

Estimating quality of template-based protein models by alignment stability

Hao Chen¹ and Daisuke Kihara^{1,2,3,4*}

¹Department of Biological Sciences, College of Science, Purdue University, West Lafayette, Indiana 47907

²Department of Computer Science, College of Science, Purdue University, West Lafayette, Indiana 47907

³Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, Indiana 47907

⁴The Bindley Bioscience Center, College of Science, Purdue University, West Lafayette, Indiana 47907

ABSTRACT

The error in protein tertiary structure prediction is unavoidable, but it is not explicitly shown in most of the current prediction algorithms. Estimated error of a predicted structure is crucial information for experimental biologists to use the prediction model for design and interpretation of experiments. Here, we propose a method to estimate errors in predicted structures based on the stability of the optimal target-template alignment when compared with a set of suboptimal alignments. The stability of the optimal alignment is quantified by an index named the Suboptimal Alignment Diversity (SPAD). We implemented SPAD in a profile-based threading algorithm and investigated how well SPAD can indicate errors in threading models using a large benchmark dataset of 5232 alignments. SPAD shows a very good correlation not only to alignment shift errors but also structure-level errors, the root mean square deviation (RMSD) of predicted structure models to the native structures (i.e. global errors), and local errors at each residue position. We have further compared SPAD with seven other quality measures, six from sequence alignment-based measures and one atomic statistical potential, discrete optimized protein energy (DOPE), in terms of the correlation coefficient to the global and local structure-level errors. In terms of the correlation to the RMSD of structure models, when a target and a template are in the same SCOP family, the sequence identity showed a best correlation to the RMSD; in the superfamily level, SPAD was the best; and in the fold level, DOPE was best. However, in a head-to-head comparison, SPAD wins over the other measures. Next, SPAD is compared with three other measures of local errors. In this comparison, SPAD was best in all of the family, the superfamily and the fold levels. Using the discovered correlation, we have also predicted the global and local error of our predicted structures of CASP7 targets by the SPAD. Finally, we proposed a sausage representation of predicted tertiary structures which intuitively indicate the predicted structure and the estimated error range of the structure simultaneously.

Proteins 2008; 71:1255–1274.

© 2007 Wiley-Liss, Inc.

Key words: protein structure prediction; quality assessment of protein models; threading; suboptimal alignment; template-based modeling; error estimation.

INTRODUCTION

Protein tertiary structure prediction from its amino acid sequence is one of the most actively studied research topics in computational biology.^{1–3} It is now possible, if not always, to build a very accurate model whose error is close to the resolution of experimentally determined structures when an appropriate template structure is available for modeling.^{4–6} Recently, it has been also reported that high resolution atomic detailed models within the root mean square deviation (RMSD) of 1.5 Å to the experimentally determined native structure were successfully produced even in template-free modeling.⁷ These high resolution atomic detailed models are practically useful in a variety of structure-based protein engineering, including drug design,⁸ redesigning enzyme specificity,⁹ and designing new fold of proteins.^{10,11} However, high resolution protein structure prediction is not always possible. A very accurate model can be expected if a highly homologous template structure to a target protein is available for modeling, partly because the template itself will be fairly close to the target protein structure.^{12,13} On the other hand, if there is no template structure which covers a significant part of a target protein, template-free modeling methods or “ab initio” methods^{14–16} need to be employed, whose accuracy is not as high as template-based methods on average.¹⁷ Although low resolution computational models, which are typically generated by template free modeling methods or threading methods,^{18–20} may not be appropriate for applications that need atomic detailed resolution, they are still useful in different purposes, for example, designing site-directed mutagenesis experiments,^{21,22}

Grant sponsor: National Institute of General Medical Sciences of the National Institutes of Health; Grant number: R01GM075004; Grant sponsor: National Science Foundation; Grant number: DMS 0604776; Grant sponsor: Purdue Research Foundation.

*Correspondence to: Daisuke Kihara, Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN 47907.

E-mail: dkihara@purdue.edu

Received 17 May 2007; Revised 3 September 2007; Accepted 5 September 2007

Published online 27 November 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21819

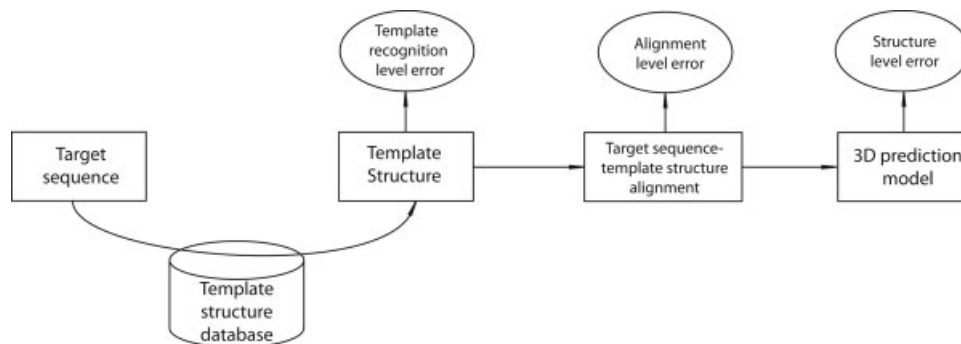


Figure 1

A procedure of template-based structure prediction. The procedure consists of three component steps: template identification, query-template alignment, and structure refinement. Error can be introduced in each of the steps.

small ligand docking prediction,^{23,24} and function prediction from predicted tertiary structures.^{25–27} Therefore, for structural prediction methods to be routinely used by biologists, it is urgent to establish quality assessment methods for predicted structures, so that predicted structures can be used wisely by knowing limitations of the model.

At this juncture it would be appropriate to briefly examine where errors in prediction models originate. We focus this discussion on template-based modeling (comparative modeling^{28,29} and threading^{20,18}), because it can routinely provide practical models^{4,25} of a good resolution. It should also be noted that most of current template-free modeling methods³⁰ somewhat rely on template-based methods by using fragment structures found by local sequence matching to fragments in the database.^{31,32} Figure 1 illustrates main steps in a template-based modeling. Errors in main chain orientation can be introduced in each step in this procedure: A severe error at the global fold level may be introduced if a structure of a different fold is misrecognized as a template (template recognition level error). A structure with the same order of secondary structure elements placed in a different spatial arrangement (i.e. the same architecture but not the same topology in the terminology by CATH classification³³) to a query protein is often wrongly recognized. An error introduced at this stage is usually too severe to be fixed in the subsequent steps. Next, even if a correct template is recognized, an error can be introduced in the alignment level (alignment level error).³⁴ A small error of a couple of residue shifts in an alignment may be able to correct in a later optimization step,³⁵ but a larger alignment shift will be source of considerable incorrectness of a resulting model. Finally, optimization of the main chain conformation for a query from the best available template is often not trivial if they are not a close homolog to each other^{36,37} (structure level error).

Different strategies have been employed to estimate the reliability of predicted models in each level of error clas-

sified above. The reliability of a recognized template for a query (the template recognition level error) can be estimated by the sequence identity (%) or the statistical significance of the sequence match between the target protein and the template protein (e.g. an *E*-value of a BLAST search³⁸). When a threading method is used for template recognition, the *Z*-score is usually used to indicate a statistical significance of recognized templates. Verify3D developed by Eisenberg *et al.* uses a threading score, namely a three-dimensional profile, for validation of protein models.³⁹ This approach could also identify a local structure level error.

The basic idea of estimating the alignment level error is to evaluate how significant or consistent the optimal alignment is relative to a group of alternative alignments.⁴⁰ Vingron and Argos initiated a method to compute suboptimal alignments⁴¹ by an extension of Needleman-Wunsch type⁴² dynamic programming (DP) algorithm. It is followed by several other methods which are also based on DP algorithm.^{43,44} These methods essentially compute different suboptimal alignments by specifying a certain residue pair to be included/not to be included in an alignment. Simply fluctuating parameters for computing alignment will also generate a set of alternative alignments.⁴⁵ Several methods employ a partition function which is borrowed from thermodynamics to compute the probability of suboptimal alignments.^{46–49} Yu and Smith used a Hidden Markov Model to estimate the importance of each aligned residue pair by posterior decoding probabilities,⁵⁰ which is later used as one of input scores in a neural network based prediction method for quality assessment.⁵¹ The other proposed ideas include considering conservation of local regions in a multiple sequence alignment.^{52,53} Although these previous works proposed reasonable ideas for computing reliability of alignments, all but two,^{52,54} have applied their methods only on a few proteins. Moreover, none of them attempted to directly predict global and local error

of predicted tertiary structures in terms of the physical distance (Å) by their reliability measure of alignments. Therefore, their applicability in the context of protein tertiary structure prediction remains unclear.

The structure level error can be predicted by examining local structures of a prediction model. Stereochemical quality of a structure is also routinely checked in the process of solving a protein structure by X-ray crystallography.⁵⁵ Programs used for this purpose, such as PRO-CHECK⁵⁶ and WHAT_CHECK,⁵⁷ essentially compare torsion angles, bond lengths, and distances between atoms⁵⁸ of a model to a standard distribution of the values sampled from a database. These programs may be applicable for homology models built on a very close homologous protein structure, thus a very precise atomic detailed position is expected, but not necessarily suitable for threading models, where precise atom positioning is an aim but not expected in an initial model.

Alternatively, statistical potentials of residue or atomic contacts⁵⁹ have been recently used for assessing the quality of structure models.^{60–63} Statistical contact potentials have been widely used in structure prediction methods,^{15,36,64–66} therefore it would be natural to employ them for quality assessment of predicted structures. However, if an aim of the quality assessment of predicted structure is predicting error of a model in terms of the global and local physical distance (Å) to the native structure, the performance of statistical contact potentials should be reexamined, because in most of the cases contact potentials were evaluated by their ability of ranking structures to select the native structure out of a pool of decoy structures, not by the ability of predicting the real value of errors of a given predicted structure. The other physical properties used for quality assessment of predicted models include preference of burial/exposure of amino acids,⁶⁷ solvation of atoms in amino acid residues,⁶⁸ and packing density of amino acids.⁶⁹

Finally there exist methods which combine several scoring terms, including structure-based terms and sequence-alignment based terms to capture different aspects of models.^{71–73}

In this study, we focus on estimating the alignment level error of template-based prediction models. Our strategy to quantify the reliability of the optimal alignment between a target sequence and a template structure is to compare the optimal alignment with a set of suboptimal alignments and compute how well the suboptimal alignments converge to the optimal alignment. We define a novel and intuitive score of the Suboptimal Alignment Diversity (SPAD) and demonstrate that the SPAD score has a significant correlation to the global and local error of threading alignments. Advantages of using SPAD to estimate the reliability of a structure model in the alignment level as opposed to the structure level (Fig. 1) are that errors are estimated in an early stage so that different templates or strategies of alignments can be sought from the beginning and also that

alternative almost equally reliable alignments can be provided from the pool of suboptimal alignments.

Because SPAD is based on sequence-alignment information, the primary aim of SPAD is to have a significant correlation to alignment errors so that alignment errors can be predicted by SPAD. Moreover, unlike all the previous works on estimating alignment level error, we go one step further to show that SPAD by itself can predict the real value of physical distance error, namely, the global coordinate RMSD of the model and local error (Å). This is the first time that a suboptimal alignment derived score is shown to be able to predict global/local structure quality of prediction models. To thoroughly understand ability and limitation of the SPAD score, we have tested it on a large alignment benchmark dataset of 5232 alignments classified into the family, the superfamily, and the fold level similarity. We have also compared performance of SPAD with the other sequence-alignment based measures and also with a statistical atomic contact potential, Discrete Optimized Protein Energy (DOPE),⁶¹ to investigate characteristics of SPAD. Furthermore, we used SPAD to predict global/local structural error of our threading prediction models in the CASP7 competition (<http://predictioncenter.org/casp7/>).

MATERIALS AND METHODS

Benchmark database

Primarily Lindahl and Elofsson's dataset (L-E dataset)⁷⁴ is used for benchmark. This dataset consists of 1130 representative proteins taken from PDB,⁷⁵ each classified into a family, a superfamily, and a fold in a hierarchical manner according to the SCOP database.⁷⁶ Following this hierarchical classification, a set of pairwise alignments were constructed in each above similarity level, that is family, superfamily, and fold level by a protein tertiary structure alignment program, LGA.⁷⁷ Aligned protein pairs in the family level set share the same family in SCOP; pairs in the superfamily level set share the same superfamily but not the same family; and pairs in the fold level set share the same fold but not the same superfamily. This resulted in 1076 target-template pairs at the family level, 1395 at the superfamily level, and 2761 in the fold level. The average sequence identity (and the standard deviation) of alignments in these three levels are 21.3% (8.06%), 15.2% (3.3%), and 15.2% (3.8%) for the family, the superfamily, and the fold level, respectively [see Fig. 1(A) in the paper by Tan *et al.*³⁴ for the distribution of the sequence identity]. The sequence identity ranges of the L-E dataset are much smaller than several other existing alignment benchmarks, such as BaliBASE⁷⁸ and HOMSTRAD,⁷⁹ because the L-E set is originally designed for threading benchmark. These structure-based pairwise alignments are the golden standard dataset in this study against which "predicted" alignments are compared.

