Ab Initio Protein Structure Prediction Using Chunk-TASSER

Hongyi Zhou and Jeffrey Skolnick
Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, Georgia

ABSTRACT We have developed an ab initio protein structure prediction method called chunk-TASSER that uses ab initio folded supersecondary structure chunks of a given target as well as threading templates for obtaining contact potentials and distance restraints. The predicted chunks, selected on the basis of a new fragment comparison method, are folded by a fragment insertion method. Full-length models are built and refined by the TASSER methodology, which searches conformational space via parallel hyperbolic Monte Carlo. We employ an optimized reduced force field that includes knowledge-based statistical potentials and restraints derived from the chunks as well as threading templates. The method is tested on a dataset of 425 hard target proteins ≤250 amino acids in length. The average TM-scores of the best of top five models per target are 0.266, 0.336, and 0.362 by the threading algorithm SP^3, original TASSER and chunk-TASSER, respectively. For a subset of 80 proteins with predicted α-helix content ≥50%, these averages are 0.284, 0.356, and 0.403, respectively. The percentages of proteins with the best of top five models having TM-score ≥0.4 (a statistically significant threshold for structural similarity) are 3.76, 20.94, and 28.94% by SP^3, TASSER, and chunk-TASSER, respectively, overall, while for the subset of 80 predominantly helical proteins, these percentages are 2.50, 23.75, and 41.25%. Thus, chunk-TASSER shows a significant improvement over TASSER for modeling hard targets where no good template can be identified. We also tested chunk-TASSER on 21 medium/hard targets <200 amino-acids-long from CASP7. Chunk-TASSER is ~11% (10%) better than TASSER for the total TM-score of the first (best of top five) models. Chunk-TASSER is fully automated and can be used in proteome scale protein structure prediction.

INTRODUCTION

Protein structure is important for understanding protein function as well as being useful in drug design (1,2). To keep pace with current genome sequencing projects as well as to narrow the gap between structure determination and sequence data, computational structure prediction methods are indispensable (3). Protein structure prediction methods can be classified into three categories (4): comparative modeling, fold recognition, and ab initio methods. Comparative modeling and fold recognition methods predict protein structures based on already solved structures (5–13). These template-based methods depend strongly on the recognition of homologous/analogous templates in the Protein Data Bank (14). On the other hand, ab initio methods being template-free can, in principle, predict protein structures without the necessity of identifying a structurally related, solved protein structure.

Ab initio protein structure prediction is not only useful for providing low-resolution structures that help the annotation of protein function (13,15,16), but is also fundamentally important for understanding the mechanism of protein folding (17). While there have been many efforts dedicated to ab initio protein structure prediction, no consistently reliable algorithm is currently available (18–30). In practice, ab initio methods fall into two groups: physics-based and knowledge-based. Methods in the first group fold proteins using only physical principles (30–32). The main obstacles that physics-based ab initio protein structure prediction methods face are the lack of accurate energy functions and the requirement of extensive computational power to find the global minimum of the energy function. Despite their conceptual appeal, this method is currently not as successful as knowledge-based approaches that make use of information from solved protein structures—in particular, knowledge-based potentials (18).

Over the past several years, we have developed the Threading ASSEmbly Refinement (TASSER) methodology (13,33) for automated tertiary structure prediction that generates full-length models by rearranging the continuous fragments identified by threading. It is a kind of hybrid method that has the capacity to do template-free as well as template-based modeling. TASSER can significantly refine the initial template alignment structures provided by threading methods (33). Furthermore, TASSER has some success in modeling template-free targets of small sizes when decoupled from threading templates (34). In this work, we develop a variant of the TASSER methodology called chunk-TASSER that utilizes consensus contacts and distance restraints from ab initio folded protein chunks of supersecondary structure in addition to information extracted from threading templates. Modeling of protein chunks is much more efficient than full-length modeling in that the sampling space is much smaller, and thus large proteins can be handled. However, it does tend to favor global topologies of lower contact order. The method is assessed on a large set of 425 effectively hard targets ≤250 residues in length where the structure of the closest identified template is at best weakly related to that of the target. The results are compared to the SP^3 threading method (35,36) and the original TASSER approach (13,33). Significant improvement of chunk-TASSER over both SP^3 and TASSER is observed for

Submitted March 30, 2007, and accepted for publication May 9, 2007.
Address reprint requests to Jeffrey Skolnick, E-mail: skolnick@gatech.edu.
Editor: Jose Onuchic.

0006-3495/07/09/1510/09 $2.00
doi: 10.1529/biophysj.107.109959

these hard targets. We also tested chunk-TASSER on 21 CASP7 targets <200 residues in length that our threading algorithm classified as medium/hard targets.

