

Sub-AQUA: real-value quality assessment of protein structure models

Yifeng David Yang¹, Preston Spratt¹, Hao Chen¹,
Changsoon Park² and Daisuke Kihara^{1,3,4,5}

¹Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN 47907, USA, ²Department of Statistics, College of Natural Science, Chung-Ang University, Seoul, Korea, ³Department of Computer Science, College of Science, Purdue University, West Lafayette, IN 47907, USA and ⁴Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, IN 47907, USA

⁵To whom correspondence should be addressed.
E-mail: dkihara@purdue.edu

Received February 9, 2009; revised April 21, 2010;
accepted May 3, 2010

Edited by Jeffery Saven

Computational protein tertiary structure prediction has made significant progress over the past years. However, most of the existing structure prediction methods are not equipped with functionality to predict accuracy of constructed models. Knowing the accuracy of a structure model is crucial for its practical use since the accuracy determines potential applications of the model. Here we have developed quality assessment methods, which predict real value of the global and local quality of protein structure models. The global quality of a model is defined as the root mean square deviation (RMSD) and the LGA score to its native structure. The local quality is defined as the distance between the corresponding C α positions of a model and its native structure when they are superimposed. Three regression methods are employed to combine different types of quality assessment measures of models, including alignment-level scores, residue-position level scores, atomic-detailed structure level scores and composite scores. The regression models were tested on a large benchmark data set of template-based protein structure models of various qualities. In predicting RMSD and the LGA score, a combination of two terms, length-normalized SPAD, a score that assesses alignment stability by considering suboptimal alignments, and Verify3D normalized by the square of the model length shows a significant performance, achieving 97.1 and 83.6% accuracy in identifying models with an RMSD of <2 and 6 Å, respectively. For predicting the local quality of models, we find that a two-step approach, in which the global RMSD predicted in the first step is further combined with the other terms, can dramatically increase the accuracy. Finally, the developed regression equations are applied to assess the quality of structure models of whole *E. coli* proteome.

Keywords: error estimation/homology modeling/model quality assessment/protein structure prediction/regression analysis/threading

Introduction

The number of experimentally solved protein structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000) is growing at a rapid pace. Recently, this growth has been boosted by a number of structural genomics projects launched around the world (Chandonia and Brenner, 2006; Levitt, 2007). Structural genomics projects have been making significant contribution in expanding the coverage of known superfamilies and folds, as one of their long-term goals is to provide representative structures for the entire protein folds (Todd *et al.*, 2005; Levitt, 2007). Each of these new structures would enable template-based modeling of dozens of homologous proteins.

On the other hand, the protein structure prediction community has made steady progress as evidenced in the biennial critical assessment of techniques for protein structure prediction (CASP) experiments (Kryshtafovych *et al.*, 2005). However, constructing an accurate model, for example, one with a root mean square deviation (RMSD) of <1.5 Å to the native structure, is still not always possible even with state-of-the-art methods. The accuracy of a structure model depends on various aspects, with the quality of template structures available for a target being among the most influential factors in the case of template-based modeling, i.e. homology modeling (Al-Lazikani *et al.*, 2001; Ginalski, 2006; Xiang, 2006) and threading (Skolnick and Kihara, 2001; Zhou and Zhou, 2005a; Qu *et al.*, 2009). Indeed it is not uncommon that the quality of a structure model becomes very poor, say, an RMSD of over 10 Å to the native, even if a structure of a homologous protein is used as the template of modeling. Such situation can happen because evolutionary-related proteins often have a large structural diversity despite of their sequence similarity (Reeves *et al.*, 2006) and also because current methods are still weak in modeling unaligned regions where the target sequence is aligned with gaps in the target–template alignment (Eswar *et al.*, 2008). In threading, it happens frequently that a model captures only the global fold of a target protein correctly and hence has an RMSD of around 6 Å.

Importantly, not only highly accurate models but also models of a moderate accuracy are still useful for many purposes (Baker and Sali, 2001). Models of an atomic-detailed accuracy with an RMSD of 1–1.5 Å are useful for almost any application where structure information can be useful, including studying enzymatic mechanism and protein design (Ashworth *et al.*, 2006; Jiang *et al.*, 2008; Rothlisberger *et al.*, 2008). Models with a correct backbone orientation (e.g. an RMSD of 4–6 Å to the native structure) can be used for applications that need residue position level accuracy, such as designing and interpreting site-directed mutagenesis experiments (Skowronek *et al.*, 2006; Wells *et al.*, 2006). Models with a slightly higher RMSD but have almost correct

overall fold may be used for predicting function from their global fold (Kihara and Skolnick, 2004) or for identifying local functional sites (Arakaki et al., 2004; Laskowski et al., 2005). Most of the structure prediction methods do not provide estimated error of produced structure models. Thus, users of predicted models cannot even know if the model is totally wrong with an RMSD of over 10 Å or not. Applications of a predicted model listed above are only possible when users know the accuracy of the model. Therefore, it is crucial to establish methods, which assess the quality of protein structure models so that they can be applied for suitable purposes for an estimated accuracy (Kihara et al., 2009).

Existing model quality assessment programs (MQAPs) can be roughly classified into three categories by their primary purpose. First, in experimental structural biology, it is a routine to examine stereochemical property of solved protein structures. PROCHECK (Laskowski et al., 1993) and WHATCHECK (Hooft et al., 1996), which are among popular software for this purpose, examine regularity of the bond lengths, angles, torsion angles and atom contacts. Second, in the structure prediction field, especially for *ab initio* prediction, which produces thousands of alternative models for a target protein, various methods are developed to rank models so that the best models can be selected from the pool (Zhou and Zhou, 2002; Skolnick, 2006; Lu et al., 2008). Reranking and selecting plausible models are also needed in a meta-server approach, which combines different models to construct a final model (Kosinski et al., 2005; Terashi et al., 2007). The third type of quality assessment is to predict the absolute value of the quality of a model, such as the RMSD to the native structure (Wallner and Elofsson, 2003; Eramian et al., 2006; Eramian et al., 2008; Mereghetti et al., 2008; Pawlowski et al., 2008).

Among more than two dozens of MQAPs proposed previously, most of them are designed to rerank models (i.e. the second type of MQAPs) (Kihara et al., 2009). Structural features, which those MQAPs evaluate vary from the quality of target-template alignments (Zhang et al., 1999; Tondel, 2004; Lee et al., 2007; Chen and Kihara, 2008), knowledge-based residue level statistical potentials (Hooft et al., 1996; Pontius et al., 1996; Eisenberg et al., 1997; Melo et al., 2002; Pettitt et al., 2005; Tosatto and Battistutta, 2007), atomic-detailed structures (Laskowski et al., 1993; Lu and Skolnick, 2001; Zhou and Zhou, 2002; Davis et al., 2004; Shen and Sali, 2006; Lu et al., 2008), to physics-based atomic-detailed potentials (Feig and Brooks, III, 2002; Lu and Skolnick, 2003; Wroblewska et al., 2008). Finally, there are several MQAPs, which combine several terms to construct a composite score (Wallner and Elofsson, 2006; McGuffin, 2007). Refer to our recent review article for further discussions (Kihara et al., 2009).

Compared with well-populated MQAPs of the second type, not many works have been done in developing methods for real-value quality prediction (i.e. the third type). The real-value quality prediction would be more difficult than ranking models because most of structure features previously used for structure prediction and quality assessment do not have clear rationale to have direct indication of the real-value quality. Exceptions would be measures that capture target-template sequence similarity, since there is a well-established relationship between the sequence identity and the RMSD of

protein structures (Chothia and Lesk, 1986). Despite its potential difficulty, MQAPs of the third type are of practical importance since the expected real-value accuracy of models enables end users of structure prediction to decide whether to use the models for a certain purpose. In the era that more template-based modeling become possible due to the increase of solved protein structures, MQAPs that can bridge computational protein structure prediction to experimental applications will become very important.

In this work, we develop MQAPs, which predict real value of the global and local quality of computational structural models by combining multiple quality assessment scores. The main focus of our methods is template-based protein models because template-based modeling is the most accurate and practical structure prediction technique and also because it is very important in the context of the structural genomics projects. We define the global quality of a structure model as the RMSD value of the model to its native structure. The local quality of a model is defined as the distance between C α atoms of corresponding residues in the model and its native structure when the two structures are superimposed. We employ regression analysis to combine different types of quality assessment measures, including ones for evaluating alignment-level errors, the residue environment, atom contacts, torsion angles, and composite scores. Combinations of scores are tested on a large benchmark data set of template-based models of a wide range of quality. We find the SubOptimal Alignment Diversity (SPAD) score, which assesses the stability of target-template alignment by considering suboptimal alignments, with log transformation shows the best linear correlation to the global RMSD of models among individual scores tested. In predicting RMSD, linear regression with a combination of two terms, length-normalized SPAD and Verify3D normalized by the square of the model length shows a significant performance, achieving 97.1 and 83.6% accuracy in identifying models with an RMSD of less than 2 and 6 Å, respectively. To the best of our knowledge, this is the first MQAP, which uses sub-optimal alignment information as a scoring term. As for predicting the local quality of models, we find that a two-step approach, in which the global RMSD predicted in the first step is further combined with the other terms, can dramatically increase the accuracy.

Methods

Benchmark data set of template-based protein structure models

A data set of template-based models are constructed using the Lindahl and Elofsson's data set (L-E data set) (Lindahl and Elofsson, 2000). 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 (Andreeva et al., 2008). Following the SCOP classification, a set of pair-wise alignments are constructed in each similarity level, using a protein tertiary structure alignment program, LGA (Zemla, 2003). 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 in the same family; pairs in the fold level set belong to the same fold but not in 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. The pairwise alignment of the two proteins is constructed by a profile-based threading algorithm (Chen and Kihara, 2010, submitted for publication). Using the alignments, a tertiary structure model of the target protein is constructed by a homology modeling software, Modeller (version 8v2) (Sali and Blundell, 1993) with the default parameter setting.

The quality of the predicted tertiary structure of the target protein is evaluated by calculating the RMSD between the predicted structure and the experimentally determined structure of the target protein using the LGA program. The distance of corresponding α carbon atoms between the predicted and the experimentally determined structure is used as the local quality of the predicted structure (the C α distance).

Quality assessment methods and scores

We combine scores from different methods ranging from sequence-based scores to those which evaluate atomic contacts. The scores used are PRSS (Pearson and Lipman, 1988), SPAD (Chen and Kihara, 2008), Verify3D (Bowie *et al.*, 1991; Luthy *et al.*, 1992), ERRAT (Colovos and Yeates, 1993), PROCHECK (Morris *et al.*, 1992; Laskowski *et al.*, 1993), ANOLEA (Melo *et al.*, 1997; Melo and Feytmans, 1997), GA341 (John and Sali, 2003), ProQ (Wallner and Elofsson, 2003), Discrete Optimized Protein Energy (DOPE; Shen and Sali, 2006) and TAP (Tosatto and Battistutta, 2007). Below brief explanation of each scoring term is provided.

PRSS. PRSS is a program included in the FASTA package (Pearson and Lipman, 1988). It evaluates the significance of the raw alignment score of the optimal target–template alignment by comparing it with a distribution of alignment scores of the target sequence and shuffled template sequences. The significance is expressed as a Z-score (Zs), which is denoted as PRSS_Zscore in this study. The larger positive value, the more significant.

SPAD. SPAD (Chen and Kihara, 2008) quantifies the consistency of a template–target alignment with a set of suboptimal alignments. Suboptimal alignments are generated by an algorithm proposed by (Vingron and Argos, 1990). The SPAD score of an alignment position (local SPAD or ISPAD) is an averaged distance from that position on a dynamic programming path matrix to the equivalent position in each suboptimal alignment. The global SPAD score (gSPAD) of a whole alignment is the average of ISPAD at all the alignment positions. Thus, SPAD is a non-negative value score. A small SPAD value means suboptimal alignments are consistent and indicates a more accurate predicted structure.

