PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Evaluating the absolute quality of a single protein model using structural features and support vector machines

Zheng Wang, Allison N. Tegge, and Jianlin Cheng*

Computer Science Department, Informatics Institute, University of Missouri, Columbia, Missouri

ABSTRACT

Knowing the quality of a protein structure model is important for its appropriate usage. We developed a model evaluation method to assess the absolute quality of a single protein model using only structural features with support vector machine regression. The method assigns an absolute quantitative score (i.e. GDT-TS) to a model by comparing its secondary structure, relative solvent accessibility, contact map, and beta sheet structure with their counterparts predicted from its primary sequence. We trained and tested the method on the CASP6 dataset using cross-validation. The correlation between predicted and true scores is 0.82. On the independent CASP7 dataset, the correlation averaged over 95 protein targets is 0.76; the average correlation for template-based and *ab initio* targets is 0.82 and 0.50, respectively. Furthermore, the predicted absolute quality scores can be used to rank models effectively. The average difference (or loss) between the scores of the top-ranked models and the best models is 5.70 on the CASP7 targets. This method performs favorably when compared with the other methods used on the same dataset. Moreover, the predicted absolute quality scores are comparable across models for different proteins. These features make the method a valuable tool for model quality assurance and ranking.

Proteins 2009; 75:638–647.

Key words: protein structure prediction; protein model evaluation; protein model quality assurance; machine learning; support vector machine.

INTRODUCTION

Predicting protein structure from a sequence is one of the most important problems in structural bioinformatics. An important task in both structure prediction and application is to evaluate the quality of a structure model.[1] Accurate model quality evaluation is useful for ranking models, refining models, and using models. Model evaluation tools or model quality assurance programs (MQAP) are receiving considerable attention in the field of protein structure prediction. Model evaluation methods were evaluated in the Seventh Edition of Critical Assessment of Techniques for Protein Structure Prediction (CASP7),[2,3] 2006.

There are two kinds of model evaluation methods: local and global methods.[3] The former is to predict the quality of local regions such as the distance between the position of a residue in a protein model and its native structure. The latter is to predict an overall score of a model. Some methods, such as Pcons,[4] can predict both local and global quality. Here we focus on global quality evaluation methods.

A number of global model evaluation methods were developed in the last 20 years. Traditional model evaluation methods use energy function and statistical potentials derived from known protein structures to assess models.[1,5–23] Most of these methods are relative scoring methods designed to discriminate near-native structures from decoys. These methods are different from methods which produce an absolute score that measures the similarity between a model and the native (or experimental) structure.[24] Recently, machine learning methods[25–29] use neural networks and support vector machines that are trained on structural models to predict model quality. Assuming a model that is more similar to the rest of models has higher quality, clustering or consensus methods[4,30–33] can score or rank models by comparing a model with all other models associated with the same protein. Pcons,[4] a consensus method, performed best in terms of average correlation between predicted and true quality scores during CASP7.[2] Spicker,[33] a clustering method, clusters a group of models by pairwise structure comparison, and selects the model that is the closest to the centroid of each cluster. To combine the strength of different methods, meta methods[34–36] integrate different model assessment methods to score models.

Grant sponsors: MU Faculty startup Grant, MU Research Board Grant, MU Bioinformatics Consortium.
*Correspondence to: Jianlin Cheng, Computer Science Department, Informatics Institute, University of Missouri, Columbia, MO 65211. E-mail: chengji@missouri.edu
Received 13 May 2008; Revised 26 August 2008; Accepted 12 September 2008
Published online 25 September 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22275

Model evaluation methods can be classified into single-model approaches such as ProQ,[25] ProQ-LG, ProQ-MX, and MODCHECK[1] and multiple-model approaches such as clustering methods based on the number of input models. Single-model approaches[25,36–40] can assign a score to a single model. In contrast, multiple-model approaches, such as clustering and consensus methods, require a large pool of models as inputs. These methods cannot be used to assess the quality of a single model. They may not reliably evaluate the quality of a small number of input models.[36]

In terms of input information, some methods[6,7,25,37,41] only need 3D coordinates as inputs, whereas many model evaluation methods[26,27] need extra information such as sequence alignments and templates. The latter methods cannot be used in many cases when only 3D models are available to end users. Other information such as sequence alignments, alignment scores, and template information is either not available (e.g., *ab initio* models) or not provided in a form required by these kinds of methods. Therefore, developing an accurate structure-based MQAP method is an important task in the field of model quality assurance (MQA). Recently, two structure-based methods (ABIpro-H and Circle-QA) were evaluated in CASP7.[3]

In terms of outputs, most methods, particularly energy-based methods, produce a relative score for model selection instead of an absolute quality score. A relative score can only select or rank models, but does not tell how good a model is, which is critical for using models. Relative scores are often protein specific and not comparable between models associated with different proteins. To date, very few methods aim to evaluate the absolute quality of a model, which is the similarity between a model and the corresponding native structure.

With wider and wider applications of protein models in the genomic era, it is critical to develop methods that can evaluate the absolute quality of a single model from only its 3D coordinates and primary sequence according to the last CASP7 assessment.[3] However, only a few such methods (e.g., Undertaker[41]) are available.

Here we developed a machine learning model evaluation method (ModelEvaluator) to predict the quantitative *absolute* quality score of a *single* model using *structural features* extracted from its 3D coordinates and predicted from its primary sequence. The method compares structural features generated from a 3D model with those predicted from its primary sequence by 1D and 2D structural feature predictors as in Ref. 42. These features include secondary structure, solvent accessibility, contact map, and beta-sheet topology. This comparison method results in a number of fitness scores used as input features for a support vector machine (SVM) to evaluate the quality of a model. This method shows good performance for both correlation between predicted and true quality scores and model ranking when applied to CASP6 and CASP7 datasets. The high performance makes it a valuable tool to control the quality of model generation and to guide model usage.

