

# Ranking predicted protein structures with support vector regression

Jian Qiu,<sup>1</sup> Will Sheffler,<sup>1</sup> David Baker,<sup>1,2</sup> and William Stafford Noble<sup>1,3\*</sup>

<sup>1</sup> Department of Genome Sciences, University of Washington, Seattle, Washington

<sup>2</sup> Department of Biochemistry, University of Washington, Seattle, Washington

<sup>3</sup> Department of Computer Science and Engineering, University of Washington, Seattle, Washington

## ABSTRACT

*Protein structure prediction is an important problem of both intellectual and practical interest. Most protein structure prediction approaches generate multiple candidate models first, and then use a scoring function to select the best model among these candidates. In this work, we develop a scoring function using support vector regression (SVR). Both consensus-based features and features from individual structures are extracted from a training data set containing native protein structures and predicted structural models submitted to CASP5 and CASP6. The SVR learns a scoring function that is a linear combination of these features. We test this scoring function on two data sets. First, when used to rank server models submitted to CASP7, the SVR score selects predictions that are comparable to the best performing server in CASP7, Zhang-Server, and significantly better than all the other servers. Even if the SVR score is not allowed to select Zhang-Server models, the SVR score still selects predictions that are significantly better than all the other servers. In addition, the SVR is able to select significantly better models and yield significantly better Pearson correlation coefficients than the two best Quality Assessment groups in CASP7, QA556 (LEE), and QA634 (Pcons). Second, this work aims to improve the ability of the Robetta server to select best models, and hence we evaluate the performance of the SVR score on ranking the Robetta server template-based models for the CASP7 targets. The SVR selects significantly better models than the Robetta K\*Sync consensus alignment score.*

Proteins 2008; 71:1175–1182.

© 2007 Wiley-Liss, Inc.

**Key words:** protein structure prediction; scoring function; support vector regression; machine learning; consensus-based feature.

## INTRODUCTION

Protein structure prediction with computational methods is an important research area that has the potential to dramatically accelerate the determination of protein structures. The availability of high quality predicted protein structures would greatly enhance our ability to understand the functions of the proteins, redesign proteins of interest, and develop drugs interacting with protein targets.<sup>1,2</sup> Protein structure prediction can be classified into two categories: template-based modeling and ab initio modeling. In template-based modeling, one or more template structures in the PDB are identified that promise to be structurally similar to the protein target, whose structure we are interested in predicting. Predicted models are then constructed based on the identified template structures. In the ab initio approach, no template structures are used, and models are constructed from scratch by efficient sampling of the conformational space.

Most protein structure prediction approaches, either template-based or ab initio approaches, first generate a large number of candidate models by either constructing models from different alignments of different templates<sup>3</sup> or by sampling different regions of the conformational space.<sup>4–6</sup> A scoring function is then needed to discriminate between high quality models and misfolded models. An ideal scoring function should have perfect correlation with the quality of a structural model, which is measured by the closeness of the model to the native structure. Scoring functions can be derived with one of the following three approaches: (1) physical potentials, (2) probability distribution-based potentials, and (3) machine learning-based scores. A physical potential computes the energy of a structure by modeling the interactions between different components of the protein or between the protein and the solvent based on physical laws.<sup>7–9</sup> A probability distribution-based potential extracts the energy parameters from the probability distribution functions of environment types in known native structures.<sup>10–13</sup> Finally, machine learning-based scores utilize machine learning techniques such as artificial neural networks<sup>14,15</sup> and support vector machines<sup>16</sup> to learn how to combine multiple

Grant sponsor: NIH; Grant numbers: P41 RR11823, R33 HGO03070.

\*Correspondence to: William S. Noble, Foege Building S-250, Box 355065, Seattle, WA 98195.

E-mail: noble@gs.washington.edu

Received 7 June 2007; Revised 10 August 2007; Accepted 24 August 2007

Published online 14 November 2007 in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/prot.21809

features, such as physical or probability distribution-based potentials, from a training set including structures of different quality.

