Structure
Article

Improving Protein Structure Prediction Using Multiple Sequence-Based Contact Predictions

Sitao Wu,[3,4] Andras Szilagyi,[3,5] and Yang Zhang[1,2,3,*]
[1]Center for Computational Medicine and Bioinformatics
[2]Department of Biological Chemistry
University of Michigan, Ann Arbor, MI 48109, USA
[3]Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, Lawrence, KS 66047, USA
[4]Center for Research in Biological Systems, University of California San Diego, La Jolla, CA 92093, USA
[5]Institute of Enzymology, Hungarian Academy of Sciences, Karolina ut 29, H-1113 Budapest, Hungary
*Correspondence: zhng@umich.edu
DOI 10.1016/j.str.2011.05.004

SUMMARY

Although residue-residue contact maps dictate the topology of proteins, sequence-based ab initio contact predictions have been found little use in actual structure prediction due to the low accuracy. We developed a composite set of nine SVM-based contact predictors that are used in I-TASSER simulation in combination with sparse template contact restraints. When testing the strategy on 273 nonhomologous targets, remarkable improvements of I-TASSER models were observed for both easy and hard targets, with p value by Student's t test <0.00001 and 0.001, respectively. In several cases, template modeling score increases by >30%, which essentially converts "nonfoldable" targets into "foldable" ones. In CASP9, I-TASSER employed ab initio contact predictions, and generated models for 26 FM targets with a GDT-score 16% and 44% higher than the second and third best servers from other groups, respectively. These findings demonstrate a new avenue to improve the accuracy of protein structure prediction especially for free-modeling targets.

INTRODUCTION

The topology of protein three-dimensional (3D) structures can be specified by their interresidue distance and contact maps. Thus, the structure of a protein molecule can be readily reconstructed by computer if all the native contacts are known. Using the power of contemporary protein structure prediction algorithms, which utilize various structural regularities such as predicted secondary structure and advanced force fields (Liwo et al., 1999; Roy et al., 2010; Sali and Blundell, 1993; Simons et al., 1997; Zhang and Skolnick, 2004a), the structure of a protein can be built based on just a small fraction of native contacts. For example, Li et al. (2004) showed that only one native contact (from nuclear magnetic resonance nuclear Overhauser enhancement data) for every eight residues is sufficient to guide the structure prediction tool TOUCHSTONE-II (Zhang et al., 2003) to construct a correct topology for single-domain proteins up to 200 residues. This is particularly encouraging because requiring fewer native contacts for structure reconstruction allows a significant reduction in experimental data collection efforts and thus makes possible the structure determination of a wide range of proteins for which obtaining a full set of native contacts is difficult.

For most proteins in nature, however, not even sparse experimental contact data are available, and the interresidue contacts must be generated by computer-based predictions. Contact prediction methods can be largely classified into two types. The first is the template-based method (Misura et al., 2006; Shao and Bystroff, 2003; Skolnick et al., 2004; Wu and Zhang, 2007), i.e., collecting contacts from solved homologous proteins whose structures can be found in the Protein Data Bank (PDB) by sequence similarity search (Altschul et al., 1997) or threading algorithms (Bowie et al., 1991; Karplus et al., 1998; Soding, 2005; Wu and Zhang, 2008b). The accuracy of the contact prediction can be very high when closely homologous templates are identified, which has been shown to be extremely useful for high-resolution template-based protein structure prediction (Raman et al., 2009; Sali and Blundell, 1993; Zhang, 2009). Zhang et al. (2003) showed that contact predictions with an average accuracy >22% should have an overall positive effect on ab initio protein folding simulations. However, one limitation of template-based contact prediction is that the accuracy highly depends on the availability of templates. For hard protein targets, i.e., those without homologous templates, template-based contact prediction usually has a low accuracy and therefore becomes useless for protein structure prediction.

The second type of contact prediction methods does not depend on protein template structures. Instead, contact predictions are derived from the primary sequence by identifying correlated mutations (Gobel et al., 1994) or machine learning methods (Cheng and Baldi, 2007; Shackelford and Karplus, 2007; Wu and Zhang, 2008a). For the hard free-modeling (FM) protein targets, it has been shown in the CASP experiments that the purely ab initio sequence-based contact predictions have a higher accuracy than those collected from the best template-based models (Ezkurdia et al., 2009). Despite its appealing feature of nondependence on template structures, the usefulness of sequence-based contact prediction in 3D structure prediction has not yet been systematically assessed and demonstrated in the literature. Considering the still low accuracy (typically ~20%–30%)

of the state-of-the-art contact prediction algorithms (Cheng and Baldi, 2007; Shackelford and Karplus, 2007; Wu and Zhang, 2008a), it is particularly important to know when the ab initio contact predictions should be used, whether and how they should be combined with the template-based contact information, and how they can be best geared into the conventional template assembly algorithms such as I-TASSER (Roy et al., 2010; Wu et al., 2007), MODELER (Sali and Blundell, 1993), and ROSETTA (Simons et al., 1997).

In this work, we aim to provide a systematic examination of these open questions in the framework of I-TASSER (Roy et al., 2010; Wu et al., 2007; Zhang and Skolnick, 2004a) that was designed to construct protein structures by assembling template structure fragments identified by threading (Wu and Zhang, 2007). To address the major weakness of high false positive rate in single sequence-based contact predictors (Cheng and Baldi, 2007; Shackelford and Karplus, 2007; Wu and Zhang, 2008a), we developed a composite set of nine contact predictors, each trained on different atom types with different distance cutoffs, using support vector machines (SVMs). The combination of ab initio contacts with sparse template-based restraints in I-TASSER is carried out differently for easy and hard targets, and improvements are demonstrated for both groups. Notably, encouraging examples have been found where nonfoldable targets can be converted into foldable ones owing to the use of ab initio contact predictions.

RESULTS

The ab initio contact predictions are generated by an extended version of SVMSEQ (Wu and Zhang, 2008a), with individual predictors trained on contacts defined by C_α, C_β atoms, and side-chain centers of mass, each with three distance cutoffs (7 Å, 8 Å, 9 Å), as described in Experimental Procedures. The contacts are then used as restraints in I-TASSER simulations (Wu et al., 2007). A total of 273 nonhomologous proteins have been collected as our benchmark test proteins, which includes 253 proteins collected from the PDB by PISCES (Wang and Dunbrack, 2003), 8 FM targets from CASP7, and 12 FM targets from CASP8. All the proteins are single domain proteins. Because I-TASSER starts with threading templates identified by LOMETS (Wu and Zhang, 2007), for fair testing, all templates having >30% sequence identity with the target were excluded from the LOMETS template library. As SVMSEQ was trained on 500 training proteins, to avoid overtraining, we have confirmed that the benchmark proteins all have a sequence identity <25% to the SVMSEQ training proteins.