METHODS

Data sets

We used CASP6 and CASP7 protein models in our experiments. The structure models associated with 64 CASP6 targets were used to train and validate ModelEvaluator. The models were predicted by three structure prediction methods: Sparks3,[19,43–45] Robetta,[46–49] and FOLDpro.[42] Both Sparks3 and FOLDpro are template-based method. Robetta includes both template-based and *ab initio* modeling. We chose models generated by the three methods because they used different structure prediction techniques. Theoretically, our machine learning method can be trained on any structure models. In fact, we trained and tested our methods on all CASP6 models generated by tens of different structure predictors, and got similar results (data not shown).

The difficulty of the CASP6 targets ranges from easy comparative modeling to hard *de novo* folds. We downloaded 64 of the top models (one model per target) predicted by Sparks3 and Robetta from the CASP web site (http://predictioncenter.org/download_area/). We used FOLDpro to generate five models for each CASP6 target using the template library that only includes proteins prior to the CASP6 experiment. We compared models with the corresponding experimental structures using LGA's[50] GDT-TS score, which is a standard CASP evaluation measure. A GDT-TS score is in the range [0,100]. We normalized GDT-TS scores into the range [0,1] to make machine learning more efficient. The CASP6 data was used to train and validate ModelEvaluator in cross-validation.

We blindly evaluated ModelEvaluator on CASP7 models. We downloaded all the models and their associated GDT-TS scores for 95 valid CASP7 protein targets. We benchmarked our method on the models of each target, each predictor, and all targets using several complementary measures.

Feature extraction

The goal of the experiment is to train SVM to learn a function to accurately map input features extracted from a model to the GDT-TS score of the model, which is a regression problem. The key is to extract a set of informative features.

We generated a set of input features by comparing structural features extracted from 3D coordinates of a model against those predicted from its primary sequence by SCRATCH[51–55] when 1D and 2D structural feature predictors are used, similarly to a fold recognition

method.[42] SCRATCH is a pure *ab initio* structure feature predictor based on neural networks. It was trained on protein structures extracted from the PDB prior to CASP6 (2004).

The predicted 1D and 2D structural features include secondary structure (3-class: alpha helix, beta sheet, and loop), relative solvent accessibility (2-class: exposed or buried at 25% threshold), contact probability map at 8 Å and 12 Å, and probability map of beta-strand residue pairings. The 1D and 2D predicted structural features are compared with the structural features directly extracted from a model to generate a set of input features for SVMs as follows.

1D features

The predicted secondary structure (SS) and relative solvent accessibility (RSA) of each residue was compared with those of the model parsed by DSSP.[56] The fractions of identical matches for both SS (as in Refs. 57 and 58) and RSA were used as two features. The two SS and RSA composition vectors (% helix, % strand, % coil, % exposed, and % buried) were compared and transformed into four similarity scores by cosine, correlation, Gaussian kernel, and dot product, respectively. So the 1D feature subset has 6 features in total.

Because the prediction accuracy of secondary structure and relative solvent accessibility is above 76%,[52,59–61] 1D features extracted from a good 3D model should generally match the predicted ones better than a bad model. So 1D features are useful indicators of model quality.

2D features

For residue pairs in a model which have sequence separation $\geq$6 residues and are in contact at an 8 Å threshold (resp. 12 Å), we computed the average contact probability of their counterparts in the predicted 8 Å (resp. 12 Å) contact probability map.[54] Similarly, for each paired beta-strand residues in the model, we computed the average pairing probability of their counterparts in the predicted beta-strand paring probability map.[55] The three probabilities measuring the compatibility between the model and its predicted 2D features are used as input features. The underlying assumption is that general (or beta-strand) residue contacts in a good model should have a higher average contact probability in the predicted contact map (or beta-strand pairing map) than a bad model.[62]

Moreover, we computed the contact order (the sum of sequence separation of contacts) and the contact number (the number of contacts) for each residue from a 3D model and the predicted contact map. The information is easy to derive for the model because the coordinates of each residue are known. For a contact map, we let the contact order for residue i to be $\sum_{|i-j|>=6} C_{ij}|i-j|$, where C_{ij} is the predicted contact probability for residues i and j. The contact number for residue i in the query is defined as the sum of the contact probabilities $\sum_{|i-j|>=6} C_{ij}$. The contact order and contact number vectors generated from the model and contact map were not used directly as features. Instead, they were compared and transformed into pairwise similarity scores using the cosine and correlation functions. For both 8 Å and 12 Å contact maps, 8 pairwise features of contact order and contact number were extracted. So the 2D feature subset has 11 features in total. The entire 1D and 2D structural feature subset has 17 features in total.

Support vector machine regression

Given a set of data points (input feature vectors) S and the GDT-TS scores of the models associated with the data points, SVMs[63–66] learn a regression function $f(x)$ in the form of

$$f(x) = \sum_{x_i \in S} (\alpha_i - \alpha_i^*) K(x, x_i) + b$$

The value computed by $f(x)$ is the estimate of the GDT-TS score of a model associated with an input feature vector x. α_i or α_i^* are nonnegative weights assigned to the training data point x_i during training by minimizing a quadratic objective function. b is the bias term. K is the kernel function, which can be viewed as a function for computing the similarity between two data points. Thus, the function $f(x)$ can be viewed as a weighted linear combination of similarities between training data points x_i and target data point x. Only data points with positive weight α or α^* in the training dataset affect the final solution—they are called support vectors.

We used SVM-light (http://svmlight.joachims.org)[67,68] to train and test our methods. We experimented with several common kernels including linear kernel, Gaussian radial basis kernel (RBF), polynomial kernel, and sigmoid kernel. The RBF ($e^{-\gamma\|x-y\|^2}$ or $e^{-\frac{\|x-y\|^2}{\sigma^2}}$) worked best. With the RBF kernel, $f(x)$ is actually a weighted sum of Gaussians centered on the support vectors. Almost any separating boundary or regression function can be obtained with this kernel,[69] thus it is important to tune the parameters of SVMs to achieve good generalization performance and to avoid overfitting.