In this study, we use a machine learning approach to develop a scoring function to select the best models for a protein. The task is to learn to rank the predicted structures of each target correctly according to their quality, i.e., their closeness to the native structure. Several widely-used metrics to measure the similarity between two protein structures have been proposed. We use the GDT\_TS measure,<sup>17</sup> which is used in the CASP evaluation,<sup>18</sup> as the criterion to judge the quality of a structure. The learning task can be formulated as a regression problem, where a function is learned that predicts the quality of a structure from multiple features. Our features include consensus-based features that measure the closeness between a structure and all the other predicted structures for the same target,<sup>19,20</sup> and structural features that measure properties of a structure directly.<sup>21,22</sup> We use support vector regression (SVR) to solve this regression problem. SVR is a modification of the support vector machine (SVM) classifier for the regression problem. Both methods are derived from statistical learning theory,<sup>23</sup> attempting to control the complexity of the learned function while minimizing training errors. This approach allows SVR to achieve good generalization performance. SVMs have been used successfully in numerous bioinformatics applications,<sup>24</sup> including remote protein homology detection,<sup>25</sup> protein function classification,<sup>26</sup> protein interaction prediction,<sup>27</sup> protein secondary structure prediction,<sup>28</sup> and microarray data analysis.<sup>29</sup>

Using SVR, we have developed a scoring function that is a linear combination of twelve features. We test this scoring function on two data sets. First, we test on the task of selecting the best models among the CASP7 server predictions. The top-ranked models with this scoring function have comparable quality as the best server (Zhang-Server), and significantly better performance than all the other server submissions. Next, to evaluate the feasibility of using this scoring function to improve the performance of the Robetta server, we test the SVR score on the ranking of template-based models of CASP7 targets generated by the Robetta server. On this task, the SVR score performs significantly better than the Robetta consensus alignment score.<sup>3</sup>

## METHODS

### Training data set

The training data set consists of predicted structures submitted to CASP5 and CASP6 by participating prediction groups, and a selection of native protein structures from the PDB. Only the first model submitted to CASP by each group is included in the training set. All the

structures are first subject to local minimization with Rosetta to remove steric clashes, build missing sidechain atoms and optimize side chain rotamers before computing the structural features. The training set only include structures that remain similar to the original structures after minimization, with a MAMMOTH E value less than 0.001.<sup>30</sup> The data set contains 7280 predicted structures for 73 CASP targets satisfying this criterion. We also include 50 native structures corresponding to the targets in the training set whose structures were determined with X-ray crystallography.

The CASP5+CASP6 targets represent a small subset of all the proteins present in the PDB. To increase the coverage of protein folds in PDB, we include additional native PDB structures from a representative data set. A list of nonredundant representative high-quality protein structures was downloaded from the PISCES server<sup>31</sup> at [http://dunbrack.fccc.edu/Guoli/pisces\\_download.php](http://dunbrack.fccc.edu/Guoli/pisces_download.php). This list contains 1060 proteins of at most 20% pairwise sequence identity, resolution of better than 1.6 Å and R factor cutoff of 0.25. The R factor measures the disagreement between the original X-ray diffraction data and the expected diffraction pattern from the crystallographic coordinates, and it reflects the quality of the crystal structure. Proteins of fewer than 50 residues or more than 500 residues are removed from the list. We remove from the list proteins sharing significant sequence similarity with any of the CASP targets in the training set, using a BLAST E value threshold of 1. Protein structures with multiple chains can have large quarternary interaction interfaces that affect the folding of each individual chain dramatically. To avoid this scenario, we removed 402 structures with multiple chains from the list. The final list contains 424 proteins.

The final training set contains 7754 structures including 7330 structures based on 73 CASP5+CASP6 targets and 424 additional native structures from PDB. The full training set can be found in the online supplement at <http://noble.gs.washington.edu/proj/decoy>.

### Features

Several studies have shown that consensus-based features are powerful predictors of the quality of a structure.<sup>19,20</sup> In other words, a correctly-folded structure is more likely to be similar to the other predicted structures for the same protein target than an incorrectly-folded structure. We include consensus-based features by measuring the median structural similarity between a structure and all other predicted structures of the same target:  $f_i = \text{median}(\text{sim}(i,j)), \forall j \neq i$ . Compared with the mean, the median has the advantage that it is insensitive to the presence of outliers. The native structures not in CASP have the consensus-based features set to the mean of the median similarity scores for all CASP structures. We include four different structural similarity measures: root

mean square deviation (RMSD), MaxSub,<sup>32</sup> GDT\_TS<sup>17</sup> and TM.<sup>33</sup> The simple RMSD measure is highly influenced by large errors in parts of the model and poor at detecting models that are partially correct. MaxSub, GDT\_TS, and TM all aim to overcome this problem by identifying maximum well-predicted substructures. MaxSub considers a substructure to be well-predicted if distances between equivalent residues in the substructure after superposition are below a constant threshold, 3.5 Å. GDT\_TS uses four thresholds, 1, 2, 4, and 8 Å, and computes the average across the four thresholds. TM takes into account that smaller proteins tend to have lower RMSD, and varies the distance threshold according to the size of the protein. The consensus-based features derived from the four different similarity scores are highly correlated. As described later, the learned model chooses to assign only the median TM feature a nonzero weight.

