# Chapter 13
# Error Estimation of Template-Based Protein Structure Models

**Daisuke Kihara, Yifeng David Yang, and Hao Chen**

**Abstract** The tertiary structure of proteins provides rich source of information for understanding protein function and evolution. Computational protein tertiary structure prediction has made significant progress over more than a decade due to the advancement of the techniques and the growth of sequence and structure databases. However, methods for assessing quality of predicted structure models are not well established. Quality assessment of structure models is important for reranking and selecting the best possible models from a pool of models as a post-processing step in structure prediction, and thus many methods are developed in this direction. Recently, it is also recognized that the model-quality assessment is crucial for practical use of a model such as design and interpretation of biochemical experimental data. For such practical application of a computational model, the real-value quality of the model should be predicted, which is different from reranking alternative models. The quality (error) of a model determines its potential practical application, ranging from protein design, designing site-directed mutagenesis experiments, ligand–protein docking prediction, to function prediction from structure. In this chapter, we discuss importance of the real-value error estimation and overview the existing methods.

## 13.1 Introduction

Protein tertiary structure prediction from amino acid sequence has made steady progress due to advances in techniques as well as the increase in the number of known solved structures available for templates for modeling. Now it is often possible to build atomic-detailed models which can be used for redesigning protein

D. Kihara (✉)
Department of Biological Sciences, College of Science; Department of Computer Science, College of Science; Markey Center for Structural Biology, Purdue University, West Lafayette, IN, USA
e-mail: dkihara@purdue.edu

A. Kolinski (ed.), *Multiscale Approaches to Protein Modeling*,
DOI 10.1007/978-1-4419-6889-0_13, 

function (Ashworth et al. 2006). However, predicting an accurate atomic-detailed model is still not always possible. Depending on the techniques employed and available template structures for modeling, the accuracy of a model ranges from an root mean square deviation (RMSD) of 1.5–2 Å (typically in comparative modeling using a closely related template to the target) to 6 Å (threading) to its native structure and even over 10 Å in case when the prediction goes significantly wrong. In the latest Critical Assessment of Techniques in Structure Prediction (CASP) experiment that was documented on *Proteins* (CASP7 held in the summer 2006), the average GDT-HA score of models submitted by the top five groups in the template-based modeling (TBM) category was 50.3 (Kopp et al. 2007). In the free-modeling (FM) category, the average GDT-TS score of the top groups was around 30–40, as read from Fig. 13.2B of the assessors' report (Jauch et al. 2007). The GDT-HA and the GDT-TS scores are defined as the average of the percentage of residues which do not deviate from the target structure by more than four threshold values, 0.5, 1.0, 2.0, and 4.0 Å for GDT-HA and 1.0, 2.0, 4.0, and 8.0 Å for the GDT-TS. Thus, very roughly speaking, the CASP results imply that on average about half of

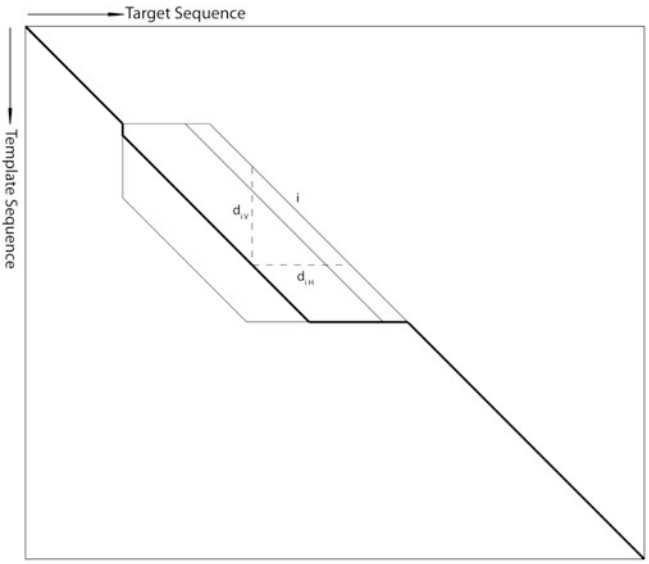


**Fig. 13.1** Definition of the SPAD score. SPAD quantifies the diversity of suboptimal alignments around the optimal alignment of a target and a template protein. The difference (distance) of a position in an alignment as compared with an alternative alignment is computed on a dynamic programming (DP) matrix. The target protein sequence is placed on the horizontal axis while the template protein is aligned on the vertical axis. The *thick line* represents the DP path of the optimal alignment of the two proteins. The distance from a position in the optimal alignment to the *i*-th suboptimal alignment is defined as the average of the horizontal ($d_{i\mathrm{H}}$) and the vertical ($d_{i\mathrm{V}}$) distance to the *i*-th suboptimal alignment. The local SPAD score for a position in the optimal alignment is the average of the distance to the set of suboptimal alignments considered. Then, the global SPAD score of the optimal alignment is defined as the average of the local SPAD score of each position in the alignment. In principle, the computation of SPAD is applicable for any threading method that uses the dynamic programming

residue position in a model is "correct" in the TBM, while only about one-third of residues are predicted correctly in the FM. Therefore, it should be realized that a computational model may have significant errors even if it is constructed by current state-of-the-art methods.

Methods for error estimation or quality assessment of structure models have drawn much attention in recent years (Kihara et al. 2009). Roughly speaking, there are three purposes or types of quality assessments of protein structure models. First, in structural biology, stereochemical properties of experimentally solved structures are routinely examined. The second type is to rerank predicted models, i.e., evaluating relative accuracy among the models, such that the most native-like structure can be selected from a pool of constructed models. The third purpose is to predict the real-quality value of a model, such as an RMSD value of the model to the native structure. These three purposes have significant overlap between each other but they are not identical. Thus, methods developed for one purpose is not necessarily suitable for the other purposes.

In experimental structural biology, validation of tertiary structures built from experimental data is an important step. Thus, earlier works on protein structure validation focused on identifying potential errors in models built from X-ray diffraction patterns of protein crystals. Tools developed for that purpose include PROCHECK (Laskowski et al. 1993), MOLPROBITY (Davis et al. 2004), protein volume evaluation (PROVE) (Pontius et al. 1996), and WHATCHECK (Hooft et al. 1996). These methods compare stereochemical properties of a protein structure, such as the bond length, bond angles, hydrogen bonds, and atom clashes, with their regular values sampled from a set of representative protein structures of a good resolution. The same methods could be applied to assess the quality of predicted structures (Bhattacharya et al. 2008). However, it may not always be suitable to use these validation tools for analyzing predicted structures because these tools concern small deviations of distances or angles, which is the level of the accuracy that may not be meaningful to expect for predicted structures of a moderate accuracy.