We adjusted three critical parameters: width γ of the RBF kernel, regularization parameter C, and regression tube width ϵ. γ is the inverse of the variance (σ^2) of the RBF and controls how peaked the Gaussians are centered on the support vectors. The bigger γ is, the more peaked are the Gaussians, and the more complex are the resulting decision boundaries.[69] C is the maximum value that weights (α or α^*) can have. C controls the trade-off between training errors and the smoothness of $f(x)$.[63–66] A larger C corresponds to less training errors and a more complex (less smooth) function $f(x)$ which can overfit

training data. ϵ controls the sensitivity of the cost associated with training errors ($f(x)$–the real GDT-TS score). The training error within range $[-\epsilon, +\epsilon]$ does not affect the regression function.

The three parameters were optimized on the training data during cross-validation.

Training, test, and evaluation

We trained and evaluated SVM methods on the structural models generated for 64 CASP6 targets by three predictors: Robetta,[46–49] Sparks3,[19,43–45] and FOLDpro.[42]

We split the models of the CASP6 dataset into 64 folds corresponding to 64 target proteins. We used 63 folds as training dataset and the remaining one as test dataset. We repeated the process 64 times to test all models once in a 64-fold cross-validation. During each round of cross-validation, we optimized the parameters on training data. We chose a set of parameters with the highest accuracy to build a SVM model; and blindly tested it on the testing dataset. The predicted GDT-TS scores for all models were collected. We compared the predicted scores with the real GDT-TS scores to compute Pearson correlation and root mean square errors (RMSE).

The CASP7 models were used as an independent test set. As in Ref. 22, the models converted from the predicted alignments were not used. We used the SVM trained on CASP6 models to predict the GDT-TS scores of CASP7 models. The predicted GDT-TS scores were compared with true GDT-TS scores. We analyzed results on all the models, models associated with each target, and models produced by each predictor. The performance was evaluated by three complementary measures: Pearson correlation, root mean square error, and loss—the difference between the GDT-TS scores of the top ranked model and the best model for a target. We compared ModelEvaluator with ProQ and Circle-QA because they share some similar properties. ProQ uses neural networks to predict an absolute RMSD-based quality score, but it uses both structural features and other information. Circle-QA is a method that uses only structural features in CASP7,[3] but it requires a group of models as inputs.

RESULTS AND DISCUSSIONS

We first report the cross-validation results on CASP6 models.[70] We then evaluate ModelEvaluator on independent CASP7 models.

Cross-validation on CASP6 models

Using 64-fold cross-validation, we computed the correlation and root mean square error (RMSE) between predicted and real GDT-TS scores on CASP6 models generated by FOLDpro, Sparks3, and Robetta. The correlation and RMSE are 0.82 and 0.127, respectively. If only 2D features are used as inputs, the correlation and RMSE are 0.65 and 0.170, respectively, indicating that the combination of 1D and 2D features improves performance.

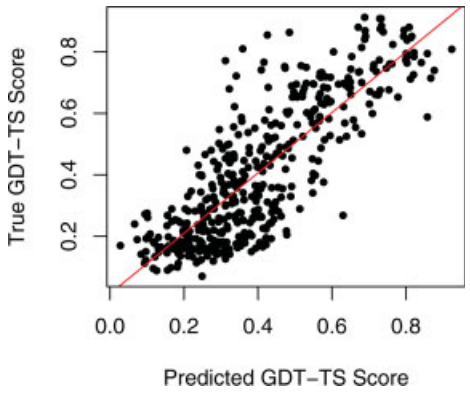

Figure 1

Predicted GDT-TS scores versus real GDT-TS scores on CASP6 models using cross-validation. The red line is the linear regression line between predicted scores and real scores. The correlation and root mean square error between predicted and real scores is 0.82 and 0.127, respectively. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Figure 1 is the plot of the predicted GDT-TS scores versus the true ones. It shows that the predicted scores have a strong linear relation with the true GDT-TS scores. The strong correlation and reasonably low RMSE (0.127) indicate that the method can accurately measure model quality.

Target-specific evaluation on CASP7 models using Pearson correlation, root mean square error, and loss

We evaluated ModelEvaluator by three complementary measures (correlation, root mean square error, and loss) for each CASP7 target. Correlation is the Pearson correlation between predicted GDT-TS scores and true GDT-TS scores of the models associated with a target. RMSE is the square root difference between predicted GDT-TS scores and true GDT-TS scores. Loss is the difference between the real GDT-TS score of the top model ranked by predicted GDT-TS scores and the highest real GDT-TS score of the best model for a target. The loss assesses the ability of a method to rank models. Ideally, loss is zero if a method always produces a score to rank the best model at the top. The Pearson correlation and the loss measure two complementary properties of a model evaluation method. A high correlation does not necessarily mean a low loss.

Table I reports the average correlation, RMSE, and loss for each CASP7 target. The highest correlation is 0.98 for target T0305. The loss and RMSE for T0305 are 0 and

Table I
The Correlation, RMSE, and Loss for Each CASP7 Target and the Average Values on All 95 Targets