We also include several additional features that measure properties of a structure directly. These features include T32S3, a distance-dependent pairwise atomic potential,<sup>22</sup> and several Rosetta-generated features,<sup>21</sup> which are listed in the online supplement. The Rosetta features capture the overall shape and burial, packing, solvation effects, hydrogen bonding patterns, torsion angle preferences, pairwise interactions, and so on. All the features are standardized so that they have a mean of 0 and standard deviation of 1.

Two scenarios in ranking decoy structures exist in practice. In the relative ranking scenario, we are interested in ranking decoy structures correctly for each target, but we are not concerned with how decoy structures from different targets are ranked relative to one another. Relative ranking is particularly important when selecting the best predicted models from a set of candidate decoy structures. On the other hand, in the absolute ranking scenario, we are interested in ranking decoy structures from all targets correctly. This scenario is important when we want to provide a confidence score for a particular predicted structure.

In this study, we concentrate on the relative ranking scenario. Therefore, to allow different scales between different proteins, we add additional protein identity features during training. Each CASP target is represented by a binary feature (flag) with value of 0 or 1. A CASP structure has the flag corresponding to its protein identity set to 1 and all the other protein identity features set to 0. The native protein structures not corresponding to a CASP target have all the protein identity features set to 0. The inclusion of these protein identity features allows the predicted output values for structures from one CASP protein to shift arbitrarily with regard to structures from another CASP protein. This in effect allows the training to focus only on the ranking of structures from the same protein, as in the relative ranking scenario.

## Support vector regression

To learn the function that maps the feature values to the predicted GDT\_TS, we use SVR. With SVR, an  $\epsilon$ -insensitive loss function is used where only errors greater than a pre-defined parameter  $\epsilon$  are considered in the loss function. We use a linear kernel to learn a linear decision function, where the predicted GDT\_TS is a weighted sum of the features. The mathematical formulation of the linear SVR is as follows<sup>34</sup>:

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^l C_i (\xi_i + \hat{\xi}_i) \\ \text{subject to} \quad & (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \leq \epsilon + \hat{\xi}_i \\ & \xi_i, \hat{\xi}_i \geq 0 \quad \forall i = 1, 2, \dots, l \end{aligned}$$

In the formulation,  $\mathbf{w}$  and  $b$  are the weights of the features and the bias term in the decision function that we want to learn,  $\mathbf{x}_i$  and  $y_i$  are the features and the target GDT\_TS value of the  $i^{\text{th}}$  training example, and  $C_i$  is a parameter that weights the error associated with the  $i^{\text{th}}$  training example.

We can adjust parameters  $C_i$  to put different weights on different training examples. In general, we are more interested in the correct ranking of the good models than the bad models. Thus we define  $C_i$  to be dependent on  $r_i$ , the rank of structure  $i$  among all the predicted structures of the same protein:

$$C_i = \begin{cases} C_{\text{nat}} & \text{if structure } i \text{ is a native protein structure} \\ \frac{5 * C_{\text{nat}}}{\sqrt{r_i}} & \text{if structure } i \text{ is a predicted structure} \end{cases}$$

$C_{\text{nat}}$  was chosen to be 10. In this scheme, a native structure gets the same weight as a predicted structure with rank 25. We did not experiment with other values of  $C_{\text{nat}}$ .

We choose the signs of all features except for the protein identity features such that they have a positive correlation with the target output, GDT\_TS. Table I lists the Pearson correlation coefficients of the features with GDT\_TS averaged over all CASP targets in the training set. The consensus-based feature, median TM, has a much larger correlation coefficient than the other features derived from individual structures. In principle, the learned function should therefore assign positive weights to all the nonidentity features. However, because the features are not independent of each other, the SVR tends to learn a decision function with both positive and negative weights. Therefore, to derive a scoring function that is more interpretable and generalizable, we introduce additional constraints requiring that all nonidentity features have non-negative weights. The Mosek toolbox<sup>35</sup> in Matlab is used to solve the resulting SVR optimization problem.

**Table I***The Weights of the Features in the Learned Scoring Function*

Feature	Description	Weight	Correlation
median TM	Consensus score	20.057	0.87
T32S3	Pairwise atomic potential	2.444	0.52
sasapack	Packing quality score	2.180	0.07
fa_prob	Torsion angle preference	1.732	0.30
numres	Number of residues	1.405	0.19
env	Residue burial preference	1.376	0.60
hb_sc	Side chain hydrogen bonds	1.167	0.004
pair	Residue-based pairwise potential	0.946	0.40
hb_lrbb	Long range backbone hydrogen bonds	0.565	0.33
co	Contact order	0.526	0.17
sasa	Solvent accessible surface area	0.475	0.41
hb_srbb	Short range backbone hydrogen bonds	0.280	0.10

The third column lists the nonzero weights of the features in the trained SVR model. The last column lists the average Pearson's correlation coefficients between these features and GDT\_TS in the training set. The correlation coefficients are computed separately for each target and then averaged over all CASP targets.