METHODS

Fig. 1 shows the flow chart of chunk-TASSER. It consists of threading and fragment library generation by the SP^3 method (35,36), ab initio folding of chunks and chunk model selection, TASSER (33) full-model assembly using contact potentials and distance restraints extracted from selected chunks, threading templates that also provide the initial starting structure, and final model selection.

SP^3 method for threading and fragment library generation

The details of SP^3 were published elsewhere (35). It took part in CASP7 as well as CASP6 and is among the best single servers (36). Here we reoptimized the parameters with a full grid search on the five-dimensional parameter space. The new optimal solution (w_0, w_1, w_{2ndary}, w_{struc}, s_{shift}) is (3.5, 0.1, −1.50, 0.5, 0.7). *This resulted in the 1:1 match alignment accuracy of* 66.1% against the ProSup structure alignment benchmark (37) compared to the original accuracy of 65.3%.

Another change made to SP^3 that increases its sensitivity is the inclusion of profiles generated by PSIBLAST with a looser *e*-value cutoff of 1.0. To the target sequence, the sequence profile is replaced by the average of two profiles with *e*-value cutoffs 0.001 and 1.0, and to the templates, the structurally derived profile is replaced by the average of original and the PSIBLAST profile with an *e*-value cutoff of 1.0.

We extend the SP^3 threading method to compute local sequence similarity by computing and recording the alignment score at each query sequence position aligned to each template during threading. The position-dependent score is then smoothed by averaging over a nine-residue sliding window. For each position, nine-residue-long fragments of the top 25 scored templates are selected to form the fragment library for the ab initio folding of chunks (18).

Ab initio folding of chunks and chunk model selection

Chunks are defined as three consecutive regular secondary structure (helix or strand) segments including their enclosed two loops. The total number of *chunks for a given target is* $N_{segment} - 2$, *where* $N_{segment}$ *is the number of* regular secondary structure segments. Chunk structures are predicted independently by a fragment insertion method as in Simons et al. (18) but with our own implementation and force field. Each residue is described by its main chain atoms (N, C_α, C, O), C_β atom, and side-chain center of mass. The force field contains the following terms:

1. The DFIRE-all atom distance-dependent pairwise statistical potential for main chain and C_β atoms (38).
2. The distance-dependent pairwise statistical potential DFIRE-SCM for the side-chain center of mass (39).
3. The TASSER hydrogen-bond term based on C_α coordinates (29).
4. An excluded volume term for the main chain and C_β atoms.

The description of DFIRE-based terms 1 and 2 are published elsewhere (38,39). *The relative weight of terms 1 and 2 is set to one, since they are* based on the same principle and procedure.

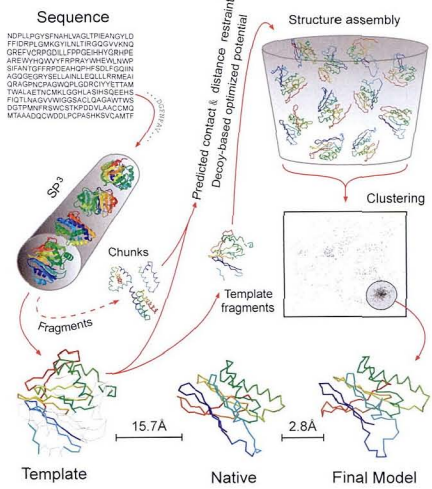

FIGURE 1 Flowchart of chunk-TASSER.

For the TASSER hydrogen-bond term, only main-chain hydrogen bonds are considered, and they are dependent on the C_α coordinates by (29)

$$E_{HB} = -\sum_{j>i} \lambda(\mathbf{u}_i \cdot \mathbf{u}_j)|\mathbf{v}_i \cdot \mathbf{v}_j|\,\Theta(i,j), \tag{1}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are unit vectors defined by the C_α coordinates: $\mathbf{l}_i = \mathbf{r}_{i,i+1}/|\mathbf{r}_{i,i+1}|$, $\mathbf{u}_i = (\mathbf{l}_{i-1} - \mathbf{l}_i)/|\mathbf{l}_{i-1} - \mathbf{l}_i|$, $\mathbf{v}_i = \mathbf{u}_i \times \mathbf{l}_i/|\mathbf{u}_i \times \mathbf{l}_i|$, where $\mathbf{r}_{i,i+1}$ is the C_α -C_α bond vector from residue i to $i+1$. The terms $\mathbf{u}_i \cdot \mathbf{u}_j$ and $|\mathbf{v}_i \cdot \mathbf{v}_j|$ impose a bias to a specific C_α -C_α bond vector orientation of regular H-bonds. The expression $\Theta_{i,j}$ defines the conditions when residue i is hydrogen-bonded to residue j:

$$\Theta = \begin{cases} 1 & \text{if } |\mathbf{r}_{ij}| < 5.8\,\text{Å}, \mathbf{u}_i \cdot \mathbf{u}_j > 0, |\mathbf{v}_i \cdot \mathbf{v}_j| > 0.43, |\mathbf{r}_{ij} \cdot \mathbf{v}_i|/|\mathbf{r}_{ij}| > 0.9, |\mathbf{r}_{ij} \cdot \mathbf{v}_j|/|\mathbf{r}_{ij}| > 0.9, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

