proposed the frozen approximation of the amino acid contacts, which assumes that a residue position in the target protein has the same contacting residues as the equivalent position in the template structure, and, thus, the contacting residues are not dependent on the alignment. They further proposed partly thawed and defrosted approximation in the program PROSPECTOR where interacting residue pairs are taken from the target protein using a target–template alignment generated without using a contact potential.<sup>4</sup> In this work, SUPRB goes one step further to consider interaction pairs in a probabilistic fashion by counting residue interactions in suboptimal alignments. We demonstrate that our approach, the probabilistic residue contacts, outperforms the partly thawed approximation. Performance comparison with existing threading methods shows that SUPRB is very competitive with the others.

## METHODS

### Benchmark data sets

We use the SALIGN alignment benchmark data set<sup>58</sup> in the analysis of suboptimal alignments and also for optimizing weighting factors for scoring terms of the threading algorithm, SUPRB. The SALIGN data set has 200 pairwise structure-based sequence alignments. These pairs of proteins are structurally aligned and have no more than 40% sequence identity, have at least 100 aligned residues and at least 50% of the residues aligned, and at least 90% of the residues of one chain are covered in the alignment. Another structure alignment program, MAMMOTH,<sup>59</sup> is further used to ensure that at least 50% of the residues in the shorter chain are aligned. The sequence identity of the pairs ranges from 4% to 40%, with an average of 16.1%. The average RMSD of the

pairs is 2.73 Å. This data set was downloaded from <http://salilab.org/suppmat/mamr03/>.

Another data set, the Lindahl and Elofsson's data set (L-E data set),<sup>60</sup> is used for benchmarking SUPRB. The L-E data set consists of 1130 representative proteins, each of which is assigned with a SCOP hierarchical classification of a family, a superfamily, and a fold.<sup>61</sup> Following the SCOP classification, a set of pairwise alignments are constructed in each similarity level, using a protein tertiary structure alignment program, LGA.<sup>62</sup> Aligned protein pairs in the family-level set belong to the same family in SCOP; pairs in the superfamily level belong to the same superfamily but not to the same family; pairs in the fold level set belong to the same fold but not to the same superfamily. This resulted in 1076 target–template pairs for the family level, 1395 for the superfamily level, and 2761 for the fold level. The average sequence identity (and the standard deviation) of alignments are 21.3% (8.06%), 15.2% (3.3%), and 15.2% (3.8%) for the family, the superfamily, and the fold level, respectively. For each protein pair in the same structural similarity level, one protein is considered as the target and another one is considered as a template.

### SUPRB threading algorithm

The novel threading algorithm, SUPRB, uses a sequence–structure compatibility score, which linearly combines five different scoring terms to evaluate fitness of a query sequence to template structures. The five terms are a sequence profile score, a secondary structure matching score, a solvent accessibility score, a main-chain angle propensity, and an amino acid contact potential. Target–template alignments are computed by the global-local DP.<sup>63</sup> A novel feature of SUPRB is that the residue contacts are handled in a probabilistic fashion by considering suboptimal alignments between a target and a template. The detailed implementation of the residue contact term and each scoring terms are described below.

### Sequence profile term

A PSI-BLAST profile is generated following the method described in a Wang and Dunbrack's article.<sup>24</sup> A query sequence is searched against the nonredundant protein sequence database,<sup>36</sup> with an *E*-value threshold of 0.002 for both printing and inclusion and with the maximum iteration of five passes. Low-complexity segments are masked. From the output multiple alignment, too similar sequences (mutual sequence identity  $\geq 98\%$ ) and too distant sequences (sequence identity to the target  $\leq 15\%$ ) are removed. The remaining sequences are weighed by the position-specific independent counts (PSIC) scheme.<sup>64</sup> Sequence *s* at profile position *m* with amino acid type  $a(s_m)$  is weighted with

$$W_s^m = \frac{n_{a(s_m)}^m}{N^m} \quad (1)$$

where  $N^m$  is the total count of sequences in the profile that have the same amino acid type,  $a(s_m)$ , as at the column *m*.  $n_{a(s_m)}^m$  is called the effective count of the amino acid type of  $a(s_m)$  at position *m*, determined by

$$n_{a(s_m)}^m = \frac{1}{\ln(1 - \frac{1}{20})} \ln\left(1 - \frac{F_{a(s_m)}^m}{20}\right) \quad (2)$$

where  $F_{a(s_m)}^m$  is the average number of different amino acid type per column in sequences in the profile that have the same amino acid type as  $s_m$ . Note that the PSIC weighting scheme is different from the Henikoff weighting method,<sup>65</sup> which is used in PSI-BLAST. The Henikoff weighting assigns the same weights to positions in a sequence that form a same subset of sequences.<sup>24</sup> Here, a subset of sequences in a profile is the set of sequences that have a residue at a particular position (i.e., not a gap) of a profile. In contrast, PSIC, assigns lower weights if many sequences have the same amino acid at a particular position. It was shown in that PSIC performed better than the Henikoff weighting in terms of the alignment accuracy and database search sensitivity and specificity.<sup>24</sup>

Based on the weight,  $W_s^m$ , [Eq. (1)], the observed frequency of amino acid type *u* at column *m* in the profile,  $f_u^m$ , is

$$f_u^m = \frac{\sum_{s, a(s_m)=u} W_s^m}{\sum_s W_s^m} \quad (3)$$

The pseudo-count,  $g_u^m$ , for column *m* and amino acid type *u*, is computed as<sup>66</sup>

$$g_u^m = \sum_{v=1}^{20} P_u f_v^m e^{0.32S_{uv}} \quad (4)$$

where  $P_u$  is the background frequency of amino acid type *u*, and  $f_v^m$  is the observed frequency of amino acid type *v* at column *m* [Eq. (3)].  $S_{uv}$  is the BLOSUM62 score between amino acid types *u* and *v*. The observed frequency and the pseudo-counts is combined to the estimated frequency of column *m* and amino acid type *u*,  $Q_u^m$ :

$$Q_u^m = \frac{(n_u^m - 1)f_u^m + 10g_u^m}{n_u^m + 9} \quad (5)$$

where  $n_u^m$  is the effective count of amino acid type *a* at column *m*.

The profile term  $S_{\text{profile}}$  follows the COMPASS scoring scheme.<sup>67</sup> The score between position (column) *i* in the target profile, *a*, and position (column) *j* in the template, *b*, is

$$S'_{\text{profile}}(a_i, b_j) = c_i \sum_{u=1}^{20} n_u^i q_u^j + c_j \sum_{u=1}^{20} n_u^j q_u^i \quad (6)$$

where  $n_u^i$  and  $n_u^j$  are the effective count of amino acid type  $u$  in column  $i$  and column  $j$ .  $q_u^i$  and  $n_u^i$  are called as the log-odds values of  $Q_u^i$  and  $Q_u^j$ , respectively, and calculated by the next equation:

$$q_u^i = \ln \frac{Q_u^i}{P_u}, \quad q_u^j = \ln \frac{Q_u^j}{P_u} \quad (7)$$

Here,  $P_u$  is the background frequency of amino acid type  $u$ , counted from the L-E data set.  $Q_u^i$  and  $Q_u^j$  come from Eq. (5). In Eq. (6),  $c_i$  and  $c_j$  are weighting factors, which are derived as follows:

$$c_i = \frac{\sum_{u=1}^{20} n_u^i - 1}{\sum_{u=1}^{20} n_u^i + \sum_{u=1}^{20} n_u^i - 2}, \quad c_j = \frac{\sum_{u=1}^{20} n_u^j - 1}{\sum_{u=1}^{20} n_u^j + \sum_{u=1}^{20} n_u^j - 2} \quad (8)$$

The score [Eq. (6)]  $S'_{\text{profile}}(a_i, b_j)$  is standardized to  $S_{\text{profile}}(a_i, b_j)$  as follows before computing alignments:

$$S_{\text{profile}}(a_i, b_j) = \frac{S'_{\text{profile}}(a_i, b_j) - \mu}{\sigma} \quad (9)$$

where  $\mu$  and  $\sigma$  are the mean and the standard deviation of scores,  $S'_{\text{profile}}(a_i, b_j)$ , for all possible  $(i, j)$  pairs between the target and the template.

### Secondary structure term

The secondary structure of a target is predicted by SABLE<sup>68</sup> and that of a template is defined by DSSP,<sup>69</sup> both of which have a three-state classification, that is, helix, sheet, and coil (the other regions). The score for the secondary structure matching term  $S_{\text{SS}}$  are adopted from the Wang and Dunbrack's article<sup>24</sup> (Table I). This is an asymmetric matrix, which provides scores for matches between predicted secondary structures in a target sequence with experimentally determined secondary structures in a template. Using Table I, the secondary structure term for an amino acid pair  $(a_i, b_j)$  is

$$S'_{\text{SS}}(a_i, b_j) = \text{Sec}(\text{SS}_p(a_i), \text{SS}_t(b_j)), \quad (10)$$

where  $\text{SS}_p(a_i)$  is the predicted secondary structure of  $a_i$  and  $\text{SS}_t(b_j)$  is the secondary structure of  $b_j$  in the template structure, and  $\text{Sec}(i, j)$  is a score read from Table I. Then,  $S'_{\text{SS}}$  is normalized to  $S_{\text{SS}}$  in the same way as Eq. (9).

### Solvent accessibility term

The relative solvent accessible area (SAA) of the target is also obtained by SABLE. The absolute SAA of the template read from DSSP is normalized using Chothia's extend state accessible area<sup>70</sup> to obtain the relative SAA. A two-state classification is applied for the relative SAA, that is, buried (relative SAA  $\leq 25\%$ ) and exposed (relative SAA  $> 25\%$ ).

**Table I**

The Scoring Matrix for the Secondary Structure Term

Predicted structure in a target ( $SS_p$ )	Experimental structure in a template ( $SS_t$ )		
	Actual helix	Sheet	Coil
Predicted helix	1.38	-3.83	-1.86
Predicted sheet	-3.40	1.54	-1.21
Predicted coil	-1.19	-0.70	0.81

The values are adopted from Wang and Dunbrack (2004).

$$S'_{\text{SAA}}(a_i, b_j) = \begin{cases} \frac{S_{\text{SABLE}}}{10} (2\delta_{\text{SAA}(a_i), \text{SAA}(b_j)} - 1), & \text{if } S_{\text{SABLE}}(a_i) \geq 5 \\ 0, & \text{if } S_{\text{SABLE}}(a_i) < 5 \end{cases} \quad (11)$$

where  $S_{\text{SABLE}}(a_i)$  is the confidence score of SABLE, which ranges from 0 to 9,  $\text{SAA}(a_i)$  and  $\text{SAA}(b_j)$  are Boolean variables, which equal to either buried or exposed, and  $\delta$  equals to 1 when  $\text{SAA}(a_i) = \text{SAA}(b_j)$  and 0 otherwise.  $S'_{\text{SAA}}$  is normalized to  $S_{\text{SAA}}$  in the same way as Eq. (9).