For each protein pair in the same similarity level in the L-E set, we predicted the optimal pairwise alignment of the two proteins by a profile-based threading algorithm (see below) and also the tertiary structure of one of the aligned proteins (called a target protein) based on the predicted optimal alignment against the another protein (called a template protein). We used a homology modeling software, Modeller⁸⁰ (version 8v2) with the default parameter setting for building the tertiary structure of the target protein from a given target-template alignment using the tertiary structure of the template protein. The quality of the predicted optimal (i.e. the top scoring) alignment is assessed by comparing with the golden standard alignment in terms of the ALignment Distance, ALD, described in the next section. In addition to the optimal alignment, we also generate a set of suboptimal alignments from which the reliability of the optimal alignment is computed. The quality of the predicted tertiary structure is evaluated globally by the coordinate RMSD (Å) to the experimentally determined structure. LGA was used to superimpose two protein structures. To evaluate the quality of the local structure of a predicted model, we also computed the distance between a corresponding α carbon of the main chains of a predicted model and that of superimposed experimentally determined structure.

Alignment distance

To evaluate the accuracy of a predicted pairwise alignment, we introduce the ALD, which measures a distance of a predicted alignment to the correct alignment on the DP matrix [Fig. 2(A)]. In the process of computing a pairwise alignment by DP algorithm,^{42,81} an alignment is represented as a path in the DP matrix. The (local) distance from a position in a path, m , to another path, termed local ALignment Distance, $lALD_m$, can be computed as the average residue shifts of the position:

$$lALD_m = \frac{d_V + d_H}{2}, \quad (1)$$

where d_V/d_H is the vertical/horizontal distance from the predicted alignment to the correct alignment, respectively. Note here that the position in the path, m , may represent either a pair of aligned residues or a residue aligned with a gap. Therefore, $lALD_m$ is assigned also to a gap in an alignment.

The global Alignment Distance for the whole predicted alignment, $gALD$, is the average of $lALD_m$ over all the positions in the path:

$$gALD = \frac{\sum_{m=1}^l lALD_m}{l}, \quad (2)$$

where l is the length of the predicted alignment.

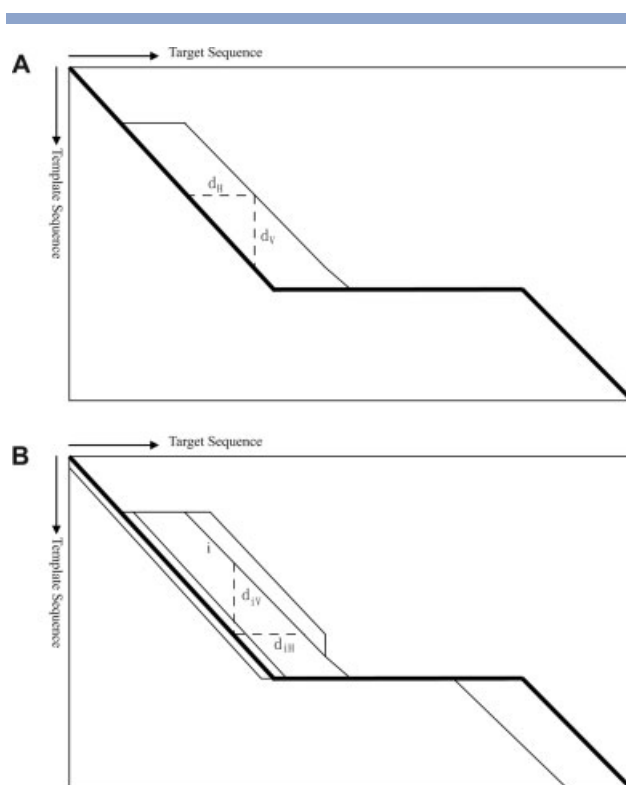


Figure 2

Definition of ALD and SPAD. The pairwise alignment space by DP algorithm is represented by a two dimensional matrix (DP matrix) with the size of $M \times N$, where M and N is the length of two aligned sequences. A pairwise alignment is represented by a path on this matrix. (A) Definition of ALD. Suppose the thick line in the DP matrix represents the correct global alignment between a pair of a target and a template proteins, and the thin line represents a predicted alignment of the two (in this example the predicted alignment is different from the correct alignment only at an N-terminal region). The local ALignment Distance (lALD) from a position in the predicted alignment is the average of the horizontal and the vertical distance (d_H , d_V), defined as the number of cells in the matrix from that position to the correct alignment. The global Alignment Distance (gALD) is the average of lALD over every position in the predicted alignment. (B) Definition of SPAD. Assume that the thick line represents the optimal (i.e. the highest scoring) alignment between a target and a template protein, and thin lines are suboptimal alignments. d_V and d_H in the figure denote the vertical and the horizontal distance from a position in the optimal alignment to the i th suboptimal alignment. The local SubOptimal Alignment Diversity (lSPAD) of a position in the optimal alignment is defined as the average to lALD over all the suboptimal alignments considered. The global SubOptimal Alignment Diversity (gSPAD) is the average of the lSPAD over all the positions of the optimal alignment.

Reliability of the optimal alignment

By extending the idea of ALD further, we define the diversity of a set of suboptimal alignments compared to the optimal alignment for a given pair of target-template proteins. This SPAD quantifies the mathematical stability of the optimal alignment, thus indicates how reliable the optimal alignment is. Essentially, SPAD is the average distance from the optimal alignment to each suboptimal alignment considered. Similar to ALD, SPAD can be computed at both local and global level.

The local level SPAD at the alignment position m , ISPAD_m in the optimal alignment is defined as follows:

$$\text{ISPAD}_m = \frac{\sum_{i=1}^n \text{IALD}_m^i}{n}, \quad (3)$$

where IALD_m^i is the local alignment distance at the position m in the optimal alignment to the suboptimal alignment i defined by the Eq. (1), thus $\text{IALD}_m^i = \frac{d_N + d_H}{2}$ [Fig. 2(B)]. And n is the number of suboptimal alignments considered. Now averaging ISPAD_m over all the positions in the optimal alignment gives the global level SPAD, gSPAD:

$$\text{gSPAD} = \frac{\sum_{m=1}^l \text{ISPAD}_m}{l}, \quad (4)$$

where l is the length of the optimal alignment.

Suboptimal alignments are computed by the algorithm proposed by Vingron and Argos.⁴¹ In their algorithm, the maximal number of possible suboptimal alignments of a pair of sequences of the length M and N is $M \times N$, because for each cell in the DP matrix (i.e. each pair of residues from the two sequences), it computes the optimal alignment which goes through the cell. In Eq. (3), the number of suboptimal alignments considered, n , was set to $n = 0.1 \times M \times N$ (i.e. top 10% high-scoring suboptimal alignments).

Threading algorithm

We have implemented a profile-to-profile alignment algorithm essentially following the paper by Wang and Dunbrack.⁸² It uses two scoring terms, a profile alignment term and a secondary structure matching term. PSI-BLAST⁸³ is used to search homologous sequences in the nonredundant protein sequence database⁸⁴ to be included in a profile with an E -value cutoff of 0.002 for both outputting and inclusion in the position specific scoring matrix. The seg program⁸⁵ is used to mask low-complexity regions in sequences. Retrieved sequences which share more than 98% sequence identity or less than 15% to the query sequence are discarded. We used the PSIC weighting and the symmetric log-odds multinomial score for aligning two profiles.⁸² Hanging gaps at the beginning and at the end of an alignment are not penalized. The opening and extension gap penalty used was 4.71 and 0.37, respectively. The secondary structure of a query protein is predicted by SABLE⁸⁶ and that of a template structure is assigned by DSSP⁸⁷ using its experimentally determined structure. The score of matching secondary structures from a query and a template is given in Eq. (26) in the Wang and Dunbrack's paper.⁸² The score from profile matching and that from the

secondary structure matching are linearly combined with the weight value of 0.7 and 0.3, respectively.

Comparison with the other quality measures of predicted structures

To understand the characteristics of SPAD for predicting the quality of predicted structures, we investigated seven other measures of different types which are expected to have the ability to indicate the model quality. These seven measures are classified into three classes, as described below.

The first class of measures is based on alignment information, which will be able to predict the global error of prediction models. These are the sequence identity, the Z -score from the threading search, and the Z -score from the PRSS program.⁸⁸ The sequence identity between a target and a template protein is simple but known to indicate the quality of the prediction model,^{4,5} thus it draws the bottom line of a reliability measure which is based on sequence analysis. The sequence identity is calculated from the optimal alignment of a given protein pair (i.e. the number of identical residues aligned divided by the length of the alignment). The Z -score from the threading search is the Z -score of the raw threading score of the optimal alignment of a target-template pair computed from the distribution of the raw threading scores of optimal alignments between the target protein to all the other proteins in the L-E set. PRSS is a program included in FASTA package which evaluates the significance of the optimal target-template alignment score by comparing it with the score distribution of alignments of target sequence to shuffled template sequences. The correlation of these measures to the global RMSD of structure models of target proteins built by Modeller (see the previous Benchmark database section) is computed and compared with the correlation with gSPAD to the global RMSD.

The second class is three sequence alignment-based measures to predict local errors. These are the degree of residue conservation of a position in the target-template profile-profile alignment, the gap density of the alignment, and the average BLOSUM45⁸⁹ score of a position in the profile-profile alignment. For an aligned position of a target-template profile-profile alignment, the residue conservation is the fraction of the most abundant residue at the position among all the residues aligned. Thus 1.0 is the perfect conservation of the residue at the position. Gaps in the position are discarded. The gap density is the fraction of gaps at the position. The average BLOSUM45 score for an alignment position is literally the average of BLOSUM45 score between all the pairs of amino acids at the position in a profile of a target and a profile of the template. Gaps are discarded. The correlation to these three measures to the local $C\alpha$ distance error is computed and compared with that of ISPAD.

Table I
Execution Time of the Programs

Target protein			Template protein			Preparing input files (s) ^b	Computing SPAD (s)
PDB ID	Length (aa)	Size of the profile ^a	PDB ID	Length (aa)	Size of the profile		
1aab	83	250	1hme	77	250	494	2
1gky	186	250	2aky	218	250	1549	25
1afrA	345	196	1mmoB	384	250	1678	229

^aThe number of sequences included in the profile. All proteins here except for 1afrA have 250 sequences in their profiles, which is the maximum number of sequences to be taken from a PSI-BLAST output file set in the current parameter setting.

^bThis calculation time includes running PSI-BLAST and SABLE (the secondary structure prediction program used), post-processing of PSI-BLAST output to make a profile.

As a measure in the last class, we use an atomic distance-dependent statistical potential, DOPE,⁶¹ which computes the energy of the tertiary structure of a protein as the sum of the energy of pairs of atoms in the protein. DOPE is a main part of Modeller's target function, which is optimized in the process of constructing a structure model from a target-template protein pair. Modeller was run with the default parameter setting. Because a DOPE score obviously depends on the number of atom pairs considered, we normalize a DOPE score by the number of all the possible pairs of heavy atoms from the first C α atom to the last C α atom in a protein. Note that the tertiary structure information DOPE captures is totally different from sequence alignment information which the alignment-based measures in the first class capture. This difference of characteristics of DOPE and the other measures is expected to appear as different performance against the three levels of the benchmark dataset. For example, in the fold level set where virtually no detectable sequence similarity exists, DOPE is still expected to perform reasonably well but not the alignment-based measures categorized in the first class. Knowing the difference between DOPE and the alignment-based measures, the reason why we still compare them is to understand the differences of them in terms of the practical predictive power of structure error of predicted structures.

Predicted structures of CASP7 targets

Using the observed correlation between g/SPAD to the global/local error of predicted structures in the L-E dataset, we made prediction of the quality of 383 prediction models of 76 CASP7 targets submitted by our group (Chen-Tan-Kihara, group number: TS536). CASP (Critical Assessment of Techniques for Protein Structure Prediction) is a world-wide protein structure prediction experiment where participants are supposed to make blind predictions from the amino acid sequence of targets posted on the CASP's official website (<http://predictioncenter.org/casp7/>).⁹⁰ We have submitted models for all the 97 targets, but here we used our models of 76 targets which are not cancelled for the competition and whose native structures were available at the time of

writing this manuscript. Our models are available at the official website (http://predictioncenter.org/download_area/CASP7/).