H-bond formation also depends on the predicted secondary structure: H-bonds between residues in strand and helix are prohibited. Here, λ is a stiffness modulation factor that is used to enhance the H-bond in the better-assigned secondary structure regions. It is set to 1.5 if a regular helix or strand structure is predicted; otherwise, it is set to 1.0. We shall optimize the relative weight w_{HB} of this hydrogen-bond term relative to the other three terms.

The excluded volume term 4 is defined as

$$E_{rep}(r) = \begin{cases} 0 & r >= r_0, \\ (r_0 - r)^2/r & r < r_0, \end{cases} \tag{3}$$

where r is the distance between two atoms and r_0 is an atom-type-dependent minimal-allowed distance of two atoms taken from Ramachandran and Sasisekharan (40). *Its relative weight to DFIRE terms is set to one, since its* weight is not as important as whether it is included or not in the force field.

Monte Carlo (MC) simulated annealing is used to sample chunk structure conformations with a fragment insertion method (18) for ab initio chunk model prediction. Initially, a structure with random main-chain torsional angles is built. At each step in the MC procedure, a residue is randomly selected and then a nine- or three-residue-long fragment corresponding to the picked residue is randomly selected from the 25 fragments in the library obtained by the above-described chunk fragment generation procedure and inserted into the position by substituting the backbone torsional angles with those from the fragment. A new conformation is accepted or rejected according to the canonical Metropolis protocol.

The usual way of selecting models that are closer to native from a set of decoys generated by ab initio approaches is to cluster the models and select the most populated clusters. The success of a clustering method depends on the fact that lower free energy conformations are closer to native than higher ones. Since chunk models are not full protein models, this free energy condition may not be satisfied. Therefore, we developed and tested an alternative way of selecting chunk models that uses the information from the fragment library used for conformational sampling. The score E_{chunk} for ranking chunk models is

$$E_{\text{chunk}} = E_{\text{frg}} + w_{\text{d}} \cdot E_{\text{dfire}}/N_{\text{r}}, \tag{4}$$

where E_{frg} is calculated by the following fragment comparison method: For each residue position in the chunk model, a nine residue fragment with the given residue in the middle (less in the N- or C-terminus, for example: the fragment for the first residue will be residues 1–5, the fragment associated with the second residue considers residues 1–6, . . . , the fifth residues 1–9, . . . , and the last residue, residues $(N_r-4)-N_r$ with N_r being the last residue) is compared with the 25 corresponding fragments in the fragment library by their root mean-square deviation (RMSD). E_{frg} is the average RMSD over the 25 fragments and over all chunk-residue positions. The value w_{d} is the relative weight of the two terms and will be optimized. E_{dfire} is the DFIRE energy (38). N_r is the residue length of the chunk model. Models are ranked by their E_{chunk} score, and those with lower scores are selected.

Full-length model assembly by TASSER and final model selection

TASSER (13) represents a protein by a C_α and side-chain center of mass representation in both off- and on-lattice space. The initial full-length model is built by connecting the continuous template-provided fragments (off-lattice building blocks) by a random walk confined to lattice bond vectors. If the specified number of unaligned residues cannot span the gap, a long C_α-C_α bond remains, and a springlike force draws sequential fragments together until a physically reasonable bond length is achieved. Parallel hyperbolic Monte Carlo (MC) sampling (41) with replica exchange explores conformational space by rearranging the continuous fragments excised from the template. During assembly, the template fragments are kept rigid and off-lattice to retain their geometric accuracy; unaligned regions are modeled on a cubic lattice by an ab initio procedure and serve as linkage points for rigid body fragment rotations. Conformations are selected using an optimized force field which includes knowledge-based statistical potentials describing short-range backbone correlations, pairwise interactions, hydrogen bonding, secondary structure propensities, consensus C_α and side-chain center of mass contacts, and short and long distance restraints for C_α atoms. TASSER obtains these sequence-specific contact potentials and distance restraints from threading templates and/or fragments. The complete details of the TASSER force field are given in the literature (13,29) and references therein. Here, we give the contact potential and distance restraint terms that are relevant to this study. The contact potential between the C_α atoms or side-chain centers of mass is calculated as

$$E_{\text{contact}} = w_{r3}\Theta_5(p_{ij}-p^0)\sum_{j>i}\Theta_5(r_{ij}-6\text{Å}) + w_{r4}\Theta_6\Big(\Theta_5(p_{ij}-p^0)\sum_{j>i}\Theta_6(r_{ij}-6\text{Å})-N_{\text{cp}}\Big), \tag{5}$$