### Main chain angle potential

A coarse-grained main-chain angle propensity of amino acids is considered, which is based on a C $\alpha$ -model of proteins.<sup>8</sup> The torsion angle is defined with four consecutive amino acid (C $\alpha$ ) positions. The torsion angle space is divided into 10 bins (i.e., 36 degrees each), and the statistical potential is computed for each amino acid type by considering torsion angles located before and after the amino acid residue. The angles are sampled from 3704 representative protein structures. Thus, the main-chain angle potential provides  $10 \times 10$  values for each amino acid type. See our previous article for details.<sup>8</sup> The main chain angle term for position  $i$  in the target aligned with position  $j$  in the template is

$$S'_{\text{angle}}(a_i, b_j) = \ln \frac{p(\text{AA}(a_i), \tau_1(b_j), \tau_2(b_j))}{p(\tau_1(b_j), \tau_2(b_j))}, \quad (12)$$

where  $\text{AA}(a_i)$  is the amino acid type of  $a_i$ ,  $\tau_1(b_j)$ ,  $\tau_2(b_j)$  are the bins of preceding and the succeeding torsion angle of residue  $b_j$  in the template, respectively.  $p(a, \tau_1, \tau_2)$  is the probability that the amino acid type  $a$  is observed to have angles  $\tau_1$  and  $\tau_2$ , and  $p(\tau_1, \tau_2)$  is the probability that any amino acid type is observed to have angles  $\tau_1$  and  $\tau_2$ . The table of the score values is made available at the Supporting Information website, as described below.  $S'_{\text{angle}}$  is normalized to  $S_{\text{angle}}$  in the same way as Eq. (9).

### Amino acid contact potential

A two-body statistical amino acid contact potential is used. The potential is computed using the same set of representative protein structures as used for the main-chain angle potential. The quasi-chemical approximation

is used for the reference state,<sup>71</sup> and the contact of a residue pair is defined as 4.5 Å between any side-chain heavy atoms from the residue pair.<sup>8</sup> The raw value is normalized to have the average of 0 and the standard deviation of 1.

For a target protein sequence  $A$  ( $a_1, a_2, \dots, a_{L_A}$ ) aligned with a template protein  $B$  ( $b_1, b_2, \dots, b_{L_B}$ ), residues that are in contact with an amino acid  $a_i$  in the target  $A$  are determined based on the residue contact pattern of the template  $B$ .  $L_A$  and  $L_B$  are the length of the target  $A$  and the template  $B$ , respectively. To state this more precisely, let us introduce two functions,  $T$ , which represents a target–template alignment, and,  $\delta$ , which represents amino acid contacts in a protein structure.

Suppose residues  $a_i$  and  $a_j$  in the target protein  $A$  are aligned with  $b_s$  and  $b_t$  in the template  $B$ , respectively:

$$\begin{aligned} b_s &= T(a_i), & b_t &= T(a_j) \quad \text{or} \\ a_i &= T^{-1}(b_s), & a_j &= T^{-1}(b_t) \end{aligned} \quad (13)$$

The function  $T$  corresponds a residue in a target with a residue in a template, which is aligned with the target residue in a given target–sequence alignment.  $T^{-1}$  is the inverse function of  $T$ .

Let  $\delta$  denotes residue contacts in a protein structure:

$$\delta_{b_s, b_t} = \begin{cases} 1, & \text{when } b_s \text{ and } b_t \text{ are in contact} \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

Then, the residue contact score for a target residue  $a_i$ , which is aligned with a template residue  $b_j$  ( $= T(a_i)$ ), that is,  $S_{\text{contact}}(a_i, b_j)$ , is defined as

$$S_{\text{contact}}(a_i, b_j) = \sum_{l=1}^{L_B} \delta_{b_j, b_l} C(\text{AA}(a_i), \text{AA}(T^{-1}(b_l))) \quad (15)$$

Here,  $L_B$  is the length of the template protein  $B$ , the function  $\text{AA}$  denotes the amino acid type (e.g., Ala, Trp, etc.) of the specified amino acid in the protein, and  $C(a, b)$  is the contact potential value for the amino acid pair  $(a, b)$ . Hence, the contacting residues for  $a_i$  are taken from the target, under the assumption that the residue contact pattern is conserved between the target and the template. Thus, the residue contacts are taken from the template  $B$ , whereas the types of contacting residues are taken from the target itself.

Equation (15) provides the contact score given a target–template alignment. However, we need to design a way to implement the contact potential in the DP-based threading algorithm because the DP cannot optimize the alignment for the two-body contact potential,<sup>4,5</sup> which considers long-range interactions. We test the following two strategies, both of which take advantage of the use of suboptimal alignments:

### Re-ranking strategy

In this strategy, we first compute optimal and suboptimal alignments for a template and a target using the

compatibility score excluding the amino acid contact term. Then, the alignments are re-ranked by the score with the contact term. The compatibility score,  $S(a_i, b_j)$ , without the contact term used for the initial alignment is as follows for a target residue  $a_i$  and a template residue  $b_j$ :

$$\begin{aligned} S(a_i, b_j) &= w_{\text{profile}} S_{\text{profile}}(a_i, b_j) + w_{\text{SS}} S_{\text{SS}}(a_i, b_j) \\ &\quad + w_{\text{SAA}} S_{\text{SAA}}(a_i, b_j) + w_{\text{angle}} S_{\text{angle}}(a_i, b_j) \end{aligned} \quad (16)$$

where  $S_{\text{profile}}(a_i, b_j)$ ,  $S_{\text{SS}}(a_i, b_j)$ ,  $S_{\text{SAA}}(a_i, b_j)$ , and  $S_{\text{angle}}(a_i, b_j)$  are the terms for the profile, the secondary structure matching, the solvent accessibility, and the main-chain angle potential, respectively. All of these scores are rescaled to have the standard normal distribution as proposed by Wang and Dunbrack.<sup>24</sup>  $w_{\text{profile}}$ ,  $w_{\text{SS}}$ ,  $w_{\text{SAA}}$ , and  $w_{\text{angle}}$  are the weighting parameters for the corresponding scoring terms. The values for the weighting parameters range from 0 to 1, and they sum up to 1:

$$w_{\text{profile}} + w_{\text{SS}} + w_{\text{SAA}} + w_{\text{angle}} = 1 \quad (17)$$

In the re-ranking stage, the optimal and suboptimal target–template alignments are re-ranked by the compatibility score with the contact term:

$$S_{\text{rerank}}(a_i, b_j) = (1 - w_{\text{contact}}) S(a_i, b_j) + w_{\text{contact}} S_{\text{contact}}(a_i, b_j) \quad (18)$$

where  $0 < w_{\text{contact}} < 1$ . To compute the  $S_{\text{contact}}$  term, the contacting amino acids for each residue,  $a_i$ , in the target are taken from the target itself according to the precomputed alignment [Eqs. (13) and (14)]. We use Eq. (18) [integrated with Eq. (15)] to recompute the score of each residue  $a_i$  in the target aligned with  $b_j$  in the template, without changing the alignment itself. Thus, the compatibility score for each of the optimal and suboptimal alignments is updated, and, thus, the order of the alignments changes. Finally, we choose the alignment with the best score as the new optimal alignment for the target and the template.

### Probabilistic handling of residue contacts

This strategy uses suboptimal alignments for considering residue contacts in a probabilistic fashion. In the first pass, optimal and suboptimal alignments are generated by Eq. (16), which does not contain the contact term. Because these initial alignments provide a tertiary structure of the target protein, we can extract contacting residues for a given amino acid position from the target protein itself.

Using the set of optimal and suboptimal alignments, the contact term, Eq. (15), is modified as follows:

$$S_{\text{contact}}(a_i, b_j) = \frac{1}{N} \sum_{n=1}^N \sum_{l=1}^{L_B} \delta_{b_j, b_l} C(\text{AA}(a_i), \text{AA}(T_n^{-1}(b_l))), \quad (19)$$

where  $N$  is the total number of suboptimal alignments considered. A new index,  $n$ , is introduced, which specifies the rank of each suboptimal alignment. Thus, residue contacts are considered in a probabilistic way. For example, the number of times two amino acids  $a_i$  and  $a_j$  in the target are considered to be in contact with the probability of

$$\frac{1}{N} \sum_{n=1}^N \delta_{T_n(a_i), T_n(a_j)} \quad (20)$$

In the second pass of computing target–template alignments, Eq. (18) with the probabilistic contact [Eq. (19)] is applied to update optimal and suboptimal alignments by DP algorithm. Because the score of aligning each residue pair in the target and the template,  $(a_i, b_j)$ , for  $1 \leq i \leq L_A$  and  $1 \leq j \leq L_B$  to be used in the DP has changed by adding the probabilistic contact term from the original score [Eq. (16)], the resulting optimal and suboptimal alignments will change. Now the updated optimal and suboptimal alignments change contacting residues for target residues, and, thus, the contact term [Eq. (19)] is updated by the new optimal/suboptimal alignments at each iteration. Therefore, a subsequent iteration will further change the alignments. The alignment computation is iterated until the fourth pass is completed or the optimal alignment converges. The convergence of the alignments is measured by the ALD score (0.05 or lower), which compares two alignments based on the paths in the DP matrix.<sup>46</sup>

### Suboptimal alignments

The suboptimal alignments are computed with the algorithm proposed by Vingron and Argos.<sup>34</sup> The total number of  $0.01 * l_{\text{target}} * l_{\text{template}}$  suboptimal alignments are computed.  $l_{\text{target}}$  and  $l_{\text{template}}$  are the length of target and the template proteins, respectively. The algorithm first computes alignments in the forward direction as usual and also in the backward direction using the DP algorithm. By design, the alignment score stored at each cell in the forward DP matrix (i.e., the DP matrix used in the forward alignment) indicates the best score to align the two proteins up to that position in the two proteins. In the same way, each cell in the backward DP matrix contains the best score to align the two proteins backward up to that position. Next, a combined matrix is computed by summing up the scores at each cell,  $(i, j)$ , in the two DP matrices and then subtracting the pairwise score for aligning residue  $i$  and  $j$ . Now the score at each position  $(i, j)$  in the combined matrix is the best possible score for aligning the two proteins with the condition that residue  $i$  and  $j$  should be aligned with each other. Finally, the scores in the combined matrix are sorted in the descendent order, and the whole alignment that trespass each selected position is constructed by tracing back

the forward and backward alignments in the forward and backward DP matrices, respectively.

### Parameter optimization and the benchmark on the testing data sets

Seven parameters (five weighting factors, an open gap penalty, and an extension gap penalty) are optimized based on the ProSup alignment data set.<sup>72</sup> Combinations of the values of parameters are essentially explored exhaustively for most of the combinations of the parameter values with an interval of 0.1. The combination that gives the highest alignment accuracy that is computed as the total number of correctly aligned amino acid pair is chosen.

For testing SUPRB with the optimized parameters, we use two independent data sets from the training set (the ProSup set): the SALIGN data set<sup>58</sup> is used for examining the alignment accuracy, and the Lindahl and Elofsson (L-E) data set<sup>60</sup> is used for testing both the alignment accuracy and the template recognition accuracy. The alignment accuracy on the SALIGN set is defined as the fraction of the correctly aligned residue pairs, which are determined by structural alignments by TM-Align.<sup>73</sup>

In the template recognition, the raw sequence–structure compatibility score,  $S_{\text{raw}}$ , is normalized by the length of the template and the target:

$$S_{\text{norm}} = \frac{2 * S_{\text{raw}}}{\frac{l_{\text{target}}}{l_{\text{template}}} + \frac{l_{\text{template}}}{l_{\text{target}}}}, \quad (21)$$

which are then ranked by the Z-score:

$$Z = \frac{S_{\text{norm}} - \mu}{\sigma} \quad (22)$$

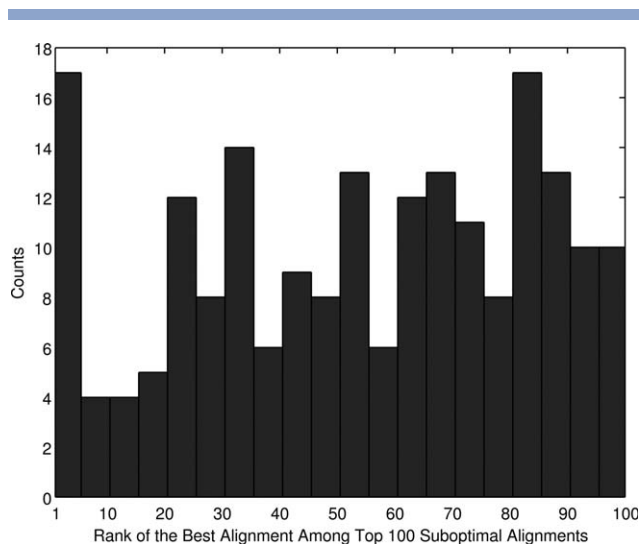
where  $\mu$  is the mean and  $\sigma$  is the standard deviation computed for all the templates in the L-E set.