Target ID	Correlation	Loss	RMSE
T0283	0.42	12.50	0.12
T0284	0.89	5.67	0.10
T0285	0.48	9.85	0.12
T0286	0.83	1.73	0.09
T0287	0.46	6.52	0.13
T0288	0.91	9.07	0.16
T0289	0.87	3.09	0.06
T0290	0.95	0.87	0.11
T0291	0.88	9.79	0.14
T0292	0.91	5.30	0.09
T0293	0.85	5.33	0.08
T0295	0.89	14.64	0.13
T0296	0.41	5.86	0.22
T0297	0.88	4.62	0.09
T0298	0.93	1.12	0.08
T0299	0.57	9.16	0.16
T0300	0.34	0.00	0.09
T0301	0.83	1.98	0.20
T0302	0.85	1.52	0.15
T0303	0.93	3.13	0.07
T0304	0.65	7.71	0.13
T0305	0.98	0.00	0.11
T0306	0.30	27.37	0.19
T0307	0.45	0.38	0.11
T0308	0.93	0.76	0.08
T0309	0.18	9.68	0.13
T0311	0.60	4.02	0.15
T0312	0.81	0.00	0.14
T0313	0.95	16.77	0.08
T0314	0.48	9.20	0.15
T0315	0.95	5.63	0.08
T0316	0.75	1.51	0.08
T0317	0.95	2.06	0.07
T0318	0.96	2.76	0.10
T0319	0.49	8.52	0.09
T0320	0.75	3.14	0.12
T0321	0.59	2.97	0.12
T0322	0.94	0.70	0.06
T0323	0.75	0.00	0.10
T0324	0.92	7.37	0.08
T0325	0.81	1.05	0.14
T0326	0.97	1.73	0.10
T0327	0.69	6.16	0.21
T0328	0.97	1.79	0.13
T0329	0.90	6.38	0.08
T0330	0.90	2.22	0.08
T0331	0.83	3.60	0.12
T0332	0.91	8.81	0.09
T0333	0.84	6.47	0.14
T0334	0.98	0.90	0.13
T0335	0.55	2.98	0.21
T0338	0.68	7.13	0.10
T0339	0.93	1.05	0.09
T0340	0.96	0.83	0.07
T0341	0.96	0.00	0.08
T0342	0.90	0.74	0.09
T0345	0.94	1.08	0.16
T0346	0.98	1.02	0.08
T0347	0.70	3.57	0.19
T0348	0.72	14.76	0.09
T0349	0.57	22.33	0.16
T0350	0.62	10.09	0.11
T0351	0.18	25.89	0.11
T0353	0.68	10.59	0.11
T0354	0.50	5.12	0.13
T0356	0.09	12.90	0.22
T0357	0.74	19.51	0.12
T0358	0.60	14.02	0.10
T0359	0.92	10.22	0.08
T0360	0.27	16.16	0.19
T0361	0.52	5.42	0.13
T0362	0.96	1.91	0.10
T0363	0.88	3.75	0.07
T0364	0.97	2.29	0.06
T0365	0.85	10.75	0.10
T0366	0.92	5.36	0.15
T0367	0.90	1.40	0.20
T0368	0.66	12.58	0.16
T0369	0.60	0.00	0.17
T0370	0.85	0.97	0.13
T0371	0.95	1.41	0.14
T0372	0.86	3.19	0.13
T0373	0.78	4.82	0.12
T0374	0.91	11.09	0.08
T0375	0.95	4.22	0.08
T0376	0.98	0.24	0.05
T0378	0.93	3.70	0.09
T0379	0.87	1.59	0.08
T0380	0.94	1.94	0.06
T0381	0.90	7.80	0.11
T0382	0.35	4.55	0.11
T0383	0.86	2.80	0.12
T0384	0.97	0.00	0.10
T0385	0.72	3.17	0.26
T0386	0.62	5.07	0.12
Average	**0.76**	**5.70**	**0.12**

0.09, respectively. The lowest correlation is 0.09 for target T0356—a hard *ab initio* target. The loss and RMSE for T0356 are 12.90 and 0.22, respectively. The average correlation, loss, and RMSE on all targets are 0.76, 5.70, and 0.12, respectively. ModelEvaluator works well on most targets, indicating that it can generalize well to predict the quality of models of different proteins.

Similarly as in Ref. 3, a correlation score is related to the difficulty of a target, or more precisely the quality of models. Figure 2 plots correlation scores against median real GDT-TS scores of models of each CASP7 target. The lower the median is, the poorer is the quality of models. The plot shows that correlation tends to increase as the median GDT-TS score increases. This indicates that it is easier to evaluate the quality of models of easy targets than models of hard targets because the model quality of the easy targets is generally better than hard ones. The variation of correlation decreases as median GDT-TS scores (i.e. quality) of models increase. Roughly speaking, the predicted quality score of a model with real GDT-TS score below 45 is not reliable.

We calculated the accuracy on template-based models (TBM) and free-modeling (FM, i.e. *ab initio*) models as in Ref. 3. A model is considered a FM model if the target

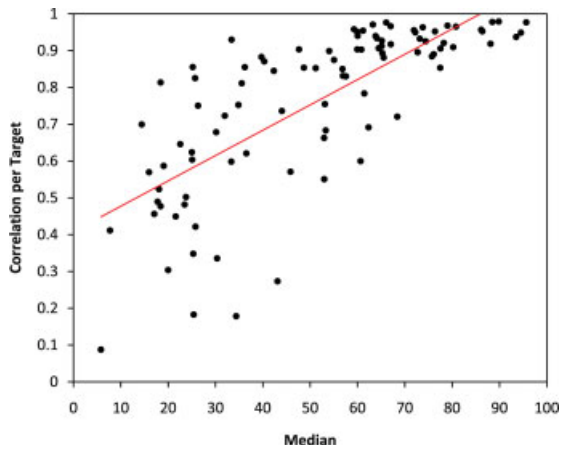

Figure 2

Correlation against median true GDT-TS score per target. A median true GDT-TS score is an indicator of the difficulty of a target.[3,4] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

protein contains at least one *ab initio* domain. According to this criteria, the models of 18 targets (T0287, T0296, T0300, T0304, T0307, T0309, T0314, T0316, T0319, T0321, T0347, T0348, T0350, T0353, T0356, T0361, T0382, T0386) are considered FM models. The models of the remaining 77 targets are considered TBM models. The quality of TBM models is generally better than FM models. The average correlation on TBM models is 0.82, substantially higher than 0.50 on FM models. However, the average loss on TBM models is 5.48, only slightly better than 6.63 on FM models. This suggests that ModelEvaluator can pick out good models from a set of generally low-quality