Verify3D. Verify3D assesses the fitness of a residue in a model to its structural environment, which is defined as the

burial area, fraction of side-chain area covered by polar residues and the secondary structure. The structural environment is classified into 18 classes. The idea of using structural environment was originally applied to the inverse folding problem (Bowie *et al.*, 1991). Later it was successfully applied to verify tertiary protein structures solved by X-ray crystallography, nuclear magnetic resonance, and computational methods (Luthy *et al.*, 1992). A large positive value indicates that a residue fits well to its structural environment.

ERRAT. ERRAT evaluates a structure by considering non-bonded interactions between carbon, nitrogen and oxygen atoms (Colovos and Yeates, 1993). The number of such contacts in a nine-residue window is counted and the counts are compared with a reference distribution. A residue in a model is evaluated by the percentage confidence limit within which the raw score of the residue locates when compared with the score distribution of residues in correct protein structures. Thus, the value ranges from 0 to 100% (the lower, the better). The percentage of residues in a model, which are not included within 95 percentile is provided as the overall quality (the smaller, the better).

PROCHECK. PROCHECK examines stereochemical parameters, such as the bond length, hydrogen bonds, atom contacts, bond angles, and torsion angles as compared with those derived from well-refined high-resolution structures (Laskowski *et al.*, 1993). In this study, two parameters from the PROCHECK output are used. The PROCHECK1 score is the percentage of residues that fall in the disallowed region in the Ramachandran plot. The PROCHECK2 score is the percentage of residues with ‘bad’ contacts.

ANOLEA. ANOLEA (Atomic Non-Local Environment Assessment) evaluates the ‘non-local environment’ of heavy atoms in a model (Melo and Feytmans, 1997). The energy of a heavy atom consists of two knowledge-based terms, a distant dependent pair-wise atom-contact potential and an accessible surface mean force potential that considers the number of atoms within spheres of different sizes. In this study, two parameters from the ANOLEA outputs are used: the percentage of residues with high energy, which is denoted as ANOLEA1 and the non-local normalized energy Zs, denoted as ANOLEA2.

GA341. The GA341 score integrates a Z-score (Zs) for combined (distance-dependent contact and accessibility) residue-level statistical potentials (Melo *et al.*, 2002), the target–template sequence identity (Si), and a measure of structural compactness (Sc) (Melo *et al.*, 2002; Pieper *et al.*, 2006). Using these scores, the GA341 score is defined as $GA341 = 1 - [\cos(Si)]^{(Si+Sc)/\exp(Zs)}$. The GA341 score ranges from 0 for models that are likely to have an incorrect fold to 1 for models that have a comparable quality to low resolution X-ray structures (John and Sali, 2003). In addition to the Z-score of the GA341 score, we also use the pG score provided by the GA341 program, which is the probability that the model is good (more than 30% of its C α atoms locate within 3.5 Å from their correct position). pG is computed using a Bayesian classifier, which uses the GA341 score and the length of the model as input. The pG score ranges from 0 to 1.0 (the larger value, the better).

DOPE. DOPE is an atomic distance-dependent statistical potential (Shen and Sali, 2006). DOPE is the target function to be optimized in the MODELLER program. The larger negative value of the score, the better. Since the DOPE score is obviously dependent on the number of atoms, we normalized the score by the square of the number of heavy atoms in a model.

TAP. The TAP score measures the fitness of a sequence to a structure based on torsion angle propensities of amino acids. The score given to each amino acid is summed over the length of the protein and normalized by the maximum and minimum possible score of the protein. TAP becomes close to 1 for the native structure and close to 0 for the models with largely incompatible sequences (Tosatto and Battistutta, 2007).

ProQ. ProQ is a neural network-based method for predicting the quality of a protein model from five structural features, atom–atom contacts, residue–residue contacts, solvent accessibility surface, agreement with predicted secondary structure and the fatness, which is defined as the ratio of the longest and the shortest axis of the ellipsoid fitted to the model. Two neural networks were trained, one for predicting the LGscore (Levitt and Gerstein, 1998) of models (ProQ-LG) and another one for the MaxSub score (Siew et al., 2000) (ProQ-MX). The LGscore and the MaxSub score indicate structure similarity of a model to its native structure. MaxSub score gives 1 for the native structure and a value close to 0 for a significantly wrong model. LGscore of 1.5 or higher indicates a correct model.

Regression analysis

We used three regression analysis methods to combine scoring terms to predict the RMSD of models (the global quality) and the C α distance of residues in models (the local quality).

Linear regression. Linear regression models the relationship between a dependent (response) variable Y and the independent (predictor) variables X_i , $i = 1, \dots, P - 1$. The model has the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p-1} X_{p-1} + \varepsilon, \quad (1)$$

where β_0 is the intercept, β_i s are regression coefficients for the respective independent variables and ε is a random term. In this study, the RMSD of the models or the C α distance of residues in the structure models are the response variable Y , while the various quality assessment scores and some of their transformations are the predictor variables. To estimate the regression coefficients, the least squares method is used, which minimizes the sum of the residuals (i.e. the difference of the predicted and observed values for the response variable).

Logistic regression. Logistic regression computes the probability of occurrence of a discrete event based on several independent variables. For binary response, where $Y_i = 0$ or

1 and $p(Y_i = 1) = \pi_i$, a logistic model is written as:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad (2)$$

where π_i denotes the probability of $Y_i = 1$. The function $\log(\pi/1 - \pi)$ is called the *logit* function. The regression coefficients are estimated by the maximum likelihood method. Using logistic regression, we predict if a model is correct (i.e. an RMSD below a cutoff value, 2, 4, 6, or 8 Å) or not.

LOESS regression. LOESS (locally weighted scatter plot smoothing) regression (Cleveland, 1979; Cleveland and Devlin, 1988) fits a low-degree polynomial to each local subset of predictor variable values, for which response is estimated. The polynomial is fit using weighted least squares, giving more weight to points near the point whose response is estimated and less weight to points further away. Two parameters control the behavior of the LOESS regression. The first one is called bandwidth or smoothing parameter denoted as α . Another parameter is the degree of the local polynomials, λ . In this study, α is set to 0.5 and λ is set to 2.

Variable selection in regression model

We use the forward stepwise regression, which sequentially selects and deletes predictor variables from a regression equation in constructing the linear and the logistic regression models. The stepwise regression procedure is more computationally efficient than an exhaustive search of combination of variables when there are many possible predictor variables. A potential drawback of the stepwise regression is that the best combination of variables could be missed, however, in practice, it can find the best combination most of the cases, especially when there are several variables with strong predictive power. The forward stepwise regression procedure first fits a simple linear regression model for each of the potential variables. The model with the largest t^* value is the candidate for the first addition. The second step is to add another variable to the regression model and again perform the t^* test for the combination of variables. The model with the largest t^* with the newly added variable is selected, however, if the largest t^* is less than the previous level, the procedure is terminated. The third step is to obtain t^* for every variables that are already in the model in order to decide which variable can be dropped. The fourth step is to continue adding new variables and checking whether any variables can be dropped repeatedly until no variables can be added or deleted.

Receiver operating characteristic and the area under the curve

The receiver operating characteristic (ROC) curve plots the sensitivity relative to $(1 - \text{specificity})$, both of which are computed for retrieving structure models by varying the threshold value of plausibility of the models being correct. In the case of logistic regression, the plausibility value used is the probability of models being correct, while the predicted RMSD is used in the case of the linear regression model and the LOESS regression model. As described in Results,

correct models are defined as ones whose RMSD to the native is below 2, 4, 6 or 8 Å. The sensitivity is defined as $TP/(TP + FN)$ and the specificity is defined as $TN/(TN + FP)$, where TP stands for the number of true positives, i.e. the number of correct models retrieved, FP is the number of false positives, i.e. the number of retrieved incorrect models, TN is the number of true negatives, i.e. the number of incorrect models, which are not retrieved using the threshold value of model plausibility. The area under the curve (AUC) is the area under the ROC curve. The AUC value ranges from 0 to 1 with 1 representing a perfect prediction model. Random prediction would yield an AUC value of 0.5.

Structural models for the *E. coli* proteome

A genome-scale structural modeling for proteins in *E. coli* K-12 W3110 is performed using two different procedures. In the first procedure, structure models are built by Nest (Petrey *et al.*, 2003) based on target–template alignments taken from the GTOPI database (Kawabata *et al.*, 2002). GTOPI provides computational sequence analyses for predicting function and structure of protein genes in over 790 genomes (<http://spock.genes.nig.ac.jp/~genome/>), which include alignments with homologous proteins with PDB entries. Homologous PDB entries for a protein are found by PSI-BLAST search (Altschul *et al.*, 1997) using an *E*-value threshold value of 0.001. In case multiple template structures are found for a protein, we use the one with the most significant *E*-value. Nest is run with the default setting. In the second procedure, we build structure models by SPARKS2 (Zhou and Zhou, 2002; Zhou and Zhou, 2005b). SPARKS2 is a threading program, which identifies compatible template structures for a target sequence and computes target–template alignments by considering the sequence similarity and structural propensity of amino acids. The SPARKS2 package was obtained from the Zhou group webpage, <http://sparks.informatics.iupui.edu/hzhou/download.html>. Using the target–template alignments, the SPARKS2 package further builds the tertiary structure models by Modeller (version 6.1). For each protein, we only used up to two top-scoring models if they have the threading Z-score of 5.6 or higher. The template structure database used is included in the SPARK2 package, which contains 14 225 structures. The structural models are available at the EcoliProtein database (<http://kiharalab.org/ecolpredict2/localsearch/index.html>) included in the website of the Protein Function Elucidation Team (<http://www.prfect.org/>).

Availability of the software and data

The quality assessment program developed in this study, Sub-AQUA (Suboptimal Alignment-based QUality Assessment method), is available at <http://kiharalab.org/SubAqua/>. Users can submit a PDB file of a structure model to the server to obtain prediction of the global RMSD value of the model to the native. The global RMSD is predicted by applying Eqn. 1, which linearly combines $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$. When the template structure used in the modeling is not specified, it uses PSI-BLAST search against the PDB database to identify the potential template. As shown later in the subsection in Results ‘Comparison with existing model quality assessment methods’, Sub-AQUA performs well with BLAST identified templates. For users who prefer to use the server anonymously, we created a dummy

email account: testsubaqua@gmail.com with password: testforserver. Sub-AQUA standalone program is also available at <http://kiharalab.org/SubAqua/package.html>. Besides, the program named SUBWAI, which computes the SPAD score from suboptimal alignments of a target–template alignment, is available for download at <http://kiharalab.org/subalignment/> for those who are interested in computing the SPAD score locally on their own computers. We also made the L–E data set available at http://kiharalab.org/SubAqua/LE_DB.tar.gz.

Results

Real-value prediction for global quality

To begin with, we examine the correlation coefficient of individual quality assessment scores and the global RMSD of a benchmark data set of structure models (Table I). The data set contains 5232 template-based models that are constructed using protein pairs classified into three hierarchical similarities, the family, the superfamily and the fold level (the L–E set). Because of the diverged sequence similarity between target and template pairs, the models in the L–E set cover a wide range of the global RMSD values from 0.36 to 25.6 Å. A variety of scores of different natures are examined, including scores that assess alignment-level quality (SPAD, the sequence identity, the PRSS Z-score), a score of structural compactness (Sc), a score of residue-level structural preference (Verify3D, TAP), scores which are based on atomic contact potentials (DOPE, ERRAT, ANOLEA1, ANOLEA2), scores of stereochemical parameters (PROCHECK1, PROCHECK2) and composite scores (GA341, pG, PROQ-LG, PROQ-MX). Correlation of the length of proteins with RMSD is also examined.