### Availability of the software

The executable file of SUPRB, the main-chain angle potential, the contact potential, and the template-based models analyzed in this study are made available at our website: <http://kiharalab.org/suprb>. In addition, the source code for computing the SPAD score is also made available at <http://kiharalab.org/subalignment/>.

## RESULTS

First we show that suboptimal solutions of target–template alignments contain significant number of correctly aligned residues and correct residue contacts. Then, we show that suboptimal alignments also improve accuracy of structure models when applied to a homology modeling program, Modeller.<sup>49</sup> Finally, we examine the performance of the SUPRB threading algorithm that uses



**Figure 1**

The rank of the most correct alignments within the top 100 scoring alignments. The  $x$ -axis is the rank of the most correct alignment in the top 100, and the  $y$ -axis is the count of the most correct alignment occurred at that rank in the 200 target–template alignments in the SALIGN data set. The accuracy of an alignment is measured by the fraction of correctly aligned residues in the alignment.

suboptimal alignments for handling a residue contact potential.

### Correctly aligned residues and contacting pairs in suboptimal alignments

The accuracy of template-based prediction depends on the amount of residue pairs from the target and the template, which are correctly matched in the alignment. Previous works<sup>33,34,43,44</sup> showed that there are often cases that a suboptimal alignment of a sequence pair have more correctly aligned residues than the optimal (i.e., top-scoring) alignments. Rather than a sequence alignment score (e.g., a BLOSUM matrix<sup>74</sup>) used in the previous works, we use a sequence-structure compatibility score [Eq. (16)] to align target sequences to template structures in the SALIGN data set. We count correctly aligned residues and correct contacting pairs found in up to the top 100 scoring target–template alignments.

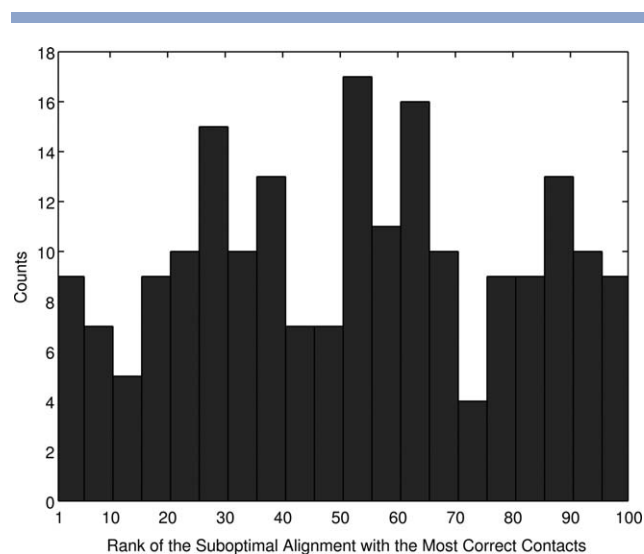
On average, only 54.1% of residues in the target are correctly aligned in the optimal alignment ( $x = 1$ ) (Fig. S1, Supporting Information). Considering more suboptimal alignments increases the number of correctly aligned residues. When 100 suboptimal alignments are taken into account, the fraction of the correctly aligned residues reach to 67.3%, which is a 13.2% point increase compared with the optimal alignment.

In Figure 1, we ask a different question regarding the alignment accuracy of suboptimal alignments: After all,

which alignment among the top 100 scoring alignments is the most correct? To our surprise, overall, the accuracy of the suboptimal alignments seems to be irrelevant to their ranks. The highest count<sup>17</sup> is observed both at the rank of 1–5, but it ties with the count at the ranks of 80–85. Thus, correctly aligned residue pairs frequently occur in low-ranked suboptimal alignments, and, moreover, the accuracy and the rank of the alignments do not show correlation. Alignments at the rank of 1–5 tend to be relatively accurate (Fig. 1), but accumulating the top five alignments only yields a marginal gain of the accuracy of 55.7% (Fig. S1, Supporting Information). Therefore, not only the top few suboptimal alignments but also lower-ranked suboptimal alignments will contribute in improving the alignment accuracy.

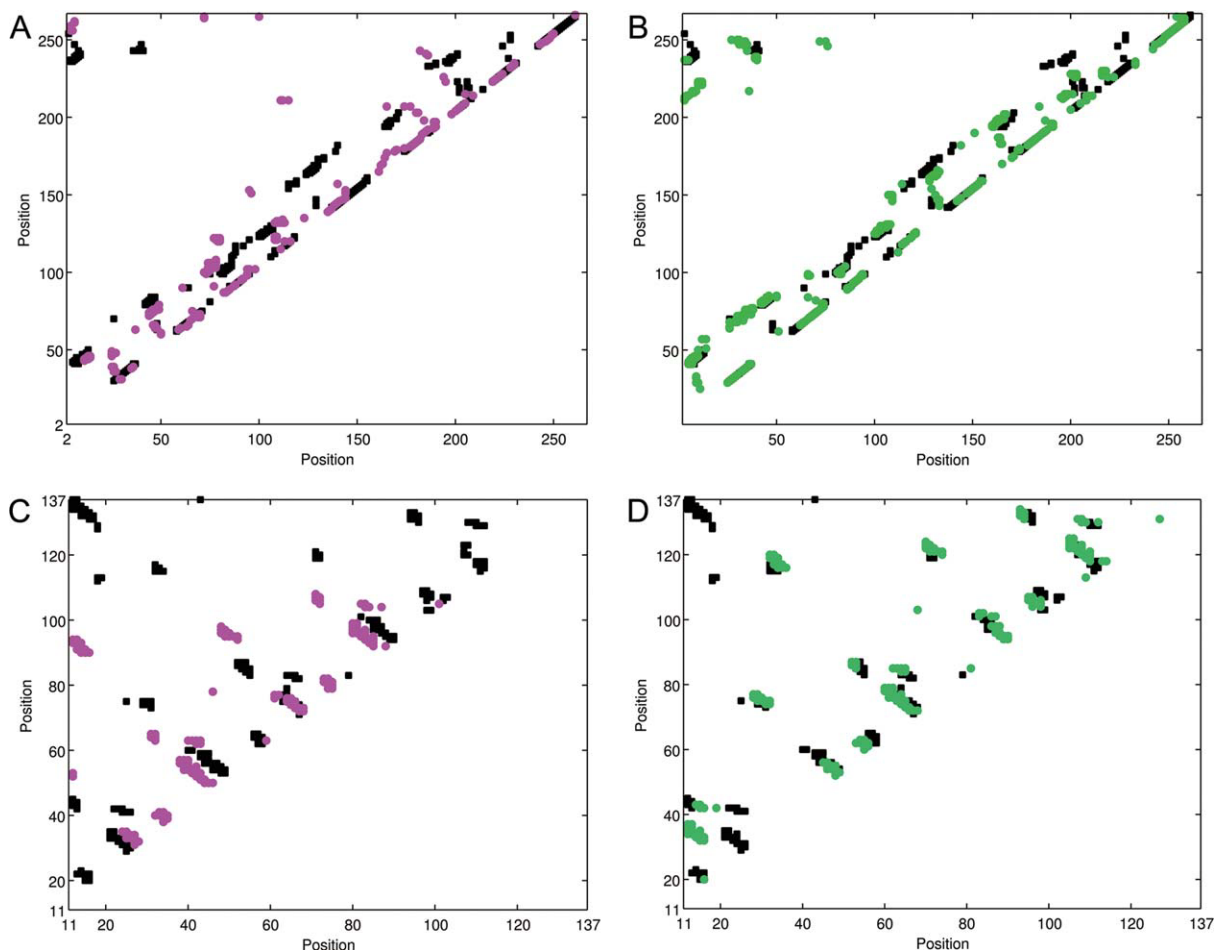
Next, we examine the accuracy of the optimal and suboptimal alignments in terms of the residue contacts. Contacting residue pairs observed in a template protein structure are transferred to a target protein based on the target–template alignment, and the accuracy of the contacts in the target protein is evaluated. On average, the optimal alignment contains 37.0% of actual contacts in target proteins and considering up to the top 100 scoring alignments increases the fraction to 49.0% (Fig. S2, Supporting Information). Figure 2 shows the histogram of the rank of the alignment from which the largest number of correct residue contacts is transferred to the target. This histogram again illustrates that the accuracy of alignments (in terms of the contacting residue pairs) has no correlation to the rank at all.

Figure 3 shows two examples that a suboptimal alignment has more correct residue contacts than in the opti-



**Figure 2**

The rank of the alignment that has the largest number of correct residue contacts. The 200 target–template alignments in the SALIGN data set are used.

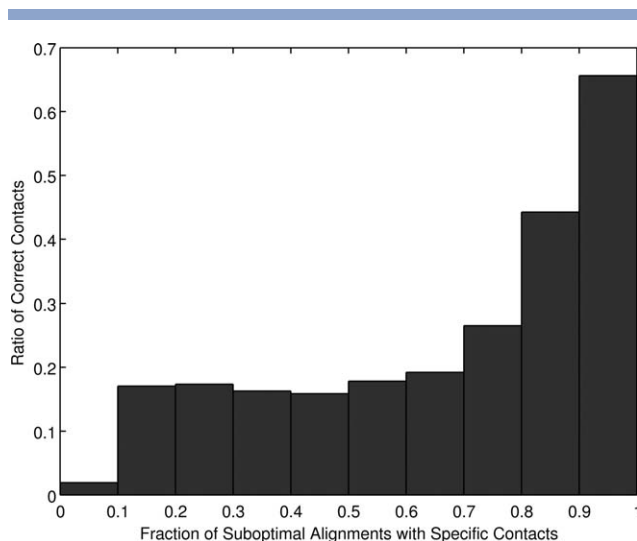


**Figure 3**

Residue contact maps of target proteins indicated by target–template alignments. Two contact maps are compared for a target protein, one indicated from the top-scoring alignment and another one from a suboptimal alignment, which is the most correct in terms of residue contacts. **A:** The contact map of the target protein, 1ad1B, indicated by the top-scoring alignment with a template, 1dioA. Black, the actual contacting residues of 1ad1B; purple, contacting residues indicated from the alignment with 1dioA. **B:** The contact map of 1ad1B indicated by the 100th suboptimal alignment with the 1xyfA (green). **C:** The contact map of 1barA; black, the actual contacting residues of 1barA; purple, contacts indicated by the top-scoring alignment with a template, 1xyfA. **D:** Residue contacts of 1barA indicated by the 70th suboptimal alignment (green).

mal alignment. The first example is the optimal alignment and the 100th alignment between 1ad1B and 1dioA, both of which have the TIM barrel fold. The optimal alignment [Fig. 3(A)] only covers 11.1% of the residue contacts, whereas the suboptimal alignment found at the 100th [Fig. 3(B)] captures a larger number of correct contacts in  $\alpha$ -helices and between parallel  $\beta$ -strands, resulting in the increase of the correct contact coverage to 30.5%. The second example is 1barA aligned with 1xyfA ( $\beta$  trefoil fold). The optimal alignment [Fig. 4(C)] has a shift in the alignment that causes a small coverage of 5.1% of the actual contacts. On the other hand, the 70th suboptimal alignment correctly captures most of the contacts between antiparallel  $\beta$ -strands, resulting in the correct contact coverage of 23.2%.