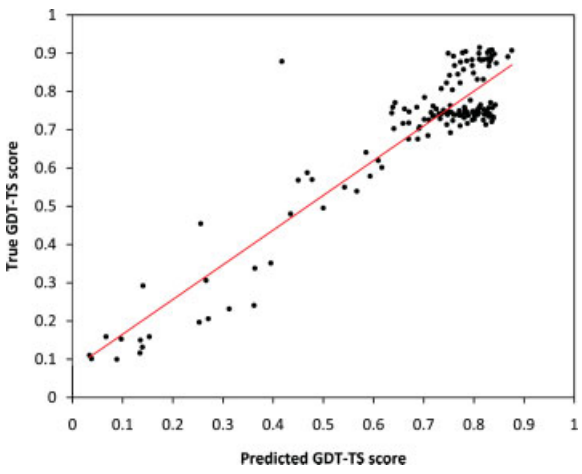

Figure 3

The predicted GDT-TS scores versus the true GDT-TS score of the models of an easy target T0308. The correlation and RMSE are 0.93 and 0.08, respectively. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

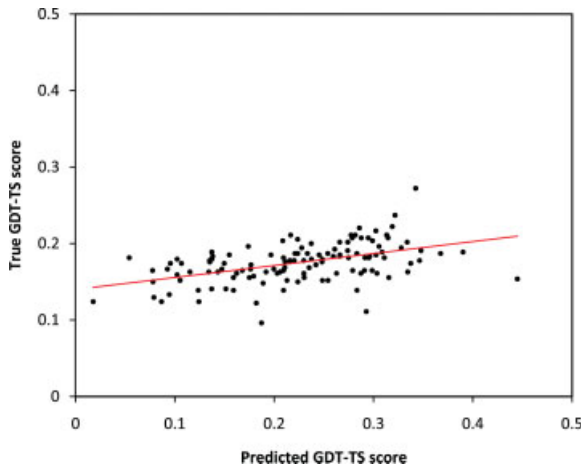

Figure 4

The predicted GDT-TS scores versus the true GDT-TS scores of the models of a hard target T0319. The correlation and RMSE are 0.49 and 0.09, respectively. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

FM models even though the overall correlation is not high. Specifically, Figures 3 and 4 plot real GDT-TS scores against predicted GDT-TS scores of an easy TBM target (T0308) and a hard FM target (T0319), respectively.

Because correlation, loss, and RMSE are complementary, it is interesting to investigate their relationships. We plotted them against each other for each of 95 targets in Figures 5–7, respectively. Figure 5 (resp. 6) shows that correlation and loss (resp. correlation and RMSE) only have weak overall correlation. However, when correlation is close to 1, loss dramatically decreases, and gets close to

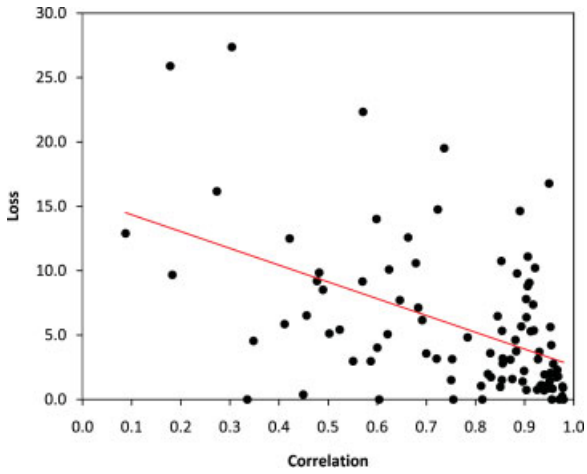

Figure 5

Correlation versus loss of 95 CASP7 targets. The red line is the linear regression line. It is interesting to note that the loss for a few targets is close to, or equal to, zero even though the correlation is low. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

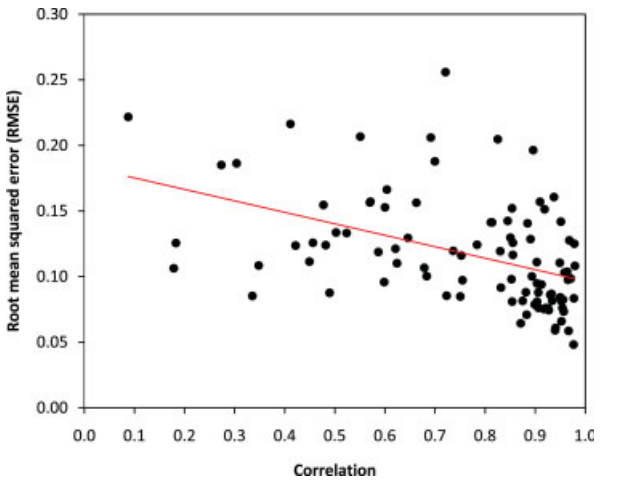

Figure 6

Correlation versus RMSE of 95 CASP7 targets. The red line is the linear regression line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

0. Figure 7 shows that RMSE and loss do not seem to have an obvious correlation.

Predictor-specific evaluation on CASP7 Models using Pearson correlation and root mean square error

We evaluated if ModelEvaluator can generalize well to models predicted by different predictors. Table II reports correlation and RMSE on the CASP7 models predicted by 51 server predictors. Several servers that predicted a small number of models or generated only alignments were not included. The results show that ModelEvaluator performs well on a variety of server predictors. The average correlation and RMSE on all the predictors are 0.82 and 0.13, respectively. Thus, ModelEvalutor can be used as an independent quality assurance component of any protein 3D structure predictor.