The structures of the target proteins were predicted by I-TASSER, either with or without sequence-based predicted contacts. The accuracy of the models was evaluated using root-mean-square deviation (rmsd) and template modeling score (TM-score) (Zhang and Skolnick, 2004b). TM-score measures the similarity of two structures in a chain length independent way, which is more sensitive than rmsd to the topological similarity of structures especially when rmsd is high. TM-score has a value in the [0, 1] interval; a TM-score <0.17 indicates a random similarity and a TM-score >0.5 corresponds to protein structures having the same global fold as defined in SCOP and CATH (Xu and Zhang, 2010).

Table 1. Contact prediction accuracy and coverage for 164 hard targets

	Contacts Combined from SVMSEQ and LOMETS	Contacts from LOMETS Only
$ACC_{C\alpha\_short}(cov)$[a]	0.285 (24.0%)	0.182 (30.1%)
$ACC_{C\alpha\_medium}(cov)$[a]	0.212 (12.3%)	0.171 (17.1%)
$ACC_{C\alpha\_long}(cov)$[a]	0.141 (6.5%)	0.131 (9.1%)
$ACC_{C\alpha\_all}(cov)$[a]	0.261 (14.1%)	0.193 (17.1%)
$ACC_{SG\_short}(cov)$[b]	0.425 (20.1%)	0.228 (36.2%)
$ACC_{SG\_medium}(cov)$[b]	0.301 (16.2%)	0.167 (30.2%)
$ACC_{SG\_long}(cov)$[b]	0.282 (9.1%)	0.180 (18.7%)
$ACC_{SG\_all}(cov)$[b]	0.362 (14.4%)	0.194 (20.7%)
$ACC_{overall}(cov)$[c]	0.310 (13.2%)	0.179 (23.1%)

ACC, accuracy; cov, coverage.
[a] C_α contact prediction for short-range ($6 \leq |i-j| < 12$), medium-range ($12 \leq |i-j| < 24$), and long-range ($|i-j| \geq 24$) contacts, and all ranges ($|i-j| \geq 6$), respectively.
[b] Side-chain center contact predictions.
[c] Overall results of all range ($|i-j| \geq 6$).

A systematic comparison of the C_α contact prediction by SVMSEQ with that by template-based predictions was conducted earlier (Wu and Zhang, 2008a). According to this study, the template-based contact prediction typically outperforms the sequence-based, ab initio contact prediction for "easy" and "medium" targets, i.e., when homologous templates are available. But for "hard" targets where no reliable templates can be identified, the accuracy of the SVMSEQ prediction is ~12%–25% higher than that produced by LOMETS threading. It was also shown that the accuracy of the SVMSEQ prediction is close to or slightly higher than that of other state-of-the-art ab initio contact predictors, including SAM server (Shackelford and Karplus, 2007) and SVMCON (Cheng and Baldi, 2007). In this study, we will not repeat these comparisons and focus instead on finding the best combination of the ab initio and template-based contact predictions that can be optimally used in protein structure assembly for both hard and easy targets.

Sequence-Based Contact Predictions Used for Hard Targets

First, we test the usefulness of sequence-based contact predictions in protein structure prediction for hard targets. We selected 164 nonhomologous proteins with lengths ranging from 41 to 207 residues that were classified as hard targets by LOMETS (Wu and Zhang, 2007) because the Z-scores of all threading templates are lower than the predefined thresholds, meaning that no threading program can identify a good template. The hard targets include 59 α, 22 β, and 83 α+/β proteins.

As shown in Table 1, the average accuracy (=17.9%) of the template-based contact predictions on the hard targets is low, compared to 22% that was found necessary to improve ab initio modeling (Zhang et al., 2003). In particular, for the long-range contacts ($|i-j| > 24$), the C_α and side-chain contact accuracy is 13.1% and 18.0%, respectively. We therefore combined the contact predictions of LOMETS and SVMSEQ using a weighted sum of the confidence scores (see Experimental Procedures).

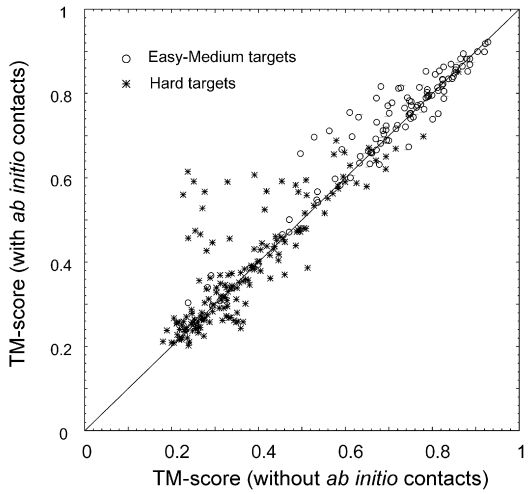

Figure 1. Structure Modeling Results with and without Using Contact Predictions

TM-score of the first-ranking models generated by the normal implementation of I-TASSER on 164 nonhomologous hard targets (stars) and 109 nonhomologous easy/medium targets (circles) versus that of I-TASSER when using ab initio contact predictions.

The combined contacts, which have a much higher accuracy (14.1% and 28.2%, respectively), were then used in I-TASSER as constraints in the structural assembly simulations.

The average TM-score of the first-ranking models is 0.386, which is consistent with the difficulty of structure prediction for these targets. For 36 of 164 hard targets, the first-ranking models have a good quality with TM-score >0.5, indicating successful prediction of the global fold. If the best of the top five models is considered for each target, the average TM-score increases to 0.410, and 41 targets have predicted models with TM-score >0.5.

When structures are predicted with the original I-TASSER procedure that only uses template-based distance and contact predictions (''old I-TASSER''), the average TM-score of the first-ranking models is 0.369. Thus, the ''new'' I-TASSER achieves a ~4.6% higher average TM-score than the ''old'' one. The p value by the paired Student's t test for the two sets of models is below 0.001, which demonstrates that the TM-score improvement by SVMSEQ is statistically significant. Figure 1 shows a head-to-head comparison of the TM-scores obtained with and without ab initio contact predictions. There are a number of targets that show significant TM-score increase. For example, there are 15 proteins that have a TM-score increase by >0.12, 10 proteins with a TM-score increase by >0.2, and 6 proteins with a TM-score increase by >0.25. Most of these targets conduct a TM-score transition from far below 0.5 to above 0.5, indicating that the use of sequence-based contacts converted these targets from ''nonfoldable'' to ''foldable'' if we consider TM-score >0.5 as a quantitative criterion for assessing whether two protein structures have a similar fold (Xu and Zhang, 2010).