As Figure 1 shows, the rank of individual suboptimal alignment does not tell the accuracy of residue contacts implied from the alignment. In Figure 4, we investigate whether residue contacts that occur more frequently among the suboptimal alignments tend to be more accurate than less frequent ones, which turned out to be true. It is shown in Figure 4 that the number of occurrences of residue contacts in suboptimal alignments correlates well with the accuracy. 65.6% of the residue contacts, which occur in more than 90% of the suboptimal alignments, are correct. This value is larger than the average accuracy of the residue contacts indicated by the optimal alignments (i.e., the number of correctly predicted residue contacts among the residue contacts implied from the optimal alignments), which is 48.2%.



**Figure 4**

Fraction of the actual contacting residue pairs of target proteins relative to the occurrence of the contacts among the suboptimal alignments. For each target–template pairs, top  $0.01 \times M \times N$  alignments ( $M$  and  $N$  are the length of the target and the template, respectively) are computed, and the occurrence of residue contacts among the suboptimal alignments are counted. The values are computed for each target–template pair in the SALIGN data set and averaged.

### Using suboptimal alignments for template-based modeling

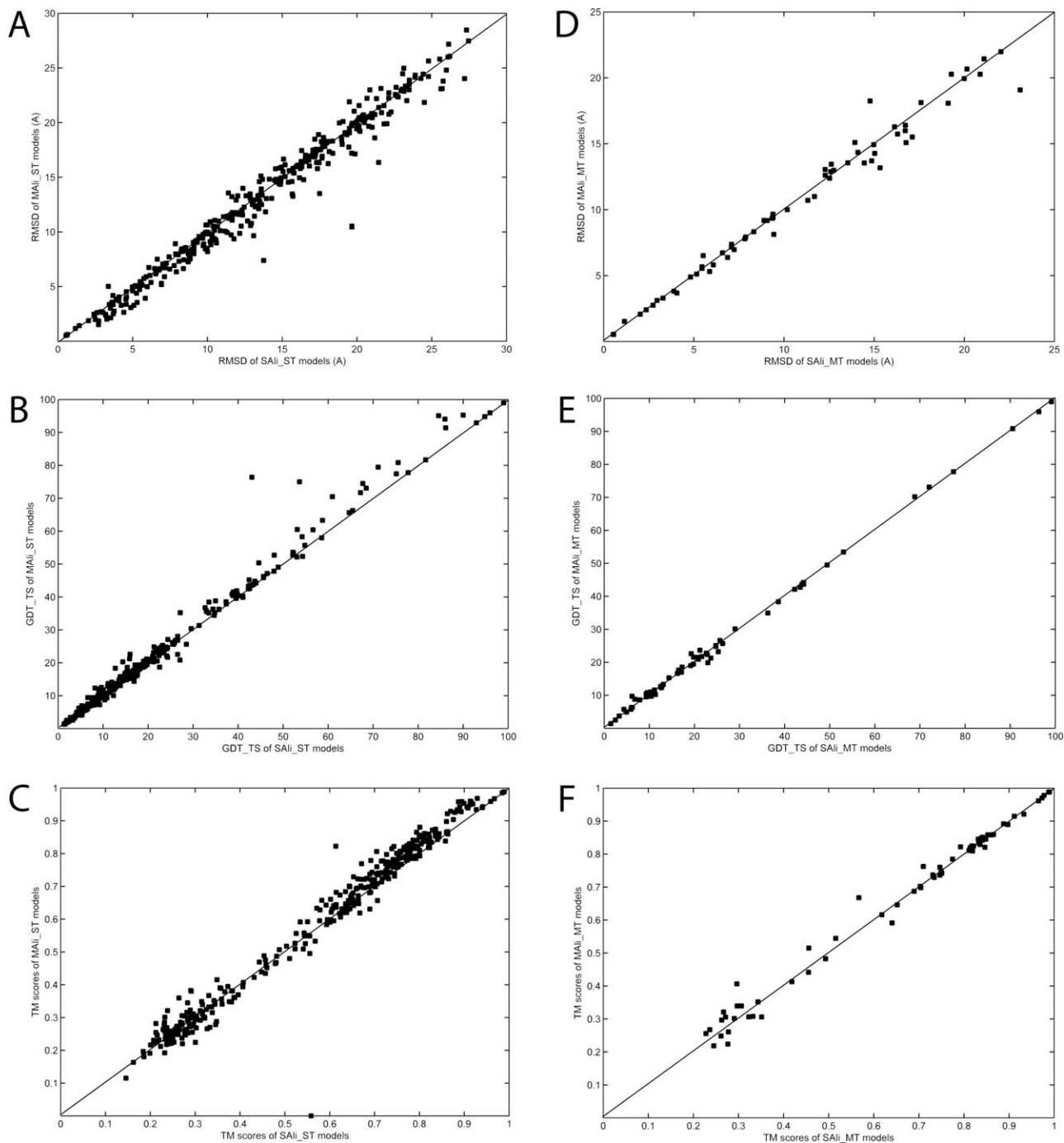
In this section, we show practical usefulness of the suboptimal alignments by using them in a popular homology modeling tool, Modeller.<sup>49</sup> Conventionally, homology modeling procedure uses only the optimal alignment between a target and a template to build the final three-dimensional coordinates model. On the other hand, several recent studies investigated the use of multiple templates to improve the model.<sup>75–77</sup> However, improving the model accuracy with multiple templates is not trivial because incorporation of additional templates can lead to deterioration of the model quality.<sup>75</sup> We compare models that are built based on the single (optimal) alignment between the target and a single template, termed SAli-ST (single alignment-single template) models, with models based on multiple alignments (optimal and suboptimal alignments), termed MAlI-ST (multiple alignment-single template) models. We also examine models that are based on multiple templates, that is, comparison between SAli-MT (single alignment-multiple template) models (i.e., the optimal alignment for each of multiple templates) and the MAlI-MT (multiple alignments for each of multiple templates) models. For the MAlI-ST and MAlI-MT models, four suboptimal alignments are added on top of the optimal alignment for each template as the input to Modeller (i.e., total of five alignments are used). For this testing, we used target

proteins used in the CASP7 (critical assessment of techniques for protein structure prediction).<sup>78</sup> Five template structures for a target are selected by considering consensus among those selected by our own threading method, a prior version of SUPRB, SP426, RAPTOR\_ ACE,<sup>79</sup> and FOLDPRO,<sup>80</sup> of which the last four are the server predictions that are made available through the CASP7 official website (<http://www.predictioncenter.org/casp7/Casp7.html>). Such a consensus approach is commonly used and some of them have been successful in the past CASP experiments.<sup>81,82</sup> The alignment between a target and a template are computed with the probabilistic residue contact strategy [Eq. (19)]. Unaligned regions at both ends of the targets are removed if they are more than 20 residues long because Modeller tends to generate unrealistic floppy conformation for such regions. All the models analyzed are made available at our website (<http://www.kiharalab.org/suprb>). The models are evaluated by the RMSD, the GDT-TS score,<sup>62</sup> and the TM-score<sup>73</sup> to the native structure. The RMSD are computed by LGA.<sup>62</sup>

First, we examine the effect of suboptimal alignments in structure prediction with a single template by comparing the SAli-ST and the MAlI-ST models [Fig. 5(A–C) and Table II]. Using suboptimal alignments (MAlI-ST models) improves the RMSD over the SAli-ST models in 68.1% (261 of 383) cases. Although the average RMSD value of the MAlI-ST models, 13.33 Å, may not seem largely improved from that of the SAli-ST models (13.73 Å), significant improvement of the MAlI-ST model over the corresponding SAli-ST model, for example, an improvement larger than 2 Å, is often observed [Fig. 5(B)]. When models are evaluated by the GDT-TS and the TM-score, improvement by MAlI-ST models over SAli-ST models is observed in 65.0% and 68.9% of the cases, respectively (Table II). Figure 5(B,C) shows that improvement by MAlI-ST is larger than deterioration in general both in the GDT-TS [Fig. 5(B)] and the TM-score [Fig. 5(C)].

Two examples in Figure 6 illustrate how suboptimal alignments improve the RMSD in MAlI-ST models. Models of two CASP7 targets, T0308 (PDB code: 2h57) and T0345 (PDB code: 2he3) are shown. In the first example, the MAlI-ST model [Fig. 6(A)] of T0308 has an RMSD of 5.2 Å, whereas the RMSD of the SAli-ST model is of 6.9 Å [Fig. 6(B)]. Compared with the structure-based alignment between the target and the template (PDB code: 2fol), the optimal alignment has a larger shift in the region around the position 105–140 [Fig. 7(A)], which corresponds to the loop region indicated with a circle in Fig. 6(A).

For the second example, the MAlI-ST model for T0345 [Fig. 6(D)] shows a large RMSD improvement from 13.8 Å by the SAli-ST model [Fig. 6(C)] to 7.3 Å. The large unstructured loop region at around residue 105–145 built by Modeller in the SAli-ST model [Fig. 6(C)] is due to



**Figure 5**

Comparison of suboptimal alignment-based models and single optimal alignment-based models. **A:** RMSD; **B:** GDT-TS score; **C:** TM-score of Sali-ST models and the Mali-ST models; **D:** RMSD; **E:** GDT-TS score; **F:** TM-score of Sali-MT models and the Mali-MT models.

the large alignment shift of that region in the optimal alignment [Fig. 7(B)]. However, interestingly, suboptimal alignments have a shift to the opposite direction in that region, which cancelled out with the shift by the optimal alignment in the modeling.

We have also examined application of suboptimal alignments to structure models using multiple templates by comparing Sali-MT and Mali-MT models [Fig. 5(D-F) and Table II]. We did not observe significant improvement by the Mali-MT models. The average RMSD value of the

**Table II**

Comparison of the Single Alignment and Multiple Alignment Models for CASP7 Targets

	Wins/average (Å)	
Models built on a single template	SAlI-ST <sup>a</sup>	MAli-ST <sup>a</sup>
RMSD	122/13.73	261/13.33
GDT-TS	134/19.97	249/20.81
TM-score	119/0.570	264/0.585
Models built on multiple templates	SAlI-MT <sup>b</sup>	MAli-MT <sup>b</sup>
RMSD	32/10.92	32/10.80
GDT-TS	23 <sup>c</sup> /25.16	38/25.40
TM-score	32/0.644	32/0.650

<sup>a</sup>SAlI-ST, structure models built on the optimal alignment with a single template structure; MAli-ST, models built on five (optimal and four suboptimal) alignments with the same single template structure as used in the SAlI-ST models. Wins count the number of cases either SAlI-ST or MAli-ST model has a smaller RMSD, a larger GDT-TS, or a larger TM-score, than the other. The number of models examined is about five times more than the SAlI-MT and MAli-MT models since here models based on each template (thus five models) for a target are examined.