The second type of the model-quality assessment, reranking of predicted models, is most well studied recently in the context of protein structure prediction. In a procedure of structure prediction, particularly in ab initio structure prediction, which usually generates a large number of alternative models, selecting a few most native-like models is essential. Of course each structure prediction method has its own scoring function that guides building models, but an outside measure specific for quality assessment could often offer useful evaluation. For example, clustering of models to find out most populated folds is proven to be an effective metric of selecting near-native models (Betancourt and Skolnick 2001; Shortle et al. 1998). In addition, reranking of models is also an important step in a meta-server approach, where models generated by different methods are compared and often combined (Kolinski and Bujnicki 2005). To rerank models, various aspects of the models are evaluated, ranging from structures and energetic terms in atomic-detailed levels, in residue levels, in global-fold levels, and often in sequence-alignment levels. In the next section we briefly discuss structural characteristics used as scoring terms in quality assessment.

The third type of quality assessment, i.e., prediction of real value of quality of models (e.g., the RMSD value to the native structure), is relatively less studied. Methods for the previous two types are not necessarily suitable for predicting a real-quality value of a model, because the accuracy of detailed stereochemical properties (i.e., the first type) does not guarantee global structural similarity of a model to the native (Wroblewska and Skolnick 2007; Melo and Sali 2007). Interestingly, it is demonstrated that structure models of a wrong fold can have good detailed stereochemical structures (Wroblewska and Skolnick 2007). Moreover, the best-ranked model among a given pool, which is identified by the second type of the quality assessment method, may not have a certain value of the RMSD, e.g., less than 3 Å. To predict a real value of quality of models, the same structural characteristics used in model reranking methods could be employed. However, real-value quality prediction would be more difficult because most of the quality assessment measures do not have rationale to have a good correlation to the absolute value of the model quality. Exceptions would be terms that evaluate target–template alignment, a typical one being the sequence identity between the target sequence to be modeled and a template protein. The relationship between the sequence similarity of and the structural divergence of proteins is well established (Chothia and Lesk 1986; Wilson et al. 2000).

The real-value quality assessment methods would be most useful for biologists who would like to practically use structure models to aid their wet-lab experiments. If the accuracy of a model can be predicted, it can be used for an appropriate purpose according to its estimated accuracy. It is important to note that a low-resolution model is still useful for certain purposes (Baker and Sali, 2001). High-resolution models with an RMSD of 1–1.5 Å are useful for almost any application where a tertiary structure of a protein can be useful, including for studying catalytic mechanism of enzymes, for structure-based protein engineering, and for drug design. A model of an RMSD of around 4 Å is still useful, for example, for designing site-directed mutagenesis experiments (Wells et al. 2006; Skowronek et al. 2006), chemical labeling, and for performing small-ligand-docking predictions (Wojciechowski and Skolnick, 2002; Vakser 1996). If the fold of a model is expected to be correct (an RMSD of about 6 Å), function of the protein could still be predicted using the predicted tertiary structures (Baker and Sali 2001; Skolnick et al. 2000; Hawkins and Kihara 2007; Kihara and Skolnick 2004). Therefore, it is important to establish real-value quality estimation methods, so that a model can be used wisely by knowing the limitations of the model. In this chapter, we will focus our review on the quality assessment methods of this type, as the methods of the other two types are recently thoroughly reviewed in another article (Kihara et al. 2009).

## 13.2 Overview of Quality Assessment Measures

We first briefly overview the structural characteristics used as scoring terms in quality assessment. For more details please refer to the recent review (Kihara et al. 2009).

### 13.2.1 Physics-Based Score

One of the most straightforward ways for assessing model quality would be to employ physics-based all-atom force field. Previous works range from using molecular mechanics energy (Kmiecik et al. 2007) to free-energy computation of structure models, including the molecular mechanics–Poisson–Boltzmann surface area (MM-PBSA) free-energy (Lee et al. 2001; Feig and Brooks III 2002), MM-Generalized Born implicit solvation model with a surface area-dependent term (MM-GBSA) (Wroblewska and Skolnick 2007; Feig and Brooks III 2002), the Explicit simulation/implicit solvent (ES/IS) method, which computes the solvation free energy from short molecular dynamics simulations with explicit solvent (Vorobjev and Hermans 2001), and the colony energy approach combined with the MM-PBSA, which assesses conformational entropy by explicitly sampling the conformational space in the vicinity of a reference structure (Fogolari and Tosatto 2005). Wroblewska and Skolnick showed that reoptimization of relative weights of energy components of the AMBER force field yielded significant improvement for scoring and refinement of protein models (Wroblewska et al. 2008).

### 13.2.2 Knowledge-Based Potential

Alternatively, structure models can be evaluated using knowledge-based statistical potentials. A knowledge-based potential considers preference of a certain structural property of atoms or amino acid residues in protein structures by counting the number of observed cases of the property, which is then normalized by the expected number of counts. Many different types of knowledge-based potentials are developed, including atom- or residue-contact potentials, main-chain torsion angles (Betancourt 2008; Tosatto and Battistutta 2007), atom/residue-level burial/exposure preference (Holm and Sander 1992), atom/residue-packing preference (Gregoret and Cohen 1991; Melo and Sali 2007), and the accessible surface area of residues of atoms/residues (Melo and Feytmans 1997; McConkey et al. 2003; Melo et al. 2002). Atom- or residue-contact potentials would be among the most well studied. Pioneer works were carried out by Sippl using knowledge-based contact potentials for identifying errors in protein crystal structures (Hendlich et al. 1990; Sippl 1993). In principle, knowledge-based atom-contact potentials are designed to evaluate structures with an atomic-level accuracy, but they are shown to have good performance on predicted structures of a moderate accuracy as well (Melo and Feytmans 1997; Pettitt et al. 2005; McConkey et al. 2003). When models are not expected to have atomic detail-level accuracy, examining residue contacts could be advantageous for quality assessment (Melo et al. 2002). Verify3D assesses structural environment of residues which are defined by a combination of the secondary structure, burial status, and polarity of positions in a structure (Eisenberg et al. 1997).