## RESULTS

### The learned scoring function

A collection of 21 structural features and 4 consensus-based features are used to learn to predict the target output, GDT\_TS to the native structure. A list of these features can be found in the online supplement. During training, the weights of these 25 features are constrained to be non-negative. These additional constraints in effect play the role of feature selection, and the resulting solution assigns positive weights to only 12 features. Table I lists the 12 selected features and their weights. Although we compute four different consensus-based features, median RMSD, GDT\_TS, MaxSub, and TM scores, because these four features are highly correlated the learned scoring function assigns positive weight only to the median TM feature. Furthermore, the median TM feature receives a much larger weight than all the other features. This observation is consistent with the superior power of consensus-based features in predicting protein structure quality, as indicated by the high correlation between median TM and GDT\_TS in the training set, 0.87.

### The CASP7 server prediction

To evaluate the ability of the SVR scoring function to predict structure quality, we test its performance in ranking the server predicted models for the CASP7 targets. All the models are first subject to local minimization with Rosetta before computing the structural features, as was done in the training set. We evaluate the performance of the SVR score in selecting the best models by four metrics: GDT\_TS1 raw and Z scores, and GDT\_TS5 raw and Z scores. GDT\_TS1 measures the GDT\_TS score of the first model ranked by a quality assessment method

or submitted by a server group, and GDT\_TS5 measures the best GDT\_TS score among the top five models. GDT\_TS scores from different targets have different distributions, and the differences in GDT\_TS may have different scales for different targets. Therefore, we also compute the Z scores of GDT\_TS1 and GDT\_TS5. Table II compares the performance of the SVR with the servers having the highest average GDT\_TS1 Z scores: Zhang-Server,<sup>36</sup> Pmodeller6,<sup>19</sup> ROBETTA,<sup>37</sup> and MetaTasser,<sup>38</sup> and the two best performing Quality Assessment (QA) groups in CASP7, QA556 (LEE),<sup>39</sup> and QA634 (Pcons).<sup>19</sup> For both GDT\_TS1 and GDT\_TS5 raw and Z scores, the SVR score performs better than both quality assessment methods, at least as well as Zhang-Server and better than all the other servers. To evaluate the statistical significance of the differences, we perform pairwise Wilcoxon signed rank tests and compute the resulting *p* values. Table III shows that the performance of the SVR is not statistically different from that of Zhang-Server, and both the SVR and Zhang-Server perform significantly better than all the other methods. Table II also lists the mean GDT\_TS score of the best model of each target. This score is the theoretically optimal score. The large margin between the best model GDT\_TS and the SVR results indicates that the SVR scoring function still has plenty of room for improvement.

Considering the outstanding performance of the Zhang server in Table II, we are interested in how much the performance of the SVR score depends on the presence of the Zhang-Server models. Therefore, we also compute the performance of the selection by the SVR score of models from server submissions other than the Zhang server, represented as “SVR (no Zhang).” Tables II and

**Table II***The Average Structural Quality of Top-Ranked CASP7 Server Models by SVR*

Method	No of targets predicted	Mean GDT_TS1 (Z score)	Mean GDT_TS5 (Z score)
Best model	98	0.636 (1.81)	0.636 (1.81)
SVR	98	0.589 (1.17)	0.614 (1.47)
Zhang-Server	98	0.589 (1.11)	0.613 (1.47)
SVR (no Zhang)	98	0.576 (1.02)	0.603 (1.32)
QA634	98	0.556 (0.88)	0.593 (1.22)
Pmodeller6	98	0.553 (0.87)	0.584 (1.16)
QA556	96	0.564 (0.83)	0.580 (1.02)
ROBETTA	97	0.550 (0.82)	0.582 (1.22)
MetaTasser	98	0.545 (0.76)	0.562 (0.96)

The first column lists the names of the server groups or the methods used to select the model. For the “best model,” we select the best model according to the observed quality of a structure, its GDT\_TS. SVR ranks the models according to the scoring function developed in this study. “SVR (no Zhang)” is the same as SVR except that all Zhang Server submissions are excluded. Zhang-Server, Pmodeller6, ROBETTA and MetaTasser are the top performing servers according to mean GDT\_TS1 Z score, and QA634 (Pcons) and QA556 (LEE) are the two best performing Quality Assessment groups in CASP7. The numbers in the parentheses represent the average Z scores based on the distributions of GDT\_TS scores of each target.