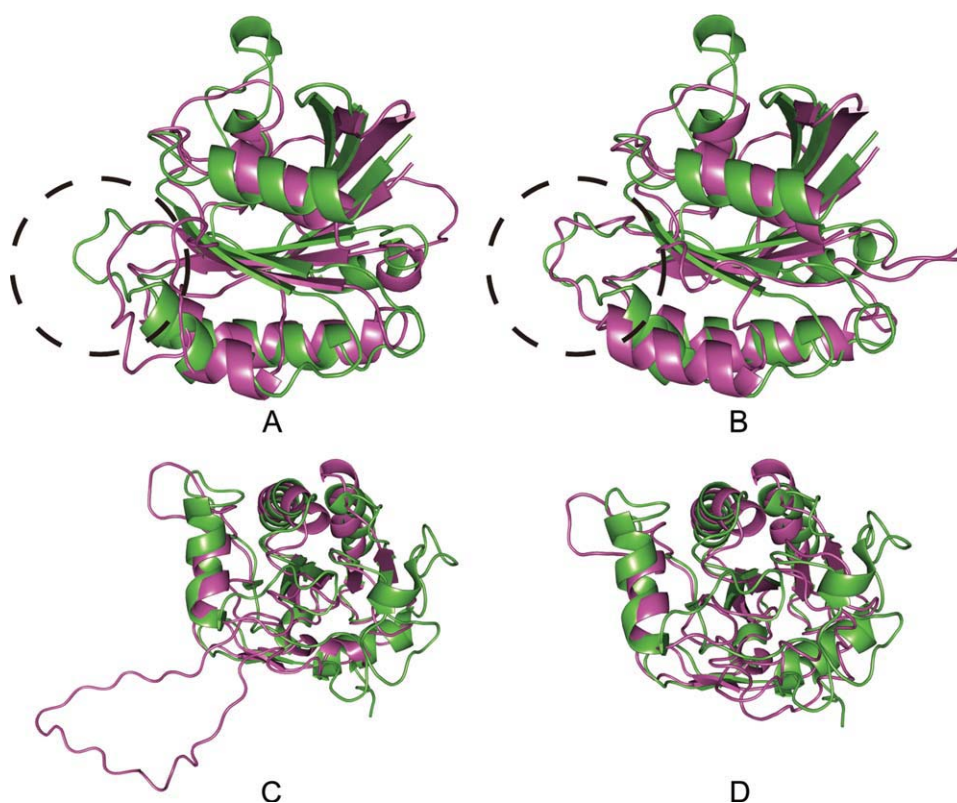
<sup>b</sup>SAlI-MT, structure models built on the optimal alignment with five template structures; MAli-MT, models built on five alignments with the same five template structures as used in the SAlI-MT models. Targets were discarded when Modeller could not compute MAli-MT models.

<sup>c</sup>There are three ties between SAlI-MT and MAli-MT models in terms of the GDT-TS score.

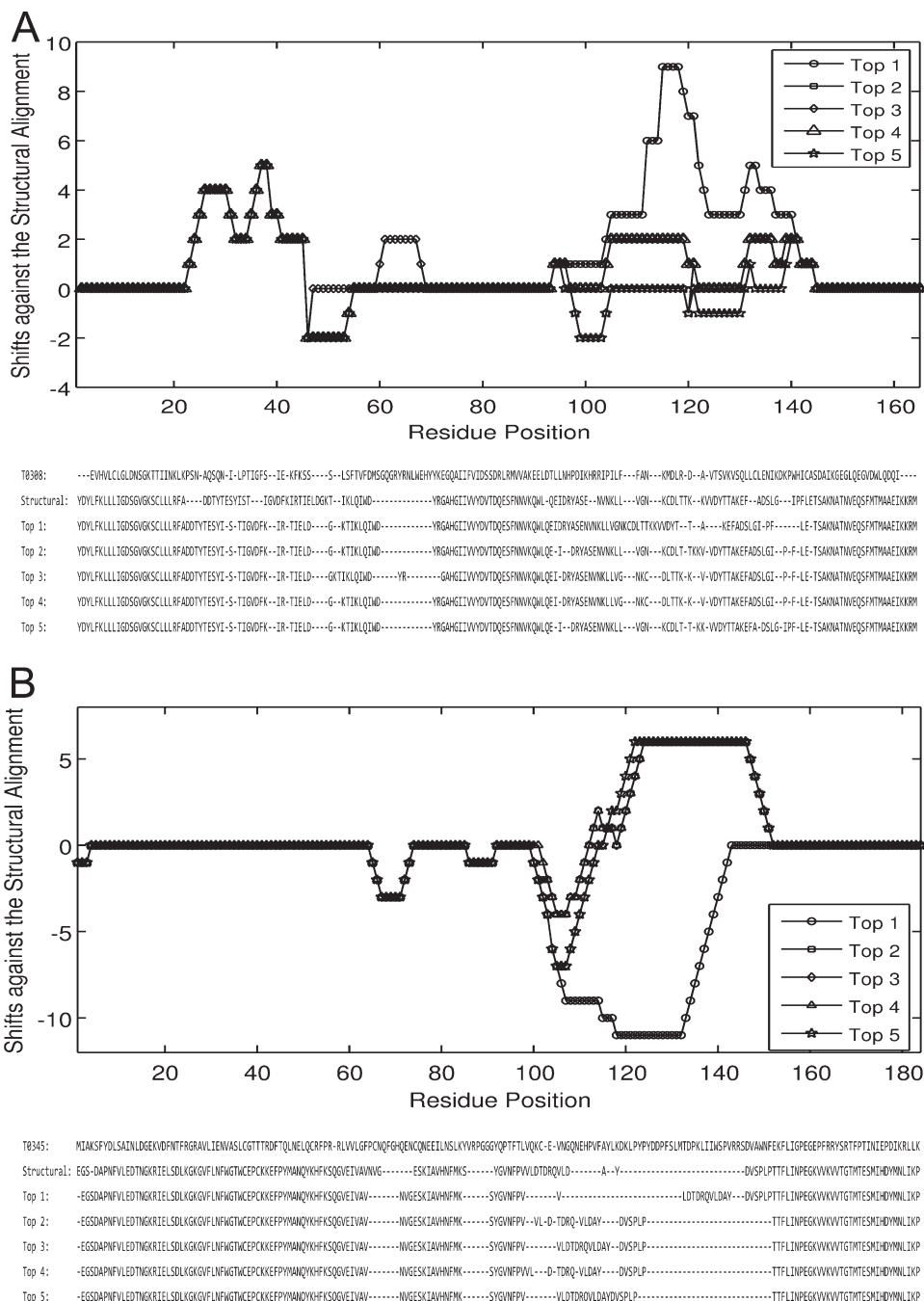
MAli-MT shows a marginal improvement over SAlI-MT; however, the number of improved cases by the MAli-MT ties with deteriorated cases (32 cases each). The average GDT-TS score and the TM-score also show marginal improvement. The quality of the MAli-MT and SAlI-MT models depends certainly on the quality of templates used and how the templates are combined. Investigation of a better way of integrating multiple templates and suboptimal alignments is left as a future study.

### Performance of SUPRB in the alignment accuracy

Finally, we examined the effect of the residue contact term for SUPRB implemented by using suboptimal alignments. This section describes results of the alignment of accuracy of SUPRB, whereas the template recognition accuracy is reported in the subsequent section. As described in Methods, two strategies are used to incorporate the contact term: In the re-ranking strategy, target–template alignments (including optimal and suboptimal alignments) computed with the local scoring terms [Eq.

**Figure 6**

Optimal alignment-based models and suboptimal alignment-based models of two CASP targets. Structural models are generated by Modeller. Models (shown in magenta) are superimposed to the native structure (green). **A:** The model of T308 based on the optimal alignment with the template structure, 2fol. The PDB code for the native structure of T308 is 2h57. **B:** The model of T308, which was computed based on five alternative target–template alignments. The same template, 2fol, was used. **C:** The optimal alignment-based model of T0345. The PDB code for T0345 is 2he3. 1st9 was used as the template. **D:** The multiple-alignment based model for T0345.



**Figure 7**

Optimal and suboptimal alignments of the two CASP7 targets and their template proteins. These alignments are used as for generating the models using Modeller shown in Figure 6. **A:** The shift in the target–template alignments (the Top1, i.e., the optimal alignment to the Top5) of T308 relative to the structural alignment (upper panel) and the actual sequence alignments (lower panel). A residue  $a_i$  in the target aligned with a template residue  $b_{j+1}$  is considered to have a shift of +1 when it should be aligned with  $b_j$  in the correct alignment. **B:** the alignment shifts (upper panel) and the actual sequence alignments (lower panel) for T0345.

(16)] are then re-ranked by the score that includes the contact term [Eq. (18)]. The second strategy is to handle residue contacts in a probabilistic fashion taking advantage of the suboptimal alignments [Eqs. (18) and (19)].

Table III lists the weighting factors trained on the Prosup data set. Terms are added one by one to the compatibility score starting from the sequence profile term. Then, the weighting factors and the opening/extending

**Table III**  
Weighting Factors Trained on the Prosup Data Set

Prof	Weighting factor values <sup>a</sup>						Alignment accuracy (%) <sup>b</sup>	
	SS	SAA	Ang	Cont	Gap_o	Gap_e	Exact ( $\pm 0$ )	$\pm 4$ Residues
1.0	—	—	—	—	-4	-0.6	53.2	71.4
0.8	0.2	—	—	—	-2	-0.2	61.0 ( $4.21 \times 10^{-13}$ )	76.1 ( $2.31 \times 10^{-10}$ )
0.7	0.2	0.1	—	—	-2	-0.2	62.9 ( $2.56 \times 10^{-5}$ )	76.9 (0.09)
0.6	0.1	0.2	0.1	—	-2	-0.2	63.3 (0.146)	78.5 (0.048)
0.5	0.2	0.2	0.1	0.2	-2	-0.2	I <sup>c</sup> : 64.9 ( $1.20 \times 10^{-3}$ ); II: 65.4 ( $2.67 \times 10^{-4}$ )	I: 80.7 ( $1.8 \times 10^{-3}$ ); II: 81.5 ( $4.8 \times 10^{-4}$ )

<sup>a</sup>Values are weighting factors for (from left to right): sequence profile, secondary structure, main-chain angle potential, solvent accessibility, residue contact potential, and the last two values are the opening and extension gap penalties. A dash (—) means that the term is not included in the combination.

<sup>b</sup>The average percentage of correctly aligned residues in the top-scoring alignments. The column for  $\pm 4$  residues counts a residue pair as correct when a residue in target is aligned within four residues to the correct partner in the template. In the parentheses, the  $P$ -value of the  $P$ -pair test is shown. The test evaluates the statistical significance of the average value compared with the one in the previous row.

<sup>c</sup>The contact potential is used in the re-ranking strategy (I) or in the strategy that consider residue contacts in a probabilistic fashion (II).

gap penalties are trained each time a new term is added to the score. It is shown that adding more terms consistently improves the alignment accuracy (considering the exact agreement): 7.8% points, 1.9, 0.4, and 1.6 (for the re-ranking strategy, I, in the table) or 2.1 (for the probabilistic contacts, II) improvement is observed by adding the secondary structure (SS) term, the solvent accessibility (SAA) term, the main-chain angle (Ang) term, and the residue contact (Cont) term, respectively. Particularly, it is noteworthy that adding the contact potential improves the results both by the re-ranking strategy (I) and the probabilistic handling of contact potential (II). Comparing the two strategies, the probabilistic contacts (II) performs better than the re-ranking strategy (I) (accuracy improvement: 1.6% points and 2.1% points for I and II, respectively), indicating that the probabilistic contacts use suboptimal alignments more effectively to capture residue contact information. We conducted the paired  $t$ -test<sup>83</sup> to examine the statistical significance of the improvement observed at each time a new term is added. Using the  $P$ -value threshold value of 0.05, the improvements are considered to be significant for almost all the cases, except for two cases (the  $P$ -value of 0.146 for adding the angle potential when the exact matches are counted and 0.09 for adding the solvent accessibility term when matches  $\pm 4$  residues are counted as correct).

The trained parameters are further tested on the SALIGN data set (Table IV). The scoring terms are added one by one with the weighting factors trained on the Prosup data set (Table III), and the alignment accuracy is evaluated at each time. The results are essentially consistent with Table III, that is, adding terms improves the accuracy, and the improvements are statistically significant in all the cases other than adding the angle potential term evaluated by the exact matches ( $P$ -value 0.098).