Both the global RMSD of the models and the C α distance error of residue positions of the models are predicted. The global RMSD of a model is predicted from its gSPAD score by applying the regression line computed with Bisquare weights from the plots of Figure 7, while the C α distance error of each residue positions can be predicted from lSPAD score from the regression line of the plots of Figure 8.

$$\text{RMSD} = \exp(0.3576 \times \ln(\text{gSPAD}) + 1.882) \quad (5)$$

$$\text{C}\alpha \text{ distance error} = \exp(0.3294 \times \ln(\text{lSPAD}) + 1.645) \quad (6)$$

Note that this regression line is computed for the correlation of all the data points in the family, the superfamily, and the fold level [Fig 7(A–C) for RMSD; Fig. 8(A–C) for C α distance error] combined, because the regression lines separately drawn for the three similarity levels are actually similar to each other and also because in an actual blind prediction it is not possible to assign the relationship of a target to a template into one of the three levels.

LGA is used for superimposition of a prediction model to its released native structure, from which the actual RMSD and the C α distance error of residues of the prediction model are computed.

Implementation of the algorithm

The execution time of computing the SPAD score is shown in Table I. Given two profiles of target and template proteins of a reasonable size (approximately 80–380 amino acid residue long), computing the SPAD score took 2 s to 4 min on a Linux machine equipped with an Intel Pentium 4 3.2 Ghz processor and 1 Gb memory. As shown in Table I, actually preparing input profiles, including running PSI-BLAST, selecting sequences from a PSI-BLAST output, computing a profile from the sequences, and running SABLE for secondary structure prediction, took

much more time than computing the SPAD score. Our main programs are implemented in C++, and available for download from our website, <http://dragon.bio.purdue.edu/subalignment/>.

RESULTS

Validation of the alignment distance

For this study we have introduced the alignment distance, ALD, as a measure of the difference between a predicted alignment and the correct alignment. ALD essentially indicates a shift of two alignments using the path representation of alignments on a DP matrix. In this section, we compare global ALD with the fraction of correctly aligned residue pairs in an alignment (i.e. the number of correctly aligned pairs divided by the length of structure alignment⁵¹; referred as the correct fraction in the following part of this section), because the correct fraction is an intuitive quality measure which has been used in some previous works.^{34,47,54} Figure 3 shows comparison between gALD and the correct fraction on the L-E dataset. Note that the greater the better for the correct fraction and the smaller the better for gALD. Overall, gALD has a good negative correlation with the correct fraction in each of the three similarity level, the family, the superfamily, and the fold. However, a discrepancy between evaluation by gALD and the correct fraction is observed in some of the alignments with a lower correct fraction (y -axis), namely, some alignments with almost zero correct fraction still have a small (good) gALD. These are alignments with a small global shift resulting in the correct fraction of zero. A typical example is an alignment of two helical proteins, where a small shift in the alignment along helices does not affect to the tertiary structure. gALD would be a better quality measure for target-template alignments for structure prediction compared to the correct fraction, because it can pick up alignments with a small global shift which do not affect the implied structure of target proteins from alignments with the zero correct fraction.

To further examine the relevance of the alignment quality measures to structure prediction, we built tertiary structures from the target-template alignments by using MODELLER and computed the RMSD of the predicted structures to their native structures. The RMSD are plotted against gALD and the correct fraction in Figure 4. Overall, gALD showed a good correlation to the RMSD of predicted structures [Fig. 4(A)]. The correlation coefficient between gALD and RMSD in the L-E set is 0.660, 0.693, and 0.612 for the family, the superfamily, and the fold, respectively. On the other hand, the correlation of the correct fraction against the RMSD significantly deteriorates in the range of a smaller value of the correct fraction [Fig. 4(B)]. That is, some alignments with a small correct fraction were still able to produce predicted

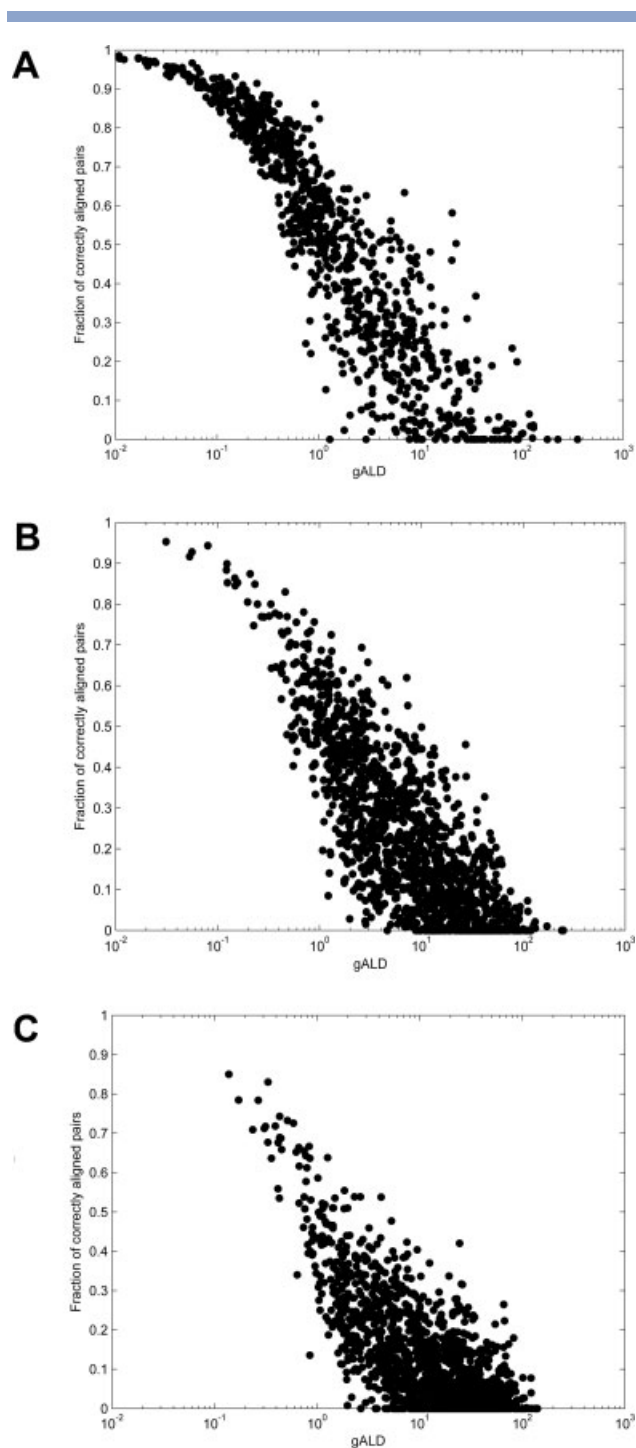


Figure 3

The correlation between gALD and the fraction of correctly aligned pairs in an alignment. The three sets of alignments in a different similarity level in the L-E dataset are used. (A) family level; (B) superfamily level; (C) fold level.

structures of a small RMSD. The correlation coefficient between the correct fraction and the RMSD is -0.696 , -0.630 , and -0.331 , for the family, the superfamily and

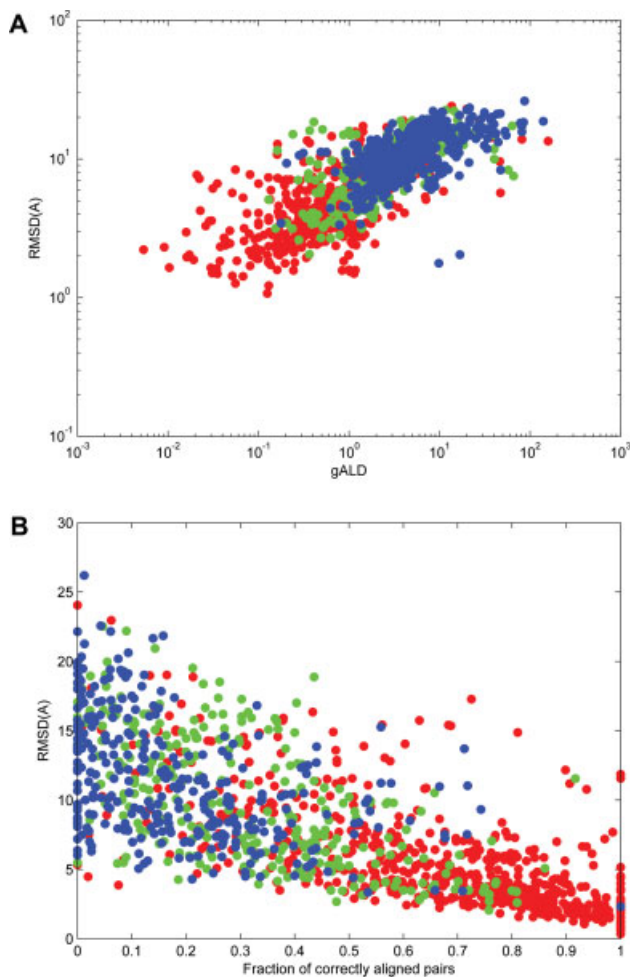


Figure 4

The correlation between the alignment quality measures against the RMSD of predicted structures. MODELLER is used for the building a tertiary structure from a target-template alignment. (A) RMSD relative to gALD; (B) RMSD relative to the fraction of correctly aligned pairs. The color shows the three similarity levels in the L-E set: red, the family; green, the superfamily; and blue, the fold level.

the fold, respectively. Because gALD is shown to be more informative for structure prediction (Fig. 4), we use ALD in the rest of the manuscript for the measure of the quality of alignments.

Correlation of the SPAD to the alignment quality

In this section and in the next, we investigate how well the SPAD score correlates with the quality of predicted structures. First, we examine here the correlation against the alignment shift error defined by ALD. Here the question is how well SPAD correlates with the alignment quality (ALD) and thus is useful to predict the alignment quality.

Figure 5 shows the correlation between gSPAD and gALD, that is the global alignment level correlation

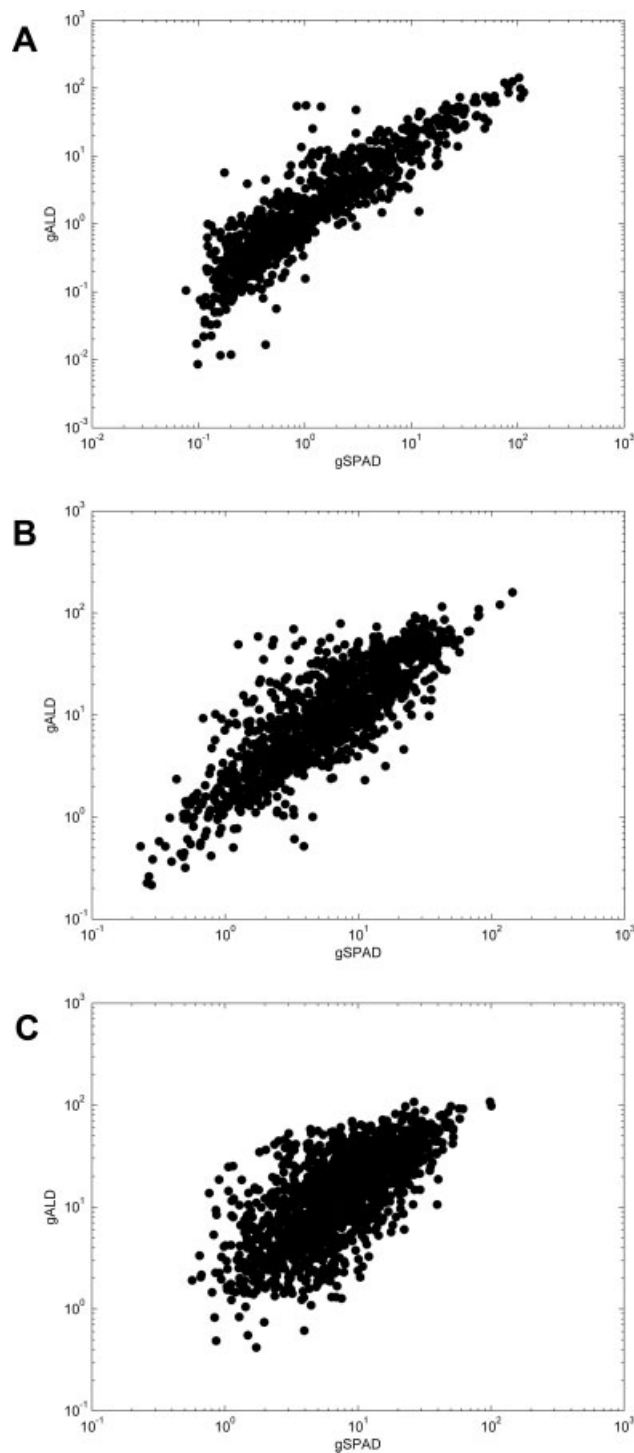


Figure 5

The correlation between the gSPAD and the global Alignment Distance (gALD). gSPAD shows a mathematical stability of the optimal alignment in comparison with suboptimal alignments. gALD shows the error of a predicted alignment when compared with a structural alignment computed by the LGA program. Alignments in three similarity levels in the L-E benchmark set are used: (A) the family level. The correlation coefficient (CC) between gSPAD and gALD is 0.899. (B) The superfamily level. CC: 0.830. (C) The fold level. CC: 0.695.

between SPAD and ALD. To examine the influence of the sequence similarity of aligned sequence pairs to the correlation, the results of the three similarity levels of the L-E dataset is separately plotted. Both axes in the plots are shown in a log scale. In all of the three similarity levels, the correlation between gSPAD and gALD is very strong: The correlation coefficient is 0.899, 0.830, and 0.695 in the family [Fig. 5(A)], the superfamily [Fig. 5(B)], and the fold level [Fig. 5(C)], respectively. The range of the gSPAD score (the x -axis) observed in the family level ranges approximately from 1×10^{-1} to 2×10^2 and the gSPAD range shifts to a larger side in the superfamily and the fold level. However, the three plots fit almost onto the same regression line. It is remarkable that a significant correlation is observed even in the fold level. Probably, this is because the profile and the secondary structure information are used in these threading alignments, not just simple pairwise sequence alignments.