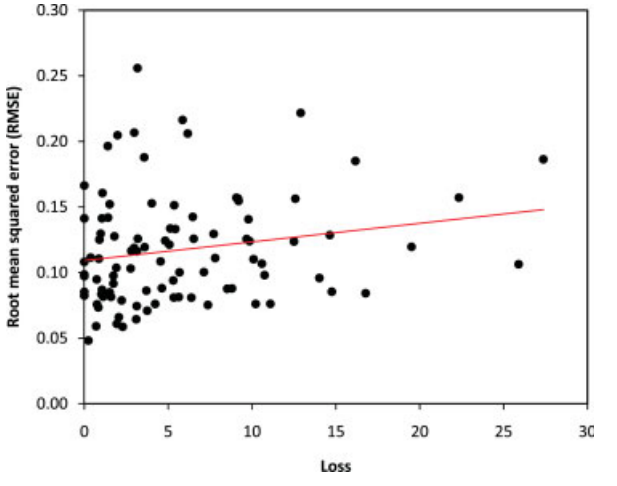

Figure 7

Loss versus RMSE of 95 CASP7 targets. The red line is the linear regression line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table II

The Correlation and RMSE for the Models Predicted by Each Server Predictor Participating in CASP7

Predictor name	Correlation	RMSE
Bilab-ENABLE	0.89	0.12
Zhang-Server	0.78	0.14
ABIpro	0.45	0.20
SP4	0.83	0.13
FAMS	0.84	0.13
CIRCLE	0.85	0.13
FAMSD	0.86	0.12
MetaTasser	0.76	0.14
ROBETTA	0.80	0.14
Pcons6	0.83	0.12
Pmodeller6	0.80	0.13
SAM-T06-server	0.79	0.14
RAPTOR-ACE	0.85	0.12
SP3	0.85	0.12
FUNCTION	0.90	0.11
RAPTOR	0.82	0.13
SPARKS2	0.86	0.12
RAPTORESS	0.82	0.13
BayesHH	0.90	0.11
HHpred1	0.89	0.11
keasar-server	0.88	0.11
karypis.srv.2	0.91	0.11
FUGMOD	0.85	0.12
HHpred3	0.88	0.11
POMYSL	0.57	0.08
Ma-OPUS-server	0.87	0.12
PROTINFO	0.87	0.11
CaspIta-FOX	0.91	0.10
karypis.srv.4	0.24	0.08
FOLDpro	0.91	0.11
Beautshot	0.90	0.12
3Dpro	0.89	0.12
3D-JIGSAW	0.92	0.10
HHpred2	0.88	0.11
nFOLD	0.90	0.17
PROTINFO-AB	0.86	0.12
Shub	0.87	0.10
Frankenstein	0.92	0.09
Karypis.srv	0.94	0.08
ROKKY	0.83	0.10
FPSOLVER-SERVER	0.26	0.08
GeneSilicoMetaServer	0.90	0.12
NN-PUT-lab	0.85	0.15
3D-JIGSAW-POPULUS	0.85	0.10
3D-JIGSAW-RECOM	0.90	0.12
LOOPP	0.88	0.14
Beautshotbase	0.94	0.12
UNI-EID-expm	0.59	0.32
Phyre-2	0.90	0.11
Huber-Torda-Server	0.85	0.28
Ma-OPUS-server2	0.84	0.13
Average	**0.82**	**0.13**

Specifically, Figures 8–10 are the plots between the predicted and real GDT-TS scores of all models generated by three predictors: HHpred2,[71] ROBETTA, and FOLD-

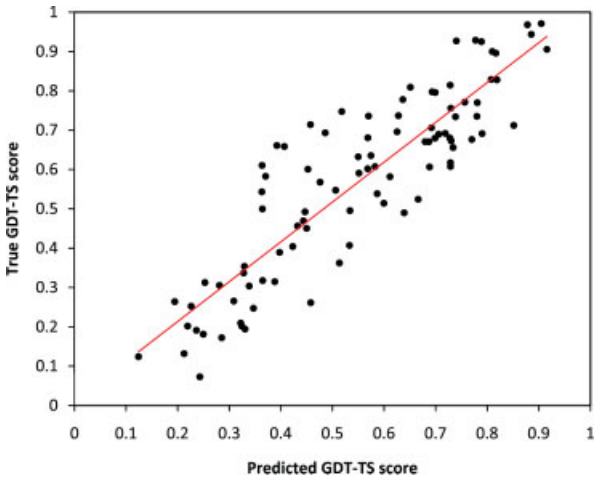

Figure 8

Predicted GDT-TS score versus the true GDT-TS score of CASP7 models generated by HHpred2. The red line is the linear regression line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

pro, respectively. In all three cases, predicted and real GDT-TS scores have a strong correlation.

Overall evaluation on all CASP7 models

It is desirable for a model evaluation method to predict quality scores that are comparable between models of different proteins. Similar scores for the models of two proteins mean similar quality. However, many existing methods, particularly relative scoring methods, produce scores that are not very comparable across different targets.

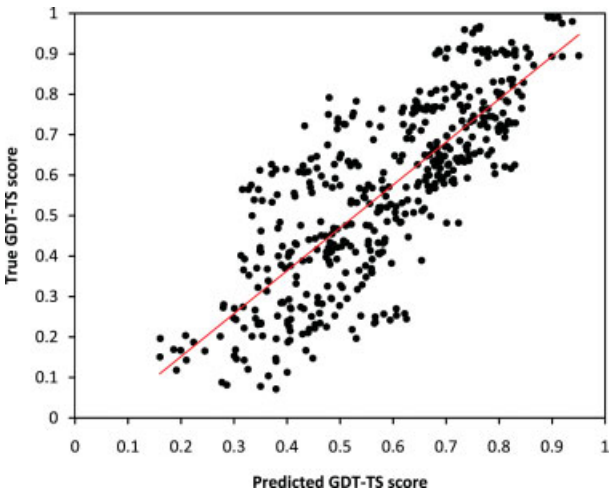

Figure 9

Predicted GDT-TS score versus the true GDT-TS score of CASP7 models generated by ROBETTA. The red line is the linear regression line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