On the contrary, there is only one protein, the γ subunit of the dissimilatory sulfite reductase (PDB ID: 1sauA), where the ab initio contact prediction reduces the TM-score of the I-TASSER model by >0.12. For this target, the SVMSEQ contact predictions in the N-terminal have low accuracy that distracted the N-tail (1P-17F) flip away from the core; this results in an overall TM-score deterioration from 0.513 to 0.386, a reduction of 0.127. However, the global topology in core region of the protein (39S-114V) remains unchanged.

What is the reason for the improvement of the predicted structures? We take a detailed look at Table 1, which summarizes the accuracy (number of correct predictions/total number of predictions) and coverage (number of correct predictions/number of contacts in target) of contact predictions, comparing the consensus contacts obtained from combining the sequence- and template-based contacts with the template-based contacts from LOMETS only. When short-, medium-, and long-range contacts are combined ($|i - j| \geq 6$), the average accuracy of the consensus contact predictions is 0.261 (with a coverage = 14.1%) for the C_α-C_α contacts, which is 35% higher than that of the template-based contacts (accuracy = 0.193, coverage = 17.1%). For the contacts between side-chain centers, the average accuracy of the consensus contact predictions is 0.362 (coverage = 14.4%), almost twice as much as that of the template-based ones (accuracy = 0.194, coverage = 20.7%). Combining the C_α and side-chain center contacts, the overall consensus contact predictions achieve an accuracy of 0.310 (coverage = 13.2%), which is 73% more than that of the template-based ones (accuracy = 0.179, coverage = 23.1%).

We want to mention that we did not perform a ''fair'' comparison of the two sets of contact predictions according to the conventional standard that compares the accuracy of predictions with the same coverage (Wu and Zhang, 2008a). However, we found that the structure prediction results were more sensitive to the correctness of contacts (the accuracy) than the number of predictions (the coverage), although both are important (Zhang, 2009). The balance of accuracy and coverage of the consensus contacts was optimized on an independent set of training proteins with the purpose of maximizing the TM-score of the final model (see Supplemental Experimental Procedures available online). In fact, even for a single set of contact predictions, the accuracy can be slightly increased by reducing the coverage because there must be a positive correlation between the confidence score and the accuracy of the prediction in any reasonable contact predictor. However, we have previously shown that ab initio contact prediction yields substantial novel contacts that are added to the template-based contacts, thus the accuracy increase attained by taking a consensus is significantly beyond what can be achieved by simply reducing the coverage (Wu and Zhang, 2008a). As shown in the following examples, most of the improvement of the final models, especially for the proteins with a >0.25 TM-score increase, is indeed due to the increase in contact prediction accuracy.

As shown in Table 1, the accuracy of the short-range contacts is higher than that of the long-range ones. The improvement of the structural models, however, is mainly due to the long-range contacts. In fact, when we removed the long-range ab initio contacts, there was almost no difference between the average TM-scores of the final models generated by old and new I-TASSER. However, the overall improvement of the I-TASSER models when both long- and short-range contacts were used was more pronounced than with using only long-range contacts,

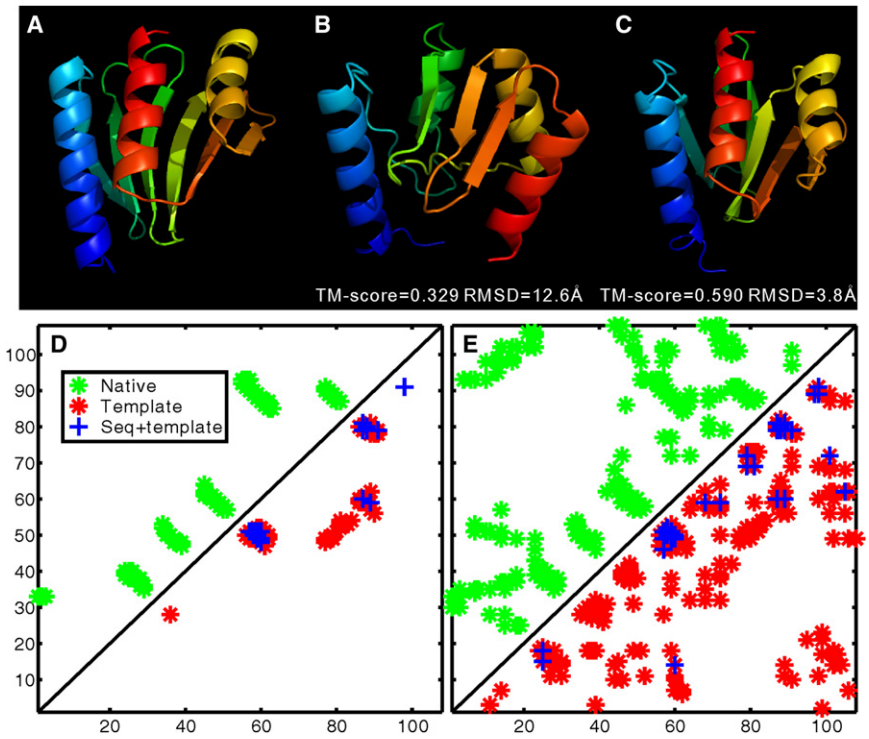

Figure 2. Illustrative Example of I-TASSER Modeling for the Target Protein 1kafA

(A) Experimental structure.

(B) Model generated by I-TASSER without using ab initio contact prediction.

(C) Model generated by I-TASSER with ab initio contact prediction.

(D) Map of native C_α contacts (*, green), template-based predicted C_α contacts (*, red), and consensus sequence- and template-based C_α contact predictions (+, blue).

(E) Map of native side-chain center contacts (*, green), template-based predicted side-chain center contacts (*, red), and consensus sequence- and template-based side-chain center contacts (+, blue).

indicating that the short-range contacts are still necessary for fine-tuning the packing of local structures.

Examples for Successful Prediction on Hard Targets

We now take a closer look at the targets where a striking improvement in model accuracy occurs. We choose three typical examples that all have a TM-score improvement >0.25. The first such protein is "1kafA" the DNA binding domain of the phage T4 transcription factor MotA (Li et al., 2002). It is an α/β protein (three α helices and six β strands) with 108 residues. If I-TASSER is used with only template-based contact predictions, the first-ranking model has a wrong fold with TM-score = 0.329 and rmsd = 12.6 Å, as seen in Figure 2B. The model is not improved much when compared with the best identified template (PDB ID "2rsl" chain "A") that has a TM-score = 0.330 to 1kafA as reported by the structural alignment program TM-align (Zhang and Skolnick, 2005). The structure of an N-terminal part, which is similar to the template, is correctly predicted but the remaining C-terminal segment is at a large distance from the native structure, where three helices are on the opposite side of the β sheet. When the sequence-based contact predictions are used in I-TASSER, the C-terminal segments are drawn closer to native owing to the correctly predicted contacts between the helices and β strands. This places the helices to the correct side of the β sheets, and the TM-score of the first-ranking model increases to 0.590 with an rmsd = 3.8 Å (Figure 2C).