When structure models in all the three similarity levels are included in the analysis, only SPAD, the sequence identity

Table I. Correlation coefficient to the global RMSD of structure models

Variable	CC all	CC family	CC superfamily	CC fold
Verify3D	−0.09	0.01	0.19	0.33
Verify3D/L	−0.45	−0.37	−0.29	−0.17
Verify3D/L ²	−0.46	−0.43	−0.51	−0.43
$\log(\text{Verify3D}/L^2)$	−0.53	−0.52	−0.52	−0.51
PRSS Z-score	−0.63	−0.55	−0.40	−0.41
Sc	−0.19	−0.07	−0.27	−0.15
SPAD	0.55	0.48	0.54	0.39
$\log(\text{SPAD})$	0.71	0.62	0.63	0.42
ERRAT	−0.35	−0.19	−0.30	−0.16
TAP	−0.34	−0.16	−0.31	−0.10
DOPE	0.29	0.23	0.48	0.43
Length	0.24	0.24	0.44	0.51
ANOLEA2	−0.02	0.01	−0.04	−0.05
PROCHECK1	−0.05	−0.08	0.05	−0.01
ProQ-MX	−0.39	−0.12	−0.36	−0.14
ANOLEA1	0.29	0.11	0.16	0.21
ProQ-LG	−0.31	−0.10	−0.22	0.06
pG	−0.01	0.01	0.19	0.15
Sequence identity	−0.58	−0.52	−0.39	−0.21
PROCHECK2	0.16	0.10	0.05	0.07
GA341	−0.17	−0.02	−0.05	−0.12

CC (all), CC (family), CC (superfamily) and CC (fold) are the correlation coefficient of each variable with the RMSD for all the structure models in the L–E data set, the family data set, the superfamily data set and, the fold data set, respectively.

and the PRSS Z-score, all of which are scores for assessing alignment-level quality, show a correlation coefficient to RMSD above 0.5 (Table I, CC All). This result confirms the fact that the quality of target–template alignments severely affects the accuracy of current template-based models. Interestingly, the correlation coefficient of the SPAD score improves from 0.55 to 0.71 when it is log-transformed (i.e. $\log(\text{SPAD})$). This is consistent with the observation we had in our previous work of developing and examining the SPAD score (Chen and Kihara, 2008). SPAD evaluates target–template alignment stability by comparing it with suboptimal alignments. The reason of a better correlation by using $\log(\text{SPAD})$ may be due to the rapid divergence of the possible alignments as the sequence similarity drops. Also we found that normalizing the Verify3D score by the model length (L) and a square of the length (L^2) improves the correlation from -0.09 to -0.45 and -0.46 , respectively. Taking the logarithm of the Verify3D score normalized by L^2 , i.e. $\log(\text{Verify3D}/L^2)$, further improves the correlation to -0.53 . It is expected that normalizing by the length of the model improves the correlation coefficient since the Verify3D score is computed as a sum of a residue-based score. The reason why the normalizing by L^2 makes the correlation better might be because Verify3D implicitly takes residue contacts into account as the environment of a residue. Among all the variables tested, $\log(\text{SPAD})$, the PRSS Z-score, the sequence identity and $\log(\text{Verify3D}/L^2)$ shows a higher correlation to the RMSD.

At the family level (CC family in Table I), the alignment-based scores show significant correlation to RMSD. When the similarity between the targets and templates are further reduced to the superfamily and the fold level, some variables start to show different levels of correlation. Evidently the sequence identity becomes much less correlated. At the fold level, $\log(\text{Verify3D}/L^2)$ shows the most significant correlation (-0.51) among all the variables. The SPAD score has a relatively weak correlation, but its log transformation, $\log(\text{SPAD})$, still maintains a correlation coefficient of above 0.40. DOPE, which is an atom-contact potential, shows higher correlation at the superfamily and the fold level than it does at the family level. It is also notable that the correlation of the length of the model becomes better as the target–template similarity level drops. Probably this result simply indicates that errors in models of larger proteins tend to result in a larger RMSD to the native. We also compute the correlation coefficients of the scoring terms to the LGA score (Table II), where the same trends are observed as discussed above.

Next, we combine scoring terms listed in Table I by linear regression to predict the global RMSD of models (Supplementary data, Table SI). The R^2 (coefficient of determination) of the linear regression including all the variables is 0.628, which means that 62.8% of the RMSD variance can be explained by this model. We find that there are many variables, which do not contribute much to the regression model. Using a P -value of 0.05, only $\log(\text{Verify3D}/L^2)$, PRSS Zs, Compactness score (Sc), $\log(\text{SPAD})$, and ERRAT are considered to be significant.

To find a more meaningful linear regression model with a reduced number of variables, we employ the forward stepwise variable selection procedure. Note that it is not appropriate to simply remove the insignificant variables since the order of

Table II. Correlation coefficient to the LGA score of structure models

Variable	CC_all	CC_family	CC_superfamily	CC_fold
Verify3D	0.12	-0.09	-0.21	-0.42
Verify3D/L	0.47	0.29	0.35	0.12
Verify3D/ L^2	0.50	0.52	0.62	0.45
$\log(\text{Verify3D}/L^2)$	0.53	0.57	0.61	0.52
PRSS Z-score	0.74	0.67	0.57	0.45
Sc	0.16	0.07	0.29	0.10
SPAD	-0.56	-0.50	-0.55	-0.45
$\log(\text{SPAD})$	-0.79	-0.68	-0.68	-0.51
Errat	0.39	0.22	0.39	0.16
TAP	0.38	0.15	0.32	0.19
DOPE	-0.34	-0.38	-0.62	-0.54
Length	-0.21	-0.31	-0.48	-0.57
ANOLEA2	0.02	-0.01	0.01	0.10
PROCHECK1	0.08	0.11	-0.08	-0.03
MaxSub	0.42	0.07	0.41	0.14
ANOLEA1	-0.37	-0.22	-0.24	-0.25
Lgscore	0.35	0.06	0.23	-0.09
pG	0.02	-0.09	-0.26	-0.14
Sequence identity	0.77	0.70	0.52	0.23
PROCHECK2	-0.22	-0.14	-0.13	-0.12
GA341	0.22	0.05	0.17	0.08

removing the variables will have a considerable effect on the significance of the remaining variables. The contribution of a variable to the regression model is represented by the partial R^2 , which indicates how much more R^2 (variance of the RMSD) can be explained by adding the variable to the original regression model. The entire L–E set and each similarity level data set are separately analyzed. The detailed results are provided in the Supplementary data, Table SII. For the entire L–E set, $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$ are selected in this order as the most contributing variables with an R^2 value of 0.586. The remaining of the variables, PRSS Zs to DOPE, are selected with a statistical significance but their contribution to the model R^2 is marginal. For the family level data set, almost the same set of variables is selected in a very similar order. For the superfamily level, $\text{Verify3D}/L^2$ is selected instead of its log transform and the ProQ-LG score and the length of models show up as the third and fourth variables. As for the fold level data set, predicting RMSD values becomes more difficult, and the regression model that incorporates all the seven variables with a P -value < 0.05 only yields an R^2 value of 0.428.

Figure 1 shows the actual and predicted RMSD of the L–E set using the linear regression model with only $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$. These two variables are used because they are selected as significantly contributing variables for the entire L–E set and for each similarity level. The coefficients are parameterized on the entire L–E set. Concretely, the following linear regression equation is obtained:

$$\text{RMSD} = -4.99 + 2.25 \times \log(\text{SPAD}) - 2.17 \times \log\left(\frac{\text{Verify3D}}{L^2}\right) \quad (3)$$

We use Eqn. 3 rather than the regression models computed individually for each similarity level because their model R^2 values are comparable. Using Eqn. 3 alone, R^2 for the family, superfamily, and the fold level are 0.510, 0.492 and

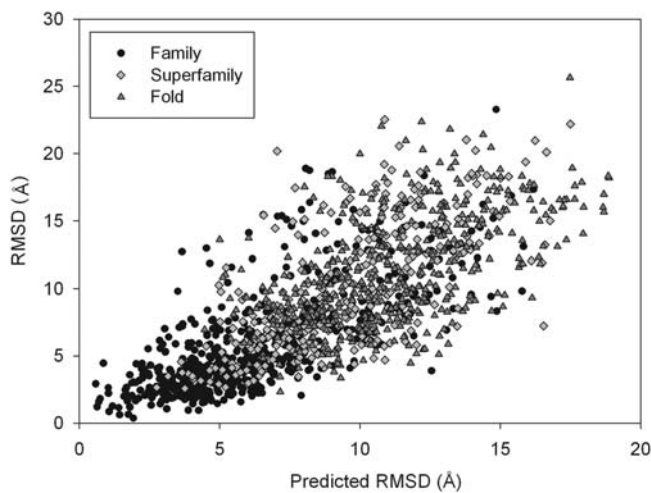


Fig. 1. Actual RMSD of structure models relative to predicted RMSD. RMSD is predicted by regression models using $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$ as predictor variables. Regression models are built on all the structural models in the L–E set. The linear regression model (Eqn. 3) is used.

0.337 while using linear regressions specific for each data set (Supplementary data, Table SII) R^2 values only have a marginal improvement resulting in 0.510, 0.493, and 0.340, respectively. Practically, using a single regression equation is more convenient for actual quality prediction since it does not need to assign a structure model to one of the three similarity levels. We have also used the LOESS regression with the same two variables, $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$ (Supplementary data, Fig. SI). LOESS regression shows a slightly better correlation (0.772) between predicted and actual RMSD values than the linear regression model (0.766). However, the linear regression model may be more convenient than LOESS in practice because that the former provides a simple analytical equation and it is computationally more efficient than the latter.

Predicting LGA score

We also constructed linear regression for predicting the LGA score by employing the forward stepwise procedure (Supplementary data, Table SIII). The LGA score (Zemla, 2003) ranges from 0 to 100, which indicates the fraction of

equivalent residues in two proteins. Consistent to the linear regression model for predicting RMSD value (Supplementary data, Table SII), the SPAD score and the Verify3D score still play important roles while the sequence identity also becomes significant for the entire L–E set (the column for All in Table III). The linear regression with the top three variables, $\log(\text{SPAD})$, $\text{Verify3D}/L^2$ and the sequence identity is as follows:

$$\text{LGA} = 25.78 - 8.98 \times \log(\text{SPAD}) + 2347.71 \times \left(\frac{\text{Verify3D}}{L^2} \right) + 91.47 \times \text{seqID} \quad (4)$$

Predicted LGA score using Eqn. 4 is plotted against the actual LGA score in Fig. 2. The correlation coefficient of the two is 0.88, which indicates that the LGA score is well predicted by the linear regression.

Predicting correct models and incorrect models

In the previous section, the actual value of global RMSD and the LGA score is predicted. Next, we predict whether a structure model has a global RMSD of less than a threshold value, i.e. if a model is ‘correct’ or ‘incorrect’. Threshold values RMSD used to define a correct model are 2, 4, 6, and 8 Å. This coarse-grained binary classification of model has a practical importance since knowing discrete classes of the quality of a model will be sufficient to tell its potential application in most of the cases. In addition to linear and LOESS regression, here we employ logistic regression, which is suitable for predicting binary classes.

The forward stepwise selection procedure is used again to select variables for constructing a logistic regression model. In this selection procedure, we consider a model is ‘correct’ if its RMSD is below or equal to 6 Å and ‘incorrect’ otherwise. The selected variables are listed in Table SIV in the Supplementary data. Among the variables selected with a P -value of <0.05 , we choose the two most significant variables, $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$, taking into account their relatively small χ^2 score and the consistency with the previous linear regression analysis. The equation of this

Table III. Variable selection of linear regression model for local quality prediction

Step	Variable	Partial R^2	Model R^2	P (F)	Correlation coefficient ^a
1	Predicted global RMSD	0.1940	0.1940	<0.0001	0.44
2	Gap ratio	0.0453	0.2393	<0.0001	0.24
3	$\log(\text{localSPAD}/\text{SPAD} + 1)$	0.0229	0.2622	<0.0001	0.21
4	LocalVerify3D	0.0066	0.2688	<0.0001	0.32
5	$\log(\text{localVerify3D}_{\text{positive}}^2 + 1)^b$	0.0048	0.2736	<0.0001	-0.14
6	Conservation ^c	0.0038	0.2774	<0.0001	-0.29
7	$\log(\text{localSPAD} + 1)/\log(\text{SPAD})$	0.0016	0.2790	<0.0001	0.09
–	$\log(\text{localSPAD} + 1)$	–	–	–	0.42
–	Mutation score ^d	–	–	–	-0.32
–	Local ERRAT	–	–	–	0.19

^aThe correlation coefficient to the $C\alpha$ distance.