where $\Theta_5(x)$ and $\Theta_6(x)$ are step functions defined as

$$\Theta_5(x)=\begin{cases}1 & \text{if } x\geq 0,\\ 0 & \text{if } x<0;\end{cases}\qquad \Theta_6(x)=\begin{cases}x & \text{if } x\geq 0,\\ 0 & \text{if } x<0;\end{cases} \tag{6}$$

and where p_{ij} is the probability of the i^{th} residue C_α or side-chain center of mass in contact with that of the j^{th} residue obtained from threading templates/fragments, and r_{ij} is the distance between residues i and j. The value p^0 defines the minimal probability that two residues are predicted to be in contact and 6 Å is the distance cutoff for contacting residues. The first term in Eq. 5 favors pairs predicted as being in contact that are within 6 Å, whereas the secondary term penalizes predicted contact pairs that are farther apart than 6 Å when the total violation exceeds a threshold value of N_{cp}.

Local distance restraints for C_α atoms with a less-than-six-residue sequence separation are collected from the threading templates/fragments. They are incorporated into the force field by

$$E_{\text{dist}} = w_{r1}\sum_{j>i}\Theta_5(|r_{ij}-d_{ij}|-\delta_{ij}) + w_{r2}\Theta_6\Big(\sum_{j>i}|r_{ij}-d_{ij}|/\delta_{ij}-N_{\text{dp}}\Big), \tag{7}$$

where $|j-i|<6$, d_{ij} is the predicted average distance between the i^{th} residue and j^{th} residues, and δ_{ij} is the root mean-square deviation of the predictions. The second term in Eq. 7 is a penalty when the cumulative normalized deviations to the predicted distance map exceeds the number of predictions N_{dp}.

Long-distance restraints for C_α atoms are calculated in the force field as

$$E_{\text{distL}} = -w_{r5}\sum_{j>i}\sum_{k=1}^{N_{ij}}1/|r_{ij}-d_{ij}(k)|, \tag{8}$$

where $|j-i|>6$, N_{ij} is the number of distance predictions between the i^{th} and j^{th} residues extracted from the templates/fragments, and $d_{ij}(k)$ is the k^{th} distance prediction of the i^{th} and j^{th} residues.

The weights w_{r1}, w_{r2}, w_{r3}, w_{r4}, and w_{r5} as well as those of other terms in TASSER not described above were optimized against a set of nonredundant decoys (29).

Chunk-TASSER uses ab initio folded chunks as well as threading templates/fragments for consensus contact potentials and distance restraint predictions. In this study, the top 10 (ranked by SP^3 Z-scores) threading templates and the top five chunk models for each chunk are included in the prediction of contact probability and distance restraints. Contacts and distance pairs from selected threading templates and chunk models are counted with equal weights in the prediction. For example, if there are n_1 distance pairs from templates and n_2 pairs from chunk models between residue i and j, the total distance pairs for residue i and j is (n_1+n_2); and if there are n_3 pairs within 6 Å (in contact), then the contact probability p_{ij} in Eq. 5 will be $n_3/(n_1+n_2)$. When $p_{ij}>p^0=0.3$, we consider residues i and j to be involved in a real contact in the native structure, and the contact potential is effective in Eq. 5. The threading templates/fragments also serve as starting structures in the chunk-TASSER simulation as they do in TASSER. Therefore, the only difference between chunk-TASSER and TASSER is that chunk-TASSER has contact and distance information from the selected ab initio folded chunks whereas TASSER does not. The final full-length models are selected by clustering the low energy trajectories (containing 16,000 energies) using SPICKER (42).

Benchmark sets and parameter optimization

Benchmarking structures were randomly picked from the Protein DataBank released between May 28, 2004 and September, 2005. The threading library used structures deposited before May 28, 2004. All structures share <35% sequence identity with each other. No resolution and domain number requirements are imposed on these sets other than they are single-chain structures and that the predicted number of chunks ≥1. In optimizing the parameter

w_{HB} (the relative weight of the TASSER hydrogen-bond term in folding chunks) and w_d (the relative weight of the DFIRE energy term used in ranking the chunk models), we used 60 structures (optimization set) that share <35% sequence identity with the 425 testing structures ≤250 amino-acid (AA)-long (testing set). Lists of these two sets can be found at http://cssb.biology.gatech.edu/chunk-TASSER.

To make the optimization set and testing set effectively hard targets, all structures with a TM-score (43) ≥0.4 to each target are removed from the threading library (44). We optimized the parameters to maximize the total TM-score of the top five selected chunk models on the 60-protein set. The optimization procedure is done iteratively by fixing one and changing the other. For example, we set an initial value for w_{HB}, then let w_d change within a range of 0–0.1; next, we set w_d to its optimal value and let w_{HB} change within 0–5, etc. The procedure is iterated until the optimal values of w_{HB} and w_d do not change. The resulting parameters are $(w_{HB}, w_d) = (0.5, 0.01)$. For a realistic small-scale test, we used 21 medium/hard targets <200 AA long from CASP7 with threading library structures and sequence database released before the CASP7 season.