The accuracy of template-based contact predictions for this target is 0.31 (or 0.21) and the coverage is 25% (or 23%) for contacts between C_α atoms (or side-chain centers). When the ab initio contact predictions are used to take a consensus with the template-based contacts, the accuracy of the C_α contacts increases by 71%, to 0.53, and the accuracy of the side-chain contacts doubles (=0.48), although the overall coverage slightly reduces. Remarkably, the contacts between helices and β strands were newly introduced by ab initio contact predictions that helped improve the overall topology.

Another reason for the improvement is the introduction of C_β contact predictions. LOMETS does not include C_β contacts, while the average accuracy of C_β contacts predicted by SVMSEQ is ~0.78. These changes are reflected in the C_α and side-chain center contact maps as shown in Figures 2D and 2E, respectively. The blue plus symbols (consensus contact predictions) in the lower triangle are much more symmetrical to the green asterisk symbols (representing the native contacts) in the upper triangle than the red asterisk symbols (representing the template-based contact predictions), clearly showing that the consensus contacts have a higher accuracy than the purely template-based ones. The contact maps also show that by taking a consensus of the sequence- and template-based contact predictions, many wrong template-based contact predictions (false positives) are filtered out.

The second example is the target "1zkeA," which is the HP1531 protein from *Helicobacter pylori*. Its function is unknown, and it consists of an 81-residue chain that folds into three α helices, with the N- and C-terminal α helices being nearly antiparallel. If the old I-TASSER procedure with default contacts and intrinsic potential is used, the first-ranking model has a TM-score of 0.252 (rmsd = 14.1 Å), with an incorrect topology containing four bent helices as shown in Figure 3B. After the introduction of ab initio contact predictions into I-TASSER, the first-ranking model has a correct topology (see Figure 3C) with TM-score = 0.591. The N- and C-terminal helices are now extended and correctly placed, although the middle helix still has some error that is mainly due to the incorrect secondary structural prediction (as a loop).

In this example, the improvement is not due to the increase in the accuracy of C_α contact predictions because the average accuracy of both the template-based and the consensus contact

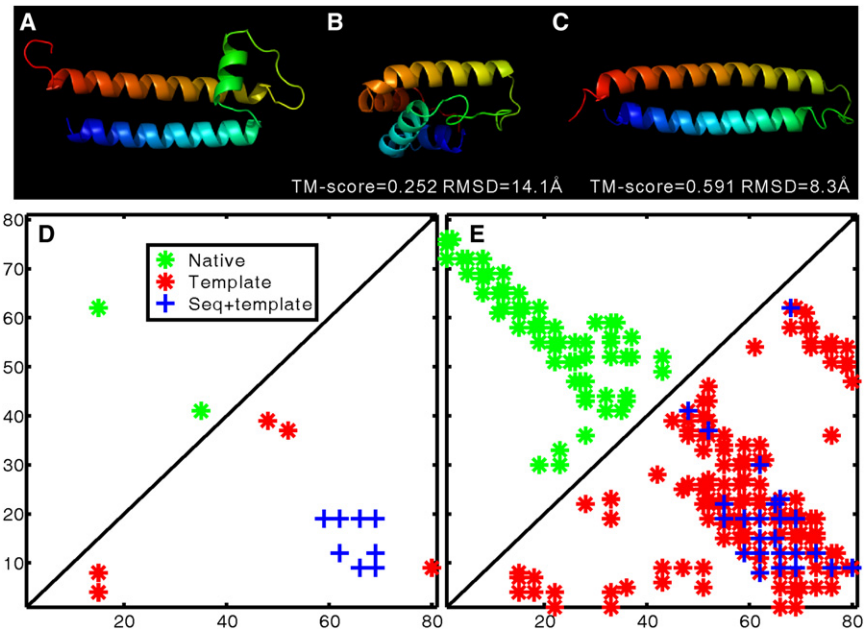

Figure 3. Illustrative Example of I-TASSER Modeling for the Target Protein 1zkeA

(A) Experimental structure.

(B) Model generated by I-TASSER without ab initio contact prediction.

(C) Model generated by I-TASSER with ab initio contact prediction.

(D) Map of native C_α contacts (*, green), template-based predicted C_α contacts (*, red), and consensus sequence- and template-based C_α contact predictions (+, blue).

(E) Map of native side-chain center contacts (*, green), template-based predicted side-chain center contacts (*, red), and consensus sequence- and template-based side-chain center contacts (+, blue).

predictions is zero, although a one-residue shift in the ab initio contacts would result in correct long-range C_α contact, as seen in the C_α contact map in Figure 3D. If we look at the side-chain center contact map in Figure 3E, however, the template-based contacts scatter all over the triangle area and most are false-positive; these contacts drove the two long helices into four bent short helices during the I-TASSER assembly process. When a consensus of sequence-based and template-based contact predictions is taken, those false positives in off-diagonal lines of Figure 3E are effectively filtered out. As a result, the average accuracy of the consensus side-chain center contact prediction doubles, from 0.20 to 0.46. Thus, the higher contact accuracy, i.e., the removal of noise, is the main reason why the model is greatly improved by the introduction of ab initio contact prediction. This example also highlights the necessity of multiple contact predictors because the single C_α SVMSEQ predictor does not help improve the modeling accuracy.

The third example is from the target "1zv1A" the dimerization domain of the doublesex protein from *Drosophila melanogaster* (Bayrer et al., 2005). It has 59 residues and its structure consists of three α helices. The "old" I-TASSER generates a first-ranking model with a TM-score = 0.454 and an rmsd = 5.4 Å, which is a considerable improvement compared to the best identified template, the second domain of "1vdu" chain A, which has a TM-score = 0.302 to 1zv1A as reported by TM-align (Zhang and Skolnick, 2005). Nevertheless, the orientation of the N-terminal and middle α helices in the I-TASSER model is incorrect. The three helices are packed in an approximately parallel and antiparallel bundle without exhibiting the subtle tilt in the native structure (Figure 4B). When the ab initio contacts are introduced into I-TASSER, the quality of the final model is improved further, with the TM-score increasing to 0.592 and rmsd reduced to 3.8 Å (Figure 4C). In this example, the intrinsic potential of I-TASSER tends to pack the helices into a more compact core with the tilted helix orientation but the strong contacts collected from template structures pushes the helices in the orientation seen in the templates. As shown in Figures 4D and 4E, the incorporation of the ab initio contact predictions eliminated most of the false-positive contacts from the templates and doubled the accuracy of the whole set of contact restraints, which resulted in an adjustment of the relative orientation of the three helices.