### *13.2.3 Assessing Alignment Quality*

In parallel to the structure-based terms, scores which assess the validity of the alignment between a target and a template are effective in the case of template-based modeling. Simply, the quality of a target–template alignment can be evaluated by considering the significance of alignment raw scores, such as the sequence identity, the Smith–Waterman alignment score, or a threading score between a target and a template (in the case of threading). More frequently, statistical significance of a raw score is considered, e.g., the *E*-value in BLAST (Altschul et al. 1990) or the *Z*-score used in threading algorithms. The *Z*-score of a raw alignment score is also obtained from a distribution of alignment scores from shuffled sequences (Pearson and Lipman 1988). Similarly, for predicting quality of local regions, raw scores local alignment regions can be used (Tress et al. 2003; Zhang et al. 1999; Lee et al. 2007; Tondel 2004).

Another strategy to estimate the reliability of an alignment is to compare it explicitly with alternative alignments of the same pair of sequences, i.e., to consider consistency with suboptimal alignments (Mevissen and Vingron 1996; Vingron and Argos 1990; Vingron 1996; Saqi and Sternberg 1991; Jaroszewski et al. 2002; John and Sali 2003). Instead of explicitly computing numerous suboptimal alignments, several other methods provide a probability (reliability) to each position in an alignment. Examples are those which use the partition function to express the probability of alternative alignments (Zhang and Marr 1995; Schlosshauer and Ohlsson 2002; Miyazawa 1995; Koike et al. 2004) and ones using hidden Markov models (Yu and Smith 1999; Cline et al. 2002). Recently, we have also developed a quality assessment score based on the consistency of a target–template alignment with suboptimal alignments, named SPAD (SuboPtimal Alignment Diversity) (Chen and Kihara 2008). SPAD quantifies divergence of suboptimal alignments around the optimal alignment on the Dynamic Programming matrix. It was shown that the SPAD score has a significant correlation not only to alignment shift-level errors but also to global and local structural-level errors (i.e., RMSD to the native structure and the distance of corresponding residues of a model and its native structure) of structure models built based on optimal alignments. We will explain SPAD in the next section in more detail.

## 13.3 The SPAD Score

### *13.3.1 Definition of the SPAD Score*

Figure 13.1 shows how the SPAD score is defined. When an alignment between a target sequence and a template structure is computed by the dynamic programming (DP) algorithm (Needleman and Wunsch 1970), the alignment is represented as a path in the DP matrix. In Fig. 13.1, the top-scoring (i.e., optimal) target–template

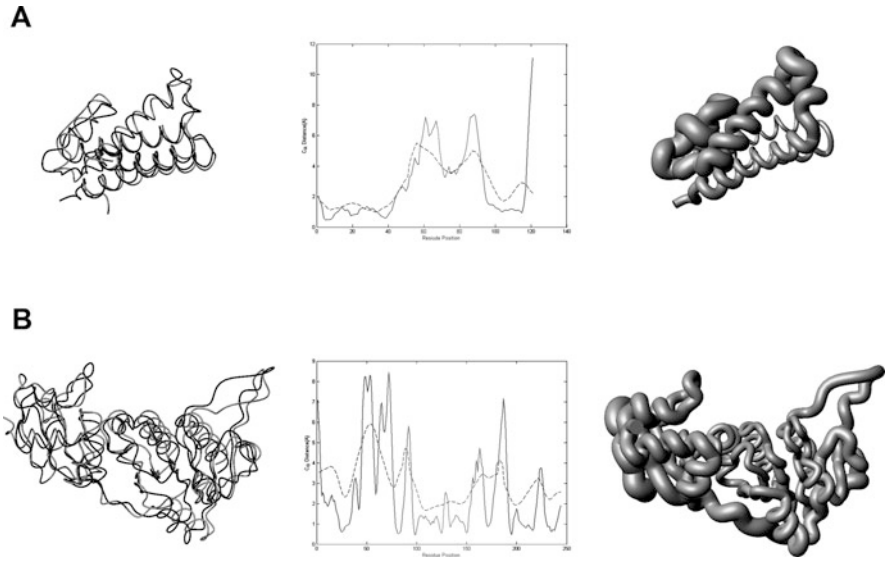


**Fig. 13.2** Examples of estimated global and local errors by the SPAD score. Prediction structures of two CASP7 targets produced by our group, Chen-Tan-Kihara, are shown. The *left panels* are superimposition of the predicted (*black*) and the native (*gray*) structures of the target. The *middle panels* show the predicted (*dotted line*) and the actual (*solid line*) Cα distance error of the models. The *right panels* show the sausage representation of models, where the radius of the tube is proportional to the estimated Cα distance error. (**a**) A model for T0367. The native structure is 2hsbA, of which the length is 125 amino acids. The model used 1ufbA as the template. The global RMSD of the model is predicted to be 4.3 Å, while the actual RMSD of the model to the native is 3.9 Å. (**b**) A model for T0378 (native: 2i6dA, the length: 254 amino acids). The template of the model is 1x7oA. The actual/predicted RMSD of the model to the native is 3.6/3.9 Å

alignment is represented as a *thick path* from the left upper corner to the right bottom corner. As a set of suboptimal alignments are also represented as their own paths, the consistency of a position in the optimal alignment as compared with suboptimal alignments can be quantified by counting the number of grid cells between the paths. Thus, the local SPAD score of a certain position $m$ ($\mathrm{lSPAD}_m$) in the optimal alignment is defined as:

$$\mathrm{ISPAD}_m = \frac{\sum_{i=1}^{n} \mathrm{IALD}^i_m}{n} \tag{13.1}$$

where $\mathrm{lALD}^i{}_m$ is the distance at the position $m$ in the optimal alignment to the suboptimal alignment $i$, which is defined as $\mathrm{IALD}^i_m = \frac{d_{iV}+d_{iH}}{2}$ (Fig. 13.1). $n$ is the number of suboptimal alignments considered. Averaging $\mathrm{lSPAD}_m$ over all the positions in the optimal alignment yields the global-level SPAD, gSPAD:

$$\mathrm{gSPAD} = \frac{\sum_{m=1}^{1} \mathrm{ISPAD}_m}{l} \tag{13.2}$$

where $l$ is the length of the optimal alignment. Suboptimal alignments are computed by the algorithm proposed by Vingron and Argos (1990). 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. The number of suboptimal alignments considered, $n$, is set to $n = 0.1 \times M \times N$ (i.e., top 10% high-scoring suboptimal alignments). A profile–profile alignment (Wang and Dunbrack Jr 2004) is employed to compute optimal and suboptimal alignments of a target and a template, which uses a profile-matching score and the secondary structure-matching score (Chen and Kihara 2008).