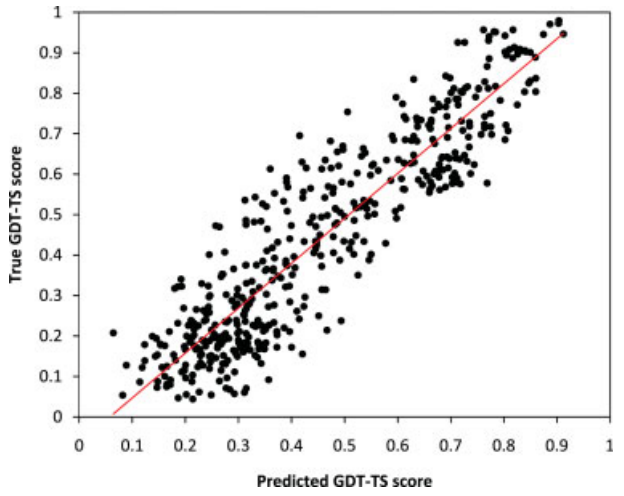

Figure 10

Predicted GDT-TS score versus true GDT-TS score of CASP7 models generated by FOLDpro. The red line is the linear regression line. FOLDpro, a template-based structure prediction method, uses similar 1D and 2D structure features as well as other sequence alignment features in template selection. Because in most cases the dominant features in FOLDpro template selection are alignment features, the secondary structure, relative solvent accessibility, contact map, and beta-sheet topology of a model produced by FOLDpro can be different from those predicted by the SCRATCH. So the correlation between predicted and true GDT-TS scores of FOLDpro is higher than most CASP7 predictors as one may expect, but not the highest (see Table II). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ModelEvalutor is an absolute-scoring method designed to overcome the problem. Evaluated on all the CASP7 models *pooled* together, the correlation and RMSE between predicted and real GDT-TS scores are 0.87 and

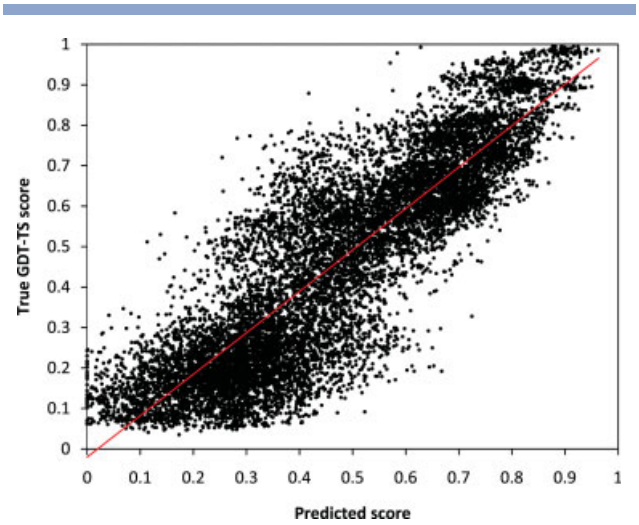

Figure 11

Predicted GDT-TS scores versus true scores of all CASP7 models. The red line is the linear regression line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table III
The Results of Three Model Evaluation Methods on CASP7 Models

Method	Ave corr	Corr (TM)	Corr (FM)	Loss	Loss (TM)	Loss (FM)	Over corr
ModelEvaluator	**0.76**	**0.82**	0.50	**5.70**	**5.48**	**6.63**	**0.87**
Circle-QA	0.75	0.79	**0.57**	6.07	5.83	7.09	0.70
ProQ	0.72	0.76	0.53	9.04	9.12	8.69	0.78

The seven columns represent average correlation on all targets, average correlation on TBM targets, average correlation on FM targets, average loss on all targets, average loss on TBM targets, average loss on FM targets, and overall correlation on all models. Bold font denotes the best results.

0.123, respectively. Figure 11 plots the real GDT-TS scores against the predicted GDT-TS scores on all CASP7 models. The high correlation shows that the GDT-TS scores predicted by ModelEvalutor are comparable among models predicted for different proteins.

Comparison with two other methods on CASP7 models

CASP7 had a comprehensive assessment of a variety of model evaluation methods including meta methods, clustering/consensus methods, energy-based methods, and machine learning methods.[3] All the methods except Circle-QA and ABIpro-H used both structural and other information in CASP7. We did not try to reevaluate these methods. Instead, we only compared ModelEvalutor with two methods: Circle-QA, a structure-feature based method; and ProQ, a machine learning method. As shown in Table III, the average correlation of ModelEvaluator on all targets and TBM targets is higher than Circle-QA and ProQ, whereas its correlation on FM targets is a bit lower; the loss of ModelEvaluator on all, TBM, and FM targets is lower than both Circle-QA and ProQ, indicating it can select good models more effectively. Its overall correlation on all CASP models is higher than Circle-QA and ProQ.

CONCLUSION

We have developed a model evaluation method (Model Evaluator) that can predict the absolute quality score of a single protein model using only structural features. The predicted GDT-TS scores correlate well with real GDT-TS scores on both CASP6 and CASP7 benchmarks. Its performance generalizes well to different proteins and structure predictors. The predicted GDT-TS scores can be used to select and rank models effectively. They are also comparable for models associated with different proteins. These features make ModelEvaluator a unique and useful quality assurance component for any protein structure prediction package and for appropriate model usage.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for helpful comments.

REFERENCES

1. Pettitt C, McGuffin L, Jones D. Improving sequence-based fold recognition by using 3d model quality assessment. Bioinformatics 2005;21:3509–3515.
2. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction – round VII. Proteins 2006;69:3–9.
3. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. Assessment of predictions in the model quality assessment category. Proteins 2007;69(S8):175–183.
4. Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 2007;69:184–193.
5. Sippl M. Calculation of conformational ensembles from potentials of mean force. An approach to the knolwdge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
6. Colovos C, Yeates T. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 1993;2:1511–1519.
7. Sippl M. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
8. Park B, Levitt M. Energy function that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258:367–392.
9. Pontius J, Richelle J, Wodak S. Deviations from standard atomic volumes as a quality measure for protein crystal structures. J Mol Biol 1996;264:121–136.
10. Park B, Huang E, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
11. Melo F, Devos D, Depiereux E, Feytmans E. ANOLEA: a www server to assess protein structures. In: Proceedings International Conference Intelligence Systems Molecular Biology. Menlo Park, CA: AAAI Press; 1997. pp. 187–190.
12. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1999;288:477–487.
13. Gatchell D, Dennis S, Vajda S. Discrimination of near-native protein structures from misfolded models by empirical free energy function. Proteins 2000;41:518–534.
14. Petrey D, Honig B. Free energy determinants of tertiary structure and evaluation of protein models. Protein Sci 2000;9:2181–2191.
15. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins 2000;38:134–148.
16. Vorobjev Y, Hermans J. Free energies of protein decoys provide insight into determinants of protein stability. Protein Sci 2001;10:2498–2506.
17. Dominy B, Brooks C. Identifying native-like protein structures using physics-based potentials. J Comput Chem 2002;23:147–160.
18. Felts A, Gallicchio E, Wallqvist A, Levy R. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the opls all-atom force field and the surface generalized born solvent model. Proteins 2002;48:404–422.
19. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
20. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
21. Wu Y, Lu M, Chen M, Li J, Ma J. OPUS-Ca: a knowledge-based potential function requiring only Ca positions. Protein Sci 2007;16:1449–1463.
22. Zhou H, Skolnick J. Protein model quality assessment prediction by combining fragment comparision and a consensus ca contact potential. Proteins 2007;71:1211–1218.
23. Zhang Y. I-TASSER server for protein 3d structure prediction. BMC Bioinformatics 2008;9:40.
24. Kosinski J, Cymerman I, Feder M, Kurowski M, Sasin J, Bujnicki J. A FRrankenstein's monster approach to comparative modeling:

Merging the finest fragmetns of fold-recognition models and iterative model refinement aided by 3D structure evaluation. Proteins 2003;53:369–379.
25. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003;12:1073–1086.
26. Tress M, Jones D, Valencia A. Predicting reliable regions in protein alignments from sequence profiles. J Mol Biol 2003;330:705–718.
27. von Grotthuss M, Pas J, Wyrwicz L, Ginalski K, Rychlewski L. Application of 3d-jury, GRDB, and Verify 3D in fold recognition. Proteins 2003;53:418–423.
28. Wallner B, Elofsson A. Can correct regions in protein models be identified. Protein Sci 2006;15:900–913.
29. Qiu J, Sheffler W, Baker D, Noble WS. Ranking predicted protein structures with support vector regression. Proteins 2007;71:1175–1182.
30. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
31. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441.
32. McGuffin L. Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics 2007;8:345.
33. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.
34. Wallner B, Fang H, Elofsson A. Automatic consensus-based fold recognition using pcons, proq, and pmodeller. Proteins 2003;53: 534–541.
35. Kajan L, Rychlewski L. Evaluation of 3D-jury on casp7 models. BMC Bioinformatics 2007;8:304.
36. McGuffin LJ. The ModFOLD server for the quality assessment of protein structural models. Bioinformatics 2008;24:586–587.
37. Luthy R, Bowie J, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–85.
38. Laskowski R, Rullmannn J, MacArthur M, Kaptein R, Thornton J. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by nmr. J Biomol NMR 1996;8:477–486.
39. McGuffin L, Bryson K, Jones D. What are the baselines for protein fold recognition? Bioinformatics 2001;17:63–72.
40. Wiederstein M, Sippl M. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 2007;35:W407–410.
41. Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, Barnes B, Soriano M, Hughey R. SAM-T04: what is new in protein structure prediction for CASP6. Proteins 2005;61:135–142.
42. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22:1456–1463.
43. Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. Proteins 2004;54:315–322.
44. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58:321–328.
45. Zhou H, Zhou Y. SPARKS 2 and SP3 servers in CASP6. Proteins 2005;61:152–156.
46. Simons K, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 1997;268:209–225.
47. Chivian D, Kim D, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss C, Bonneau R, Rohl C, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003;53(S6):524–533.
48. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim D, Meiler J, Misura K, Baker D. Free modeling with Rosetta in casp6. Proteins 2005;61:128–134.
49. Chivian D, Kim D, Malmstrom L, Schonbrun J, Rohl C, Baker D. Prediction of CASP6 structures using automated robetta protocols. Proteins 2005;61:157–166.
50. Zemla A. LGA: a method for finding 3d similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
51. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
52. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47:228–235.
53. Pollastri G, Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Bioinformatics 2002;18(Suppl 1):S62–S70. Proceeding of the ISMB 2002 Conference.
54. Cheng J, Randall A, Sweredoski M, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005;33 (web server issue):w72–76.
55. Cheng J, Baldi P. Three-stage prediction of protein beta-sheets by neural networks, alignments, and graph algorithms. Bioinformatics 2005;21(suppl 1):i75–i84.
56. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577–2637.
57. Jones D. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
58. Xu J, Xu Y, Lin G, Kim D, Li M. Protein structure prediction by linear programming. In Pac Symp Biocomput. Waterloo, Ontario, Canada, 2003.
59. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19: 55–72.
60. Jones D. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
61. Pollastri G, McLysaght a. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2005;21:1719–1720.
62. Miller C, Eisenberg D. Using inferred residue contacts to distinguish between correct and incorrect protein models. Bioinformatics 2008;24:1575–1582.
63. Vapnik V. Statistical learning theory. New York, NY: Wiley; 1998.
64. Vapnik V. The nature of statistical learning theory. Berlin, Germany: Springer-Verlag; 1995.
65. Drucker H, Burges C, Kaufman L, Smola A, Vapnik V. Support vector regression machines. In: Mozer MC, Jordan MI, editors, Advances in neural information processing systems, Vol. 9. Cambridge, MA: MIT Press; 1997: pp. 155–161.
66. Schölkopf B, Smola A. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, MA: MIT Press; 2002.
67. Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A. editors, Advances in kernel methods–support vector learning, Cambridge, MA: MIT Press; 1999;169–184.
68. Joachims T. Learning to classify text using support vector machines. Dessertation. Springer, 2002.
69. Vert J, Tsuda K, Scholkopf B. A primer on kernel methods. In: Scholkopf B, Tsuda K, editors, Kernel methods in computational biology. Cambridge, MA: MIT Press; 2004. pp. 55–72.
70. Moult J, Fidelis K, Tramontano A, Rost B, Hubbard T. Critical assessment of methods of protein structure prediction (casp)–round VI. Proteins 2005;61(Suppl 7):3–7.
71. Soeding J, Biegert A, Lupas A. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33:w244–248.