^bLocalVerify3D_{positive} is a non-negative localVerify3D score assigned to each residue; 0 is assigned when a negative localVerify3D score is replaced with 0.

^cThe conservation is the fraction of the most abundant residue at the position in the multiple sequence alignment of the target protein used for alignment with the template.

^dAverage BLOSUM45 score of a column in the profile.

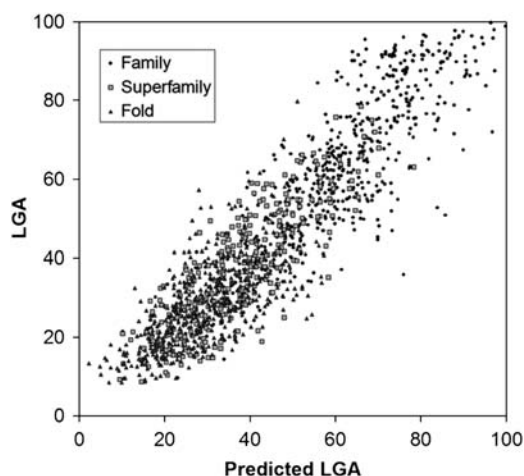


Fig. 2. Predicted and actual LGA score. LGA is predicted by a linear regression model using $\log(\text{SPAD})$, seqID and $\text{Verify3D}/L^2$ as predictor variables (Eqn. 4). Regression model is built on all the structural models in the L–E data set. Correlation coefficient between true LGA and predicted LGA is 0.88.

logistic regression is

$$\log \frac{p}{1-p} = -7.93 + 1.62 \times \log(\text{SPAD}) - 1.48 \times \log\left(\frac{\text{Verify3D}}{L^2}\right), \quad (5)$$

where p is the probability that a model is correct (i.e. an RMSD of below 6 Å). This reduced regression model is comparable to the one using all the variables (the full regression model) in terms of the performance of the classification of correct and incorrect models (Fig. 3A): the AUC value of the two ROC curves are 0.911 and 0.933, respectively, for the reduced and the full regression models. Since the reduced model is much simplified without sacrificing the AUC value much, we use the reduced regression model in the subsequent experiments.

Figure 3B shows the accuracy of predicting the correct/incorrect structure models by the logistic regression. Four different RMSD cutoff values are used. The overall prediction accuracy (All in Fig. 3B) is high; over 80% at all the RMSD cutoff values. The large difference in the accuracy for predicting accurate and inaccurate model is observed at the threshold of 2 and 4 Å due to the unequal fraction of correct and incorrect models in the data. The fraction of inaccurate models in the L–E set is larger for smaller threshold values: 96.8, 79.2, 64.78, and 49.3% at the threshold value of 2, 4, 6, and 8 Å, respectively.

We have further compared three regression models, linear, logistic, and LOESS (Supplementary data, Fig. SII). All the three regression models show excellent performance both in terms of AUC and the accuracy: the average AUC of 0.89 or higher and the accuracy of 0.80 or higher are achieved at each RMSD threshold value. LOESS regression shows slightly better performance but we conclude that linear and logistic regression are sufficient for our purpose of predicting the real-value RMSD and distinguishing correct and incorrect models, because their performance are almost the same as the results by LOESS regression.

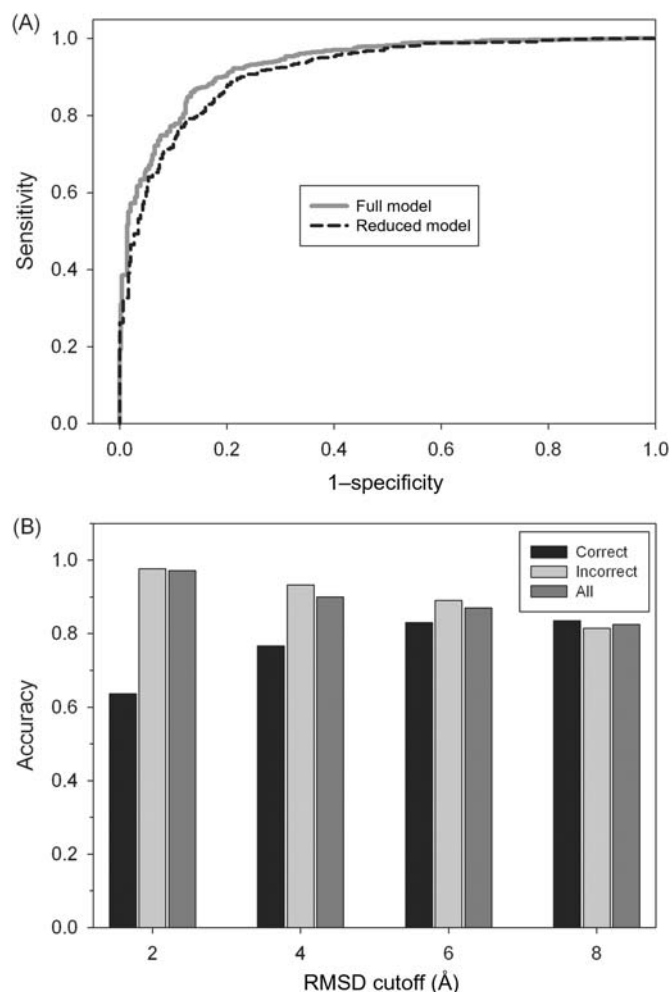


Fig. 3. Discriminating correct structure models from incorrect models by logistic regression. (A) ROC curve of a logistic regression with all the twenty variables (the full model) and one with $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$ (the reduced model). Correct models are those which have an RMSD of 6 Å or lower to native. (B) Predicting correct/incorrect models by the reduced logistic regression model. Correct models are defined as those with an RMSD of 2, 4, 6 and 8 Å or lower.

The real-value quality assessment method, which combines the SPAD and Verify3D scores by using the linear (Eqn. 3) and the logistic regression (Eqn. 5) for predicting RMSD value and also for predicting the LGA score (Eqn. 4) is named Sub-AQUA. The program is made available as a webserver and a standalone program as described in the Methods section.

Multinomial classification of model quality

In the previous section, we performed binary prediction of structure models into correct/incorrect models using several different threshold values. Here, we further employed multinomial logistic regression to predict structure model into four RMSD ranges, less than 3, 3–6, 6–9 Å and larger than 9 Å. Results of the forward stepwise selection procedure (Supplementary data, Table SV) show that the SPAD score and the Verify3D score are consistently selected as the two most significant variables. The classification results using the regression model with the two most significant variables are shown in Fig. 4. The percentage of the correctly classified structure models to each RMSD range is written in the

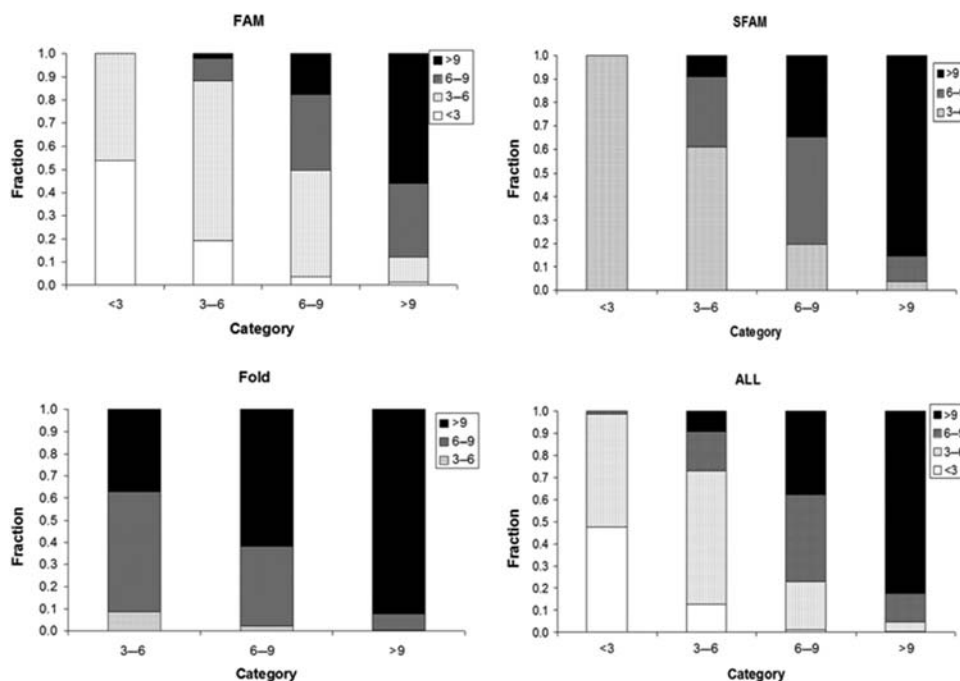


Fig. 4. Classification of structure models into four different categories. A multinomial logistic regression model is constructed to classify structure models into four different RMSD ranges, less than 3, 3–6, 6–9 or larger than 9 Å. x -axis represents the true classification of the models in terms of the four categories. y -axis shows the proportions of the models that are predicted to be the four different categories. The percentage of the correctly classified structure models to each RMSD range is as follows. The numbers are ordered for <3, 3–6, 6–9, >9 Å: The family level data set: 53.8, 69.2, 32.7, 56.1%. The superfamily level: 0, 61.2, 45.7, 85.6%. The fold level (there was no model with RMSD <3 Å), 8.6, 36.2, 92.4%. All data set: 47.5, 60.4, 39.2, 82.4%.

figure caption. Overall, the multinomial logistic model performs very well especially for detecting bad models (>9 Å) and good models (3–6 Å). The average success rate for the entire data set averaged over each RMSD range is 57.38%.

Prediction of local quality of models

Next, we perform regression analyses for predicting local quality of structure models, i.e. The $C\alpha$ distance between corresponding residues of a structure model to its native structure. Individual scores examined do not show significant correlation to the $C\alpha$ distance (Table III, the right column). We also tried several other transformations of localVerify3D and global/local SPAD scores but none of them has a correlation coefficient >0.42, which is achieved by $\log(\text{localSPAD} + 1)$. Similar to the previous sections, regression analysis is used to combine scores but the accuracy of predicting $C\alpha$ distance does not make much improvement (detailed data not shown).

However, we find that prediction performance of regression models shows significant improvement when predicted global RMSD of structure models is included as an independent variable. This is because a global RMSD is essentially an average of $C\alpha$ distance of each residue and therefore has a good correlation to the $C\alpha$ distance (0.44, Table III). Using the predicted RMSD, we present hierarchical procedures for predicting the $C\alpha$ distance. First, the global RMSD of structure model is predicted by a regression with two variables, $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$ (Eqn. 3). Second, the $C\alpha$ distance is predicted by a regression using three variables, namely, the predicted global RMSD, the gap ratio and $\log(\text{localSPAD}/\text{SPAD} + 1)$. These three variables are selected by the forward selection procedure

(Table III). R^2 for the linear regression is 0.262 and the correlation coefficient to the $C\alpha$ distance is improved to 0.51.

The hierarchical approach is implemented using two regression methods, linear regression and LOESS regression (2_Linear and 2_LOESS in Fig. 5). In 2_Linear, linear regression is used in both the first and the second steps, while LOESS regression is used in both steps in 2_LOESS. For comparison, a regular linear regression, which combines $\log(\text{localSPAD} + 1)$, the gap ratio and the mutation score (1_Linear) is also computed. These three variables are selected by the forward selection procedure when the predicted global RMSD is absent. In Fig. 5, the three approaches are compared in terms of their performance in distinguishing correctly predicted $C\alpha$ positions and incorrect ones. The two hierarchical approaches, 2_Linear and 2_LOESS, significantly outperform the one-layer linear regression model. The accuracy of 2_Linear and 2_LOESS is over 0.7 at all the cutoff values, which is about 0.1 better than 1_Linear (Fig. 5B).