### *13.3.2 Correlation of SPAD to RMSD of Models*

We prepared a large dataset of 5,232 template-based structure models and examined how well SPAD correlates with global and local qualities of the models. Models of a variety of quality are obtained by using template structures of three similarity levels to target proteins, i.e., those in the same family as targets, the same superfamily, and the same fold. Target–template alignments are computed by the in-house profile–profile alignment program mentioned above and MODELLER (Eswar et al. 2008) was used to construct the tertiary structure of the model from the alignments. The correlation coefficient computed for log–log plots of the RMSD of the models relative to SPAD (Eq. (13.2)) is 0.598, 0.630, and 0.384, for the models using templates of the family, the superfamily, and the fold-level similarity, respectively (refer to Table 13.2A in Chen and Kihara (2008)). These correlation coefficient values are shown to be much more significant than the other sequence alignment-based measures, which are the sequence identity, the threading $Z$-score, and the $Z$-score by PRSS (an alignment shuffling program) (Pearson and Lipman 1988). Moreover, interestingly, SPAD has more significant correlation than the discrete optimized protein energy (DOPE) score (normalized by the number of atom contacts), the target function of the MODELLER, to the RMSD in the models constructed with templates in the family and the superfamily-level similarity. The correlation of the normalized DOPE score to the RMSD is 0.453, 0.617, and 0.587 in the family, the superfamily, and the fold-level template models. This implies a dominant influence of the alignment quality on the final structural quality in the structure modeling process in MODELLER.

### *13.3.3 Correlation to the Local Quality of Models*

We also examined the correlation of the local SPAD score (Eq. (13.1)) to the local quality of models, which is defined as the Euclidean distance between corresponding

Cα atoms of the model to the native structure when they are globally superimposed. The correlation is 0.565, 0.509, and 0.277 for models which used templates in the family, the superfamily, and the fold-level similarity. These values are not as significant as that of the global SPAD to the RMSD, however, much higher than the other simple local alignment-based scores, such as the conservation in multiple sequence alignments, the ratio of gaps in the alignment position, and the average BLOSUM score of the alignment position (Table 13.2B in Chen and Kihara (2008)). According to the linear correlation we observed in the log–log plots of the SPAD score and the RMSD and the local Cα distance error, two equations are obtained:

$$\text{RMSD} = \exp(0.3576 \times \ln(\text{global SPAD}) + 1.882) \tag{13.3}$$

$$\text{Cα distance error} = \exp(0.3294 \times \ln(\text{local SPAD}) + 1.645) \tag{13.4}$$

Figure 13.2 shows two examples of actual prediction of the global RMSD and the local Cα distance error of structure models using the two equations. These two structure models are predicted for targets T0367 and T0378 in the CASP7. The in-house profile-based threading program was used to find the template and make the alignment, which was followed by running MODELLER to build the structural model. In these examples, the global RMSD of the models are predicted quite well. The predicted Cα distance error (the middle panels) captures poorly predicted regions, although the absolute value of the predicted and actual error did not agree well in some regions. What are shown in the *right panel* are the sausage representations of the models, which intuitively represent predicted Cα distance error as well as the overall fold of models.

## 13.4 Real-Value Quality Assessment of Structure Models

In this section, we overview three quality assessment methods which predict real-value quality. At last, we introduce a method we recently developed, named SubAqua (Subotpimal Alignment-based quality assessment method), which uses the SPAD score as a main component of its scoring term.

### *13.4.1 Tondel's Method*

Tondel showed that the global RMSD and the total residue contact area of homology models can be well predicted by a regression model which combines sequence alignment-based scores between a target and a template (Tondel 2004). The scores employed are the sequence identity, the number of non-aligned residues in the target–template alignment, and an amino acid similarity score (PAM250) of each position in the alignment. The method was tested on a set of homology models of kinase, whose RMSD to the native varies in a relatively small range from 1 to 7 Å. Since the score of each alignment position is used as variables in the regression

model, this method can be only applied to the specific protein family for which the regression model is built for. However, as the author discusses, it is possible to construct a regression model for each major protein family.

### *13.4.2 ProQ*

ProQ uses neural network to predict real value of two structure similarity scores, the LGscore (Cristobal et al. 2001) and the MaxSub score (Siew et al. 2000) to the native structure (Wallner and Elofsson 2006). They found that combining different types of quality assessment measures improves the accuracy of the correct fold prediction. They combined seven terms of different natures that range from those for describing coarse-grained to atomic-detailed structural features of a model. Two terms are for capturing overall global features of a model: the fraction of the protein modeled and the fatness, which is the ratio of the longest and the shortest axes when the structure is fit to an ellipsoid. Two other terms are for describing main-chain level features: the agreement of the model to the template structure as measured by LGscore or MaxSub and the agreement of the actual and predicted secondary structure of the model. In addition, two residue-level features are combined: the fraction of residues in four different bins of accessible surface area and the fraction of residue–residue contacts classified into four categories. Finally, an atomic-detailed structural feature is captured by the fraction of atom–atom-contact types observed in the model.

### *13.4.3 TVSMod*

Eramian et al. developed a method named TVSMod, which predicts the RMSD as well as the number of correctly predicted residues in a model (residues locating within 3.5 Å to the corresponding residues in the native structure, named No3.5 Å) using support vector machine (SVM) regression, which combines up to nine alignment and structure features (Eramian et al. 2008). The nine features are the sequence identity between a target and template, the percentage of gaps in the target–template alignment, a distance-dependent residue-contact potential, a distance-dependent atom-contact potential (DOPE), a residue-based accessible surface statistical potential score, a composite score named GA341, which combines residue-level contact and accessibility scores and a score measuring structural compactness, and the sequence identity (John and Sali 2003), and agreement scores of prediction and actual secondary structure of the model. The residue- and atom-contact potential values and the composite score are normalized by computing the *Z*-score referencing a score distribution of 200 random sequences with the same amino acid composition and the structure as the query model. Consistent with the paper of ProQ (Wallner and Elofsson 2006), they reported that combination of these scores improve the accuracy although each individual score does not have a strong correlation to the RMSD and No3.5 Å.