We have further examined the correlation between ISPAD and IALD, that is, whether the local level SPAD can predict the local level alignment shifts (Fig. 6). This is a more challenging task than prediction of global level alignment error, but can provide valuable information for practical use of prediction models if successful. A very good correlation between ISPAD and IALD is observed in the family and the superfamily level [Fig. 6(A,B)], with the correlation coefficient being 0.580 and 0.515, respectively. In the fold level [Fig. 6(C)] the correlation decays but still has a considerable correlation coefficient of 0.377. ISPAD has a great predictive power for local alignment shift error (IALD) especially when the ISPAD value is small: when an alignment position has a ISPAD score of less than 1, 96.5%, 88.1%, and 69.8% of the cases IALD of the position is less than 5 in the family, the superfamily, and the fold level, respectively. In the global level, when global alignments have gSPAD of less than 1 (Fig. 5), 98.6, 93.5, and 73.3% of them in the family, the superfamily, and the fold level, respectively, have gALD of less than 5. Therefore, that gSPAD and ISPAD of less than 1 can be a simple criterion to find reliable global alignments or local alignment regions.

Correlation of the SPAD to the structure quality

In the previous section, we discussed that SPAD can be used for estimating global and local level alignment shifts using its good correlation to global and local ALD (Figs. 5 and 6). Here, we further examine how well SPAD correlates with the quality of predicted tertiary structures. MODELLER was used to build the tertiary structure models using the target-template alignments as input. Alignments are discarded from the analyses of this section if the sum of the N-terminal gap and the C-terminal gap exceeds 20 residues, because MODELLER tends to build dangling stretched tails in these terminal gap regions, which result in a spuriously large RMSD.

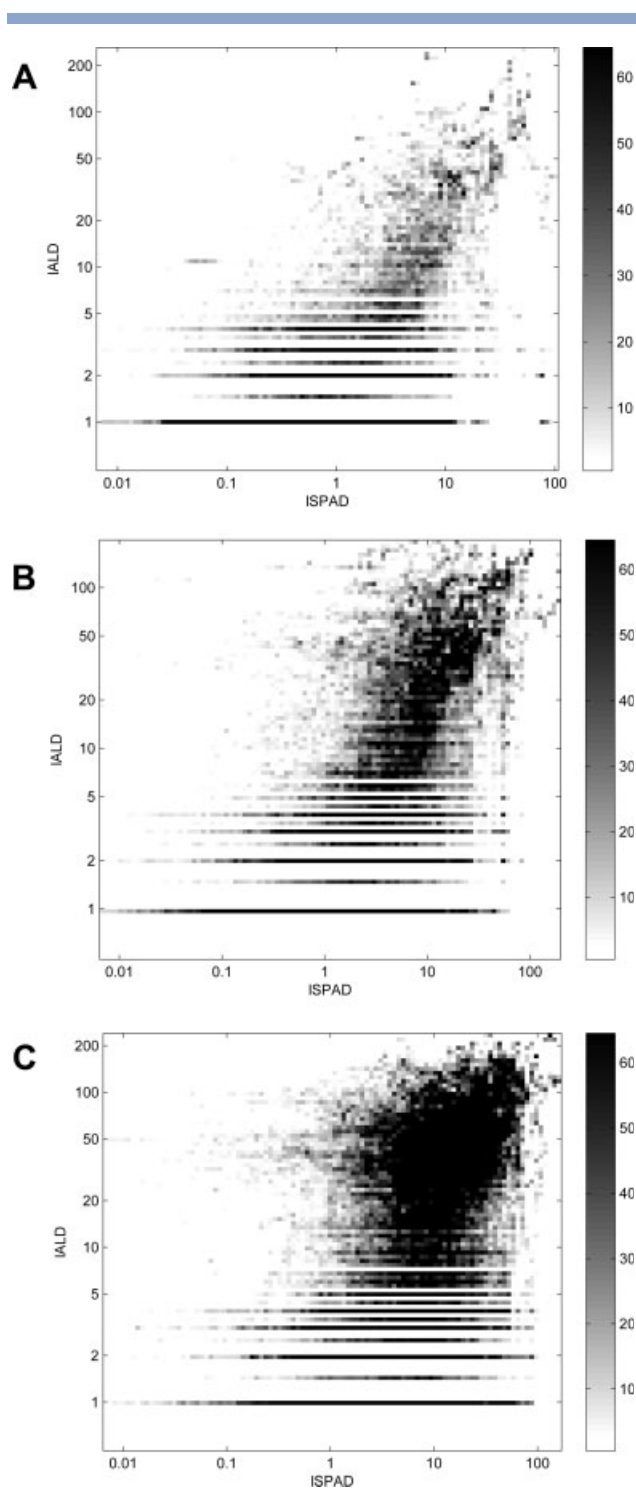


Figure 6

The correlation between the ISPAD and the local Alignment Distance (IALD) of the optimal alignment to the structure-based correct alignment. ISPAD at each position in an alignment is smoothed (averaged) by a window of a size of five residues. The gray scale represents the number of data points locating at each grid. (A) The family level alignments in the L-E set, CC: 0.580. (B) Superfamily, CC: 0.515. (C) Fold, CC: 0.377.

Figure 7 shows the correlation between the gSPAD and the RMSD of predicted structures to their native structures, that is the global structure level correlation. A significant correlation in the family and the superfamily level and a worse but still considerable correlation in the fold level are observed. The correlation coefficient is 0.598, 0.630, and 0.387 at the family [Fig. 7(A)], the superfamily [Fig. 7(B)], and the fold level [Fig. 7(C)], respectively. Furthermore, we present in Figure 8 the local structure level correlation, that is, the correlation between LSPAD of an alignment position and the prediction error of that position measured as the physical distance (\AA) between the predicted and the actual positions of the $C\alpha$ atom. The correlation coefficient of these plots is 0.565, 0.509, and 0.277 for the family [Fig. 8(A)], the superfamily [Fig. 8(B)], and the fold level [Fig. 8(C)], respectively. It might be argued that the correlation in the fold level [Fig. 8(C)] is too weak for making use of predicting local error of predicted structures. However, LSPAD is still useful even in the fold level in the small value range of LSPAD: when LSPAD is less than 0.1, 93.9, 82.4, and 71.9% of the positions in the family, the superfamily, and the fold level alignments are within the error of 4.0 \AA .

To summarize the last two sections, we demonstrated that SPAD have a quite good correlation to the quality of alignment and also predicted tertiary structures both at global and local levels thus can be used for prediction of the quality of predicted alignments and tertiary structures. The correlation is stronger in the global level (Figs. 5 and 7) than in the local level (Figs. 6 and 8). The strongest correlation was observed when two proteins are in the family level similarity and the correlation decreases as the similarity decreases into the fold level. The primary scope of the current work is the family and the superfamily level where reasonably good alignments are expected; in the other words, the sequence identity range where template-based modeling is possible. In our previous work, we have shown that sequence pairs in the fold level similarity in the L-E set have virtually no detectable sequence similarity³⁴ so that they cannot be aligned by using BLOSUM45 amino acid similarity matrix.⁸⁹ Therefore, the considerable correlation of the SPAD to the quality of alignments and predicted tertiary structures in the fold level is beyond our expectation and highly encouraging, showing the validity of applying suboptimal alignment based SPAD to a threading alignment method. It is also worthwhile to emphasize the significance of having a good correlation to the structure quality by a sequence alignment-based measure.

Comparison with the other seven quality measures

Next, we compare SPAD with the other measures in terms of the correlation to the quality of predicted structures. First, we examine the correlation between gSPAD (Fig. 7), the sequence identity (Fig. 9), the threading Z-

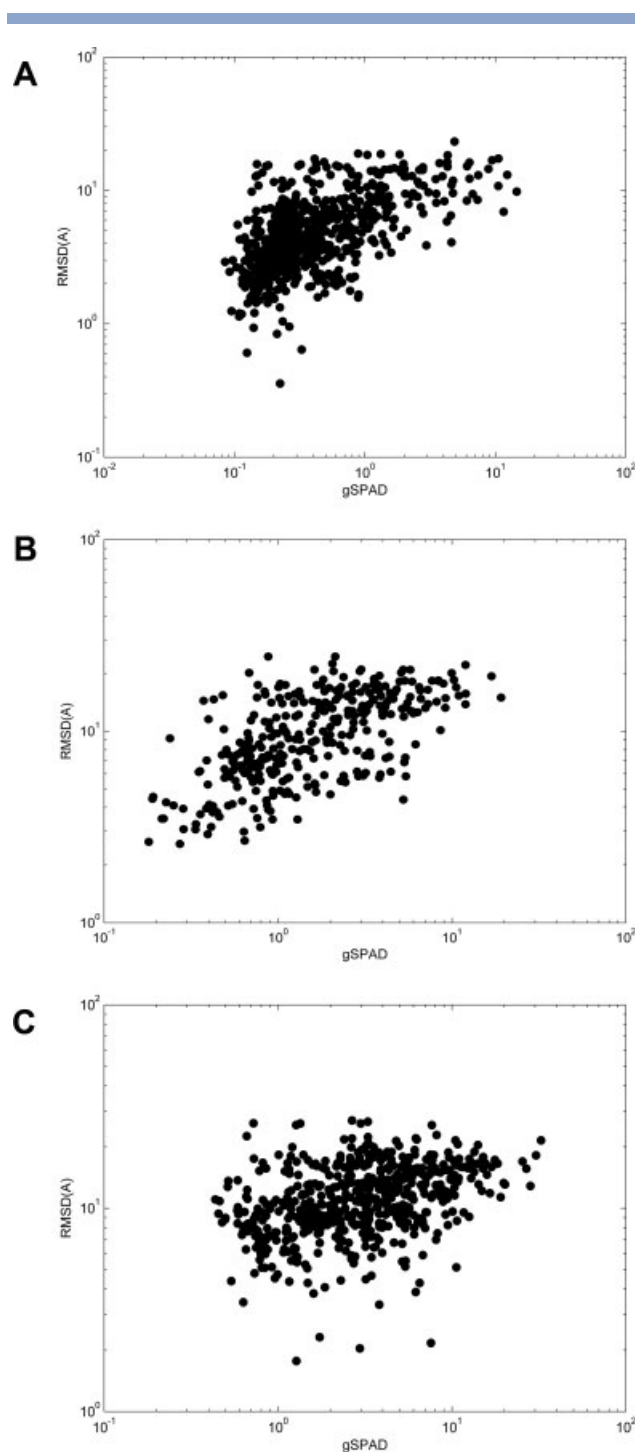
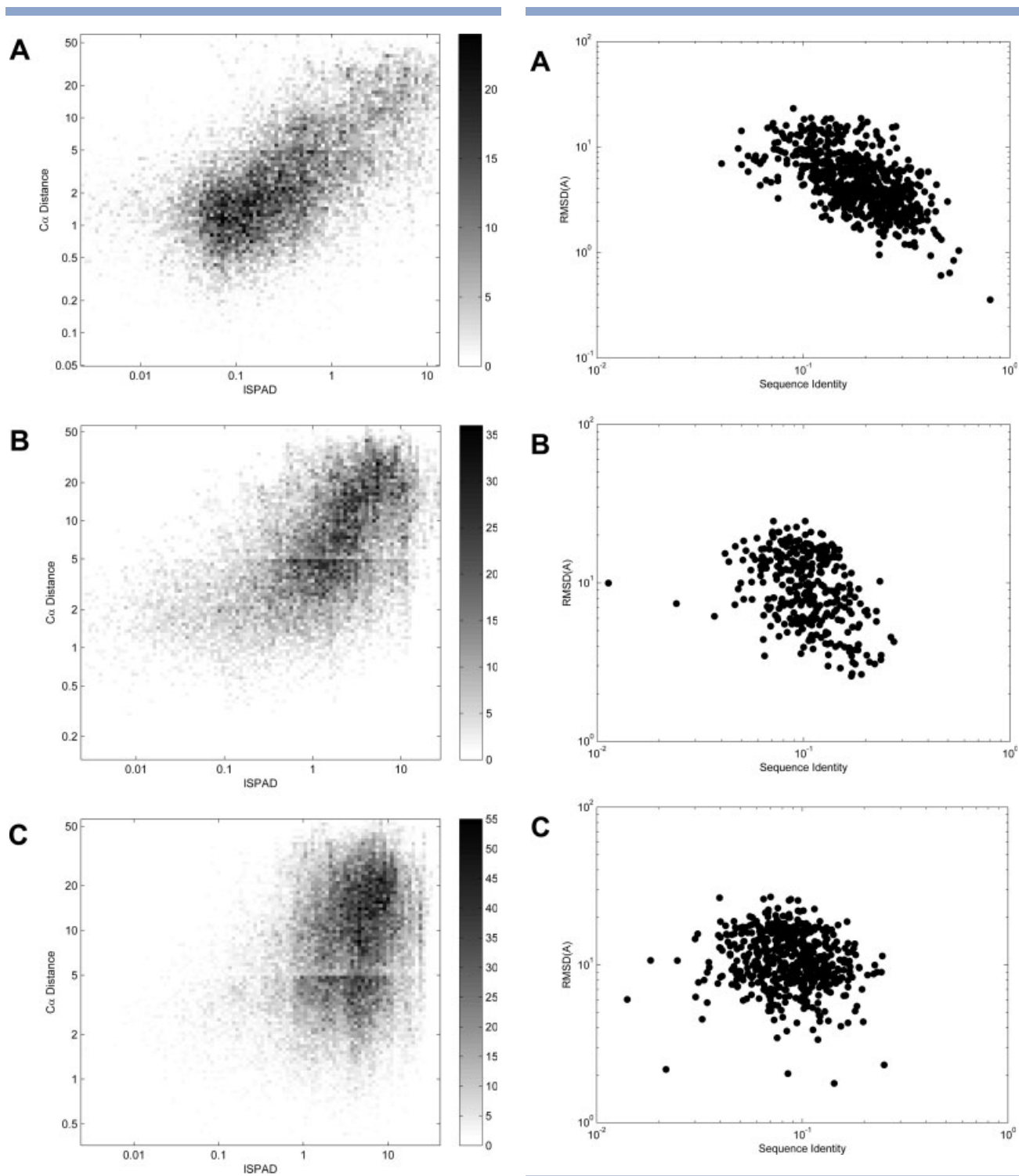