**Table III**

Pairwise Comparisons Between SVR and Best-Performing Servers and Quality Assessment Methods

	SVR (no Zhang)	SVR Zhang)	QA634	QA556	Pmodeller6	ROBETTA
Zhang-Server	—	0.002	0.0003	6e-05	3e-08	7e-09
SVR		0.0001	0.008	0.002	2e-07	2e-07
SVR (no Zhang)		—	—	—	0.0004	0.001
QA634			—	—	0.006	0.009
QA556				—	0.01	0.0009
Pmodeller6					—	—

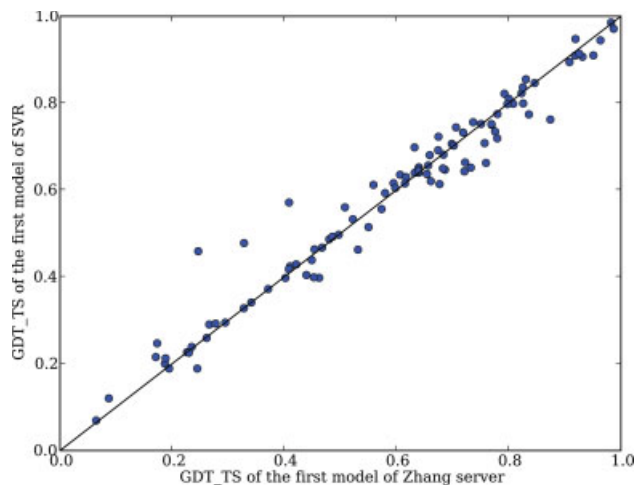
This table lists one-tailed Wilcoxon signed rank test *P*-values comparing the GDT\_TS1 ranked by SVR or “SVR (no Zhang)” and that submitted by the best-performing servers and Quality Assessment methods. The table is sorted according to the median of GDT\_TS1. When a *P*-value is less than 0.05, the *P*-value is shown indicating the significance of the row method performing better than the column method. A dash indicates that the median GDT\_TS1 for the row exceeds the median GDT\_TS1 for the column, but that the difference is not significant at a threshold of 0.05.

III show that although “SVR (no Zhang)” performs worse than both SVR and Zhang-Server, it is able to perform significantly better than the other top performing servers, Pmodeller6 and ROBETTA. “SVR (no Zhang)” also achieves better performance than QA634 and QA556 for all four metrics evaluated, but the difference is not statistically significant. Table IV shows the 11 server groups with at least two models selected by SVR as the best model for a target. The three top performing servers, Zhang-Server, ROBETTA and Pmodeller6, contribute the most models to the top 1 SVR-ranked models. In total, there are 32 servers with predictions ranked by SVR as the best for some targets. Figure 1 compares the structural quality of the first model selected by SVR with that of Zhang-Server for each target. In the lower left region, a majority of the points lie above the diagonal. This suggests that among the difficult CASP7 targets, the top-ranked models by the SVR are more often better than the first models of Zhang-Server.

**Table IV**

Distribution of Server Choices Among the First Models Ranked by SVR

Group	No of models selected
Zhang-Server	30
ROBETTA	11
Pmodeller6	8
PROTINFO	6
Pcons6	6
HHpred2	3
RAPTORESS	3
CPHmodels	3
3Dpro	3
Frankenstein	2
keasar-server	2

**Figure 1**

Comparison of the structural quality of the first models selected by SVR and submitted by Zhang-Server. Each point represents one target, with the *x* coordinate indicating the GDT\_TS of the first Zhang-Server model and the *y* coordinate indicating the GDT\_TS of the first model selected by SVR. [Color figure can be viewed in the online issue, which is available at [www.interscience.wiley.com](http://www.interscience.wiley.com).]