Sequence-Based Contact Predictions Used for Easy/Medium Targets

Although our focus was to use ab initio contact prediction to improve the structure prediction of hard targets, we have occasionally observed in our benchmark tests that the contact predictions can also have a positive effect for easy and medium targets. Here we conduct a systematic examination of the possible usefulness of ab initio contact prediction for easy and medium targets, i.e., where templates with a high Z-score are identified in threading (Wu and Zhang, 2007). Because the contacts derived from such templates usually have a higher accuracy than those generated by sequence-based methods (Wu and Zhang, 2008a), we expect that the latter becomes most useful for targets with substantial weakly aligned or unaligned regions. We will focus our analysis on these cases. Because of the imbalance of the contact accuracy of the ab initio and template-based predictions, instead of taking the consensus of different contact predictions, here we implement all contact predictions as separate restraints in I-TASSER (see Experimental Procedures). One advantage of this approach is that the restraints for the ab initio contacts will be in effect in the threading-unaligned regions, where they would otherwise be filtered out if the consensus method were used.

We collected a set of 109 targets that had been categorized by LOMETS as easy/medium targets but have >10% regions not aligned by threading; the sequence-based predictions on these proteins generated >15% new contacts that do not appear among the template-based contact predictions (Wu and Zhang, 2007). This set of proteins includes 24 α, 11 β, and 74 α+/β proteins, with lengths ranging from 31 to 273 residues. As with the hard targets, we used I-TASSER to generate models with and without including the sequence-based contacts.

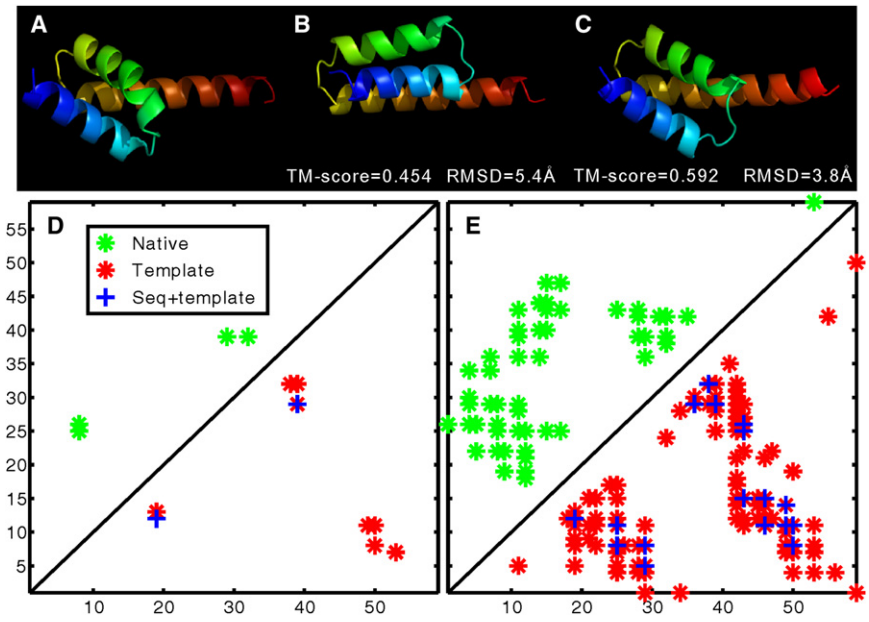

Figure 4. Illustrative Example of I-TASSER Modeling for the Target Protein 1zc1A

(A) Experimental structure.

(B) Model generated by I-TASSER without ab initio contact prediction.

(C) Model generated by I-TASSER with ab initio contact prediction.

(D) Map of native C_α contacts (*, green), template-based predicted C_α contacts (*, red), and consensus sequence- and template-based C_α contact predictions (+, blue).

(E) Map of native side-chain center contacts (*, green), template-based predicted side-chain center contacts (*, red), and consensus sequence- and template-based side-chain center contacts (+, blue).

For the 109 easy/medium targets, the average TM-score of the first-ranking I-TASSER models is 0.714 with the inclusion of the sequence-based contacts, 2.7% higher than that without including them (0.695). Because for easy/medium targets, the template-based contact predictions are usually more accurate than the sequence-based ones (Wu and Zhang, 2008a), it is not surprising that adding the sequence-based contacts yields slightly less improvement in the overall topology than that for hard targets. But clearly, it does no harm even if the sequence-based contact prediction has an obviously higher false-positive rate. Actually, the paired Student's t test p value of the two sets of models is 9.3×10^{-6}, which is statistically even more significant than what we observed in the hard proteins (p value = 0.00091). This is mainly because of the fact that in the well-aligned regions where the template-based restraints from consensus threading alignments are strong and dominant, the I-TASSER simulation is not influenced by the SVMSEQ predictions that are relatively more divergent. In the regions where threading has low confidence, the template-based restraints are usually divergent and weak, and ab initio contact predictions become dominant, which helps in improving the modeling accuracy due to the higher accuracy of predictions in these regions. Thus, overall, there are more proteins in the easy target achieving a positive TM-score improvement, which resulted in a lower p value, although the average magnitude of improvement is not as big as that in the hard targets.

To illustrate the improvement in easy/medium proteins, we take the example of "T0437" from the blind CASP8 experiment, where we tested I-TASSER with the ab initio contact predictions for the first time. This target is the yiiS protein from *Shigella flexneri*, which was categorized by CASP8 as a template based modeling-high accuracy (TMB-HA) target. In addition to an unstructured N terminus (that was ignored in the CASP8 analysis), it contains 68 residues with 2 α helices and 3 β strands. The LOMETS threading results were dominated by the template "2jz5A" for which the best threading alignment generated by HHsearch (Soding, 2005) has an rmsd = 2.30 Å and TM-score = 0.778. If we structurally align 2jz5A to the experimental structure by TM-align, the rmsd is 1.34 Å with TM-score = 0.838 (Figure 5A). Although the global topology of 2jz5A matches the target well, there is a major mismatch in the region V49-T60 (the lower part of the second β sheet). Correspondingly, there is no correct contact prediction from LOMETS in this region (Figure 5B). The sequence-based SVMSEQ contact prediction, however, generates 10 correct C_α contact predictions in this region (two others are false positive; Figure 5B). These restraints helped I-TASSER generate models with a correct β sheet structure in this region. The rmsd of the overall model is 1.13 Å, which is even closer to native than the best structural alignment (Figure 5C). In this example, although the overall accuracy of the SVMSEQ prediction is still lower than that from LOMETS, the novel contacts from the ab initio prediction improve the quality of local structures. In other regions (e.g., the N-terminal β sheet), SVMSEQ generates a number of false positive contact predictions. Because the LOMETS predictions provide strong consensus restraints, these weak false-positive predictions did not reduce the modeling accuracy in those regions.