We further investigate the alignment accuracy in the two testing data sets, the SALIGN data set (Table V) and the L-E data set (Table VI), using the parameter set optimized for the combination of all the terms (the last row

in Tables III and IV). For the probabilistic residue contact strategy, results for iterative updates of suboptimal alignments are shown. The first iteration uses alignments generated with the local scoring terms [Eq. (16)] to define residue contacts, which is then recomputed by DP using the score with the probabilistic residue contacts [Eq. (18)]. The subsequent iteration further updates the alignments by identifying contacting residues based on the alignments of the previous pass and applying the probabilistic residue contact strategy.

The iteration of updating alignments by the probabilistic residue contacts improves the accuracy consistently from 55.07% (the first iteration) to 55.56% at the fourth iteration (Table V). The re-ranking strategy is more accurate (55.15%) than the probabilistic contact strategy at the first iteration (55.07%); however, the probabilistic contact strategy overtakes the re-ranking strategy from the second iteration. When only the optimal alignment is used, the probabilistic contact strategy converges to the “partly thawed” approach proposed by Skolnick and Kihara for a threading program PROSPECTOR,<sup>4</sup> which takes the residue contact information from the optimal sequence-profile-based alignment. The comparison with

**Table IV**  
Alignment Accuracy on the SALIGN Data Set

Scoring terms combined	Alignment accuracy (%) <sup>a</sup>	
	Exact ( $\pm 0$ )	$\pm 4$ Residues
$S_{\text{profile}}$	46.5	67.4
$S_{\text{profile}}, S_{\text{SS}}$	51.3 ( $9.03 \times 10^{-12}$ )	73.9 ( $2.73 \times 10^{-12}$ )
$S_{\text{profile}}, S_{\text{SSr}}, S_{\text{SAA}}$	53.7 ( $1.49 \times 10^{-7}$ )	77.0 ( $3.22 \times 10^{-8}$ )
$S_{\text{profile}}, S_{\text{SSr}}$	54.1 (0.098)	77.8 (0.025)
$S_{\text{SAA}}, S_{\text{angle}}$		
$S_{\text{profile}}, S_{\text{SSr}}, S_{\text{SAA}}$	I: 55.2 ( $1.9 \times 10^{-3}$ );	I: 79.2 ( $1.4 \times 10^{-3}$ );
$S_{\text{angle}}, S_{\text{contact}}$	II: 55.6 ( $6.9 \times 10^{-4}$ )	II: 80.7 ( $9.7 \times 10^{-5}$ )

The weighting factor values and gap penalty values trained on the Prosup data set (Table III) are used.

<sup>a</sup>In the parentheses, the  $P$ -value of the  $P$ -pair test is shown. The test evaluates the statistical significance of the average value compared with the one in the previous row.

**Table V**  
Improvement of Alignment Accuracy by Iterations Using the Probabilistic Residue Contacts

	Re-ranking strategy	Probabilistic handling of residue contacts				Partly thawed <sup>a</sup>
		First	Second	Third	Fourth	
Accuracy (%)	55.15 (partly thawed: 0.040)	55.07	55.41 (first: 0.060)	55.53 (second: 0.095)	55.56 (third: 0.76) (re-ranking: 0.048) (partly thawed: 0.0017)	54.83

Benchmarked on the SALIGN data set. The average value over all target–template alignments is shown. In the parentheses, the *P*-value of the paired *t*-test is shown. The test is performed to compare the value in the column against the one indicated before the colon in the parenthesis.

<sup>a</sup>The partly thawed approach proposed by Skolnick and Kihara (2001).

the partly thawed approach (54.83%) and the probabilistic contact strategy with the first iteration (55.07%) (Table V) illustrates the positive contribution of the sub-optimal alignments in improvement of the accuracy. The re-ranking strategy (55.15%) also performs better than the partly thawed approach. Note that this comparison is intended only to show the positive effect of using sub-optimal alignments but not to make performance comparison between SUPRB and PROSPECTOR because there are many practical technical differences between the two threading methods. We decided to set the maximum number of iterations to 4 because the gain in the accuracy decreases monotonically, resulting in a marginal improvement of 0.03% point for the fourth over the third iteration. Using the *P*-value cutoff of 0.05, the improvement at each iteration over the previous pass does not show statistical significance (the *P*-value of 0.06, 0.095, and 0.76). However, both the re-ranking and the probabilistic contact strategy using four iterations show higher accuracy than the partly thawed approach with statistical significance. Also, the improvement by the probabilistic contact over the re-ranking strategy is statistically significant.

The results on the L-E data set (Table VI) are qualitatively same as in Table V: The re-ranking strategy shows a higher accuracy than the probabilistic contact strategy at the first iteration; however, the latter makes consistent improvement by the subsequent iterative updates of residue contacts and the alignments. The results of the fourth iteration of the probabilistic contact strategy are

better than the re-ranking strategy and the partly thawed approach, with the statistical significance for all three similarity levels, the family, the superfamily, and the fold.

### Performance in the template recognition accuracy

Next, we examine the performance of SUPRB in terms of the template recognition accuracy in comparison with the other existing methods. The benchmark was performed on the L-E data set. Each query protein sequence is aligned with the rest of the proteins in the L-E data set, which are then ranked by the Z-score of the raw alignment score. Retrieved templates are evaluated in the three similarity levels between the target and templates, that is, in the family, the superfamily, and the fold levels. In evaluating the methods for the superfamily level, template hits that belong to the same family with the query are neglected in the list, and counted if a template in the same superfamily (but not in the same family) is hit at the Top1 or within Top5 or not. The same is done for the evaluation at the fold-level accuracy. SUPRB was run with four different scoring schemes: First, the score with the local terms without the residue contact terms [Eq. (16)] was used. Then, the contact term was incorporated either by the re-ranking strategy (I) or by the probabilistic contacts strategy (II). In addition, we also examined another score to rank template hits, which combines the Z-score of the raw score by the probabilistic contact strategy and the SPAD score in the following fashion:

**Table VI**  
Improvement of Alignment Accuracy by Iterations Using the Probabilistic Residue Contacts on the L-E Data Set

	Family	Superfamily	Fold
Re-ranking strategy	59.63(%) (partly: 0.018)	29.80 (partly: 0.077)	14.32 (partly: 0.057)
Probabilistic contacts			
First iteration	59.24 (re-ranking: 0.018)	29.28 (re-ranking: 0.077)	14.13 (re-ranking: 0.057)
Second iteration	59.79 (first: 0.011)	30.12 (first: 0.0046)	14.81 (first: 0.0087)
Third iteration	60.05 (second: 0.087)	30.50 (second: 0.054)	14.91 (second: 0.042)
Fourth iteration	60.12 (third: 0.650)	30.55 (third: 0.470)	14.93 (third: 0.071)
Partly thawed	(re-ranking: 0.016) (partly: 0.00072)	(re-ranking: 0.0038) (partly: 0.00025)	(re-ranking: 0.0057) (partly: 0.0010)
	59.10	29.45	14.04

The average accuracy for alignments at the family, the superfamily, and the fold level similarity is shown. In the parentheses, the *P*-value of the paired *t*-test is shown. The test is performed to compare the alignment accuracy with the one indicated before the colon. For example, (re-ranking: *x.xxx*) and (partly: *x.xxx*) are the *P*-value computed by comparing with the accuracy observed for the re-ranking strategy and the partly thawed approach, respectively.

**Table VII**  
Template Recognition on the Lindahl and Elofsson Data Set

Method	Family		Superfamily		Fold	
	Top1	Top5	Top1	Top5	Top1	Top5
PSIBLAST	71.2	72.3	27.4	27.9	4.0	4.7
HMMER	67.7	73.5	20.7	31.3	4.4	14.6
SAMT98	70.1	75.4	28.3	38.9	3.4	18.7
THREADER	49.2	58.9	10.8	24.7	14.6	37.7
FUGUE	82.2	85.8	41.9	53.2	12.5	26.8
RAPTOR	75.2	77.8	39.3	50.0	25.4	45.1
PROSPECT II	84.1	88.2	52.6	64.8	27.7	50.3
SPARKS	81.6	88.1	52.5	69.1	24.3	47.7
SPARKS-0	82.9	90.1	56.5	<b>72.4</b>	23.1	43.6
FOLDPro	<b>85.0</b>	89.9	55.5	70.0	26.5	48.3
SP1	81.3	87.2	51.8	65.0	20.2	35.8
SP2	82.5	87.0	52.5	67.1	24.9	40.2
SP2+	82.2	86.8	55.1	68.4	27.8	45.8
SP3	81.6	86.8	55.3	67.7	28.7	47.4
SP4	80.9	86.3	57.8	68.9	30.8	53.6
SP5 <sup>a</sup>	82.4 (81.6)	87.6 (87.0)	<b>59.8</b> (59.9)	70.0 (70.2)	<b>37.9</b> (37.4)	<b>58.7</b> (58.6)
SUPRB no contact term	82.9	89.0	51.4	66.8	26.8	51.4
SUPRB re-ranking	<i>84.9</i>	<b>90.7</b>	52.5	67.7	26.2	51.7
	(no cont.: 0.031) <sup>b</sup>		(no cont.: 0.064)		(no cont.: 0.120)	
SUPRB prob. contacts	<i>84.5</i>	<i>90.5</i>	<i>56.9</i>	69.8	27.4	51.7
	(no cont.: 0.035) (re-ranking: 0.22)		(no cont.: 0.0078) (re-ranking: 0.017)		(no cont.: 0.073) (re-ranking: 0.095)	
SUPRB $S_{\text{combined}}$ score <sup>c</sup>	<i>84.5</i>	<i>90.5</i>	56.2	70.5	27.7	52.6
	(no cont.: 0.035) (re-ranking: 0.23) (prob. cont.: 0.54)		(no cont.: 0.0068) (re-ranking: 0.012) (prob. cont.: 0.19)		(no cont.: 0.029) (re-ranking: 0.041) (prob. cont.: 0.086)	

The success rate (%) of recognizing a template that belongs to the same family, superfamily, or fold to the query within Top1/Top5 scoring templates are shown. The values for the methods other than SUPRB are taken from Liu *et al.* (2007). The highest success rate for each category is highlighted in bold and the second and the third best values are italicized.

<sup>a</sup>The values are for SP5<sup>f</sup> taken from Table II in the SP5 article by Zhang *et al.* (2008). In the parentheses, the values for SP5<sup>g</sup> in the table are shown.

<sup>b</sup>In the parentheses, statistical significance of the ranking of correct templates is evaluated by Mann–Whitney *U* test. The method compared against is shown before the colon: no cont., SUPRB with no residue contact term; re-ranking, SUPRB with the re-ranking strategy; prob. cont., SUPRB with the probabilistic contact term.