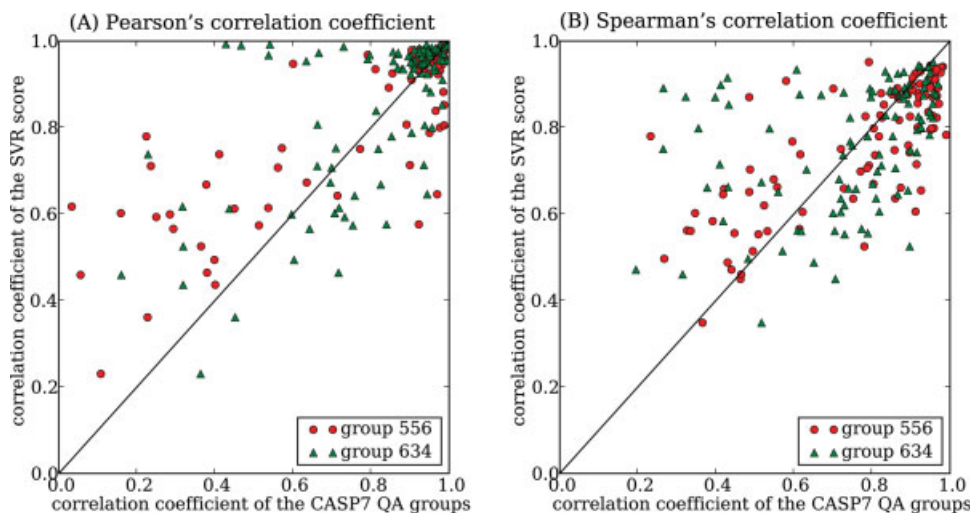
In CASP7, Pearson's and Spearman's correlation coefficients were used to evaluate the Quality Assessment performance. We also compare the Pearson's and Spearman's correlation coefficients of the SVR with those of QA556 and QA634 in Table V. The SVR yields Pearson's correlation coefficients that are significantly better than both QA556 and QA634, and Spearman's correlation coefficients that are statistically indistinguishable from QA556 and QA634. Figure 2 plots the correlation coefficient of the SVR versus that of QA556 or QA634 for each target. As shown in Figure 2(A), the majority of the points lie above the diagonal, indicating that the SVR yields better Pearson's correlation coefficients than the two QA methods in general. In figure 2B, although there are more

**Table V**

Comparison of the Correlation Coefficients Between the SVR and Best-Performing CASP7 Quality Assessment Methods

Correlation method	QA method	SVR mean	QA mean	No of win	No of lose	<i>P</i> -value
Pearson	QA556	0.852	0.806	53	43	0.02
	QA634	0.852	0.818	62	36	0.03
Spearman	QA556	0.762	0.764	37	59	0.09
	QA634	0.762	0.746	41	57	0.2

The first column lists the correlation methods used, and the second column lists the QA methods in comparison. The third and fourth columns indicate the mean correlation coefficients of the SVR and the QA method in comparison, respectively. The fifth column shows the number of times the SVR achieves a better correlation than the QA method, and the sixth column shows the number of times the SVR has a worse correlation coefficient. The last column indicates the Wilcoxon signed rank test *P*-values.



**Figure 2**

The correlation coefficients of SVR and best-performing CASP7 Quality Assessment methods.

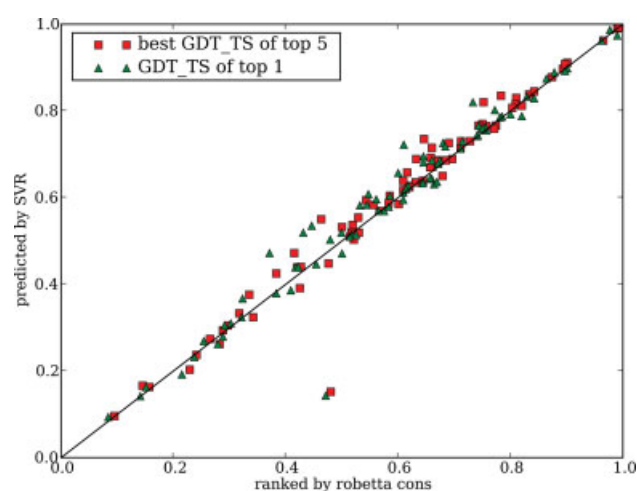
cases where the SVR yields a lower Spearman's correlation coefficient than QA556 or QA634, the difference between the SVR and the QA method tends to be smaller than in the cases where the SVR has a better Spearman's correlation coefficient.

Note, however, that the SVR is not trained to optimize the overall correlation. In particular, we assign a larger weight to training examples with better structural quality. This design is motivated by the observation that we are often more interested in the correct ranking of the good models rather than the bad models. Thus the SVR score is optimized for the task of selecting the best models rather than having good correlation with GDT\_TS among both good and bad models. Nevertheless, it is encouraging that the SVR score performs rather well in correlation coefficients compared with QA556 and QA634.

### The Robetta server prediction

One goal of this study is to develop a ranking method that improves the performance of the Robetta server by selecting better models from the sampled decoy structures. Therefore, we test the ability of the SVR scoring function to rank Robetta template-based decoy models for CASP7 targets. This dataset contains predictions for 77 targets with from 30 to 1955 predicted structures for each target. The SVR scoring function is used to rank these predicted models and select either one, or five predictions for each target. The structural quality of the best prediction in the selected subset is then compared with that selected by the Robetta consensus score based on the

Robetta K\*Sync alignment ensemble.<sup>3</sup> Figure 3 plots the GDT\_TS scores of the models selected by SVR versus those selected by the Robetta consensus score. More data points lie above the diagonal, indicating that the SVR selects better models than the Robetta consensus alignment score in most of the cases. Table VI shows that in both top 1 and top 5 selections, the SVR score selects significantly better models than the Robetta consensus



**Figure 3**

Comparison of the structural quality of the top-ranked models by SVR and by the Robetta consensus alignment score.