TVSMod also uses SVM. SVM is trained in a model-specific fashion, that is, SVM is developed using structure models of the similar size and the secondary structure content as the query model. Interestingly, for an input structure model to be evaluated, the SVM is trained on the fly using a large database of 5,790,899 template-based models with known quality: First, if the aligned target and template sequences share more than 85% identity, the system simply predicts an RMSD of 0.5 and 1.0 Å for No3.5 Å without taking any further steps. If the target and template are not so closely related, the model database is scanned to find all examples where the same region of the template was used either as a template or as the target sequence. A region in a model in the database is considered equivalent with a certain region in the query model if the starting and the ending residues are each within ten residues and the difference is within 10% of the length of the query. Subsequently, those selected models are further filtered by considering similarity of the score values of the residue- and atom-contact potentials, the residue-based accessible surface area potential, and the composite score for the distance and surface potential score. Finally, SVM is trained on the tailored training dataset. They reported a high-correlation coefficient of 0.84 between the predicted and actual RMSD and 0.86 for predicted/actual No3.5 Å. The advantage of using a query-tailored training dataset is that the structure-based potentials, e.g., the contact potentials, will be able to have better correlation to the model-quality values.

### *13.4.4 The SubAqua Method*

Recently we have extended the idea of using suboptimal alignments for model-quality assessment to develop a better quality assessment method, named SubAqua (Subotpimal Alignment-based quality assessment method) (Yang et al. 2009). The webserver is available at http://kiharalab.org/SubAqua/. SubAqua combines the SPAD score introduced above (Chen and Kihara 2008) with other structure-based scoring terms. It predicts a global and a local real-value quality measures, the RMSD value of a model to its native structure, and the error of Cα positions as compared with corresponding residues when the model is superimposed with the native structure. Since SubAqua uses suboptimal alignments as one of the scoring terms, it is more suitable for evaluating template-based models. Below we briefly describe outline of the work.

#### 13.4.4.1 Correlation of Quality Assessment Terms to RMSD

We first prepared a large dataset of template-based models with a variety of quality, whose RMSD ranges from around 1 to 20 Å. Pairs of proteins which share the family, the superfamily, and the fold-level similarity are selected from the Lindahl and Elofsson's dataset (Lindahl and Elofsson 2000), which resulted in 1,076, 1,395, and 2,761 pairs, respectively. Optimal alignments of the pairs are computed using a profile–profile alignment (Wang and Dunbrack Jr 2004), which are then fed to MODELLER (Eswar et al. 2008) to build the tertiary structure models.

Using the dataset of template-based models, first we examined a dozen of quality assessment measures in terms of the correlation coefficient to the RMSD of models (Table 13.1). Five different types of measures are examined, those based on target–template alignments, overall model fold, those concern local residue environments, stereochemistry of atoms, and composite model-quality assessment scores. Individual scores are explained in the table legend. What is roughly consistent with the observation by Eramian et al. (2008) is that none of the individual scores by itself has a significant correlation to the RMSD of the models in this diverse dataset. Among the measures examined, the alignment-based scores have relatively higher correlation over 0.5, which implies strong dependence of the modeling procedure (by MODELLER) to the quality of target–template alignments. SPAD shows a better correlation when log-transformed. This may be because the possible alternative alignments diverge rapidly as the sequence similarity drops. We also found that normalizing the Verify3D score by the model length ($L$) and a square of the length ($L^2$) improves the correlation and taking the logarithm of Verify3D/$L^2$ further improves the correlation to 0.53. It is expected that normalization by the length improves the correlation since Verify3D is the 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. The other residue-level and atomic-level scores have insignificant correlation to the RMSD. ProQ also shows insignificant correlation to the RMSD, which is consistent with Table 13.1 in Eramian et al. that reports the correlation coefficient of 0.57 and 0.44 for ProQ-MX and ProQ-LG (Eramian et al. 2008).

#### 13.4.4.2 Variable Selection for Constructing Regression Models

Next, we employed the forward stepwise variable selection procedure to build a linear regression model with a meaningful subset of the variables in Table 13.1. In this procedure, variables are added to the regression model sequentially until adding more variables makes no significant contribution. The contribution of a variable to a regression model is shown by the partial $R^2$, which indicates how much more $R^2$ can be explained by adding the variable. Two variables, log(SPAD) and log(Verify3D/$L^2$), are selected in this order as the most contributing variables to predict the global RMSD with the model $R^2$ value of 0.586 (Table 13.2). The rest of the variables, the $Z$-score by PRSS to the normalized DOPE, are selected with statistical significance but their contribution to the model judged by the partial $R^2$ is marginal. Therefore, we decided to use only the two variables and the resulting linear regression model is as follows:

$$\mathrm{RMSD} = -4.99 + 2.25 \times \log(\mathrm{SPAD}) - 2.17 \times \log(\mathrm{Verify3D}/L^2) \qquad (13.5)$$

Figure 13.3 shows the relationship between the actual and predicted RMSDs of the structure models in the entire dataset. The correlation coefficient is 0.77.

**Table 13.1** Correlation coefficient to the global RMSD of structure models

| Types of variables | Variable | Correlation coefficient |
|---|---|---|
| Alignment | Seq. identity | 0.58 |
| | Length | 0.24 |
| | PRSS *Z*-score | 0.63 |
| | SPAD | 0.55 |
| | log (SPAD) | 0.71 |
| Overall fold | Compactness (Sc) | 0.19 |
| Residue environment | ERRAT | 0.35 |
| | TAP | 0.34 |
| | Verify3D | 0.09 |
| | Verify3D/$L$ | 0.45 |
| | Verify3D/$L^2$ | 0.46 |
| | log(Verify3D/$L^2$) | 0.53 |
| Atom environment | DOPE/${N_h}^2$ | 0.29 |
| | ANOLEA1 | 0.29 |
| | ANOLEA2 | 0.02 |
| | PROCHECK1 | 0.05 |
| | PROCHECK2 | 0.16 |
| Quality assessment scores | ProQ-MX | 0.39 |
| | ProQ-LG | 0.31 |
| | GA341 | 0.17 |
| | Pg | 0.01 |