Modeling of Hard Targets in CASP9 Experiment

The SVMSEQ contact predictions were also used to assist I-TASSER modeling in the CASP9 experiments. According to the assessor's classification, there were 26 proteins/domains that had no similar structures in the PDB and belonged to the free modeling (FM) targets. In Table S1, we present a summary of the automated I-TASSER predictions, together with 19 best servers from other groups, on the 26 FM targets/domains (T0529_1, T0531_1, T0534_1, T0534_2, T0537_1, T0537_2, T0544_1, T0547_3, T0547_4, T0550_1, T0550_2, T0553_1, T0553_2, T0561_1, T0571_1, T0571_2, T0578_1, T0581_1, T0604_1, T0604_3, T0608_1, T0616_1, T0618_1, T0621_1, T0624_1, and T0629_2), which have lengths ranging from 56 to 333 residues.

The accuracy of SVMSEQ contact prediction is highly correlated with the confidence score. For example, a C_α contact

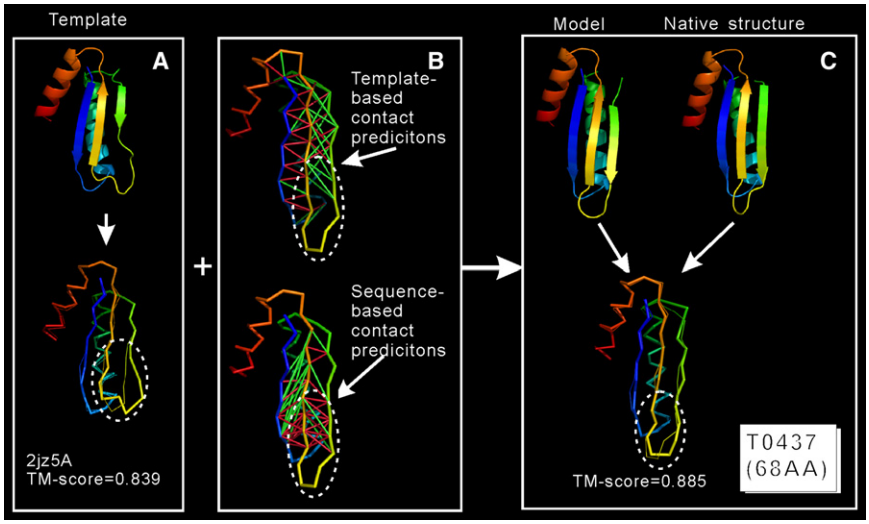

Figure 5. Analysis of I-TASSER Modeling for the CASP8 Easy Target T0437

(A) The best template protein 2jz5A (top) and its optimal structural alignment by TM-align (Zhang and Skolnick, 2005) on the experimental structure (rmsd = 1.34 Å; bottom).

(B) Template-based (top) and sequence-based (bottom) contact predictions represented by thin sticks (red color: true positive predictions; green color: false positive predictions).

(C) The I-TASSER model and its superimposition to the native structure (rmsd = 1.13 Å). In the bottom subfigures of (A) and (C), the native structure is displayed in thin lines and the template (or model) is in thick lines. Blue to red color runs from N to C terminus. The region encircled in white is where most of the improvement occurs.

prediction with a SVMSEQ confidence score >0.8 is almost always correct and that with a confidence score >0.7 has an average accuracy of 80%. However, for most hard targets, there may be very few predictions with high confidence score. For the FM target in CASP9, to cover sufficient contact predictions, we use the top 0.6**L* (*L* is the length of proteins) contacts regardless of the accuracy or more if the confidence score of additional contacts is higher than the specific confidence score cutoffs (see Supplemental Experimental Procedures). The average accuracy of the contact predictions by SVMSEQ is 27.6% with an average number of predictions = 0.606**L* for the 26 FM targets. The side-chain contacts collected from the LOMETS templates have an accuracy of only 11.9%, which confirms that the threading templates are poor for this protein set. According to the assessor's assessment (Grishin, 2010) (see also http://prodata.swmed.edu/CASP9/evaluation/domainscore_sum/human_server-best-Z.html), the total GDT-TS score of the I-TASSER server models is 39.86, which is 16% higher than the second best server and 44% higher than the third best server (Table S1).

In Figure 6, we present two representative examples: one is an α-protein and one is a β-protein. "T0553" is the PBS linker domain from *Anabaena* sp. and is 141 residues long. Although the assessor split the target into two domains ("T0553_1" and "T0553_2," both being FM domains), we analyze the target here as the whole chain because the I-TASSER server did model it as single domain and the experimental structure looks overall well packed (Figure 6, top middle). The SVMSEQ C_α contact prediction has only six contacts that have a confidence score above the threshold (see Supplemental Experimental Procedures) but all are correct. Regardless of the confidence threshold, the I-TASSER server used the top 85 (0.6**L*) contact predictions from SVMSEQ, of which 32 were correct, distributed quite evenly along the chain except for the second helix (I27-E45; Figure 6, top right). The other eight SVMSEQ predictors have a comparable accuracy and coverage. Finally, the I-TASSER server built a model with TM-score = 0.493 that is ~20% higher than the best prediction from all other servers. The TM-score of the best threading template "1k94A" identified by LOMETS is only 0.289.

Target "T0604_1" (M11-P86) is the N-terminal domain of the VP0956 protein from *Vibrio parahaemolyticus*. The I-TASSER server identified the M1-A102 stretch as a domain based on the LOMETS alignments. SVMSEQ generated 48 contact predictions of which 35 were correct, resulting in an accuracy = 72.9%, which is significantly higher than the accuracy of contacts from LOMETS (15.2%). In particular, most of the SVMSEQ contact predictions along the two β sheets are correct (Figure 6, bottom left), which drove the I-TASSER simulation to precisely pack the 3 β strands. As a result, the first-ranking I-TASSER model has a TM-score = 0.691 and rmsd = 2.66 Å, whereas the best template for this domain identified by LOMETS is "3goaA" that only has a TM-score = 0.345. In both cases of T0553 and T0604_1, the high accuracy of the composite contact predictions by SVMSEQ is essential to the success in the modeling.

DISCUSSION

Residue-residue contacts predicted purely from protein sequence have been widely regarded as being of little use in protein structure prediction due to their low accuracy, especially compared to contacts from template structures. However, the low accuracy does not imply that sequence-based contacts are useless when appropriately combined with template-derived contacts. In this study, we present ways to combine residue contacts predicted from sequence with those predicted from threading templates, and demonstrate that employing sequence-based contact predictions does improve the accuracy of the models obtained from protein structure prediction.