Comparison with existing model quality assessment methods

In this section, we compare the performance of Sub-AQUA with the other existing methods in terms of predicting global RMSD of models. To this end, we downloaded protein structure models built and submitted by automatic servers in the CASP7 experiment (from http://predictioncenter.org/CASP7/server_predictions/) and predicted the global RMSD of the models by applying linear regression (Eqn. 3) and the logistic regression (Eqn. 5). We compare our method with 43 CASP7 participants in the quality assessment category so that we can evaluate our method relative to the state-of-the-art methods. A total of 15 601 models are used (models that lacks side-chain heavy atoms are excluded).

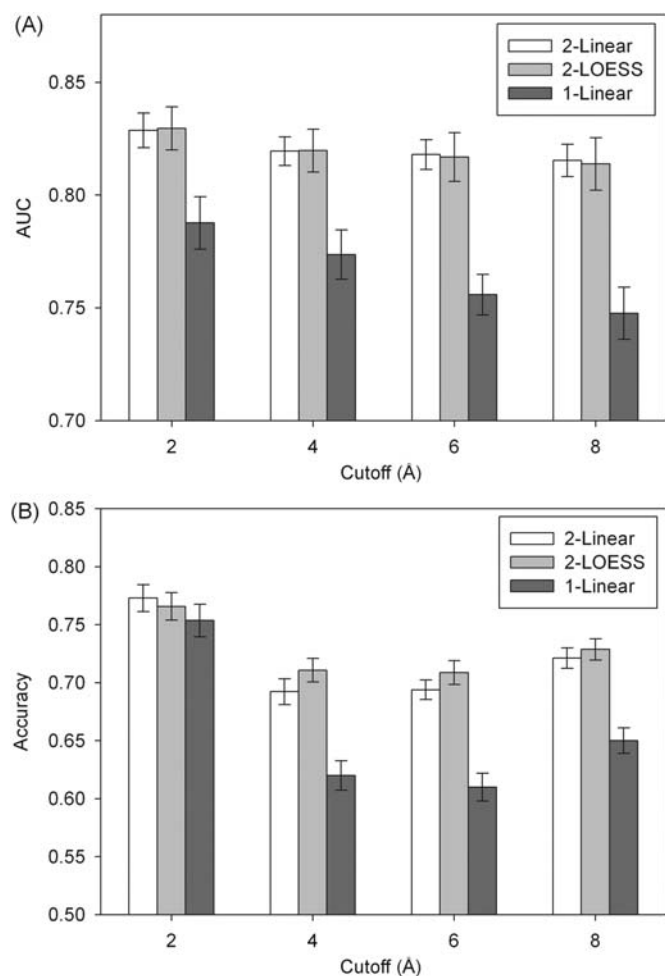


Fig. 5. Comparison of hierarchical regression approaches and regular linear regression in terms of predicting $C\alpha$ distance of residues in structural models. 2_Linear and 2_LOESS are the hierarchical approaches, in which the global RMSD is predicted first and then used in the second step as one of the independent variables. 2_Linear uses linear regression while 2_LOESS uses LOESS regression. Residues in structure models are classified into correct or incorrect, using threshold values of $C\alpha$ distance of 2, 4, 6 and 8 Å. The cross-validation using four subsets of data for training and one for testing is performed fifty times and the average and the standard deviation are shown. The total number of residues is 159,947 in 5232 structure models. The proportion of correct positioned residues with the cutoff of 2, 4, 6 and 8 Å is 27.2, 48.1, 59.4 and 66.6%, respectively. (A) AUC of the three models with different RMSD cutoffs. (B) Accuracy of the three models with different RMSD cutoffs.

Note that comparison of Sub-AQUA with the other participants in the CASP7 quality assessment category is not straightforward since the CASP7 participants did not predict the global RMSD as our Sub-AQUA does but aimed to rerank models by their own global model quality scores. Ideally, we wanted to compare with other methods that also predict the global RMSD of structure models, but as mentioned in Introduction, such methods are few and their software or data are not available for comparison. Another difficulty for us to apply Sub-AQUA to the CASP7 models is the unavailability of the original alignments that the models are built on, which are required for computing the SPAD score. Moreover, some models, especially those for targets of the free-modeling (FM) category, may not be based on a single global template. This is a significant disadvantage of Sub-AQUA. Nevertheless, we perform this

comparison to obtain a better understanding of the behavior of the Sub-AQUA method. Because target–template alignments are not available, we used BLAST to search a template from the PDB and made an alignment between the target and the template sequence using BLAST. Thus, the alignments we computed may not be the same as the ones actually used for constructing the models.

We compare the methods in two ways: First, in Table IV, we compare the correlation coefficient of the quality assessment scores and the global RMSD of models in different target categories, i.e. the high-accuracy template-based modeling (HA-TBM), the template-based modeling (TBM), the template-based modeling/free modeling (TBM-FM), the free-modeling (FM) category and models for all categories combined (All). The correlation coefficient is calculated for structure models of each target and averaged over different targets. Figure 6 compares the predicted RMSD by the linear regression of Sub-AQUA (x -axis) and the actual RMSD of the models (y -axis) in each category. Predictions for the HA-TBM models shows the highest correlation value (0.536). In comparison with the other methods in Table IV, Sub-AQUA performs best in the All and TBM-FM categories. In the HA-TBM category, the ranking of the Sub-AQUA is moderate (Table IV), although the correlation coefficient value itself is largest among the categories (Fig. 6).

Second, in Fig. 7, the average RMSD of the top scoring models of each target in the different categories is computed. The prediction groups are sorted by the average RMSD of the All category. It is shown that Sub-AQUA achieves the smallest average RMSD among those compared for each target category except for the case of the HA-TBM. Overall, Sub-AQUA is very competitive among the existing quality assessment methods.

Quality assessment of structural models of the *E. coli* proteome

Finally, to examine the actual situation that occurs in a large-scale structure modeling, we assess the global and local quality of predicted structure models of *E. coli* proteins. Two sets of structural models of proteins in the *E. coli* K-12 W3110 genome are prepared (Table V): in the first set, target–template alignments are taken from the GTOP database (Kawabata et al., 2002) and the tertiary structure models are built based on the alignments by the Nest program (Petrey et al., 2003). The second set is computed by the SPARKS2 threading program package (Zhou and Zhou, 2005b). A total of 1797 structural models are obtained by the GTOP-Nest procedure and 3764 structural models by SPARKS2 (Table IV). Among the total of 4226 protein genes, 2853 genes (67.5%) have at least one structure model constructed by either of the two procedures. The GTOP-Nest set contains models for 1797 (42.5%) genes and the SPARKS2 set has models for 2507 (59.3%) genes. Thus, roughly speaking, SPARKS2 identifies templates for 20% more proteins compared with regular PSI-BLAST search.

For evaluating the global quality, the linear regression of Sub-AQUA (Eqn. 3) is used. Here, an RMSD threshold value of 6 Å is used to define a correct prediction. Among the 1797 GTOP-Nest models, 1125 of them are predicted to have an RMSD of less than 6 Å, while 1373 models are predicted to be correct among the 3764 SPARKS2 models (Table V). Figure 8 shows the distribution of predicted

Table IV. Correlation coefficient of the MQAP scores and the true RMSD of CASP7 models

Rank	All ^a			HA-TBM			TBM			TBM-FM			FM		
	QA ^b	CC	Num	QA	CC	Num	QA	CC	Num	QA	CC	Num	QA	CC	Num
1	SA_lin ^c	0.62	96	556_1	0.86	22	556_1	0.63	52	SA_lin	0.53	8	191_1	0.32	6
2	699_1	0.61	96	634_1	0.82	23	SA_lin	0.63	53	734_1	0.47	8	038_1	0.32	9
3	SA_logit	0.61	96	713_1	0.81	23	634_1	0.62	53	699_1	0.42	8	734_1	0.29	9
4	038_1	0.59	88	699_1	0.81	23	038_1	0.62	49	SA_logit	0.41	8	SA_logit	0.29	9
5	713_1	0.58	96	704_1	0.80	23	699_1	0.61	53	713_1	0.27	8	SA_lin	0.28	9
6	556_1	0.58	94	717_1	0.80	19	SA_logit	0.61	53	013_1	0.26	8	178_1	0.25	9
7	178_1	0.57	95	633_1	0.79	23	713_1	0.60	53	633_1	0.26	8	713_1	0.24	9
8	734_1	0.57	95	038_1	0.79	20	178_1	0.59	53	276_1	0.26	6	699_1	0.24	9
9	634_1	0.57	96	692_1	0.79	23	633_1	0.58	53	038_1	0.24	7	013_1	0.19	89
10	633_1	0.55	96	178_1	0.79	22	692_1	0.58	53	704_1	0.23	8	703_1	0.18	9
11	692_1	0.55	96	SA_logit	0.78	23	734_1	0.57	53	178_1	0.21	8	338_1	0.15	9
12	704_1	0.53	96	SA_lin	0.77	23	013_1	0.55	43	556_1	0.18	8	691_1	0.12	9
13	013_1	0.52	81	091_1	0.74	22	704_1	0.54	53	692_1	0.18	8	091_1	0.12	9
14	191_1	0.49	62	013_1	0.73	19	191_1	0.52	40	191_1	0.17	7	704_1	0.11	9
15	091_1	0.47	95	016_1	0.71	19	091_1	0.47	53	016_1	0.17	8	276_1	0.11	8
16	703_1	0.46	69	734_1	0.70	22	703_1	0.44	36	657_1	0.16	8	633_1	0.10	9
17	717_1	0.42	89	691_1	0.70	23	691_1	0.43	53	091_1	0.15	8	692_1	0.10	9
18	338_1	0.41	95	703_1	0.69	18	717_1	0.41	51	338_1	0.14	8	717_1	0.08	9
19	691_1	0.41	96	338_1	0.66	22	338_1	0.40	53	634_1	0.14	8	718_1	0.06	9
20	016_1	0.38	89	657_1	0.54	23	276_1	0.37	48	718_1	0.11	8	657_1	0.04	9
21	276_1	0.34	80	026_1	0.50	23	016_1	0.36	50	717_1	0.07	7	634_1	0.04	9
22	026_1	0.31	95	276_1	0.42	16	718_1	0.35	53	691_1	0.05	8	556_1	0.03	9
23	718_1	0.30	94	718_1	0.37	21	026_1	0.32	52	026_1	0.01	8	016_1	0.02	9
24	657_1	0.29	96	066_1	0.20	20	657_1	0.24	53	–	–	–	026_1	0.02	9
25	066_1	0.20	96	–	–	–	066_1	0.20	32	–	–	–	–	–	–

^aHA-TBM, HA-TBM targets; TBM, template-based modeling targets; TBM-FM, targets, which overlap between template-based modeling and free modeling; FM, targets for FM. All include all the models in all the categories. The number of targets and the total number of predicted models for each category is:

HA-TBM, 3959 models for 23 targets; TBM, 8621 models for 53 targets; TBM-FM, 1151 models for eight targets; and FM, 1557 models for nine targets.

^bQA, the group number of predictors; CC, the correlation coefficient; num, the number of targets whose quality was assessed by the group. In the ranks of individual categories, groups who made predictions for less than 20% of the available models are excluded. In the all category, groups that are excluded at any one or more individual category are excluded.

^cSA_Lin, real-value RMSD prediction by Sub-AQUA using the linear regression; SA_logit, the logistic regression (within or without an RMSD of less than 6 Å) by Sub-AQUA.