Figure 7

The correlation between gSPAD and the RMSD of predicted structures to their native structures. (A) The family level in the L-E set, CC: 0.598. (B) Superfamily, CC: 0.630. (C) Fold, CC: 0.384.

score, the Z-score by PRSS (Fig. 10), and DOPE (Fig. 11) against the global RMSD of predicted structures. Their correlation coefficients are summarized in Table II(A).

**Figure 8**

The correlation between the ISPAD and the residue level error of predicted structures. The gray scale shows the number of data points at each grid. ISPAD at each position in an alignment is averaged by a window of a size of five residues. (A) The family level in the L-E set, CC: 0.565. (B) Superfamily, CC: 0.509. (C) Fold, CC: 0.277.

Figure 9

The correlation between the sequence identity of an alignment and the global RMSD of the structure model built by MODELLER using the alignment. (A) The family level alignments in the L-E set, CC: -0.601. (B) Superfamily, CC: -0.363. (C) Fold, CC: -0.123.

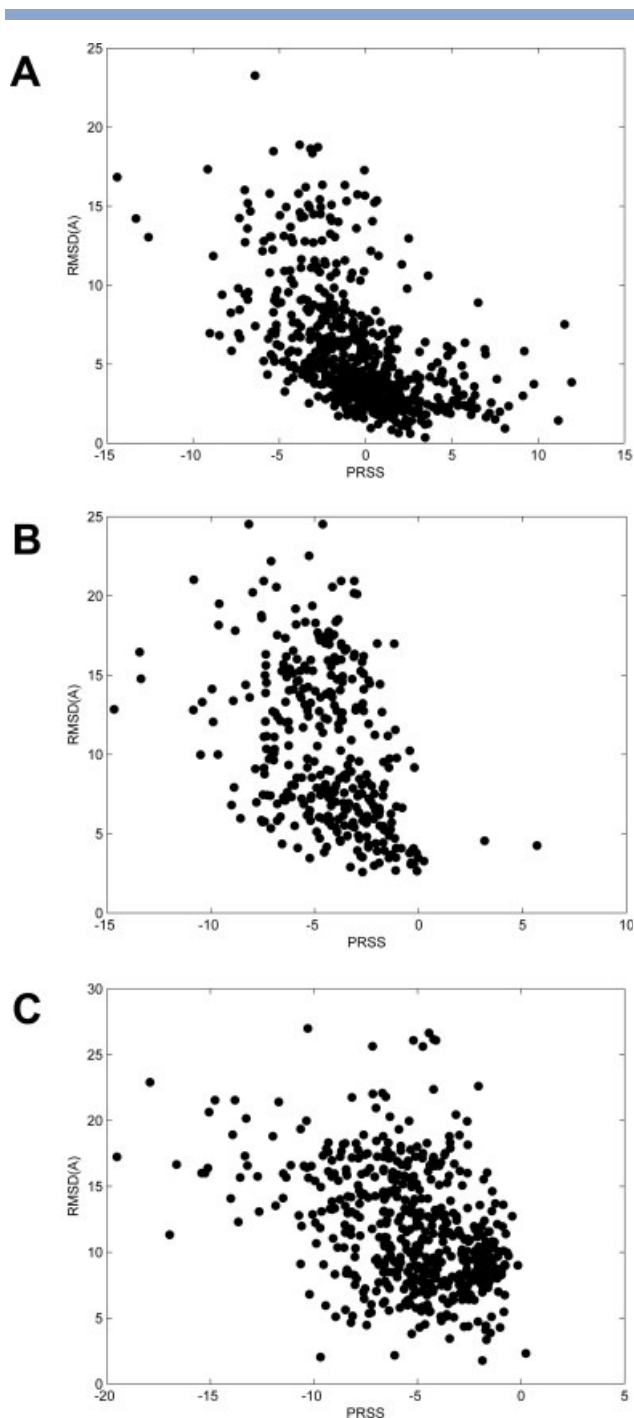


Figure 10

The correlation between the Z-score of the alignment score computed by PRSS and the global RMSD of the predicted structures. (A) Family level alignments in the L-E set, CC: -0.530 ; (B) Superfamily, CC: -0.391 ; (C) Fold, CC: -0.391 .

Note that the sequence identity, the threading Z-score, and the PRSS Z-score have a negative correlation to the RMSD. It is not an easy task for these measures except for DOPE to have a good correlation to the quality of the structure (RMSD), because they are computed from

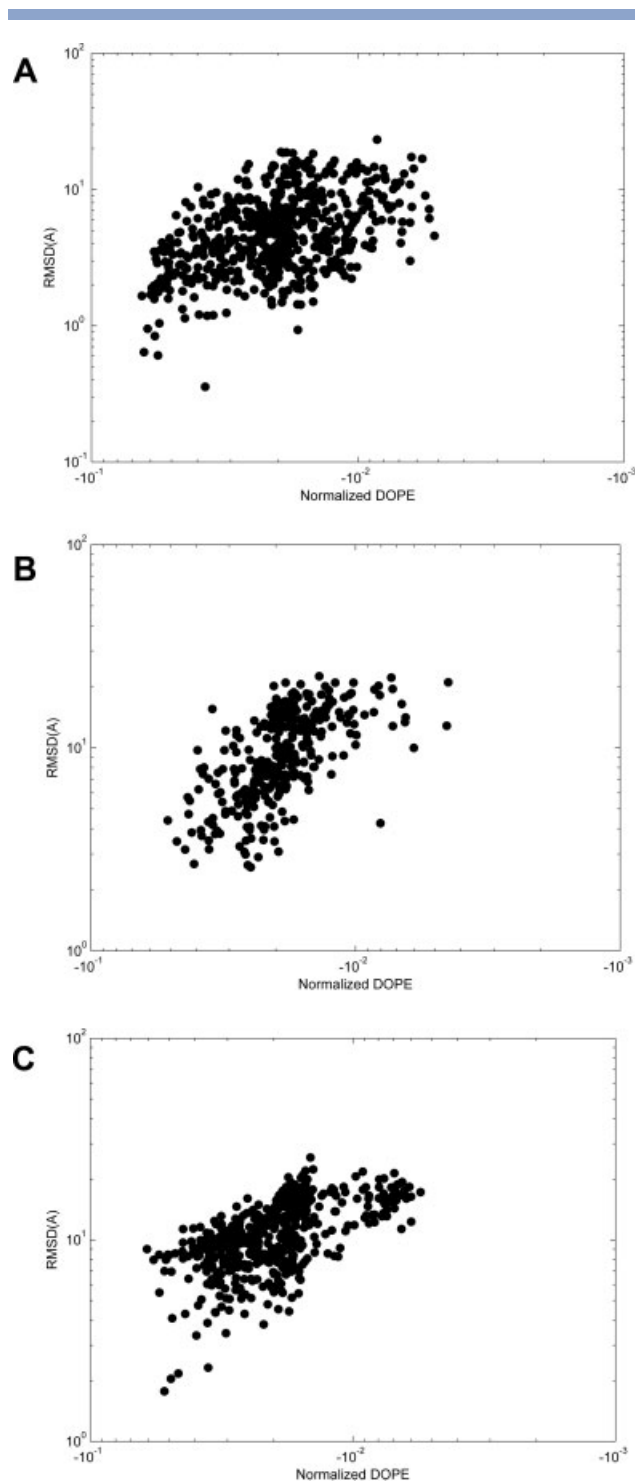


Figure 11

The correlation between the normalized DOPE score and the RMSD of predicted structures. The DOPE score given to a model is divided by the number of all the pairs of heavy atoms used for the computation. Heavy atoms from the first C α atom to the last C α atom of the protein are used, ignoring the N-terminal and C-terminal nitrogen and oxygen atoms.⁶¹ (A) Family level in the L-E set, CC: 0.453 ; (B) superfamily, CC: 0.617 ; (C) fold, CC: 0.587 .

Table II

The Correlation Coefficient between the Quality Measures and the Structure Errors in Predicted Tertiary Structures

	Family	Superfamily	Fold
A. The correlation coefficient between the measures and the global RMSD of the predicted structures to the native			
gSPAD	0.598	0.630	0.384
Sequence identity	-0.601	-0.363	-0.123
Threading Z-score ^a	-0.276	-0.216	0.107
PRSS Z-score ^a	-0.530	-0.391	-0.391
DOPE	0.453	0.617	0.587
B. The correlation coefficient between the measures and the local C α distance			
ISPAD	0.565	0.509	0.277
Conservation	-0.215	-0.092	-0.063
Gap density	0.254	0.147	0.165
Avg. BLOSUM45 ^b	-0.127	-0.040	-0.0056

The three categories of the family, the superfamily, and the fold level sets in the L-E dataset are used. A predicted structure of a target protein is built by Modeller using the optimal alignment between the target and a template structure.

^aFor gSPAD, the sequence identity, and the DOPE score, the correlation coefficient computed on the log-log plot (Figs. 7, 9, and 11) is shown, because the correlation was better than the one computed on a linear-scaling plot. On the other hand, the correlation was computed for a linear-scaling plot for the threading Z-score and the PRSS Z-score (Fig. 10), because these measures have negative values and also the log-log plot drawn by shifting the Z-score by a certain value to make them all positive gave a worse correlation.

^bFor the average BLOSUM score, the correlation computed for a linear-scaling plot is shown.

sequence alignment information. Among the four sequence alignment based measures in Table I, gSPAD and the sequence identity have almost the same correlation to the RMSD at the family level. However, at the superfamily and the fold level, the sequence identity quickly loses its correlation to the RMSD when compared with gSPAD. The PRSS Z-score has a slightly better correlation to the RMSD than gSPAD at the fold level, but worse at the family and the superfamily level, which is contrary to our expectation. We expected that the PRSS Z-score performs similarly well as gSPAD at all the three levels, because both measures essentially evaluate stability of the optimal alignment by comparing it with distribution of alternative alignments. The threading Z-score does not show a strong correlation at neither of the levels (plots not shown). This result indicates that ranking template structures in a database according to the compatibility to a target sequence, which is the common task for the threading Z-score, and predicting the absolute value of a structure model quality (i.e. RMSD) are different tasks.