Using I-TASSER as an illustrative framework, we designed new contact energy terms that allow introduction of sequence-based contact predictions in the energy function of Monte Carlo simulations. The way we modified the energy function to allow for the sequence-based contacts is, however, different for hard targets and medium/easy targets. For hard targets, to reduce the negative effect of false positive contact predictions, we

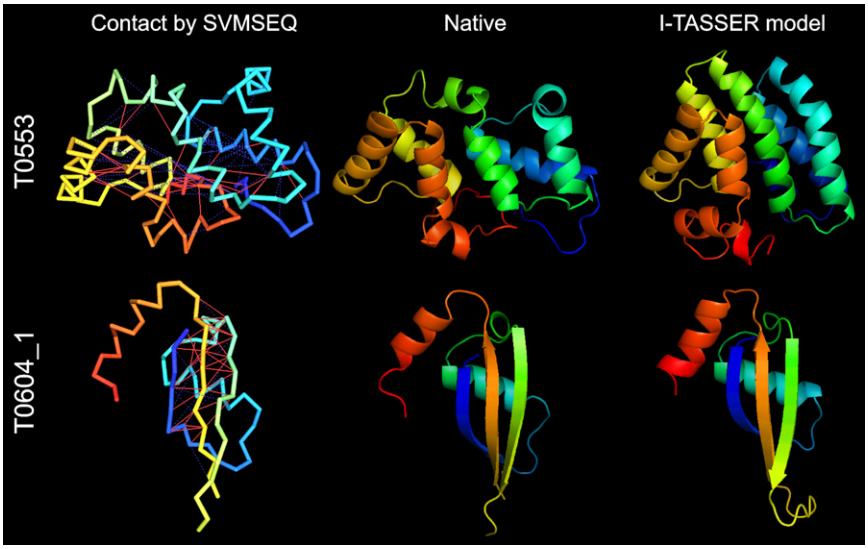

Figure 6. I-TASSER Modeling on Two CASP9 Hard Targets: T0553 and T0604_1
(Left) Backbones of the native structures with cross line representing SVMSEQ C_α contact predictions at distance <8 Å (red solid lines: true positive predictions; blue dashed lines: false positive predictions). (Middle) Experimental structures. (Right) I-TASSER models. Blue to red runs from N to C terminus. For T0553, TM-score = 0.493, rmsd = 7.3 Å; for T0604_1, TM-score = 0.691, rmsd = 2.7 Å. See also Table S1.

take a consensus of the sequence- and template-based predicted contacts so that contacts that do not have sufficient combined confidence are eliminated. The consensus method takes a weighted average of the confidence scores of predicted contacts from nine different sequence-based contacts (generated by an extended version of SVMSEQ) and two sets of template-based contacts (those for C_α and those for side-chain centers), and then uses the consensus contacts having a confidence score larger than a threshold in the I-TASSER's energy function. This solution introduces a "filtering" effect that can remove bad template-based contacts. Applying this method to the test proteins, we find that for a number of cases, I-TASSER could successfully convert a nonfoldable target with TM-score far <0.5 to a foldable one with TM-score >0.5. The overall TM-score improvement by SVMSEQ is statistically significant with the p value in Student's t test below 1.0×10^{-3}. An analysis of the CASP9 blind test performed on 26 FM targets also demonstrates the significant value of the method in the structural modeling of hard targets.

The basis for taking a consensus of the contacts predicted in different ways is that the accuracy of the template-derived and the ab initio predicted contacts is comparable for the hard targets. For easy and medium targets, however, the accuracy of the template-based contacts is, for most protein regions, higher than that of sequence-based ab initio predictions and taking a consensus might therefore significantly degrade the overall accuracy of template-based predictions. To take advantage of the ab initio contact predictions, which are mainly useful in the weakly aligned or unaligned regions for the easy/medium targets, we incorporated both sets of contact predictions into the I-TASSER assembly simulation. This way, the highly accurate template-based predictions are assigned strong weights in the well-aligned regions due to their high confidence, and can automatically eliminate the negative influence of the sparse ab initio contact predictions. In regions where template alignments are not available, sequence-based contacts become dominant and come to the rescue. Thus, the introduction of sequence-based contacts does not harm the modeling of structural regions that are sufficiently covered by template-based alignments, but is beneficial for the regions not covered by template-derived contacts. The overall TM-score improvement by SVMSEQ is shown to be statistically significant with the p value in Student's t test below 1.0×10^{-5} for the easy/medium proteins. A successful example is the easy target "T0437" in CASP8, where, using sequence-based ab initio contact predictions, the model generated by the I-TASSER server had a high accuracy (rmsd = 1.1 Å), which is closer to the native structure than even the best template in the best structural alignment.

In summary, although the accuracy of the ab initio contact prediction is generally low, it can still be used in protein structure assembly because it often complements the template-derived contacts in a way that eventually improves model accuracy. For hard targets, even though some weak templates may often be available, their number is small, and they are too distant from the target in most cases and thus may provide incorrect contact predictions. The use of sequence-based contacts, which are generated after learning from a large set of protein structures rather than a small number of possibly wrong templates, helps eliminate the false structural information coming from the templates. In the case of easier targets, there may be some regions that are not sufficiently covered by template-based contacts. Sequence-based contacts are helpful in the prediction of those regions.

Compared with the previous (less successful) attempts by us and others, the success of the procedure here is partly due to the fact that we generated nine different sets of sequence-based contact predictions, which are designed to capture the contact maps defined using various distance cutoffs and various objective atoms. The larger number and the diversity of the generated contacts allows the more reliable contacts to be selected or weighted. Meanwhile, the variation of contact predictions and integration strategies gives us a variety of options and parameter sets to optimize our approaches while keeping the training and testing proteins strictly nonredundant. However, because we generated all sets by the same SVM algorithm and based on the same training set (500 nonredundant proteins), the diversity of the contacts is probably not as large as it could be. Using a broader variety of contacts (for example, from methods relying on evolutionary information from correlated mutations) (Gobel et al., 1994; Halperin et al., 2006; Kundrotas and Alexov, 2006; Olmea and Valencia, 1997; Vicatos et al., 2005) or generated

by other machine learning methods such as neural networks (Shackelford and Karplus, 2007; Tegge et al., 2009) would probably further improve the performance of the method. Finally, although the data in this work have been presented using I-TASSER as a framework, we expect that the method can be demonstrated as a general approach to improve the accuracy of protein structure prediction in many other template-based modeling algorithms including MODELER (Sali and Blundell, 1993), ROSETTA (Simons et al., 1997), and TASSER (Zhang and Skolnick, 2004a).