Absolute values of the correlation coefficients are shown. Scores with a correlation coefficient of over 0.5 are highlighted in gray. Alignment-based scores are the sequence identity between the target and the template sequences; the number of residues of the model, the *Z*-score of the alignment score computed from a score distribution of shuffled sequences by PRSS; the SPAD score considers the consistency of the target–template alignment relative to suboptimal alignments. The compactness is defined as the total volume of each amino acids relative to that of a sphere with the diameter of the maximum distance of residue pairs (Melo and Sali 2007). The three residue-level scores, ERRAT, evaluates the number of non-bonded interactions between heavy atoms (Colovos and Yeates 1993); TAP evaluates torsion angle propensity of amino acids (Tosatto and Battistutta 2007); and Verify3D assesses the fitness of each residue in a model to its structural environment defined by the total/poplar burial area and the secondary structure (Luthy et al. 1992). In addition to the raw Verify3D score, three variations of normalized score by the model length ($L$) are also examined. Atomic-detailed structure is examined by four measures: DOPE is an atom distance-dependent statistical potential (Shen and Sali 2006). Here we normalized DOPE by square of the number of heavy atoms ($N_h$) in a model. ANOLEA1 and ANOLEA 2 come as output from the ANOLEA method that evaluates atom contact and accessible surface propensity (Melo and Feytmans 1997). PROCHECK1 is the percentage of residues in the disallowed region in the Ramachandran plot and PROCHECK2 is the percentage of residues with bad contacts, both of which come from the PROCHECK program (Laskowski et al. 1993). ProQ-MX and ProQ-LG is predicted MaxSub and the LG score, respectively, by a neural network-based program, ProQ (Wallner and Elofsson 2006). GA341 combines atom contact and solvent accessibility potentials, the sequence identity, and the compactness score (Sc), and pG is predicted probability that a model has a correct fold using a Bayesian classifier which uses Ga341 and the model length (Melo et al. 2002)

**Table 13.2** Variable selection for linear regression model

| Step | Variable | Partial $R^2$ | Model $R^2$ | $P$ ($F$)[a] |
|---|---|---|---|---|
| 1 | log (SPAD) | 0.499 | 0.499 | <0.001 |
| 2 | log (Verify3D/$L^2$) | 0.087 | 0.586 | <0.001 |
| 3 | PRSS $Z$-score | 0.014 | 0.600 | <0.001 |
| 4 | ERRAT | 0.010 | 0.610 | <0.001 |
| 5 | Sc | 0.007 | 0.617 | <0.001 |
| 6 | ProQ-MX | 0.004 | 0.622 | 0.0001 |
| 7 | DOPE/$N_h{}^2$ | 0.003 | 0.624 | 0.0015 |

[a]The $p$-value of $F$-value is another statistical metric to show the significance of the contribution of the variable to the regression model. A smaller $p$-value means more significant contribution.

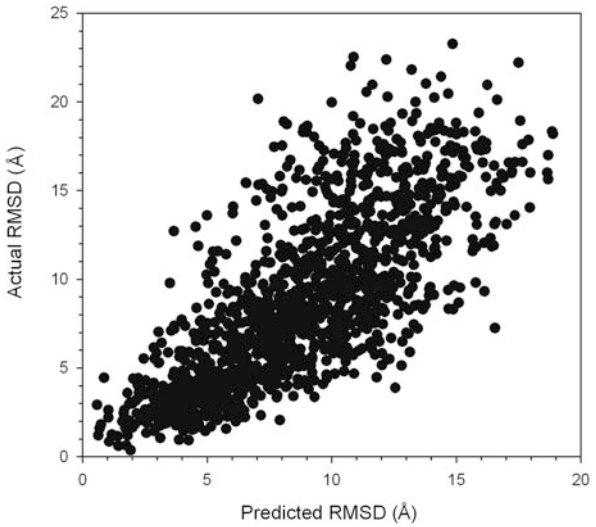


**Fig. 13.3** Predicted and actual RMSD values of template-based models. Equation (13.5) is used to predict RMSD

We have also employed logistic regression to predict if a model's RMSD is smaller than a certain value (i.e., "correct" structure) or not (i.e., "incorrect" structure). We used 6.0 Å as the threshold value of the RMSD. The forward variable selection is used with the same set of variable choices (shown in Table 13.1). Again, the same two variables, log(SPAD) and log(Verify3D/$L^2$), are selected as the most significant variables:

$$\log(p/1-p) = -7.93 + 1.62 \times \log(\text{SPAD}) - 1.48 \times \log(\text{Verify3D}/L^2), \quad (13.6)$$

where $p$ is the probability that a model is correct (i.e., an RMSD of below 6.0 Å). The model dataset has 1,843 correct models and 3,389 incorrect models. Using the logistic regression model (Eq. (13.6)), 4,376 models (83.64%) are correctly classified either to correct or incorrect models.

We conclude the followings from these results: First, as also observed by other related studies, there is no single term which is sufficient by itself to predict the global RMSD of models (Table 13.1) but combination of the terms improves prediction of RMSD (Table 13.2). Among the quality assessment terms tested, those which evaluate quality of target–template alignments rather than atom- or residue-based structural terms have higher correlation to the quality of the models (Table 13.1). Moreover, the forward stepwise variable selection procedure identified two variables for constructing linear regression as the most significant contributing terms, namely, log(SPAD) and log(Verify3D/$L^2$), both of which are evaluating coarse-grained features of models (Table 13.2). Adding more structure-based terms including those which evaluate atomic-detailed structures (ERRAT and DOPE) improves the linear regression model, however, their contribution is marginal (Table 13.2). We will further extend the discussion later while summarizing this chapter.