In Figure 11, we show the correlation between the normalized DOPE score and the RMSD of predicted structures. DOPE is a statistical potential which evaluates the quality of a predicted tertiary structure per se. In this experiment, we normalized a DOPE score by the number of atom pairs considered, because the size of predicted structures in the data set differs. Normalization of DOPE in a physically appropriate manner to be able to evaluate proteins of different sizes is not trivial, but here we simply

used the average score per atom pair. Note that the way DOPE is used here is different from the situation of model selection in ab initio structure prediction,^{61,15} where the closest structure to native is selected from a pool of decoy structures of the same size. Here the goal is different: Given a single predicted structure, can we predict the absolute value of the quality (i.e. RMSD) of the structure? At all the three levels, the normalized DOPE showed a reasonably good correlation to the RMSD of the predicted structures. By comparing with gSPAD (Table II(A)), we found a clear difference between gSPAD and DOPE in terms of the correlation against the RMSD: at the family level, gSPAD showed a better correlation to the RMSD. At the superfamily level, gSPAD showed a slightly better correlation, and finally at the fold level, DOPE's correlation is better than gSPAD's. It is natural that DOPE shows a good correlation to the RMSD in all three levels, because it is a target function to be optimized in structure modeling process by Modeller. Actually what would be really interesting is that gSPAD performs better than DOPE at the family and the superfamily level (which is also true for the sequence identity and the PRSS Z-score at the family level), because it would imply a dominant influence of the quality of a target-template alignment to the final structure quality in the structure modeling process by Modeller.

To summarize Table II(A), at the family level, the sequence identity shows the best correlation to the RMSD (-0.601), and gSPAD follows with a slightly worse absolute value of the correlation coefficient (0.598). At the superfamily level, gSPAD was the best. At the fold level, DOPE shows the best correlation to the RMSD. However, by a head-to-head comparison among the compared measures, gSPAD would be best because it is better at least at two levels when compared with any other measures in Table II(A).

Table II(B) compares the correlation of the four measures, ISPAD (Fig. 8), the residue conservation (Fig. 12), the gap density (Fig. 13), and the average BLOSUM45 score (Fig. 14) to the local structure error. It is evident that the residue conservation, the gap density, and the average BLOSUM45 score have little correlation to the local error and that ISPAD is the best indicator for the local structure error among those compared.

Quality assessment of predicted structures for CASP7 targets

Up to the previous section, we have demonstrated that gSPAD and ISPAD of threading alignments have a significant correlation to the global RMSD and the C α distance error of structure models generated from the alignments, respectively. Next, we make actual predictions of the quality of predicted structures of CASP7 targets using SPAD. For prediction, we simply use the regression line computed for the relationship between gSPAD with the global RMSD [Eq. (5)] and ISPAD with the C α distance error [Eq. (6)].

Figure 15(A) shows the performance of our global RMSD prediction to the CASP7 models relative to their actual RMSD. Here for a computed gSPAD value of a predicted structure, we read the intersection of the global RMSD (y -axis) for the gSPAD value on the regression

line [Eq. (5)]. Overall, the prediction and the actual RMSD have a very good agreement with a correlation coefficient of 0.743. About 56 of 70 models (80.0%) predicted to have an RMSD of less than 4.0 Å actually have a better RMSD than 4.0 Å, and 144 of 206 models

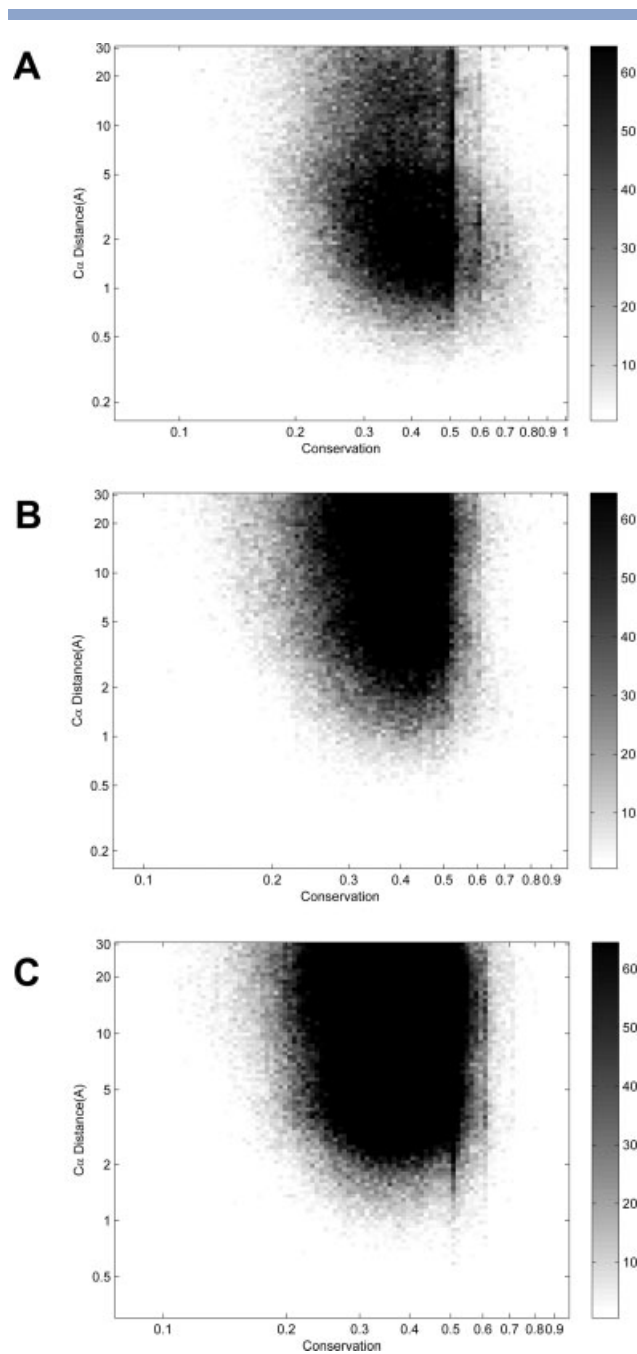


Figure 12

The correlation between the degree of residue conservation of a position and the residue level error of predicted structures. The gray scale shows the number of data points at each grid. The residue conservation at each position in an alignment is averaged by a window of a size of five residues. (A) The family level in the L-E set, CC: -0.215 . (B) Superfamily, CC: -0.092 . (C) Fold, CC: -0.063 .

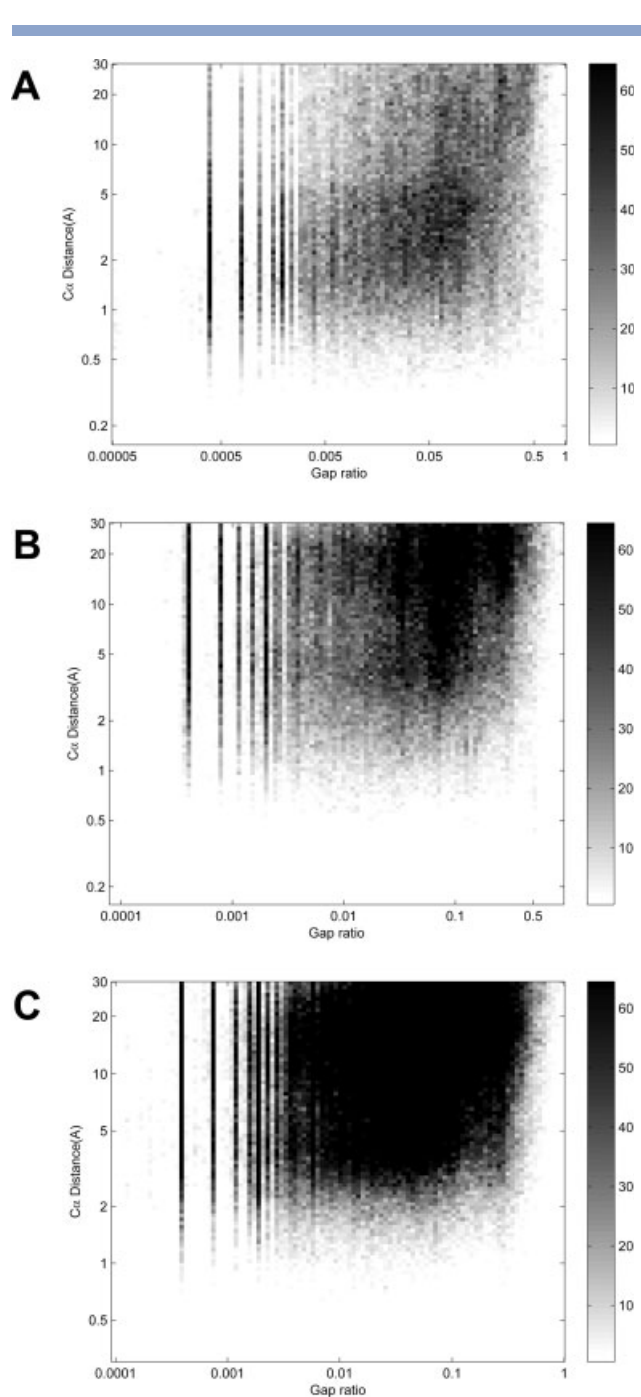


Figure 13

The correlation between the gap density of a position and the residue level error of predicted structures. The gap density at each position in an alignment is averaged by a window of a size of five residues. (A) The family level in the L-E set, CC: 0.254 . (B) Superfamily, CC: 0.147 . (C) fold, CC: 0.165 .

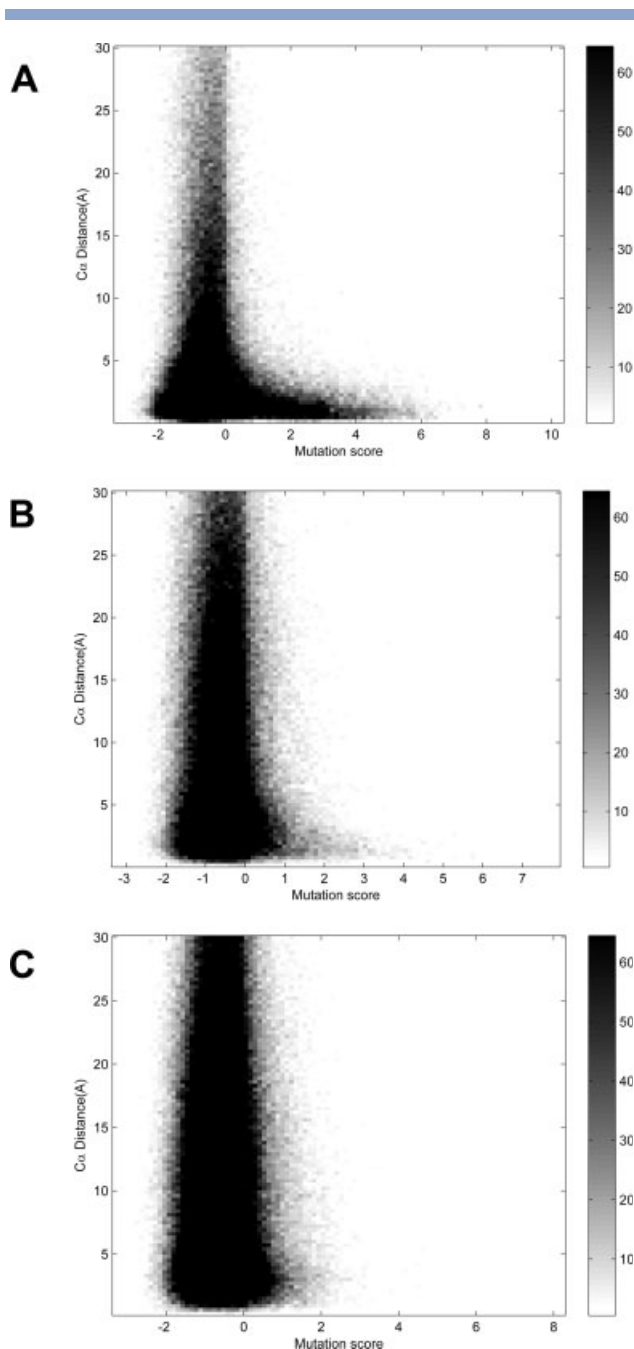


Figure 14

The correlation between the average BLOSUM45 score (the mutation score) of a position and the residue level error of predicted structures. The average BLOSUM45 score at each position in an alignment is averaged by a window of a size of five residues. (A) The family level in the L-E set, CC: -0.127 . (B) Superfamily, CC: -0.040 . (C) Fold, CC: -0.0056 .