EXPERIMENTAL PROCEDURES

A detailed description of the methods is provided in the Supplemental Experimental Procedures. Here, we provide a short summary.

The contact energy used in original I-TASSER is defined as

$$E_{contact\_temp} = w_1 \sum_{(i,j)\in list_1,\, j\geq i+6} f\left(d^{ij}_{ca}\right) + w_2 \sum_{(i,j)\in list_2,\, j\geq i+6} g\left(d^{ij}_{sg}\right), \quad (1)$$

where $f(.)$ is a contact energy term encouraging satisfaction of C_α contact restraints (distance cutoff d^{ij}_{ca} = 6 Å), $g(.)$ is a contact energy term penalizing violation of side-chain contact restraints (with distance cutoff = *cut*(A,B) between amino acids A and B; see Supplemental Experimental Procedures), $list_1$ and $list_2$ are predicted template-based C_α and side-chain center contact lists, respectively, and w_1 and w_2 are weighting factors.

The sequence-based contact predictions are generated by extended versions of SVMSEQ (Wu and Zhang, 2008a) that were trained on C_α, C_β atoms, and side-chain centers of mass with three different distance cutoffs (a total of nine types of SVMSEQ predictions). For hard targets, we first combine the sequence-based (SVMSEQ) and template-based (from LOMETS) contact predictions by taking a weighted average of their confidence scores:

$$\mathrm{conf}(i,j) = \sum_{n=1}^{10} w_n \mathrm{conf}_n(i,j), \quad (2)$$

where conf(i, j) is the consensus contact confidence score for residues i and j, $\mathrm{conf}_n(i, j)$ is the contact confidence score for the nth individual predictor (nine predictors are sequence-based, and the last one is template-based), and w_n is the weighting factor for the nth predictor. With the new sets of consensus contacts, we then use Equation 1 to apply contact restraints in the new I-TASSER simulation.

For easy and medium targets, because the template-based contact predictions have usually a higher accuracy than SVMSEQ, we do not construct a new set of contacts to replace the template-based ones. Instead, we add terms to the contact energy function of Equation 1 corresponding to the nine sets of sequence-based contacts, i.e.,

$$E_{contact\_consensus} = E_{contact\_temp} + \sum_{k=1}^{9} w_k \sum_{(i,j)\in list_k,\, j\geq i+6} f\left(d^{ij}_{k}\right). \quad [3]$$

Here, the same weight is used for all sequence-based contact predictors. In this way, the contacts predicted by a larger number of different predictors will naturally obtain a higher weight than those predicted by fewer predictors.

SUPPLEMENTAL INFORMATION

Supplemental Information includes one table and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.str.2011.05.004.

ACKNOWLEDGMENTS

The project is supported in part by the Alfred P. Sloan Foundation, NSF Career Award (DBI 1027394), and the National Institute of General Medical Sciences (GM083107, GM084222). A.S. was supported by a grant from the Hungarian Scientific Research Fund (OTKA PD73096).

Received: February 5, 2011
Revised: April 13, 2011
Accepted: May 12, 2011
Published: August 9, 2011

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Bayrer, J.R., Zhang, W., and Weiss, M.A. (2005). Dimerization of doublesex is mediated by a cryptic ubiquitin-associated domain fold: implications for sex-specific gene regulation. J. Biol. Chem. *280*, 32989–32996.

Bowie, J.U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science *253*, 164–170.

Cheng, J., and Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics *8*, 113.

Ezkurdia, I., Grana, O., Izarzugaza, J.M., and Tress, M.L. (2009). Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins *77* (*Suppl 9*), 196–209.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins *18*, 309–317.

Grishin, N.V. (2010). Assessment of Free Modeling in CASP9 Experiment. In The 9th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (Asilomar Conference Grounds, Pacific Grove, CA).

Halperin, I., Wolfson, H., and Nussinov, R. (2006). Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins *63*, 832–845.

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics *14*, 846–856.

Kundrotas, P.J., and Alexov, E.G. (2006). Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics *7*, 503.

Li, N., Sickmier, E.A., Zhang, R., Joachimiak, A., and White, S.W. (2002). The MotA transcription factor from bacteriophage T4 contains a novel DNA-binding domain: the 'double wing' motif. Mol. Microbiol. *43*, 1079–1088.

Li, W., Zhang, Y., and Skolnick, J. (2004). Application of sparse NMR restraints to large-scale protein structure prediction. Biophys. J. *87*, 1241–1248.

Liwo, A., Lee, J., Ripoll, D.R., Pillardy, J., and Scheraga, H.A. (1999). Protein structure prediction by global optimization of a potential energy function. Proc. Natl. Acad. Sci. USA *96*, 5482–5485.

Misura, K.M., Chivian, D., Rohl, C.A., Kim, D.E., and Baker, D. (2006). Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc. Natl. Acad. Sci. USA *103*, 5361–5366.

Olmea, O., and Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold. Des. *2*, S25–S32.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins *77* (*Suppl 9*), 89–99.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. *5*, 725–738.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Shackelford, G., and Karplus, K. (2007). Contact prediction using mutual information and neural nets. Proteins *69*, 159–164.

Shao, Y., and Bystroff, C. (2003). Predicting interresidue contacts using templates and pathways. Proteins *53* (*Suppl 6*), 497–502.

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. *268*, 209–225.

Skolnick, J., Kihara, D., and Zhang, Y. (2004). Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins *56*, 502–518.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics *21*, 951–960.

Tegge, A.N., Wang, Z., Eickholt, J., and Cheng, J. (2009). NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Res. *37*, W515–W518.

Vicatos, S., Reddy, B.V.B., and Kaznessis, Y. (2005). Prediction of distant residue contacts with the use of evolutionary information. Proteins *58*, 935–949.

Wang, G., and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589–1591.

Wu, S., and Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. *35*, 3375–3382.

Wu, S., and Zhang, Y. (2008a). A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics *24*, 924–931.

Wu, S., and Zhang, Y. (2008b). MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins *72*, 547–556.

Wu, S., Skolnick, J., and Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. *5*, 17.

Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889–895.

Zhang, Y. (2009). I-TASSER: fully automated protein structure prediction in CASP8. Proteins *77*, 100–113.

Zhang, Y., and Skolnick, J. (2004a). Automated structure prediction of weakly homologous proteins on a genomic scale. Proc. Natl. Acad. Sci. USA *101*, 7594–7599.

Zhang, Y., and Skolnick, J. (2004b). Scoring function for automated assessment of protein structure template quality. Proteins *57*, 702–710.

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309.

Zhang, Y., Kolinski, A., and Skolnick, J. (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys. J. *85*, 1145–1164.