<sup>c</sup>Applied to the probabilistic contacts strategy.

$$S_{\text{combined}} = 0.7 * Z_{\text{score}} - 0.3 * \ln(\text{SPAD}) \quad (23)$$

As the SPAD score<sup>46</sup> measures the consistency of the top-scoring alignment compared with the suboptimal alignments (the lower score more consistent), the  $S_{\text{combined}}$  score is intended to select reliable (i.e., a low SPAD score, which indicates that alignments are consistent) target–template alignments that have a high Z-score. The results by the other existing methods are taken from the article by Liu *et al.*<sup>26</sup> The results of the SP5 method are taken from the article by Zhang *et al.*<sup>6</sup>

We first examine the performance of the four scoring schemes for SUPRB (last rows in Table VII). The two schemes with the contact potential perform better than SUPRB without contact potential except for Top1 for the fold-level similarity. Interestingly, the re-ranking strategy shows slightly better performance than the probabilistic contacts strategy in the family level (both Top1 and Top5) and ties at Top5 for the fold level. The probabilistic contact strategy outperforms the re-ranking strategy for the rest of the categories (i.e., Top1 and Top5 for the superfamily and Top1 for the fold level). In the fold rec-

ognition, recognizing correct templates with a distinct Z-score is also very important. Table VIII shows that the probabilistic contacts strategy and the re-ranking strategy increase the Z-score of correct hits over the local score based ranking. The  $S_{\text{combined}}$  score applied to the probabilistic contacts strategy (the last row in Table VII) did not deteriorate the accuracy if not marginally improved it, except for one case, Top1 in the superfamily level. The improvement by the  $S_{\text{combined}}$  score was observed for Top5 for the superfamily and both Top1 and Top5 for the fold-level similarity. The *P*-values calculated by the Mann–Whitney *U* test shows SUPRB with the probabilistic contacts strategy and the  $S_{\text{combined}}$  score are signifi-

**Table VIII**  
The Average Z-Scores of the Correct Hits by SUPRB

	Family		Superfamily		Fold	
	Top1	Top5	Top1	Top5	Top1	Top5
SUPRB without contact	9.74	9.41	5.27	4.45	3.32	3.03
Re-ranking	10.20	9.68	5.42	4.72	3.81	3.28
Probabilistic contacts	10.56	10.06	5.51	4.82	3.68	3.39

The template recognition results on the L-E data set are used.

**Table IX**  
Comparison With the Other Template Ranking Methods

	Family		Superfamily		Fold	
	Top1	Top5	Top1	Top5	Top1	Top5
Length normalized score	82.1 (%)	87.3	48.2	62.4	21.9	50.7
SP4 method <sup>a</sup>	84.3	89.9	53.0	67.2	24.7	50.3
SUPRB Prob. contacts <sup>b,c</sup>	84.5	90.5	56.9	69.8	27.4	51.7
SUPRB $S_{\text{combined}}$ score <sup>c,d</sup>	84.5	90.5	56.2	70.5	27.7	52.6

The raw target–template compatibility score by the probabilistic residue contact strategy of SUPRB is transformed in four different schemes.

<sup>a</sup>Note that only the template ranking method used in SP4 is applied but not the entire SP4 threading method, to the raw compatibility score by the probabilistic residue contact strategy of SUPRB.

<sup>b</sup>Eq. (21) is applied.

<sup>c</sup>These values are copied from Table VII.

<sup>d</sup>Eq. (23) is applied.

cantly better than SUPRB without the residue contact term except for the accuracy measured in the fold-level recognition.

In Table IX, we examine the template ranking method we used [Eq. (21)] in comparison with the two other methods, a normalized score method and the one proposed for SP4<sup>26</sup> threading methods. In the normalized score method, the raw sequence–structure compatibility score for a template is first normalized the alignment length, which is then used for ranking templates. In SP4, templates are ranked by the difference between the raw alignment score and the reverse alignment score in which the alignment is made with the reversed query sequence.<sup>26</sup> If there is no structural similarity between the first and the second models, templates are re-ranked by the larger one of the two Z-scores, one computed for the raw alignment scores normalized by the alignment length and another one the raw scores normalized by the nongap alignment length. Here, the score from SUPRB computed with the probabilistic contact strategy is considered as the raw score, to which the two alternative methods are applied. The results (Table IX) show that both alternative methods do not work as well as Eq. (21) and the  $S_{\text{combined}}$  score [Eq. (23)]. These results may not imply general superiority of Eq. (21) and the  $S_{\text{combined}}$  score over them. Template ranking methods are developed based on the empirical observation of the raw score distribution of particular threading methods and, thus, can be specific to each threading method.

Finally, in comparison with the other existing methods (Table VII), SUPRB is ranked within the best three methods in all the cases other than Top1 for the fold level, for which the results using the  $S_{\text{combined}}$  score is ranked at fifth (27.7%). The head-to-head comparison between SUPRB using the  $S_{\text{combined}}$  score and the other methods reveals that SUPRB with  $S_{\text{combined}}$  score ties with SP4 and SP5 (SUPRB wins at Top1 and Top5 at the family level, Top5 at the superfamily level, and SP4 and SP5 win in the other three cases) and are indeed the best among all. To conclude, the use of suboptimal alignments for

incorporating the residue contact potential is effective also in improving the accuracy of template recognition. Moreover, SUPRB shows very competitive results in the template recognition compared with the other existing methods.

### Application to CASP8 targets

We apply SUPRB to targets of Critical Assessment of Techniques in Structure Prediction 8 (CASP8; <http://predictioncenter.org/casp8>) to further compare the performance of SUPRB with the other existing state-of-the-art threading methods. To this end, we selected 28 single-domain human/server targets in the template-based modeling category because assembling structures of multiple domains is beyond the scope of this work. For this test, we prepared a template database of 10,926 proteins, which are selected by the PDB-REPRDB server<sup>84</sup> with the threshold value of 40% sequence identity and an RMSD of 6 Å. For a target sequence, the template database is first scanned using SUPRB with the re-ranking strategy to obtain top 50 templates, which are then re-scanned using the probabilistic contact strategy. The tertiary structure of a target is built with Modeller using the five alignments between the recognized template and the target. Manual refinement is not performed.

The results are shown in Table X and Figure 8. The correct template is recognized as the top hit for 21 of the 28 targets, and considering top five templates increases it to 22. We also compute the average GDT-TS score of the Top1 ranked model and the best among the top five models and compared it with all the other 72 servers participated in CASP8. The average GDT-TS score of the Top1 model by SUPRB is 58.55, which would rank at the 11th among the servers on these targets (Fig. 8). As shown in the figure, the difference of the score to higher ranked servers is small; for example, the score difference to the second server, TS429 (Pcon\_multi), is 1.99. The difference to the top server (TS426, Zhang-server) is 5.07. Building an accurate tertiary structure model based on a correct template by threading needs additional key developments, which many of the top servers have implemented, such as building the structure of unaligned regions and main-chain optimization for the target sequence starting from the template structure.<sup>55</sup> Here, we simply used Modeller for those tasks. Although there are differences in the servers in the types of implemented methods and their complications, the results show the competitive performance of SUPRB among the other existing methods.

### Computational time of SUPRB

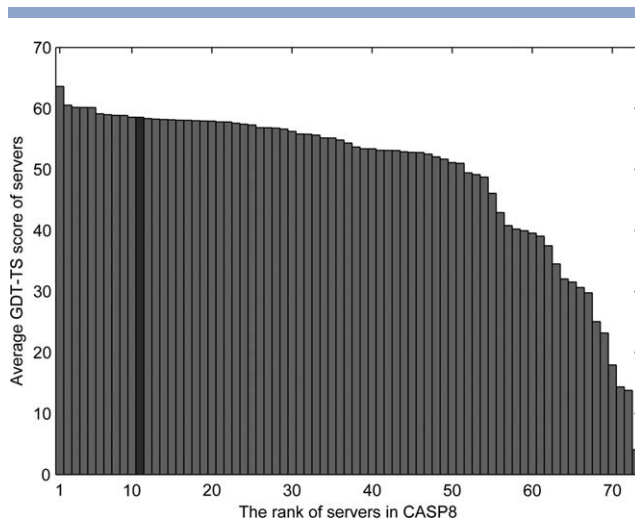
The two strategies of SUPRB, the re-ranking strategy and the probabilistic contact strategy, obviously take more time than computing only the optimal target–tem-

**Table X**  
Performance on CASP8 Template-Based Modeling Targets

Target	Recognized template	Template correctness <sup>a</sup>	GDT-TS
T0389	2oucA	X	71.08
T0393	2f1kA (2gf2A)	X	63.99 (64.78)
T0395	1surA	X	53.78
T0396	1jr8A	X	88.24
T0399	1q7fA	—	36.57
T0401	2pagA	X	60.43
T0407	2hqaA	X	57.04
T0411	1yt8A (1e0cA)	X	69.28 (72.88)
T0415	1c8cA	X	71.50
T0419	2plrA	X	45.65
T0421	2plrA	X	51.24
T0425	1xovA	X	56.98
T0427	1hr6A (1bccA)	X	50.80 (66.06)
T0440	1m65A	X	49.64
T0443	2abk	—	42.05
T0449	1snzA (2ciq)	—	55.24 (61.82)
T0451	1s5aA (1nu3A)	X	65.55 (67.13)
T0454	1pb6A	X	93.27
T0457	2qb6A	X	52.45
T0464	2jr6A (2hflA)	—	45.29 (51.81)
T0469	2fhfA	X	76.19
T0474	1p94A	X	83.54
T0476	1w5sA	—	46.84
T0480	1dl6A	X	68.33
T0489	1yqyA	X	34.64
T0495	1gefA (2p14A)	— (X)	30.76 (40.29)
T0498	1pgx (1mhxA)	—	35.56 (37.78)
T0499	2i2yA (2igdA)	X	83.48 (84.38)
Average			58.55 (60.23)

A structure model ranked at Top1 is listed. If a better model is found within Top5 ranked models, it is shown in the parentheses.

<sup>a</sup>A template is considered to be correct for a target if it appears in the template list provided by the CASP8 organizers found at [http://predictioncenter.org/download\\_area/CASP8/list\\_of\\_templates.txt](http://predictioncenter.org/download_area/CASP8/list_of_templates.txt).

**Figure 8**

The average GDT-TS score of Top1 models of CASP8 template-based model targets of the servers. SUPRB is shown with the dark gray bar and the other servers are shown in pale gray.

**Table XI**  
Computational Time of Target–Template Alignments by SUPRB

Target	Target length (aa)	Template	Template length	Regular DP (s)	SUPRB I	SUPRB II
1aab	83	1hme	77	0.25	0.66	2.70
2aky	218	1gky	186	1.48	8.83	43.95

The regular DP computes only the optimal alignment between the target and the template. Here, SUPRB I and SUPRB II compute  $0.01 \times l_{\text{target}} \times l_{\text{template}}$  suboptimal alignments, where  $l_{\text{target}}$  and  $l_{\text{template}}$  are the length of target and the template proteins, respectively. SUPRB I, the re-ranking strategy; SUPRB II, the probabilistic residue contact strategy with five iterations.

plate alignment using a regular DP because they compute suboptimal alignments. Table XI shows the execution time (user time) of SUPRB for two examples on a Linux machine with an Intel i7 2.67-GHz processor and 12 Gb memory. For both SUPRB strategies,  $0.01 \times l_{\text{target}} \times l_{\text{template}}$  suboptimal alignments are computed, where  $l_{\text{target}}$  and  $l_{\text{template}}$  are the length of target and the template proteins, respectively. For the probabilistic contact strategy, the alignments are refined for five iterations. The SUPRB re-ranking strategy took three to six times, and the probabilistic contact strategy took  $\sim 10$ – $30$  times more the regular DP. Threading scans performed for the CASP8 targets took around 36–48 h on a single CPU of the Linux machine.