RMSD of the models of the two sets. Approximately one-third of the GTOPI-Nest models have a predicted RMSD within 1 Å and the remaining of the models almost evenly distributed in the range of 2–11 Å (Fig. 8A). In contrast, the predicted RMSD of the SPARKS2 models has a peak at around 6–7 Å and there are fewer models with a predicted RMSD of <1 Å (Fig. 8B). Close examination revealed that this difference in the RMSD distribution originates from their different procedures: GTOPI uses PSI-BLAST for template search with a conservative *E*-value threshold value (0.001) against the entire PDB. Therefore, GTOPI only identifies very closely related templates for a target, and moreover, the native structure of the target will be identified if it exists in PDB. In fact among the 594 GTOPI-Nest models predicted to have an RMSD of 1 Å or less, 238 of them (40.0%) use its native structure as the template and 367 (61.8%) identified a template with the sequence identity of at least 95%. In addition, since we only used aligned regions of target–template alignments (i.e. excluding the flanking gaps on both ends) taken from GTOPI, the flanking unaligned regions did not undergo template-FM by Nest. On the other hand, SPARKS2, as usual threading programs do, uses a template database of a selected set of non-redundant structures. Thus, native structures of *E. coli* proteins are missing if they are not included in the template database. In such cases, SPARKS2 uses a homologous structure of a target in the template database rather than its native structure.

Moreover, the SPARKS2 package feeds alignments including hanging gap regions to Modeller. Those flanking unaligned regions often cause erroneous dangling conformation by Modeller. Figure 8C shows clear correlation between the length of flanking gaps in input alignments and predicted global RMSD of the SPARKS2 models. The influence of flanking gaps is also seen in the correlation between the protein sequence length and their predicted RMSD (Fig. 9). Structure models in the GTOPI-Nest set have no significant correlation between them (Fig. 9A), whereas there is an apparent positive correlation between them for the SPARKS2 models (Fig. 9B).

Both GTOPI-Nest and SPARKS2 models share a similar relationship between predicted global RMSD and the sequence identity (Fig. 10). The predicted RMSD starts to diverge when the sequence identity drops below around 30%. This plot resembles the well-known relationship of the protein sequence and structure similarity observed in experimentally determined structures (Chothia and Lesk, 1986; Wilson *et al.*, 2000).

Local quality, the C α distance, are also predicted using the linear regression, which combines the predicted global RMSD, the gap ratio and log(localSPAD/SPAD + 1). Figure 11 shows predicted C α distance relative to the predicted global RMSD. Both GTOPI-Nest and SPARKS2 show a similar trend; clearly the average local C α distance grows as the predicted global RMSD becomes larger.

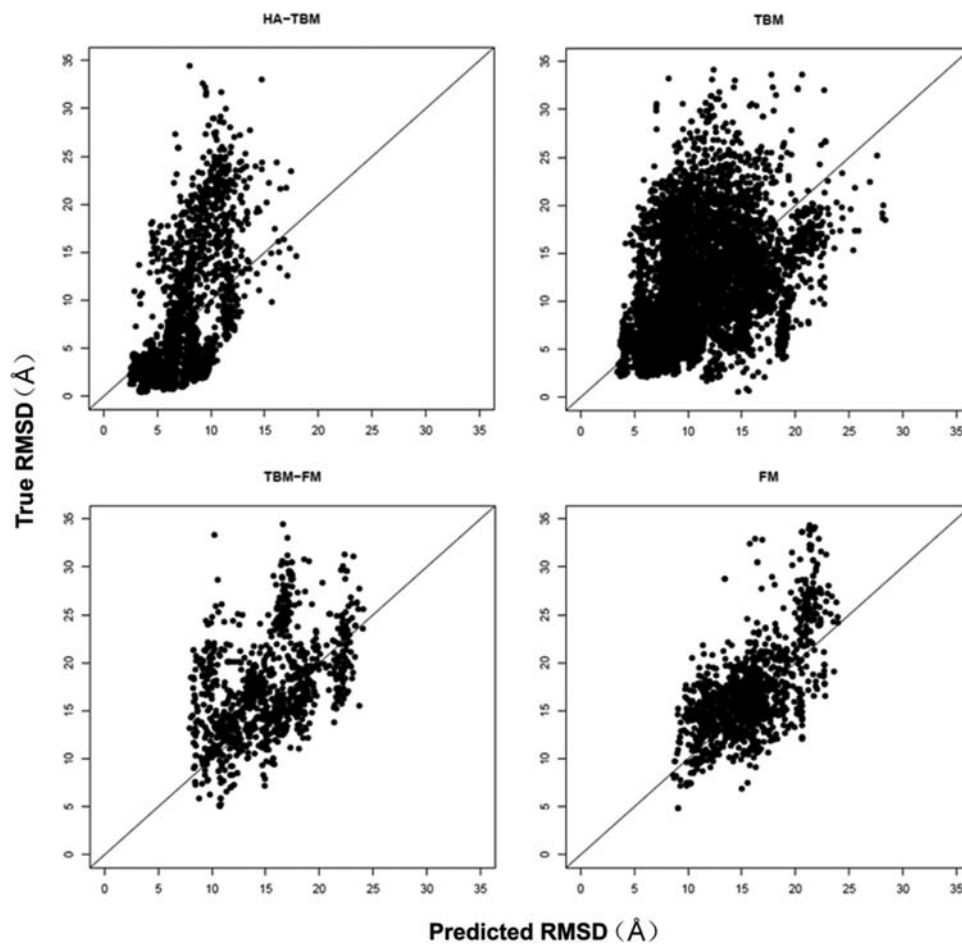


Fig. 6. Correlation between the predicted RMSD by Sub-AQUA and the true RMSD of CASP7 models. The linear regression is used. Models are divided into four categories (HA-TBM, TBM, TBM-FM and FM) according to the CASP7 criteria. The correlation coefficients of the predicted and actual RMSD are 0.536, 0.419, 0.350 and 0.438 for the HA-TBM, TBM, TBM-FM and FM category, respectively. These correlation coefficients are computed using all the structure models from all the targets in the category together.

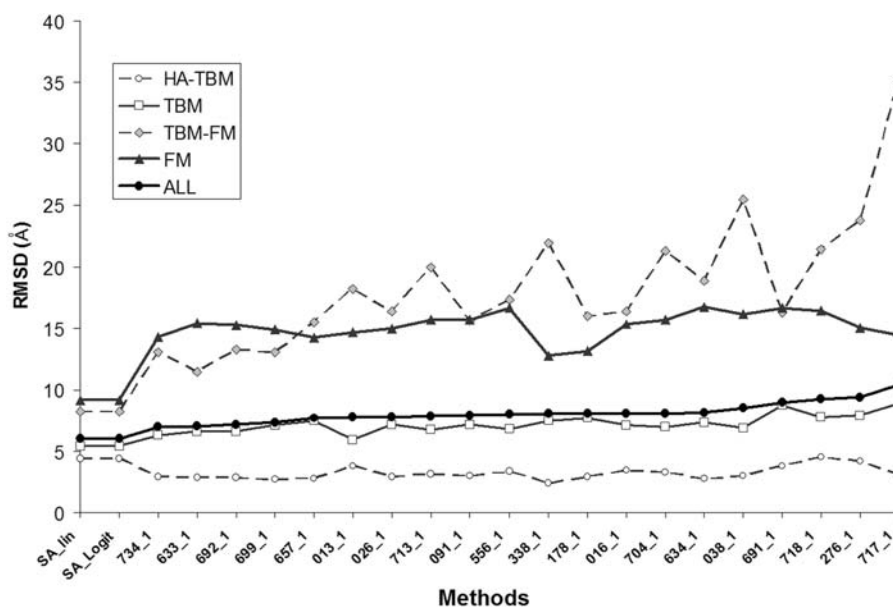


Fig. 7. Average RMSD of the top models. Different MQAP scores are used to rank all available models for each target protein in CASP7 and the top 1-ranked model is selected. The average RMSD is calculated over all the top 1 models selected by each MQAP score. Models are divided into four categories (HA-TBM, TBM, TBM-FM and FM) according to the CASP7 criteria and 'All' denotes all the models pooled together. The MQAPs are ordered by the average RMSD for the All category.

Table V. Summary of structure models of *E. coli* proteins

	GTOP-Nest	SPARKS2
Number of models	1797	3764
Number of genes with a model	1797	2507
Good models (%) ^a	1125 (62.6)	1373 (36.5)
Bad models	672	2391

^aModels with a predicted RMSD of 6 Å or less.

Figure 12 shows examples of predicted structure models of *E. coli* proteins with predicted C α distance. Three models each are selected from the GTOP-Nest set and the SPARKS2 set, whose global RMSD is predicted to be less than 2, 4 and 6 Å. The sausage representation (Chen and Kihara, 2008) used here can intuitively represent predicted global structure with its estimated local errors. Low quality local structures are often found at the two ends of proteins (Fig. 12B–F) and also at loop regions (Fig. 12A–C). A very small global RMSD is predicted for structures of *yccK* (Fig. 12A) and *rpsK* (Fig. 12B) is attributed to closely related template structures with above 35% sequence identity to the targets.

Discussion

In this work we developed quality assessment methods for template-based structure models named Sub-AQUA. The method predicts the real value of global and local quality of structure models, rather than reranking a pool of decoy structures, because the real value of quality is more meaningful for practical use of predicted structures in designing or interpreting biological experiments. For global RMSD prediction, the variable selection procedure identified $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$ for linear regression among many other scores examined. For local RMSD prediction, we designed a two-step prediction procedure, which uses predicted global RMSD as one of the variables together with the gap ratio and $\log(\text{localSPAD}/\text{SPAD} + 1)$. To the best of our knowledge, Sub-AQUA is the first MQAP, which uses sub-optimal alignment information as a scoring term. Interesting observation to note is that scores which evaluate alignment stability (i.e. SPAD and its transformations as well as PRSS) and a coarse-grained residue to structural environment compatibility score (i.e. Verify3D) show more significant correlation to global RMSD in general as compared with scores that examine detailed stereochemical properties of structures, such as PROCHECK scores. It seems like Verify3D captures key structural features of proteins well in a good balance for indicating main-chain level quality of protein models. Also, it effectively takes into account structure information that can compensate for alignment information by SPAD.

Since Sub-AQUA incorporates the SPAD score that evaluates alignment stability, RMSD value prediction by Sub-AQUA works best by design when a target–template alignment is available. However, the results on the CAPS7 models (Table IV, Fig. 7) show that the Sub-AQUA method is very competitive even when the target–template alignment is not given, by ‘predicting’ the alignment by BLAST.