#### 13.4.4.3 Two-Step Procedure to Predict Local Quality

Next, we developed a regression model for predicting local quality of models, i.e., the Cα distance between corresponding residues of a model and its native structure. Individual scoring terms examined do not show significant correlation to the Cα distance (Table 13.3) and regression models constructed by combining these scores does not show sufficient correlation, neither. However, we find that the prediction performance show significant improvement when the predicted global RMSD (Eq. (13.5)) is integrated. The predicted global RMSD correlates relatively well with

**Table 13.3** Variable selection for linear regression model for local quality prediction

| Step | Variable | Partial $R^2$ | Model $R^2$ | $P$ ($F$) | Corr. coeff.[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 (local SPAD/SPAD + 1) | 0.0229 | 0.2622 | <0.0001 | 0.21 |
| 4 | local Verify3D | 0.0066 | 0.2688 | <0.0001 | 0.32 |
| 5 | log ($\text{localVerify3D}_{\text{positive}}{}^2$ + 1)[a] | 0.0048 | 0.2736 | <0.0001 | –0.14 |
| 6 | Conservation[b] | 0.0038 | 0.2774 | <0.0001 | –0.29 |
| | Mutation score[c] | – | – | – | –0.32 |
| | Local ERRAT | – | – | – | 0.19 |

[a]The correlation coefficient against the Cα distance.
[b]$\text{localVerify3D}_{\text{positive}}$ is a non-negative local Verify3D score assigned to each residue; 0 is assigned when a negative localVerify3D score is replaced with 0.
[c]The 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.
[d]Average BLOSUM45 score of a column in the profile.

the Cα distance, with the correlation coefficient of 0.44. Therefore, we designed a two-step procedure by integrating the predicted global RMSD to predict local quality of individual residues: First, given a protein structure model, the SubAqua method predicts the global RMSD using the Eq. (13.5). Then, the Cα distance is predicted by a linear regression combining the predicted global RMSD, the gap ratio, and log(localSPAD/SPAD + 1). These three variables are selected by the forward stepwise variable selection. The resulting regression is as follows:

$$\begin{aligned} \text{C}\alpha\,\text{distance} = &-4.04 + 0.94 \times (\text{predicted RMSD}) + 16.59 \times (\text{gap ratio}) \\ &+3.55 \times \log(\text{localSPAD/SPAD} + 1) \end{aligned} \tag{13.7}$$

The gap ratio is the fraction of the gaps at the residue position in the multiple sequence alignment and the localSPAD is a measure of the divergence of suboptimal alignments at a local position, expressed by Eq. (13.1). The predicted Cα distance using the regression (Eq. (13.7)) showed the correlation coefficient of 0.51 to the actual Cα distance. Although this correlation coefficient value is not very large, it does improve that of the predicted global RMSD alone (0.44).

## 13.5 Summary

In this chapter, we introduced methods for quality assessment of protein models, which predict real-value of estimated errors. Quality assessment of protein structure models has been recently studied increasingly in the context of model selection (reranking) from a pool of prediction models, which is an important post-processing step for ab initio structure prediction. What we addressed here are methods that are closely related but aim at a different purpose – methods for real-value error estimation of single protein structure model, which is crucial information for applying structure models for practical biological purposes.

Estimating real-value error is probably more difficult than reranking models, since most of metrics, especially knowledge-based or physical potentials, do not have clear rationale to be able to indicate the degree of real-value global/local error of models. Probably only exceptions are sequence alignment-based scores, including the classical sequence identity between the target and template, which are empirically known to have a significant correlation to the RMSD of structures (Chothia and Lesk 1986). Here, we overviewed four methods for real-value quality assessment. We can learn the followings from these methods: First, as expected, the sequence alignment-based scores are useful for predicting real-value errors (Tondel, TVSMod, SubAqua). Moreover, rather than the simple sequence identity between a target and a template, metrics which evaluate the significance of an alignment in comparison with alternative alignments (the PRSS *Z*-score, the SPAD score) correlate better to the RMSD. Second, structure-based terms, in general, particularly

those which examine atomic-detailed level structures, do not have strong correlation to the real-value errors by itself, but combinations of terms can improve the correlation (ProQ, TVSMod, SubAqua). Third, to make structure-based terms correlate better to the global RMSD, Tondel and Eramian et al. (TVSMod) presented a very interesting idea of constructing a query-dependent dataset for training parameters of the predicting algorithm (Tondel used regression analysis and TVSMod is based on SVM). Obviously values of knowledge/physics-based potentials computed for different proteins cannot be directly compared since the systems are different. Thus, in order to use such potential values for predicting real-value quality, some pre-processing (e.g., normalizations) is needed. The idea of the query-dependent dataset is to only use models with known quality that use the same templates as the query model in training prediction algorithms, so that the potential values are as well correlated as what would be expected for decoys of the same protein. Fourth, predicted global RMSD of a model is a useful variable for predicting local quality at each residue position (SubAqua). Thus, predicting quality in multiple steps, from global to local or coarse-grained to finer-grained, seems to be a valid strategy.

The tertiary structure of proteins provides crucial information for elucidating function of proteins and its mechanism. Computational structure models are expected to serve to biology research in the same way, but it is only possible when errors of the model well estimated. By knowing the absolute value and location of the error in a model, it can be effectively used for practical purposes, which depend on the estimated error range. Thus, real-value error estimation is a key for bridging structure prediction to practical application, and thereby capitalizes tremendous efforts paid in the past years for developing protein structure prediction methods.

**Acknowledgments** This work was supported in part by grants from the National Institute of General Medical Sciences of the National Institutes of Health (R01GM075004 and U24GM077905) and from National Science Foundation (IIS0915801, DMS0800568 and EF0850009).

## References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ Jr, Stoddard BL et al (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 441:656–659

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

Betancourt MR (2008) Knowledge-based potential for the polypeptide backbone. J Phys Chem B 112 5058–5069

Betancourt MR, Skolnick J (2001) Finding the needle in a haystack: educing native folds from ambiguous ab initio protein structure predictions. J Comput Chem 22:339–353

Bhattacharya A, Wunderlich Z, Monleon D, Tejero R, Montelione GT (2008) Assessing model accuracy using the homology modeling automatically software. Proteins 70:105–118

Chen H, Kihara D (2008) Estimating quality of template-based protein models by alignment stability. Proteins 71:1255–1274