RESULTS

Chunk model selection procedure

For each target, a total of 5000 chunk models were generated regardless of the number of chunks (N_{chunk}) the target has, i.e., the number of models per chunk is $5000/N_{chunk}$. This makes the time needed to generate chunk models about the same for all targets. For a typical 150-AA-long target, it takes ~75 CPU hours on a dual core 2.0 GHz Opteron CPU to generate 5000 chunk models (~1 model/min). This can be easily distributed to several independent computers. To test how accurate the models are when chunks are included into TASSER, we select the top five best chunk models according to their RMSD to native in combination with the top 10 threading templates (selected according to threading Z-score) to construct contact potentials and distance restraints for chunk-TASSER to build full length models. The cumulative TM-score of chunk-TASSER on the 425 testing set proteins in this ideal scenario is shown in Table 1. The average TM-score of the best of top five SPICKER (42) cluster full-length models is 0.407. Approximately half of the targets have models with TM-scores ≥0.4 to native structure.

We next analyze the scoring function E_{chunk} used for ranking the chunk models. There are 2785 chunks for the 425 structure set, i.e., on average, each structure has 6.6 chunks and each chunk has ~760 models. The average linear correlation coefficient between E_{chunk} and chunk C_α RMSD to native is 0.266. Although this number is small, the p-value (45) associated with it together with the degrees of freedom 758 (= 760–762) is $<10^{-13}$, which means a very significant correlation exists. Fig. 2, *a* and *b*, show two examples of the correlation between the rank score E_{chunk} and the chunk RMSD. For some chunks like the first chunk of 1wiia shown in Fig. 2 *a*, the linear correlation coefficients are as high as 0.9. To test if use of the E_{chunk} score works better than a simple clustering method, we compare chunk-TASSER results with the following scenarios for constructing contact potentials and distance restraints:

1. Top five best chunk models.
2. Top five chunk clusters selected by SPICKER (42).
3. Top five chunk models selected by E_{chunk}.
4. Top 10 chunk models selected by E_{chunk}.

Table 1 shows the cumulative TM-scores of these scenarios. Scenario 3 is ~3% better than scenario 2 and 1.5% better than 4 by TM-score, whereas the ideal scenario 1 is ~12% better than scenario 3. The gap between the realistic scenario 3 and the ideal upper limit (scenario 1) is even larger in terms of the number of models having a TM-score ≥0.4 to native

TABLE 1 Cumulative and average TM-scores of first and best of five full-length models by chunk-TASSER and TASSER for the 425 set using different chunk selection criteria

	First model	Best of top five
Best top five chunk models	157.4 (140) 0.370	172.9 (202) 0.407
Top five cluster chunk models	136.4 (77) 0.321	150.9 (110) 0.355
Top five chunk models by E_{chunk}	140.1 (85) 0.330	153.7 (123) 0.362
Top 10 chunk models by E_{chunk}	138.1 (79) 0.325	152.2 (117) 0.358
TASSER	128.9 (54) 0.303	143.0 (89) 0.336

Numbers in parentheses are the number of targets having models with a TM-score to native ≥0.4.

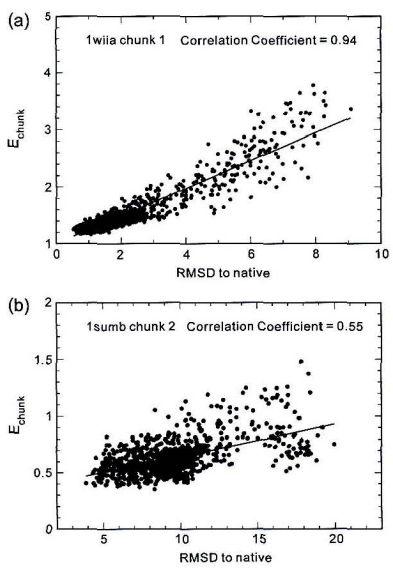

FIGURE 2 Examples of chunk ranking score E_{chunk} versus chunk RMSD to native. (*a*) 1wiia chunk 1, covering residues 34–59. (*b*) 1sumb chunk 2, covering residues 38–137.

(123 vs. 202). The cumulative TM-score by original TASSER is also shown in Table 1. Chunk-TASSER scenario 3 is ~8% better than TASSER. In what follows, we shall analyze chunk-TASSER in scenario 3 in more detail.