## DISCUSSION

In this article, we investigated the effect of using suboptimal alignments in template-based structure prediction. Aligning two protein sequences is not trivial, especially when they do not have significant sequence similarity. We showed that suboptimal alignments are often more accurate than the optimal one, and such accurate suboptimal alignments can occur even at a very low alignment score rank. Moreover, lower-ranked suboptimal alignments contain a significant number of correct amino acid residue contacts. Benefits of suboptimal alignments can be immediately enjoyed by using them as input for a template-modeling tool, Modeller, by feeding a set of alternative alignments rather than a single optimal alignment between a target and a template. Finally, we used suboptimal alignments in handling a contact potential in a probabilistic way in a threading program, SUPRB. The probabilistic contacts strategy outperforms the partly thawed approach, which only uses the optimal alignment in defining residue contacts and the re-ranking strategy, which uses the contact potential in re-ranking alignments. To the best of our knowledge, this is the first time that the effect of the suboptimal alignments in structure prediction is thoroughly investigated and that the suboptimal alignments are implemented for using a contact potential for template-based modeling. The probabilistic handling of residue contacts may also be useful

to capture changes in residue contact on protein motion. Our approach can be implemented in any DP-based template-based structure prediction methods. Although we have not addressed in this article, we would like to mention that suboptimal alignments are also useful in assessing the quality of template-based models, as we showed in related works.<sup>45–47</sup>

A residue contact potential is unique among the other commonly used scoring terms in threading methods in the sense that it directly represents long-range interaction<sup>85</sup> of amino acids in the tertiary structure. Such long-range interaction is not well captured by one-body scoring terms, such as sequence-profile terms, secondary structure terms, or residue environment terms.<sup>86</sup> Hence, a residue contact potential will be a key for success in recognizing very distantly related template structures, which share virtually no sequence similarity with a target protein. However, a recent trend in template-based structure prediction methods is to rely heavily on sequence-based information and occasionally on some one-body potentials but not two-body contact potentials. There are two main reasons for this trend. First, sequence-information has become more and more useful as the amount of sequences in databases grows rapidly and new effective approaches have been developed for using the sequence and local structure information.<sup>25</sup> Second, a residue contact potential is cumbersome to handle in template recognition because the optimal alignment for a contact potential cannot be obtained with a conventional DP algorithm. In addition, it should be also noted that a residue contact pattern is not necessarily conserved even in proteins of the same family<sup>44</sup>; thus, simply copying contacting residue pairs from a template may not be the best way to use contact information.

Despite all these challenges, we believe that it is worthwhile to revisit residue contact potentials in template-based prediction because using sequence-structure compatibility terms rather than sequence similarity-based terms is logically the only the way to find template structures with virtually no sequence similarity to a target. Such distantly related template structures are expected to be made available in an increasing pace because of the progress of experimental protein structure determinations. It is also noteworthy that the number of template structures needed for structure modeling of the entire protein space will be reduced once methods are established for detecting and using very distantly related structures for target proteins.<sup>16</sup>

Recent studies have reported technical improvement in computing alignments once an appropriate template is recognized for a target protein.<sup>24,87,88</sup> However, templates should be identified in the first place for using such advanced alignment techniques. Therefore, both techniques, that is, distant template recognition and optimizing target–templates alignments, should be developed in a good harmony between each other to further

advance our technology of template-based protein structure prediction.

## ACKNOWLEDGMENTS

The residue contact potential was provided by Yifeng D. Yang. The authors also appreciate stimulus discussion with Y. D. Yang. The authors are grateful to Rebecca Harding for proofreading the manuscript.

## REFERENCES

1. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. Assessment of CASP7 structure predictions for template free targets. *Proteins* 2007; 69(Suppl 8):57–67.
2. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. *Proteins* 2007;69(Suppl 8):38–56.
3. Zhang Y. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* 2008;18:342–348.
4. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins* 2001;42: 319–331.
5. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Proteins* 2004;56:502–518.
6. 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.
7. Qu X, Swanson R, Day R, Tsai J. A guide to template based structure prediction. *Curr Protein Pept Sci* 2009;10:270–285.
8. Yang YD, Park C, Kihara D. Threading without optimizing weighting factors for scoring function. *Proteins* 2008;73:581–596.
9. Kolinski A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* 2004;51:349–371.
10. Liwo A, Czaplowski C, Oldziej S, Scheraga HA. Computational techniques for efficient conformational sampling of proteins. *Curr Opin Struct Biol* 2008;18:134–139.
11. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci U S A* 2001;98:10125–10130.
12. Meinke JH, Hansmann UH. Free-energy-driven folding and thermodynamics of the 67-residue protein GS- $\alpha$ 3W—a large-scale Monte Carlo study. *J Comput Chem* 2009;30:1642–1648.
13. Itoh SG, Okamoto Y. Effective sampling in the configurational space of a small peptide by the multicanonical-multioverlap algorithm. *Phys Rev E Stat Nonlin Soft Matter Phys* 2007;76:026705.
14. Das R, Baker D. Macromolecular modeling with rosetta. *Annu Rev Biochem* 2008;77:363–382.
15. Zhou H, Skolnick J. Protein structure prediction by pro-Sp3-TASSER. *Biophys J* 2009;96:2119–2127.
16. Friedberg I, Jaroszewski L, Ye Y, Godzik A. The interplay of fold recognition and experimental structure determination in structural genomics. *Curr Opin Struct Biol* 2004;14:307–312.
17. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
18. Burley SK. An overview of structural genomics. *Nat Struct Biol* 2000;(7 Suppl):932–934.
19. Zhang C, Kim SH. Overview of structural genomics: from structure to function. *Curr Opin Chem Biol* 2003;7:28–32.
20. Skolnick J, Arakaki AK, Lee SY, Brylinski M. The continuity of protein structure space is an intrinsic property of proteins. *Proc Natl Acad Sci U S A* 2009;106:15690–15695.
21. Kihara D, Skolnick J. The PDB is a covering set of small protein structures. *J Mol Biol* 2003;334:793–802.

22. Petrey D, Fischer M, Honig B. Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc Natl Acad Sci U S A* 2009;106:17377–17382.
23. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci U S A* 2005;102:1029–1034.
24. Wang G, Dunbrack RL Jr. Scoring profile-to-profile sequence alignments. *Protein Sci* 2004;13:1612–1626.
25. Dunbrack RL Jr. Sequence comparison and protein structure prediction. *Curr Opin Struct Biol* 2006;16:374–384.
26. Liu S, Zhang C, Liang S, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. *Proteins* 2007;68:636–645.
27. Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* 2004;55:1005–1013.
28. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* 2003;51:504–514.
29. Li SC, Bu D, Xu J, Li M. Fragment-HMM: a new approach to protein structure prediction. *Protein Sci* 2008;17:1925–1934.
30. Xu J, Jiao F, Yu L. Protein structure prediction using threading. *Methods Mol Biol* 2008;413:91–121.
31. Fischer D. Servers for protein structure prediction. *Curr Opin Struct Biol* 2006;16:178–182.
32. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 2003;19:1015–1018.
33. Mevissen HT, Vingron M. Quantifying the local reliability of a sequence alignment. *Protein Eng* 1996;9:127–132.
34. Vingron M, Argos P. Determination of reliable regions in protein sequence alignments. *Protein Eng* 1990;3:565–569.
35. Vingron M. Near-optimal sequence alignment. *Curr Opin Struct Biol* 1996;6:346–352.
36. Saqi MA, Sternberg MJ. A simple method to generate non-trivial alternate alignments of protein sequences. *J Mol Biol* 1991;219:727–732.
37. Sommer I, Toppo S, Sander O, Lengauer T, Tosatto SC. Improving the quality of protein structure models by selecting from alignment alternatives. *BMC Bioinformatics* 2006;7:364.
38. Miyazawa S. A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng* 1995;8:999–1009.
39. Schlosshauer M, Ohlsson M. A novel approach to local reliability of sequence alignments. *Bioinformatics* 2002;18:847–854.
40. Zhang MQ, Marr TG. Alignment of molecular sequences seen as random path analysis. *J Theor Biol* 1995;174:119–129.
41. Kschischo M, Lassig M. Finite-temperature sequence alignment. *Pac Symp Biocomput* 2000;5:624–635.
42. Yu L, Smith TF. Positional statistical significance in sequence alignment. *J Comput Biol* 1999;6:253–259.
43. Cline M, Hughey R, Karplus K. Predicting reliable regions in protein sequence alignments. *Bioinformatics* 2002;18:306–314.
44. Godzik A. The structural alignment between two proteins: is there a unique answer? *Protein Sci* 1996;5:1325–1338.
45. Kihara D, Chen H, Yang YD. Quality assessment of computational protein models. *Curr Opin Pept Sci* 2009;10:216–228.
46. Chen H, Kihara D. Estimating quality of template-based protein models by alignment stability. *Proteins* 2008;71:1255–1274.
47. Yang YD, Spratt P, Chen H, Park C, Kihara D. Sub-AQUA: real-value quality assessment of protein structure models. *Protein Eng Des Sel* 2010;23:617–632.
48. Tang CL, Petrey D, Fasnacht M, Kosloff M, Alexov E, Honig B. Use of limited suboptimal alignment in homology modeling. Poster presented at 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB), March 27–31, 2004, San Diego, USA.
49. Eswar N, Eramian D, Webb B, Shen MY, Sali A. Protein structure modeling with MODELLER. *Methods Mol Biol* 2008;426:145–159.
50. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;234:779–815.
51. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci* 1995;4:2107–2117.
52. Sippl MJ. Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 1995;5:229–235.
53. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins* 1999;36:357–369.
54. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. *Nature* 1992;358:86–89.
55. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* 2001;44:133–149.
56. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. *Proteins* 1995;23:356–369.
57. Godzik A, Kolinski A, Skolnick J. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol* 1992;227:227–238.
58. Marti-Renom MA, Madhusudhan MS, Sali A. Alignment of protein sequences by their profiles. *Protein Sci* 2004;13:1071–1087.
59. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* 2002;11:2606–2621.
60. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. *J Mol Biol* 2000;295:613–625.
61. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 2008;36:D419–D425.
62. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* 2003;31:3370–3374.
63. Fischer D, Elofsson A, Rice DW, Eisenberg D. Assessing the performance of inverted protein folding methods by means of an extensive benchmark. In: *Proceedings of the First Pacific Symposium on Biocomputing*. 1996. pp 300–318.
64. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng* 1999;12:387–394.
65. Henikoff S, Henikoff JG. Position-based sequence weights. *J Mol Biol* 1994;243:574–578.
66. 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.
67. 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.
68. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* 2005;59:467–475.
69. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
70. Chothia C. The nature of the accessible and buried surfaces in proteins. *J Mol Biol* 1976;105:1–12.
71. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? *Protein Sci* 1997;6:676–688.
72. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J Mol Biol* 2000;297:1003–1013.
73. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005;33:2302–2309.

74. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;89:10915–10919.
75. Chakravarty S, Godbole S, Zhang B, Berger S, Sanchez R. Systematic analysis of the effect of multiple templates on the accuracy of comparative models of protein structure. *BMC Struct Biol* 2008;8:31.
76. Cheng J. A multi-template combination algorithm for protein comparative modeling. *BMC Struct Biol* 2008;8:18.
77. Liu T, Guerquin M, Samudrala R. Improving the accuracy of template-based predictions by mixing and matching between initial models. *BMC Struct Biol* 2008;8:24.
78. Moulton J, Fidelis K, Kryzhanovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction—round VIII. *Proteins* 2009;77(Suppl 9):1–4.
79. Xu J, Li M, Lin G, Kim D, Xu Y. Protein threading by linear programming. *Pac Symp Biocomput* 2003;8:264–275.
80. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* 2006;22:1456–1463.
81. Kolinski A, Bujnicki JM. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins* 2005;61(Suppl 7):84–90.
82. von GM, Pas J, Wyrwicz L, Ginalski K, Rychlewski L. Application of 3D-Jury, GRDB, and Verify3D in fold recognition. *Proteins* 2003;53(Suppl 6):418–423.
83. Zar JH. *Biostatistical analysis*. Upper Saddle River, NJ: Prentice Hall; 1999.
84. Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res* 2003;31:492–493.
85. Kihara D. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci* 2005;14:1955–1963.
86. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature* 1992;356:83–85.
87. Joo K, Lee J, Kim I, Lee SJ, Lee J. Multiple sequence alignment by conformational space annealing. *Biophys J* 2008;95:4813–4819.
88. Tan YH, Huang H, Kihara D. Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. *Proteins* 2006;64:587–600.