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

Cline M, Hughey R, Karplus K (2002) Predicting reliable regions in protein sequence alignments. Bioinformatics 18:306–314
Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2:1511–1519
Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A (2001) A study of quality measures for protein threading models. BMC Bioinform 2:5
Davis IW, Murray LW, Richardson JS, Richardson DC (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res 32:W615–W619
Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277:396–404
Eramian D, Eswar N, Shen MY, Sali A (2008) How well can the accuracy of comparative protein structure models be predicted? Protein Sci 17:1881–1893
Eswar N, Eramian D, Webb B, Shen MY, Sali A (2008) Protein structure modeling with MODELLER. Methods Mol Biol 426:145–159
Feig M, Brooks CL III (2002) Evaluating CASP4 predictions with physical energy functions. Proteins 49:232–245
Fogolari F, Tosatto SC (2005) Application of MM/PBSA colony free energy to loop decoy discrimination: toward correlation between energy and root mean square deviation. Protein Sci 14:889–901
Gregoret LM, Cohen FE (1991) Protein folding. Effect of packing density on chain conformation. J Mol Biol 219:109–122
Hawkins T, Kihara D (2007) Function prediction of uncharacterized proteins. J Bioinform Comput Biol 5:1–30
Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K et al (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 216: 167–180
Holm L, Sander C (1992) Evaluation of protein models by atomic solvation preference. J Mol Biol 225:93–105
Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381:272
Jaroszewski L, Li W, Godzik A (2002) In search for more accurate alignments in the twilight zone. Protein Sci 11:1702–1713
Jauch R, Yeo HC, Kolatkar PR, Clarke ND (2007) Assessment of CASP7 structure predictions for template free targets. Proteins (S 8) 69:57–67
John B, Sali A (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. Nuc Acid Res 31: 3982–3992.
Kihara D, Chen H, Yang YD (2009) Quality assessment of computational protein models. Curr Protein Pept Sci 10:216–228
Kihara D, Skolnick J (2004) Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. Proteins 55:464–473
Kmiecik S, Gront D, Kolinski A (2007) Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field. BMC Struct Biol 7:43
Koike R, Kinoshita K, Kidera A (2004) Probabilistic description of protein alignments for sequences and structures. Proteins 56:157–166
Kolinski A, Bujnicki JM (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. Proteins (S 7) 61:84–90
Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins (S 8) 69:38–56
Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) Procheck – A program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291
Lee M, Jeong CS, Kim D (2007) Predicting and improving the protein sequence alignment quality by support vector regression. BMC Bioinform 8:471

Lee MR, Tsai J, Baker D, Kollman PA (2001) Molecular dynamics in the endgame of protein structure prediction. J Mol Biol 313:417–430
Lindahl E, Elofsson A (2000) Identification of related proteins on family superfamily and fold level. J Mol Biol 295:613–25
Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356:83–85
McConkey BJ, Sobolev V, Edelman M (2003) Discrimination of native protein structures using atom–atom contact scoring. Proc Natl Acad Sci USA 100:3215–3220
Melo F, Feytmans E (1997) Novel knowledge-based mean force potential at atomic level. J Mol Biol 267:207–222
Melo F, Sali A (2007) Fold assessment for comparative protein structure modeling. Protein Sci 16:2412–2426
Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Sci 11:430–448
Mevissen HT, Vingron M (1996) Quantifying the local reliability of a sequence alignment. Protein Eng 9:127–132
Miyazawa S (1995) A reliable sequence alignment method based on probabilities of residue correspondences. Protein Eng 8:999–1009
Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448
Pettitt CS, McGuffin LJ, Jones DT (2005) Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 21:3509–3515
Pontius J, Richelle J, Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 264:121–136
Saqi MA, Sternberg MJ (1991) A simple method to generate non-trivial alternate alignments of protein sequences. J Mol Biol 219:727–732
Schlosshauer M, Ohlsson M (2002) A novel approach to local reliability of sequence alignments. Bioinformatics 18:847–854
Shen M, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524
Shortle D, Simons KT, Baker D (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 95:11158–11162
Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16:776–785
Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17: 355–362
Skolnick J, Fetrow JS, Kolinski A (2000) Structural genomics and its importance for gene function analysis. Nat Biotechnol 18:283–287
Skowronek KJ, Kosinski J, Bujnicki JM (2006) Theoretical model of restriction endonuclease *Hpa*I in complex with DNA predicted by fold recognition and validated by site-directed mutagenesis. Proteins 63:1059–1068
Tondel K (2004) Prediction of homology model quality with multivariate regression. J Chem Inf Comput Sci 44:1540–1551
Tosatto SC, Battistutta R (2007) TAP score: torsion angle propensity normalization applied to local protein structure evaluation. BMC Bioinform 8:155
Tress ML, Jones D, Valencia A (2003) Predicting reliable regions in protein alignments from sequence profiles. J Mol Biol 330:705–718
Vakser IA (1996) Low-resolution docking: prediction of complexes for underdetermined structures. Biopolymers 39:455–464
Vingron M (1996) Near-optimal sequence alignment. Curr Opin Struct Biol 6:346–352
Vingron M, Argos P (1990) Determination of reliable regions in protein sequence alignments. Protein Eng 3:565–569

Vorobjev YN, Hermans J (2001) Free energies of protein decoys provide insight into determinants of protein stability. Protein Sci 10:2498–2506

Wallner B, Elofsson A (2006) Identification of correct regions in protein models using structural alignment and consensus information. Protein Sci 15:900–913

Wang G, Dunbrack RL Jr (2004) Scoring profile-to-profile sequence alignments. Protein Sci 13:1612–1626

Wells GA, Birkholtz LM, Joubert F, Walter RD, Louw AI (2006) Novel properties of malarial *S*-adenosylmethionine decarboxylase as revealed by structural modeling. J Mol Graph Model 24:307–318

Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence structure and function through traditional and probabilistic scores. J Mol Biol 297:233–249

Wojciechowski M, Skolnick J (2002) Docking of small ligands to low-resolution and theoretically predicted receptor structures. J Comput Chem 23:189–197

Wroblewska L, Jagielska A, Skolnick J (2008) Development of a physics-based force field for the scoring and refinement of protein models. Biophys J 94:3227–3240

Wroblewska L, Skolnick J (2007) Can a physics-based all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. J Comput Chem 28:2059–2066

Yang YD, Spratt P, Chen H, Park C, Kihara D (2010) Sub-AQUA: real-value quality assessment of protein structure models. Protein Eng Des Sel 23:617–32

Yu L, Smith TF (1999) Positional statistical significance in sequence alignment. J Comput Biol 6:253–259

Zhang MQ, Marr TG (1995) Alignment of molecular sequences seen as random path analysis. J Theor Biol 174:119–129

Zhang Z, Berman P, Wiehe T, Miller W (1999) Post-processing long pairwise alignments. Bioinformatics 15:1012–1019