An important finding of this work is that different quality measures show different levels of correlation to RMSD and the LGA score and the correlation changes as the global

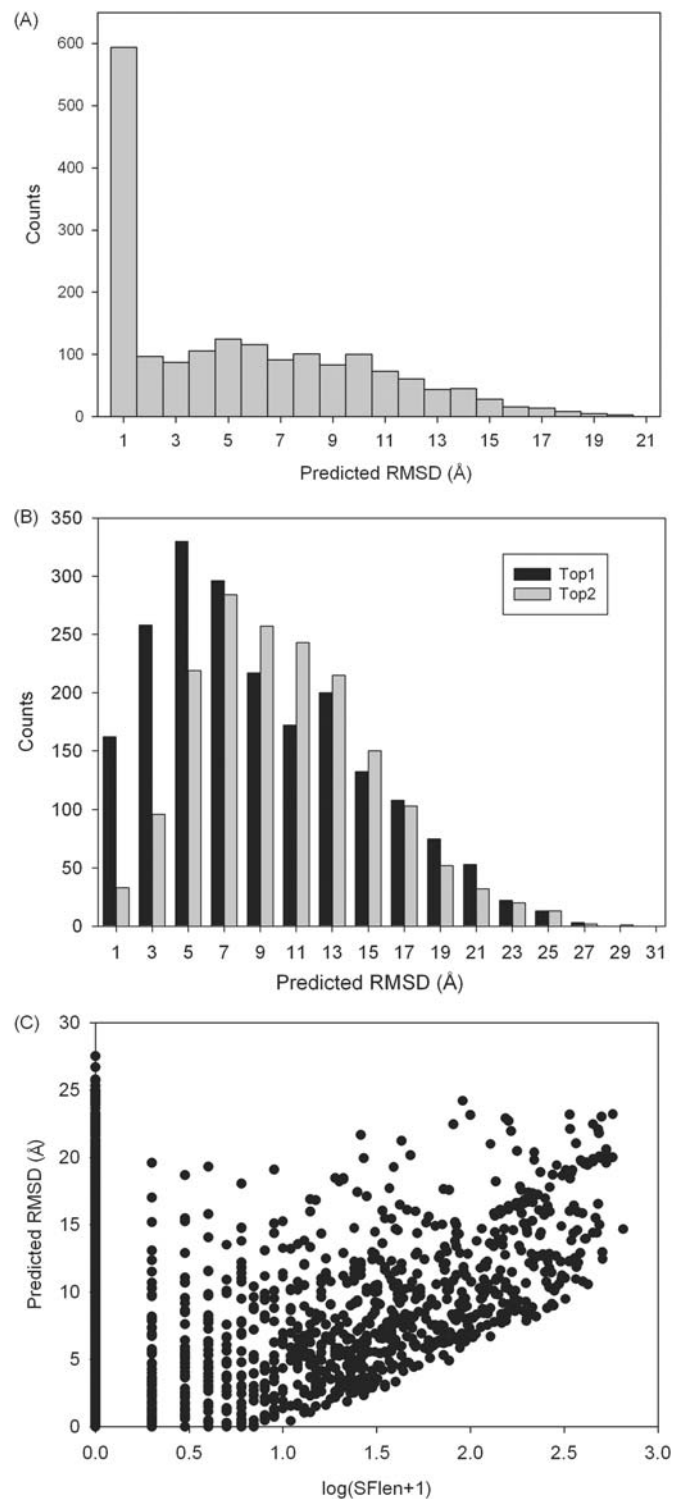


Fig. 8. Distribution of the predicted RMSD of *E. coli* protein models. RMSD is predicted using the linear regression using $\log(\text{SPAD})$ and $\log(\text{Verify3D}/L^2)$ as predictor variables. (A) the GTOP-Nest set and (B) the SPARKS2 set.

quality of structure models changes (Table I). For example, DOPE, which is an atom-contact assessment score, shows relatively strong correlation to RMSD when distantly related templates, i.e. those of the fold level similarity, are used while its correlation to RMSD is low when closely related templates, i.e. those of the family level similarity, are used. Therefore, it may be worthwhile to alter contribution of different types of

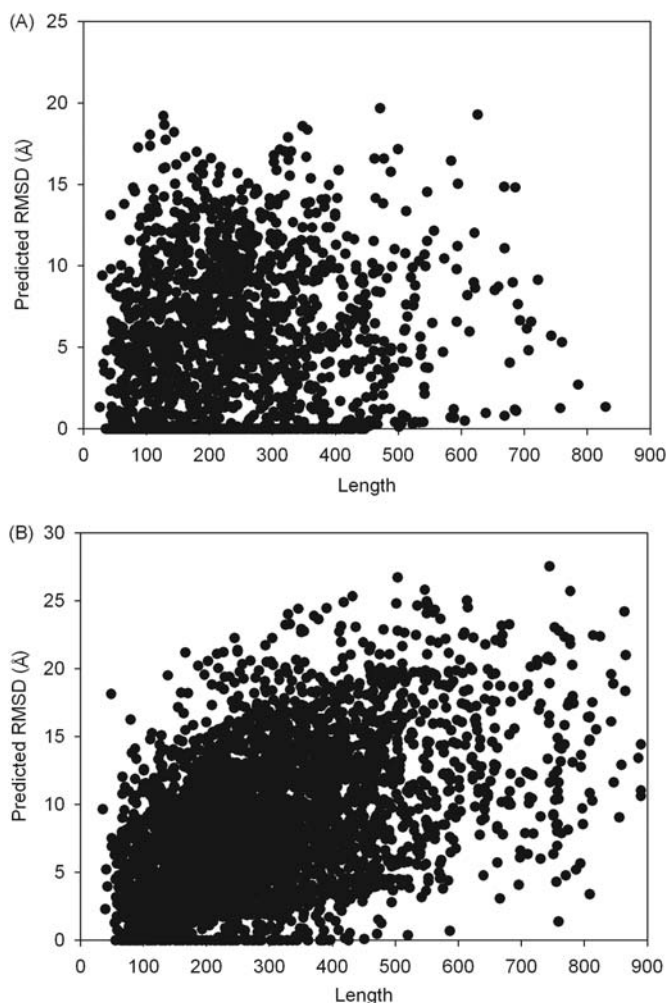


Fig. 9. Predicted RMSD relative to the sequence length of structure models. (A) The GTOP-Nest structural models. (B) the SPARKS2 set.

assessment scores, i.e. alignment-based, residue-based and atom-based scores, automatically according to the sequence similarity between the target and the template.

Now that structure prediction methods, especially template-based structure prediction methods, have become mature, it is time to make them more practical for experimental biologists. We believe quality assessment is a key for bridging computational structure prediction and experimental biology by indicating potential applications of a structure model.

Conclusions

Quality assessment methods named Sub-AQUA, which predict the RMSD and the LGA score (the global quality) and the C α distance (the local quality) of protein structure models, are developed by combining various scoring terms with regression models. Among the scoring terms tested, scores that evaluate alignment stability using suboptimal alignments (SPAD) and residue-level structure compatibility (Verify3D) showed good correlation to RMSD. Regression models, which combine these two scores are able to select correct structure models with a significant accuracy. To predict the C α distance between corresponding residues of a model and the native structure, we found a two-level

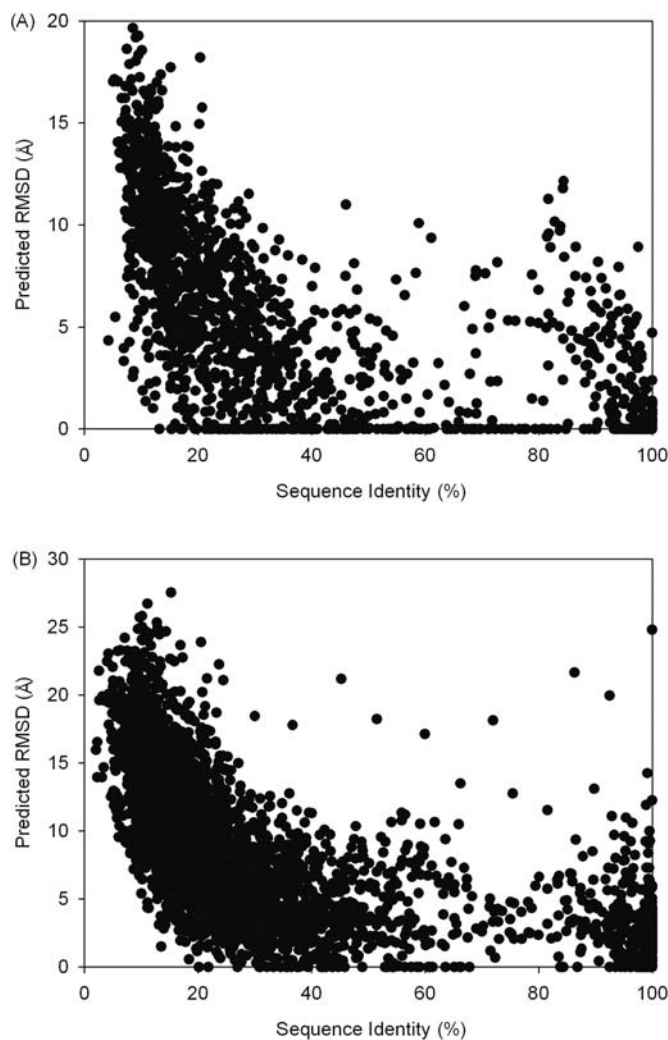


Fig. 10. Predicted RMSD of the structural models is plotted against the sequence identity between the target protein and its template. (A) The GTOP-Nest structural models. (B) The SPARKS2 set.

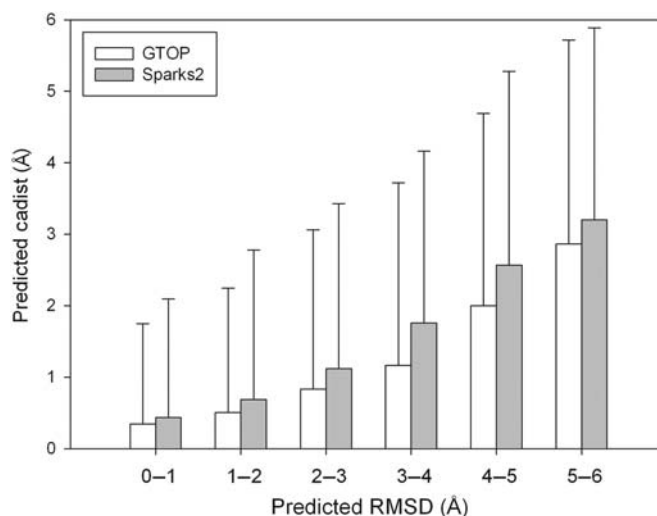


Fig. 11. The range of predicted C α distance relative to predicted global RMSD. The standard deviation is shown by the error bars. White bars, the GTOP-Nest structural models; gray bars, the SPARKS2 set.