Overall results

Table 2 shows the same data of chunk-TASSER as in Table 1 on different subsets of the 425 set in comparison to SP^3 and TASSER. The average TM-score of chunk-TASSER for the 115 proteins ≤100 AA is 0.395, and for the 80 proteins with predicted α-helix content ≥0.5 is 0.403, whereas those of TASSER are 0.360 and 0.356, respectively. For larger proteins or less α-helix content proteins, chunk-TASSER's performance is slightly worse, but it is still better than TASSER. The percentages of proteins with models having TM-score ≥0.4 are 3.76, 20.94, and 28.94% by SP^3, TASSER, and chunk-TASSER, respectively. These percentages are 2.50, 23.75, and 41.25% by SP^3, TASSER, and chunk-TASSER, respectively, for the 80 α-proteins. Fig. 3 shows the comparison of the prediction results by TASSER and chunk-TASSER on the 425 protein set with the number of proteins having models with a TM-score greater than a given threshold. Chunk-TASSER has an improvement over TASSER for all thresholds of TM-score >0.30 (the average value of the best structural alignment between a pair of randomly related structures). We compare chunk-TASSER and TASSER using the TM-score on the 80 α-protein set in Fig. 4. In 60 cases, chunk-TASSER is better than TASSER, whereas in 20 cases TASSER is better than chunk-TASSER.

We analyzed the dependence of model TM-scores on some target properties with the results compiled in Table 3. It is clear that chunk-TASSER, similar to TASSER, still has a strong dependence on SP^3 model quality although it is slightly weaker than TASSER. That is because chunk-TASSER also uses templates from threading. The correlation coefficients of both TASSER and chunk-TASSER with predicted α-helix content are small but very significant according to the corresponding p-values. This is due to the fact that, on average, an α-protein has fewer regular secondary structure segments for a given length and lower contact order than β or α/β proteins and therefore the conformational space of an α-protein is relatively smaller and easier to access. This explains why TASSER

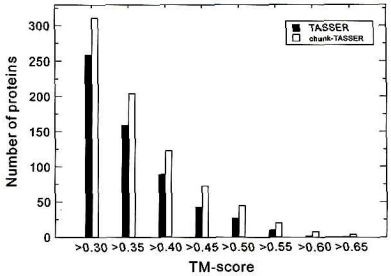

FIGURE 3 Comparison of the prediction results by TASSER and chunk-TASSER on the 425-protein set. Shown in the plot is the number of proteins having models with TM-score greater than the given threshold versus the TM-score threshold.

and chunk-TASSER perform better for α-proteins than for other types of proteins. The correlation between TASSER's performance and target size (as defined by the number of chunks) is marginally significant, whereas for chunk-TASSER, it is insignificant. As expected, both TASSER and chunk-TASSER have a small but significant dependence on contact order.

True blind predictions were made on a 21 target set from CASP7 that are <200 residues and were classified by our in-house three-dimensional jury (46) threading method as medium/hard targets (unpublished). We used library structures and a nonredundant sequence database that were released before CASP7. For this small set, we are able to compare chunk-TASSER with the ROSETTA automated approach (15,18), i.e., the ROSETTA methodology without human intervention and full atom refinement. The results are shown in Table 4. The average TM-score of the first (best) models increases from 0.272 (0.319) by the threading approach SP^3 to 0.349 (0.401) by chunk-TASSER. The average TM-scores by TASSER and ROSETTA are 0.315 (0.363) and 0.325 (0.355), respectively, for the first (best) models. TASSER and ROSETTA perform similarly on this set, whereas chunk-TASSER is 7% better than ROSETTA and 11% better than TASSER. We noted that the total TM-score 6.82 of the first models by ROSETTA in our run is very close to that of ROBETTA in

TABLE 2 Cumulative and average TM-scores of the best of top five models are shown for chunk-TASSER, SP^3, and TASSER on different subsets of the 425-protein benchmark set

Subset criteria (# of structures)	SP^3	TASSER	chunk-TASSER
≤250 AA (425)	113.0 (16) 0.266	143.0 (89) 0.336	153.7 (123) 0.362
≤200 AA (361)	97.0 (14) 0.269	121.5 (74) 0.336	131.0 (106) 0.363
≤150 AA (273)	75.5 (13) 0.277	93.4 (61) 0.342	100.6 (83) 0.369
≤100 AA (115)	34.5 (9) 0.300	41.5 (32) 0.360	45.4 (46) 0.395
Predicted α-content ≥0.5 (80)	22.7 (2) 0.284	28.5 (19) 0.356	32.3 (33) 0.403
Predicted α-content <0.5 (345)	90.3 (14) 0.262	114.5 (70) 0.332	121.4 (90) 0.352

Numbers in parentheses are the number of targets having models with a TM-score to native ≥0.4.