(69.9%) predicted to have an RMSD of 6.5 \AA or smaller have a better RMSD than 6.5 \AA . We can also predict models with a bad quality from the regression line: 94.4% of the predictions that a model has an RMSD of 8.0 \AA or higher were correct (135 of 143 models). As for

the local error prediction in CASP7 models [Fig. 15(B)], the correlation is weaker (the correlation coefficient is 0.316) but still good agreement when a small or a large error is predicted: 80.7% of the residue positions predicted to have an error of 4.0 \AA or smaller (30,090 of 37,276 positions) actually have an error of 4.0 \AA or smaller and 80.2% of predictions of residue positions with an error of 8.0 \AA or higher were correct. The same data as Figure 15(A,B) are plotted in a log-log plot to show that the prediction error grows in an exponential fashion as a predicted error value becomes larger [Fig. 15(C,D)], which can be read by the almost constant “width” of the distribution of data points along the regression line of predicted and actual RMSD/ $C\alpha$ distance. This is an outcome of using regression lines drawn on log-log plots for the prediction [Eqs. (5) and (6)]. The regression lines can predict a global/local error more accurately when the value of the predicted error is small.

In the predictions above, we simply read a RMSD/local $C\alpha$ distance value on the regression line at the intersection of a given gSPAD/ISPAD value (named as pin-point prediction). Using the prediction (confidence) bounds of the regression lines, we can go one step further to roughly estimate the accuracy of predictions of the quality by predicting the range of the RMSD/local $C\alpha$ distance value. Figure 16 shows 40% upper and lower prediction bounds of for the regression line [Eq. (5)] drawn between gSPAD and RMSD. Now using the 40% upper bound line (Fig. 16), for a gSPAD value of a predicted structure, predicting the global RMSD of the structure to be an RMSD value on the upper bound line at the intersection of the gSPAD value or lower than that RMSD value is expected to be correct in 70% of the cases. Similarly, for a given gSPAD value of a predicted structure, predicting its RMSD to be the RMSD value on the lower bound line or higher is expected to be accurate in 70% of the cases. Compared to the pin-point prediction, we call it upper/lower bound prediction. Using this strategy, we predicted 139 CASP7 prediction models to have an upper bound of RMSD that is less than 6.5 \AA , which turned out to be correct for 99 models (71.2%). We also predicted 98 CASP7 models to have an RMSD which is not better than 8.0 \AA or higher than that RMSD, and the predictions of 80 models (81.6 %) of them were correct. Of course upper/lower bound prediction can be also made for predicting the $C\alpha$ distance error of residue positions: 81.7% of the residue positions predicted to have a certain value of error which is less than 4.0 \AA or smaller than that value were correctly predicted. 78.9% of the residues predicted to have a certain value of error which is more than 8.0 \AA or worse than that value were correctly predicted.

In Figure 17, four examples of predicted structures of CASP7 targets with the estimated error are shown. These four proteins are categorized as template-based modeling

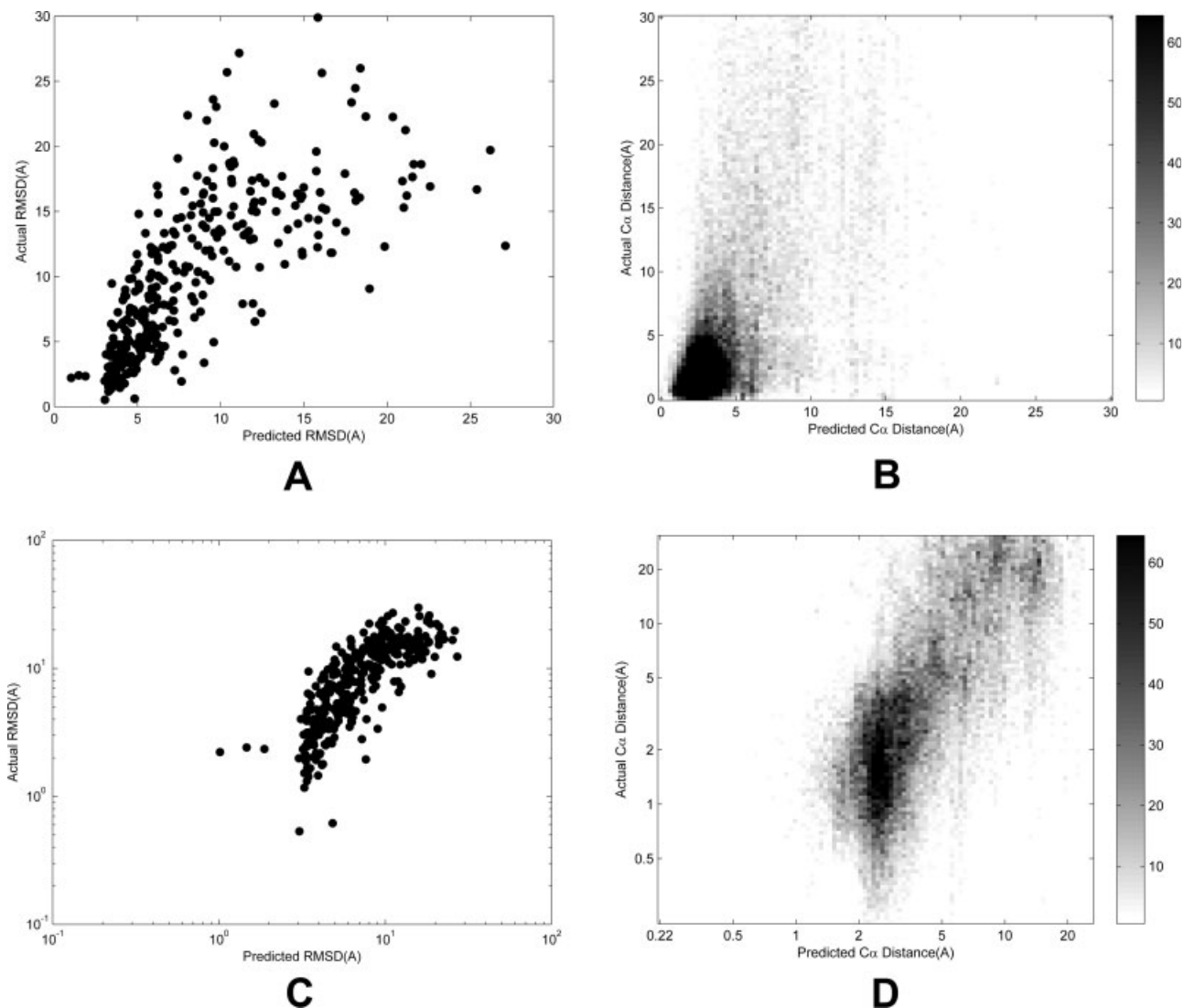


Figure 15

Predicted quality of predicted structures of CASP7 targets. All 383 prediction models submitted for 76 targets by our group are analyzed. (A) The actual RMSD of the predicted structures to the native structure is plotted relative to the predicted RMSD. The correlation coefficient is 0.743. (B) The actual error of each position of predicted structures is plotted relative to the predicted error. An error of a position of a predicted structure denotes the distance (Å) between C α atoms of the predicted and the native structure when they are superimposed. The gray scale shows the number of data points of that grid point of the plot. The correlation coefficient is 0.316. (C) The actual RMSD relative to the predicted RMSD is shown in a log–log plot. The correlation coefficient is 0.793. (D) The actual local error relative to the predicted local error is shown in a log–log plot. The correlation coefficient is 0.647.

targets by the CASP7 organizers. The sequence identity between each target and its template used is relatively low, ranging 11.4% to 27.1%. For each protein, three figures are shown: The first one on the left shows superimposition of a predicted structure (blue) of a target protein with its native structure (pink). The second figure in the middle shows the actual and predicted error (Å) at each position of the predicted structure. The last one is a “sausage” representation of the predicted structure, where the radius of the tube is proportional the estimate error of each position. The first target protein is T0369, which was stored in

PDB after the CASP7 contest with the code of 2hkv [Fig. 17(A)]. This is a 148 residue long α -helical protein with four long α helices bundled. A poor quality of the middle part of the predicted structure is well predicted by the ISPAD score. The correlation coefficient of the actual and predicted error is 0.906, and the C α distance error of 75.7% of the positions is predicted within 2.0 Å. In the case of the second target protein, T0288 [Fig. 17(B)], both the RMSD of the predicted structure and the local error were very well predicted. The actual RMSD is 3.6 Å, which is predicted to be 3.3 Å, and the C α distance error of

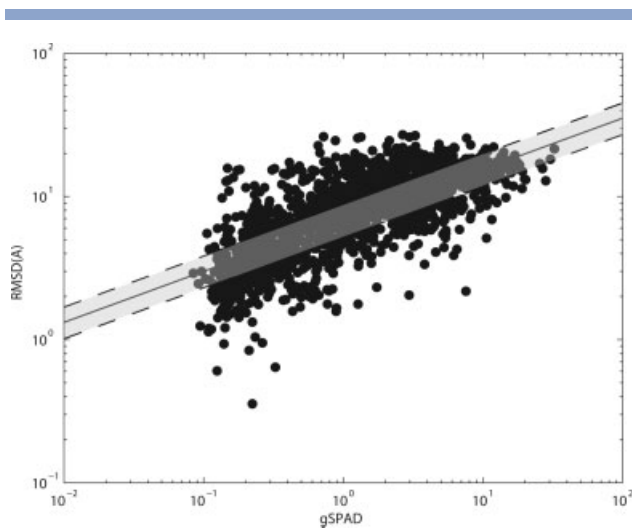


Figure 16

The regression line (solid) and the 40% upper/lower prediction bound lines (dashed lines) are drawn for the distribution of gSPAD against the RMSD of the family, the superfamily, and the fold level data combined in the L-E database (Fig. 7).

95.6% of the positions is predicted correctly within 2.0 Å. The structure of T0362 [Fig. 17(C)] was well predicted at its β -sheet region (residues 65–120) but poorly at one end of the structure which is shown on the right side of the figure [Fig. 17(C1)]. This right end of the structure consists of the C-terminal helix and a loop region with a small helix (residue 40–60). Interestingly, although the C-terminal helix and the loop region are apart on the primary sequence, both regions are accurately predicted to have a large error by the LSPAD score. T0374 [Fig. 17(D)] is an example of an α/β protein (2i6c). The error of 94.8% of the positions is predicted within 2.0 Å and the correlation coefficient of the actual and the predicted error is 0.758.

The sausage representation we used here offers intuitive understanding of a predicted structure with an estimated local error range. Alternative representation to show error would be to use a sphere with a radius proportional to the estimated error range to indicate an estimated local error of a C α position. This is similar to the displacement ellipsoids often used in dynamics study of proteins.⁹²

DISCUSSION

In this work, we have defined the SPAD score for indicating reliability of the optimal alignment by considering the diversity of suboptimal alignments. The SPAD score is shown to have an excellent correlation to global and local alignment-level errors (Figs. 5 and 6). Moreover, not only being able to indicate the alignment level error, SPAD is shown to have a significant correlation to the

quality of predicted structure both globally (i.e. the RMSD to the native structure) and locally (the C α distance error of each position). Although it has been pointed out in earlier works^{54,41} that the difference in the optimal alignment and suboptimal alignments could indicate alignment error, this is the first time that a sequence alignment based measure is applied to a threading alignment and directly compared with the global and local error in a physical distance of tertiary structures, not merely with alignment shifts. Using the established correlation between SPAD and global/local structural error, we showed that SPAD had a good performance in predicting the quality of CASP7 models. At last, we proposed the sausage representation in structure prediction, which can represent a predicted fold together with the error range of each position (Fig. 17).

Quality estimation of a predicted structure is crucial for practical use of the predicted structure in designing or interpreting biochemical experiments. It is common to report the successful prediction rate counted for a benchmark dataset in papers which describe a protein structure prediction method. However, this kind of accuracy is not always useful in a practical situation, because what a user cares is the quality of one prediction model of a protein of her/his interest. So the real question is, given a predicted structure model, what is the estimated RMSD of this model? Which region of this model is more reliable, and what is the estimated error range of that region?

At the end of this manuscript we discuss two future directions. Obviously, development of quality assessment methods of protein structures which take several different aspects of structures is of the most interest. In the current work, we focused on estimating the alignment level error because this is suitable for template-based modeling. Combining SPAD with DOPE or some other scores which directly assess the structure level error (Fig. 1) by machine learning techniques such as the support vector machine or neural network would be promising because we demonstrated that these two different types of scores are complementary to each other. Another interesting direction would be to apply the sausage representation of threading models to the generalized homology modeling, GENECOMP, developed by Kolinski *et al.*³⁵ The GENECOMP algorithm³⁵ employs an ab initio folding with a protein lattice model in the vicinity of a threading template so that the main chain of the model can move far from the template structure than conventional homology modeling algorithms.⁸⁰ In GENECOMP the vicinity of a template structure is defined by a tube of a fixed radius (called the envelope), but a variable radius along the tube depending on the estimated reliability of the template region illustrated by the sausage representation can naturally replace the fixed envelope in order to improving the accuracy of models and the sampling efficiency.