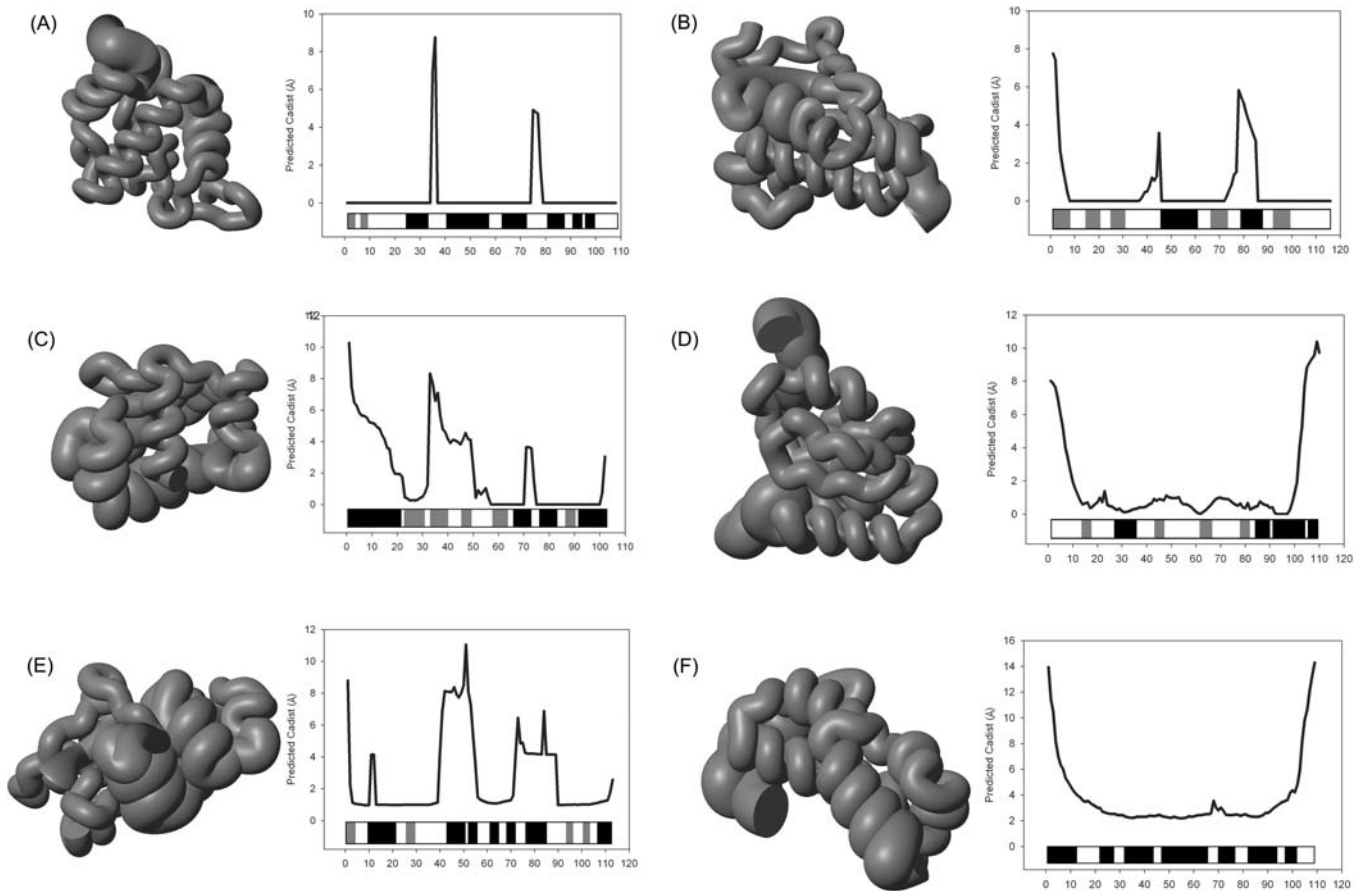


Fig. 12. Examples of estimated local structure quality (the $C\alpha$ distance) of *E. coli* proteins. The $C\alpha$ distance is represented in the ‘sausage representation’, where the radius of the tube is proportional to the estimated $C\alpha$ distance. The $C\alpha$ distance is predicted by the two-layer linear regression model (Fig. 4). (A), (C) and (E) are predicted by the GTOP-Nest procedure and (B), (D) and (F) are predicted by the SPARKS2 procedure. Associated graphs show predicted $C\alpha$ distance. Black/gray boxes in the graph show location of α helices/ β strands in the chain. (A), yccK (predicted sulfite reductase subunit), the template structure used is 1sauA. The sequence identity (seqID) between the gene and the template is 38%. The global RMSD is predicted 0.7 Å. (B), rpsK (30S ribosomal subunit protein S11), template: 1s1hK, seqID: 35%. Predicted global RMSD: 0.6 Å. (C), yiqP (function unknown), template: 1iktA, seqID: 15%, predicted RMSD: 3.1 Å. (D), holD (DNA polymerase ψ subunit), template: 1em8B, seqID: 80%, predicted RMSD: 3.14 Å. (E), yfgD (predicted oxidoreductase), template: 1rw1A, seqID: 15%, predicted RMSD: 5.3 Å. (F), rhaR (positive regulatory gene), template: 1ft9A, seqID: 7%, predicted RMSD: 5.3 Å.

hierarchical approach, which uses predicted RMSD as one of the variables, is effective. The developed quality assessment methods named Sub-AQUA should be useful for biologists who intend to use computational structure models in designing and interpreting biological experiments.

Funding

This work is partially supported by National Institute of General Medical Sciences of the National Institutes of Health (U24GM077905 and R01GM075004), National Science Foundation (DMS604776, DMS800568, IIS0915801, EF0850009) and the Purdue Research Foundation. P.S. was supported by Howard Hughes Summer Internship through Department of Biological Sciences, Purdue University.

References

Al-Lazikani, B., Jung, J., Xiang, Z. and Honig, B. (2001) *Curr. Opin. Chem. Biol.*, **5**, 51–56.
 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucl. Acids Res.*, **25**, 3389–3402.
 Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2008) *Nucl. Acids Res.*, **36**, D419–D425.

Arakaki, A.K., Zhang, Y. and Skolnick, J. (2004) *Bioinformatics*, **20**, 1087–1096.
 Ashworth, J., Havranek, J.J., Duarte, C.M., Sussman, D., Monnat, R.J. Jr, Stoddard, B.L. and Baker, D. (2006) *Nature*, **441**, 656–659.
 Baker, D. and Sali, A. (2001) *Science*, **294**, 93–96.
 Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucl. Acids Res.*, **28**, 235–242.
 Bowie, J.U., Luthy, R. and Eisenberg, D. (1991) *Science*, **253**, 164–170.
 Chandonia, J.M. and Brenner, S.E. (2006) *Science*, **311**, 347–351.
 Chen, H. and Kihara, D. (2008) *Proteins*, **71**, 1255–1274.
 Chothia, C. and Lesk, A.M. (1986) *EMBO J.*, **5**, 823–826.
 Cleveland, W.S. (1979) *J. Am. Stat. Assoc.*, **74**, 829–836.
 Cleveland, W.S. and Devlin, S.J. (1988) *J. Am. Stat. Assoc.*, **83**, 596–610.
 Colovos, C. and Yeates, T.O. (1993) *Protein Sci.*, **2**, 1511–1519.
 Davis, I.W., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2004) *Nucl. Acids Res.*, **32**, W615–W619.
 Eisenberg, D., Luthy, R. and Bowie, J.U. (1997) *Methods Enzymol.*, **277**, 396–404.
 Eramian, D., Shen, M.Y., Devos, D., Melo, F., Sali, A. and Marti-Renom, M.A. (2006) *Protein Sci.*, **15**, 1653–1666.
 Eramian, D., Eswar, N., Shen, M.Y. and Sali, A. (2008) *Protein Sci.*, **17**, 1881–1893.
 Eswar, N., Eramian, D., Webb, B., Shen, M.Y. and Sali, A. (2008) *Methods Mol. Biol.*, **426**, 145–159.
 Feig, M. and Brooks, C.L., III (2002) *Proteins*, **49**, 232–245.
 Ginalski, K. (2006) *Curr. Opin. Struct. Biol.*, **16**, 172–177.
 Hoof, R.W., Vriend, G., Sander, C. and Abola, E.E. (1996) *Nature*, **381**, 272.
 Jiang, L., et al. (2008) *Science*, **319**, 1387–1391.
 John, B. and Sali, A. (2003) *Nucl. Acids Res.*, **31**, 3982–3992.

- Kawabata, T., Fukuchi, S., Homma, K., Ota, M., Araki, J., Ito, T., Ichiyoshi, N. and Nishikawa, K. (2002) *Nucl. Acids Res.*, **30**, 294–298.
- Kihara, D. and Skolnick, J. (2004) *Proteins*, **55**, 464–473.
- Kihara, D., Chen, H. and Yang, Y.D. (2009) *Curr. Protein Pept. Sci.*, **10**, 216–228.
- Kosinski, J., et al. (2005) *Proteins*, **61**(Suppl. 7), 106–113.
- Kryshtafovych, A., Venclovas, C., Fidelis, K. and Moulton, J. (2005) *Proteins*, **61**(Suppl. 7), 225–236.
- Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Crystallogr.*, **26**, 283–291.
- Laskowski, R.A., Watson, J.D. and Thornton, J.M. (2005) *J. Mol. Biol.*, **351**, 614–626.
- Lee, M., Jeong, C.S. and Kim, D. (2007) *BMC Bioinformatics*, **8**, 471.
- Levitt, M. (2007) *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.
- Levitt, M. and Gerstein, M. (1998) *Proc. Natl Acad. Sci.*, **95**, 5913–5920.
- Lindahl, E. and Elofsson, A. (2000) *J. Mol. Biol.*, **295**, 613–625.
- Lu, H. and Skolnick, J. (2001) *Proteins*, **44**, 223–232.
- Lu, H. and Skolnick, J. (2003) *Biopolymers*, **70**, 575–584.
- Lu, M., Dousis, A.D. and Ma, J. (2008) *J. Mol. Biol.*, **376**, 288–301.
- Luthy, R., Bowie, J.U. and Eisenberg, D. (1992) *Nature*, **356**, 83–85.
- McGuffin, L.J. (2007) *BMC Bioinformatics*, **8**, 345.
- Melo, F. and Feytmans, E. (1997) *J. Mol. Biol.*, **267**, 207–222.
- Melo, F., Devos, D., Depiereux, E. and Feytmans, E. (1997) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 187–190.
- Melo, F., Sanchez, R. and Sali, A. (2002) *Protein Sci.*, **11**, 430–448.
- Mereghetti, P., Ganadu, M.L., Papaleo, E., Fantucci, P. and De, G.L. (2008) *BMC Bioinformatics*, **9**, 66.
- Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) *Proteins*, **12**, 345–364.
- Pawlowski, M., Gajda, M.J., Matlak, R. and Bujnicki, J.M. (2008) *BMC Bioinformatics*, **9**, 403.
- Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
- Petrey, D., et al. (2003) *Proteins*, **53**(Suppl. 6), 430–435.
- Pettitt, C.S., McGuffin, L.J. and Jones, D.T. (2005) *Bioinformatics*, **21**, 3509–3515.
- Pieper, U., et al. (2006) *Nucl. Acids Res.*, **34**, D291–D295.
- Pontius, J., Richelle, J. and Wodak, S.J. (1996) *J. Mol. Biol.*, **264**, 121–136.
- Qu, X., Swanson, R., Day, R. and Tsai, J. (2009) *Curr. Protein Pept. Sci.*, **10**, 270–285.
- Reeves, G.A., Dallman, T.J., Redfern, O.C., Akpor, A. and Orengo, C.A. (2006) *J. Mol. Biol.*, **360**, 725–741.
- Rothlisberger, D., et al. (2008) *Nature*, **453**, 190–195.
- Sali, A. and Blundell, T.L. (1993) *J. Mol. Biol.*, **234**, 779–815.
- Shen, M.Y. and Sali, A. (2006) *Protein Sci.*, **15**, 2507–2524.
- Siew, N., Elofsson, A., Rychlewski, L. and Fischer, D. (2000) *Bioinformatics*, **16**, 776–785.
- Skolnick, J. (2006) *Curr. Opin. Struct. Biol.*, **16**, 166–171.
- Skolnick, J. and Kihara, D. (2001) *Proteins*, **42**, 319–331.
- Skowronek, K.J., Kosinski, J. and Bujnicki, J.M. (2006) *Proteins*, **63**, 1059–1068.
- Terashi, G., Takeda-Shitaka, M., Kanou, K., Iwadata, M., Takaya, D., Hosoi, A., Ohta, K. and Umeyama, H. (2007) *Proteins*, **69**(Suppl. 8), 98–107.
- Todd, A.E., Marsden, R.L., Thornton, J.M. and Orengo, C.A. (2005) *J. Mol. Biol.*, **348**, 1235–1260.
- Tondel, K. (2004) *J. Chem. Inf. Comput. Sci.*, **44**, 1540–1551.
- Tosatto, S.C. and Battistutta, R. (2007) *BMC Bioinformatics*, **8**, 155.
- Vingron, M. and Argos, P. (1990) *Protein Eng.*, **3**, 565–569.
- Wallner, B. and Elofsson, A. (2003) *Protein Sci.*, **12**, 1073–1086.
- Wallner, B. and Elofsson, A. (2006) *Protein Sci.*, **15**, 900–913.
- Wells, G.A., Birkholtz, L.M., Joubert, F., Walter, R.D. and Louw, A.I. (2006) *J. Mol. Graph. Model.*, **24**, 307–318.
- Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) *J. Mol. Biol.*, **297**, 233–249.
- Wroblewska, L., Jagielska, A. and Skolnick, J. (2008) *Biophys. J.*, **94**, 3227–3240.
- Xiang, Z. (2006) *Curr. Protein Pept. Sci.*, **7**, 217–227.
- Zemla, A. (2003) *Nucl. Acids Res.*, **31**, 3370–3374.
- Zhang, Z., Berman, P., Wiehe, T. and Miller, W. (1999) *Bioinformatics*, **15**, 1012–1019.
- Zhou, H. and Zhou, Y. (2002) *Protein Sci.*, **11**, 2714–2726.
- Zhou, H. and Zhou, Y. (2005a) *Proteins*, **58**, 321–328.
- Zhou, H. and Zhou, Y. (2005b) *Proteins*, **61**(Suppl. 7), 152–156.