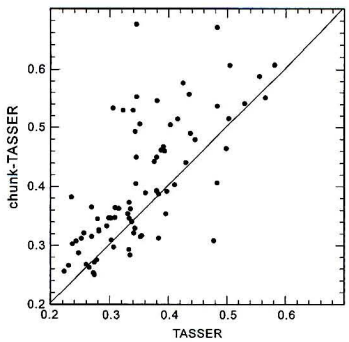

FIGURE 4 Comparison of the TM-scores obtained from chunk-TASSER and TASSER on the 80 α-proteins.

CASP7, which is 6.75. We performed a paired-samples student's t-test on the difference between chunk-TASSER and TASSER or ROSETTA predictions to see whether the difference is significant (48). The corresponding p-value of the prediction difference between chunk-TASSER and TASSER for the first (best) models is 0.025 (0.013). For both the first model and best model, the differences are significant at the 95% confidence level (p-values < 0.05). The difference between chunk-TASSER and ROSETTA for the first model on this small set is insignificant (p-value $= 0.24$), whereas chunk-TASSER is significantly better than ROSETTA for the best of top five models (p-value $= 0.025$). The reason for the insignificant difference might be due to the fact that the dataset is too small.

Representative examples

In Fig. 5, we show some examples of chunk-TASSER predictions for the best of the top five models: 1u9la is a small α-protein of 68 amino acids. The best template found by SP^3 for this protein is 11.0 Å to native. TASSER was able to refine it to 7.9 Å. Chunk-TASSER predicted the best model (the fourth model) to be 3.3 Å to native. 1u0sa is an 86-residue Å α-/β-protein with three helices packed against a four-β strand sheet. The best TASSER model has an RMSD of 7.7 Å away from native, whereas the best chunk-TASSER model (third model) is 3.7 Å to native. 1s3la is medium sized protein with 165 AA. It is a mixed α- and β-protein. The best template has a RMSD of 11.2 Å to native and the best model given by TASSER has a RMSD of 7.6 Å and a TM-score of 0.41. With the inclusion of ab initio folded chunks, chunk-TASSER refines it to a RMSD of 6.6 Å and a TM-score of 0.51. The success of chunk-TASSER for this protein is mainly due to the ability of TASSER methodology to make use of the weak signal from the threading templates, although all templates have a RMSD > 11 Å to native. 1wixa and 1sumb are all α-proteins that contain 164 and 225 residues, respectively. The best models by TASSER have RMSDs of 13.5 Å and 9.5 Å for 1wixa and 1sumb, respectively. Chunk-TASSER significantly improves the models to 7.4 Å and 6.6 Å, respectively. 1tvca is another example that shows the ability of chunk-TASSER to utilize information from templates through the TASSER methodology. 1tvca is a two-domain protein. The templates found by SP^3 all have a TM-score <0.4 to native because they cover only one of the domains. TASSER's best (ranked first) prediction for this protein is 6.1 Å and has a TM-score of 0.68. The best model (ranked first) by chunk-TASSER is almost the same in that it is 5.9 Å and has a TM-score of 0.64 to native. The relative orientation of the

TABLE 4 TM-scores of first (best of top five) models by SP^3, TASSER, ROSETTA (15,18), and chunk-TASSER for the 21 CASP7 targets

Target	SP^3	TASSER	ROSETTA*	chunk-TASSER
T0283	0.324 (0.365)	0.451 (0.451)	0.577 (0.577)	0.703 (0.703)
T0299	0.167 (0.208)	0.255 (0.324)	0.232 (0.281)	0.248 (0.255)
T0300	0.230 (0.372)	0.266 (0.405)	0.328 (0.356)	0.413 (0.420)
T0304	0.228 (0.230)	0.341 (0.343)	0.449 (0.449)	0.375 (0.387)
T0306	0.241 (0.241)	0.205 (0.267)	0.197 (0.197)	0.200 (0.342)
T0307	0.223 (0.230)	0.219 (0.318)	0.285 (0.295)	0.273 (0.317)
T0309	0.208 (0.231)	0.196 (0.248)	0.176 (0.206)	0.193 (0.293)
T0312	0.160 (0.224)	0.249 (0.249)	0.188 (0.293)	0.237 (0.353)
T0314	0.149 (0.181)	0.174 (0.281)	0.258 (0.258)	0.190 (0.275)
T0319	0.184 (0.193)	0.199 (0.231)	0.212 (0.236)	0.261 (0.276)
T0335	0.421 (0.502)	0.418 (0.436)	0.519 (0.519)	0.414 (0.454)
T0348	0.506 (0.506)	0.458 (0.508)	0.403 (0.403)	0.472 (0.499)
T0350	0.211 (0.261)	0.256 (0.350)	0.438 (0.438)	0.323 (0.365)
T0351	0.360 (0.360)	0.229 (0.277)	0.275 (0.275)	0.255 (0.385)
T0353	0.364 (0.364)	0.259 (0.266)	0.329 (0.398)	0.286 (0.293)
T0354	0.290 (0.376)	0.470 (0.482)	0.335 (0.335)	0.489 (0.495)
T0358	0.275 (0.319)	0.286 (0.358)	0.248 (0.332)	0.327 (0.368)
T0361	0.204 (0.212)	0.308 (0.308)	0.396 (0.396)	0.339 (0.363)
T0363	0.532 (0.532)	0.626 (0.628)	0.374 (0.496)	0.658 (0.678)
T0382	0.254 (0.267)	0.400 (0.400)	0.352 (0.434)	0.400 (0.436)
T0383	0.186 (0.534)	0.349 (0.502)	0.250 (0.278)	0.279 (0.473)
Total	5.717 (6.708)	6.614 (7.632)	6.820 (7.453)	7.334 (8.428)
Average	0.272 (0.319)	0.315 (0.363)	0.325 (0.355)	0.349 (0.401)