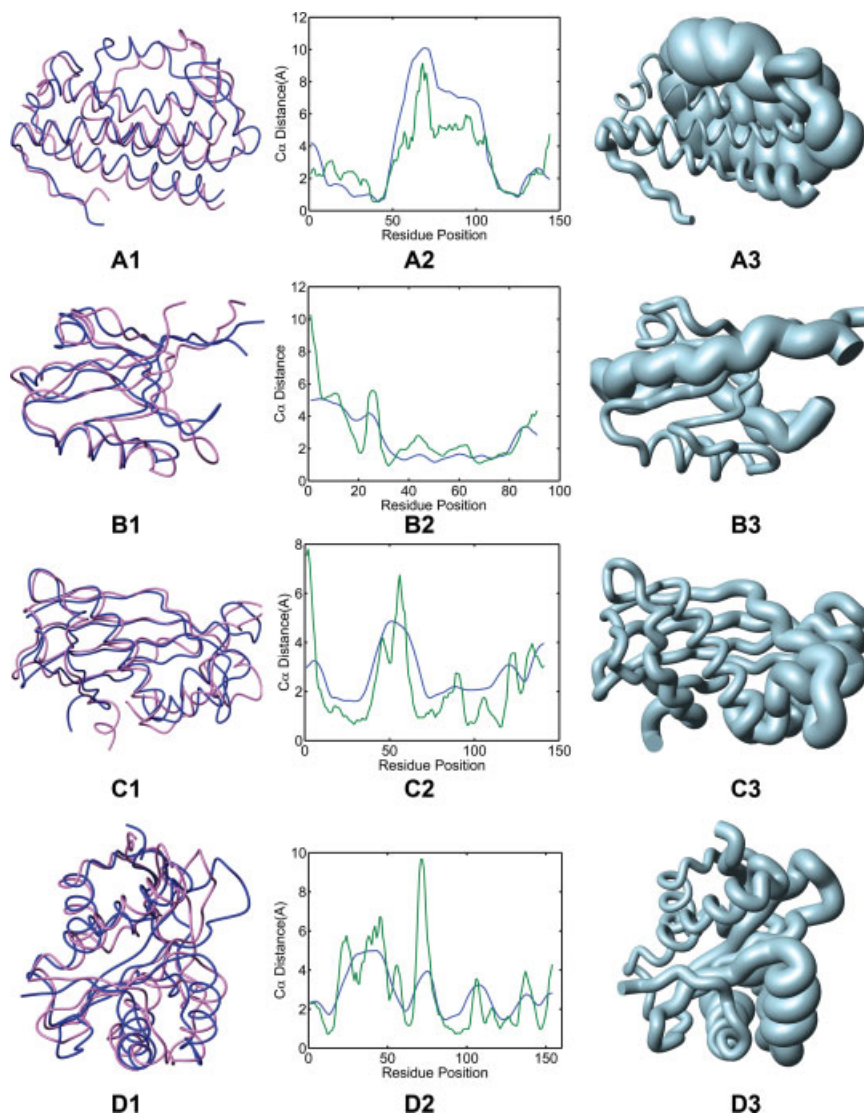


Figure 17

Examples of local error prediction for CASP7 models. For each target, three figures are shown: Left, the superimposition of a predicted structure (blue) of the target to its native structure (pink); Middle, the actual error of a predicted model (the C α distance between a predicted structure and the native structure when superimposed) (green) and the predicted error (blue) by ISPAD are plotted at each position of a target protein; right, the predicted structure is shown in a sausage representation with the radius proportional to the predicted error (Å). MolMol⁹¹ is used for producing this representation. (A) The second model submitted for Target T0369 (PDB code: 2hkv) (we denote T0369_2). The template structure used for modeling this target is 1rxqA. The sequence identity (seq. id.) between T0369 and 1rxqA is 11.4%. The global RMSD of this predicted structure to the native structure (1rxqA) is 4.0 Å, and the predicted global RMSD is 6.0 Å. (B) T0288_3 (2gzv). Template: 1t2mA; Seq. id: 27.1 %; Actual/predicted global RMSD: 3.6/3.3 Å. (C) T0362_1 (2hx5). Template: 1z54A; Seq. id: 20.5 %; Actual/predicted global RMSD: 2.9/3.1 Å. (D) T0374_3 (2i6c). Template: 1yvoA; Seq. id: 20.2 %; Actual/predicted global RMSD: 3.9/3.1 Å.

ACKNOWLEDGMENT

The authors are grateful to Preston Spratt for proof-reading the manuscript.

REFERENCES

- Jones DT. Protein structure prediction in the postgenomic era. *Curr Opin Struct Biol* 2000;10:371–379.
- Schonbrun J, Wedemeyer WJ, Baker D. Protein structure prediction in 2002. *Curr Opin Struct Biol* 2002;12:348–354.
- Petrey D, Honig B. Protein structure prediction: inroads to biology. *Mol Cell* 2005;20:811–819.
- Ginalski K. Comparative modeling for protein structure prediction. *Curr Opin Struct Biol* 2006;16:172–177.
- Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 2000;29:291–325.
- Xiang Z. Advances in homology protein structure modeling. *Curr Protein Pept Sci* 2006;7:217–227.
- Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science* 2005;309:1868–1871.

8. Jacobson M, Sali A. Comparative protein structure modeling and its applications to drug discovery. *Annu Rep Med Chem* 2004;39:259–276.
9. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ, Jr, Stoddard BL, Baker D. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* 2006;441:656–659.
10. Rosenberg M, Goldblum A. Computational protein design: a novel path to future protein drugs. *Curr Pharm Des* 2006;12:3973–3997.
11. Park S, Yang X, Saven JG. Advances in computational protein design. *Curr Opin Struct Biol* 2004;14:487–494.
12. Wilson CA, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* 2000;297:233–249.
13. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986;5:823–826.
14. Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. *Curr Opin Struct Biol* 2002;12:176–181.
15. 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 USA* 2001;98:10125–10130.
16. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. *J Mol Biol* 2001;306:1191–1199.
17. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. *Proteins* 2005;61 (Suppl):767–783.
18. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins* 2001;42:319–331.
19. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 2001;310:243–257.
20. 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.
21. Wells GA, Birkholtz LM, Joubert F, Walter RD, Louw AI. Novel properties of malarial S-adenosylmethionine decarboxylase as revealed by structural modelling. *J Mol Graph Model* 2006;24:307–318.
22. Skowronek KJ, Kosinski J, Bujnicki JM. Theoretical model of restriction endonuclease HpaI in complex with DNA, predicted by fold recognition and validated by site-directed mutagenesis. *Proteins* 2006;63:1059–1068.
23. Wojciechowski M, Skolnick J. Docking of small ligands to low-resolution and theoretically predicted receptor structures. *J Comput Chem* 2002;23:189–97.
24. Vakser IA. Low-resolution docking: prediction of complexes for underdetermined structures. *Biopolymers* 1996;39:455–464.
25. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
26. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. *Nat Biotechnol* 2000;18:283–287.
27. Hawkins T, Kihara D. Function prediction of uncharacterized proteins. *J Bioinform Comput Biol* 2007;5:1–30.
28. Forster MJ. Molecular modelling in structural biology. *Micron* 2002;33:365–384.
29. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol* 2003;374:461–491.
30. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. *Proteins* 2005;61 (Suppl 7):128–134.
31. Jones DT. Predicting novel protein folds by using FRAGFOLD. *Proteins* 2001;Suppl 5:127–132.
32. Bonneau R, Strauss C, Rohl C, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* 2002;322:65–78.
33. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. *Structure* 1997;5:1093–108.
34. Tan YH, Huang H, Kihara D. Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. *Proteins* 2006;64:587–600.
35. 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.
36. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* 2001;44:223–232.
37. Wroblewska L, Skolnick J. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. *J Comput Chem* 2007;28:2059–2066.
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
39. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* 1997;277:396–404.
40. Vingron M. Near-optimal sequence alignment. *Curr Opin Struct Biol* 1996;6:346–352.
41. Vingron M, Argos P. Determination of reliable regions in protein sequence alignments. *Protein Eng* 1990;3:565–569.
42. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;48:443–453.
43. Chao KM, Hardison RC, Miller W. Locating well-conserved regions within a pairwise alignment. *Comput Appl Biosci* 1993;9:387–396.
44. Saqi MA, Sternberg MJ. A simple method to generate non-trivial alternate alignments of protein sequences. *J Mol Biol* 1991;219:727–732.
45. 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.
46. Miyazawa S. A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng* 1995;8:999–1009.
47. Schlosshauer M, Ohlsson M. A novel approach to local reliability of sequence alignments. *Bioinformatics* 2002;18:847–854.
48. Zhang MQ, Marr TG. Alignment of molecular sequences seen as random path analysis. *J Theor Biol* 1995;174:119–129.
49. Kschischo M, Lassig M. Finite-temperature sequence alignment. *Pac Symp Biocomput* 2000;624–635.
50. Yu L, Smith TF. Positional statistical significance in sequence alignment. *J Comput Biol* 1999;6:253–259.
51. Cline M, Hughey R, Karplus K. Predicting reliable regions in protein sequence alignments. *Bioinformatics* 2002;18:306–314.
52. Tress ML, Jones D, Valencia A. Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol* 2003;330:705–718.
53. Dopazo J. A new index to find regions showing an unexpected variability or conservation in sequence alignments. *Comput Appl Biosci* 1997;13:313–317.
54. Mevissen HT, Vingron M. Quantifying the local reliability of a sequence alignment. *Protein Eng* 1996;9:127–132.
55. Laskowski RA, MacArthur MW, Thornton JM. Validation of protein models derived from experiment. *Curr Opin Struct Biol* 1998;8:631–639.
56. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. Procheck—a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 1993;26:283–291.
57. Hoof RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature* 1996;381:272.
58. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 1993;2:1511–1519.
59. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 1990;213:859–883.

60. 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.
61. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* 2006;15:2507–2524.
62. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 2002;11:2714–2726.
63. Rojnuckarin A, Subramaniam S. Knowledge-based interaction potentials for proteins. *Proteins* 1999;36:54–67.
64. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol* 2000;10:139–145.
65. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. *Proteins* 2004;56:502–518.
66. Matsuo Y, Nishikawa K. Protein structural similarities predicted by a sequence-structure compatibility method. *Protein Sci* 1994;3:2055–2063.
67. Baumann G, Frommel C, Sander C. Polarity as a criterion in protein design. *Protein Eng* 1989;2:329–334.
68. Holm L, Sander C. Evaluation of protein models by atomic solvation preference. *J Mol Biol* 1992;225:93–105.
69. Gregoret LM, Cohen FE. Protein folding. Effect of packing density on chain conformation. *J Mol Biol* 1991;219:109–122.
70. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. *Protein Sci* 2006;15:1653–1666.
71. Wallner B, Elofsson A. Can correct protein models be identified? *Protein Sci* 2003;12:1073–1086.
72. Tosatto SC. The victor/FRST function for model quality estimation. *J Comput Biol* 2005;12:1316–1327.
73. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 1999;287:797–815.
74. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. *J Mol Biol* 2000;295:613–25.
75. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–242.
76. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* 2002;30:264–267.
77. Zemla A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* 2003;31:3370–3374.
78. Thompson JD, Koehl P, Ripp R, Poch O. BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins* 2005;61:127–136.
79. Stebbings LA, Mizuguchi K. HOMSTRAD: recent developments of the homologous protein structure alignment database. *Nucleic Acids Res* 2004;32(Database issue):D203–D207.
80. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;234:779–815.
81. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;147:195–197.
82. Wang G, Dunbrack RL, Jr. Scoring profile-to-profile sequence alignments. *Protein Sci* 2004;13:1612–1626.
83. 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.
84. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2005;33(Database issue):D39–D45.
85. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 1994;18:269–285.
86. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* 2005;59:467–475.
87. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–637.
88. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;85:2444–2448.
89. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:10915–10919.
90. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round 6. *Proteins* 2005;61 (Suppl):73–77.
91. Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* 1996;14:51–32.
92. Northrup SH, Pear MR, Morgan JD, McCammon JA, Karplus M. Molecular dynamics of ferrocycytochrome c. Magnitude and anisotropy of atomic displacements. *J Mol Biol* 1981;153:1087–1109.