**Table VI***Comparison Between SVR and the Robetta Consensus Alignment Score*

Subset selected	SVR mean	Robetta mean	No of win	No of lose	<i>P</i> -value
Top1	0.595	0.587	48	28	0.003
Top5	0.609	0.600	57	18	4e-05

This table compares the GDT\_TS scores of the best predicted structures in subsets selected by SVR with those selected by the Robetta consensus alignment score. The first column indicates the sizes of the subsets selected. The second and third columns indicate the mean GDT\_TS score of the best model in the subset selected by the SVR score and by the Robetta consensus alignment score, respectively. The fourth column indicates the number of times the SVR-selected models have better GDT\_TS scores than Robetta-selected models, and the fifth column indicates the number of times SVR-selected models have worse GDT\_TS scores. The last column lists the Wilcoxon signed rank test *P*-values comparing the GDT\_TS scores of SVR-selected models with those of Robetta-selected models.

alignment score according to a Wilcoxon signed-rank test, and has a better mean GDT\_TS score.

## DISCUSSION

In this paper, we have used support vector regression to derive a scoring function for ranking predicted protein structures. The SVR scoring function is evaluated in two settings: ranking CASP7 server models and ranking Robetta server models. With the CASP7 test set, if we use the SVR scoring function to choose one model for each target, the selected models have comparable quality to the Zhang server and perform significantly better than all the other servers and the two best performing Quality Assessment groups in CASP7. Even if the Zhang server submissions are excluded, then the selected models still perform significantly better than all the other servers. With the Robetta test set, the SVR score is able to select models significantly better than the Robetta consensus alignment score.

The SVR score gives a much larger weight to the consensus-based feature, median TM, than the other features. To investigate how much this consensus-based feature contributes to the performance of the SVR score, we train another scoring function without consensus-based features, represented as “SVR (no consensus).” The features and their weights in the new scoring function are listed in the online supplement. Table VII compares the performance of the original full SVR score with the median TM score alone and “SVR (no consensus).” Both median TM and “SVR (no consensus)” perform better than all the servers except the Zhang server. The full SVR score performs better than both median TM and “SVR (no consensus),” indicating the benefit of combining features from both categories. When evaluated with the Wilcoxon signed rank test, the full SVR score performs significantly better than median TM with a *P*-value of 0.03.

Several groups have previously used SVR to learn a composite score for evaluating protein structure qual-

ity.<sup>16,40</sup> This work differs from previous work in several aspects. First, our SVR formulations are different. We modify the standard SVR formulation to specifically suit our problem in hand. To allow learning in the local ranking scenario, we introduce additional protein identity features such that models from different proteins can have different scales. We also set the *C* parameter in a structure-specific fashion, assigning larger weights to the learning of structures with better quality. The last modification is to enforce the directions of the features and disallow negative weights in the training. Second, we use different features to derive the SVR score. Xu *et al.*<sup>40</sup> derived all their features from consensus-based features, and Eramian *et al.*<sup>16</sup> derived features based on individual structures such as the DOPE potential, several potentials from MODPIPE and two PSIPRED/DSSP secondary structure agreement scores. Our SVR score instead includes both a consensus-based feature, the median TM feature, and features directly describing the properties of each individual structure, such as several Rosetta features and the T32S3 potential. Our study shows that these two kinds of features are complementary, and the inclusion of both kinds perform better than either alone. Finally, the three approaches use different metrics to measure the quality of a structure, the target function to train the regressors. Eramian *et al.*<sup>16</sup> trained their regressor to predict the RMSD of a model. RMSD has the drawback that it is highly influenced by large errors in a local region, and may not reflect the quality of the global topology. We instead choose to evaluate the quality of a structure using the GDT\_TS metric. Xu *et al.*<sup>40</sup> used the MaxSub score to measure structural quality.

Quite often we are not only interested in the overall quality of a structure, but we are also interested in how the structural quality varies across different regions. The additional knowledge of region-specific structural quality can have multiple applications. First, poor quality regions can be identified that need remodeling. Second, multiple models can be combined to construct better predicted structures by combining regions of high quality. Finally, a predicted structure can be used more wisely by a biologist with region-specific quality information. One future direction of this work is to extend the approach used in this paper to the problem of predicting residue-level structure quality.

**Table VII***The Performance Comparison Between the SVR Score and the SVR Score Without Consensus-Based Features*

Method	Mean GDT_TS1 (Z score)
SVR	0.589 (1.17)
Zhang-Server	0.589 (1.11)
median TM	0.572 (1.02)
SVR (no consensus)	0.559 (0.89)
Pmodeller6	0.553 (0.87)

## ACKNOWLEDGMENTS

The authors thank Bin Qian for valuable discussions and Dylan Chivian for help with the Robetta data set.

## REFERENCES

- Baker D. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
- Ginalski K, Grishin NV, Godzik A, Rychlewski L. Practical lessons from protein structure prediction. *Nucleic Acids Res* 2005;33:1874–1891.
- Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res* 2006;34(17):c112.
- Park B, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol* 1997;266:831–846.
- Xia Y, Huang ES, Levitt M, Samudrala R. ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol* 2000;300:171–185.
- Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* 2003;53:76–87.
- Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy minimization and dynamic calculations. *J Comput Chem* 1983;4:187–217.
- Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* 1998;288:477–487.
- Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 1990;112:6127–6129.
- Bowie FU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991;253:164–170.
- Lu H, Skolnick J. A distance dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* 2001;44:223–232.
- Samudrala R, Moulton J. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* 1998;275:895–916.
- Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 1999;34:82–95.
- Jones DT. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 1999;287:797–815.
- Wallner B, Elofsson A. Can correct protein models be identified? *Prot Sci* 2003;12:1073–1086.
- Eramian D, Shen M, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. *Prot Sci* 2006;15:1653–1666.
- Zemla A. LGA – a method for finding 3d similarities in protein structures. *Nucleic Acids Res* 2003;31:3370–3374.
- Moulton J, Kryzysztow F, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - round 6. *Proteins* 2005;61(S7):3–7.
- Wallner B, Elofsson A. Pcons5: combining consensus, structural evaluation and fold recognition scores. *Bioinformatics* 2005;21:4248–4254.
- Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 2003;19:1015–1018.
- Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using rosetta. *Methods Enzymol* 2004;383:66–93.
- Qiu J, Elber R. Atomically detailed potential to recognize native and approximate protein structures. *Proteins* 2005;61:44–55.
- Vapnik V. *The nature of statistical learning theory*. New York: Springer; 1995.
- Noble WS. Support vector machine applications in computational biology. In: Schoelkopf B, Tsuda K, Vert J-P, editors. *Kernel methods in computational biology*. Cambridge, MA: MIT Press; 2004. pp 71–92.
- Jaakkola T, Diekhans M, Haussler D. Using the Fisher kernel method to detect remote protein homologies. In the Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. AAAI Press; Menlo Park, CA; 1999. pp 149–158.
- Cai CZ, Wang WL, Sun LZ, Chen YZ. Protein function classification via support vector machine approach. *Math Biosci* 2003;185:111–122.
- Ben-Hur A, Noble WS. Kernel methods for predicting protein-protein interactions. *Bioinformatics* 2005;21(suppl 1):i38–i46.
- Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *J Mol Biol* 2001;208(2):397–407.
- Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc Nat Acad Sci USA* 2000;97(1):262–267.
- Ortiz AR, Strauss CEM, Olmea O. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Prot Sci* 2002;11:2606–2621.
- Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19:1589–1591.
- Siew N, Elofsson A, Rychlewski L, Fischer D. MaxSub: An automated measure for the assessment protein structure prediction quality. *Bioinformatics* 2000;16:776–785.
- Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins* 2004;57:702–710.
- Cristianini N, Shawe-Taylor J. *An introduction to support vector machines*. Cambridge, UK: Cambridge University Press; 2000.
- Andersen ED, Andersen KD. The mosek interior point optimizer for linear programming: an implementation of the homogeneous algorithm. *High Perform Optim* 2000;197–232.
- Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 2007;5:17.
- Chivian D, Kim DE, Malmstrom L, Schonbrun J, Rohl CA, Baker D. Prediction of CASP6 structures using automated rosetta protocols. *Proteins* 2005;61(S7):157–166.
- Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in casp6. *Proteins* 2005;61(S7):91–98.
- Lee J, Scheraga HA, Rackovsky S. New optimization method for conformational energy calculations on polypeptides: conformational space annealing. *J Comput Chem* 1997;18:1222–1232.
- Xu J, Yu L, Li M. Consensus fold recognition by predicted model quality. In the Proceedings of the 3rd Asia-Pacific Bioinformatics Conference, January 2005, Singapore, pp 73–84.