*ROSETTA was downloaded from the Baker website http://depts.washington.edu/bakerpg/ and run locally. For each target, 15,000 models were generated and the clustering procedure provided by ROSETTA with default settings was used for final model selection.

TABLE 3 Correlation coefficients of model TM-scores to native with target properties

	SP^3 model TM-score	α-content	No. of chunks	Contact order
TASSER	0.57 (5.5×10^{-38})	0.36 (1.9×10^{-14})	−0.15 (0.002)	−0.18 (0.0002)
Chunk-TASSER	0.67 (0.0)	0.26 (5.4×10^{-8})	−0.09 (0.06)	−0.17 (0.0004)

Results are based on best of top five models. Numbers in parentheses are two sided p-values (45). A p-value of <0.05 is considered significant.

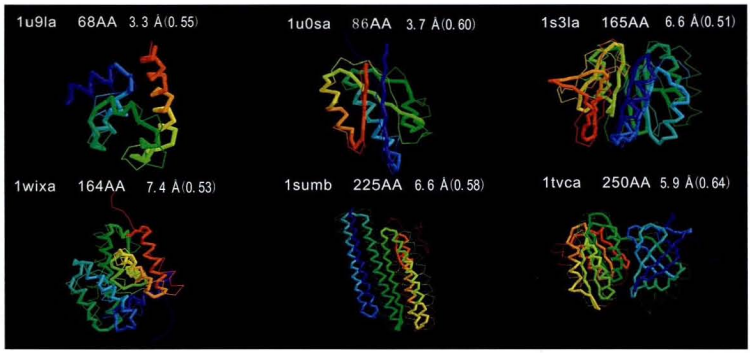

FIGURE 5 Representative predictions by chunk-TASSER. Models are superimposed onto the native structure with thick lines representing models and thin lines natives. Numbers in parentheses are TM-scores of the models to native.

two domains is correctly modeled by both chunk-TASSER and TASSER. The reason for this successful modeling is due to the 3–7 overlapping residues between templates covering different domains that provide contact and distance restraints that define the domain orientation.

DISCUSSION

We have developed a new ab initio protein structure prediction method chunk-TASSER that integrates the advantages of two of the most successful protein structure prediction methods to date: ROSETTA (18) and TASSER (33). The former utilizes the similarity of target fragments with known structures, while the latter mainly depends on the identified templates that have weak similarity with the target. The integration is realized by folding protein chunks through a fragment insertion methodology as in ROSETTA (18) and combining these chunk models with threading templates for full length modeling with the TASSER methodology (33). Chunk-TASSER is shown to perform better than the original TASSER on the 425 dataset. Furthermore, a small-scale blind test on the 21 CASP7 target set indicates that chunk-TASSER is also better than ROSETTA. We carried an informal large-scale comparison between chunk-TASSER and ROSETTA on the 425 dataset (because we do not have control over ROSETTA's database, it may contain homologous structures to the test targets for fragment generation). For a subset of 380 targets in the 425 dataset, for which ROSETTA successfully predicted structures, chunk-TASSER is ~6% (4%) better than ROSETTA as assessed by total TM-score of the first (best) models. There are several examples in the 21 CASP7 set which show that chunk-TASSER combines the advantages of both ROSETTA and TASSER. For example, chunk-TASSER performs similarly with ROSETTA for target T0283, whereas chunk-TASSER is similar to TASSER for target T0348, T0354, T0363, and T0383 for which the SP^3 template structures are also good.

Although chunk-TASSER shows significant improvement over original TASSER for hard targets, there is still much room for further improvement as indicated by the big gap between the realistic scenario 3 and the ideal scenario 1 (see Table 1). The current chunk model selection procedure is far from satisfactory. Further improvement can be achieved by more accurate selection of chunk models and generation of more near-native chunk models. As in TASSER, improvement can also be gained through more sensitive template identification from the structure library or from other modeling approaches. For example, using an approach like iterative TASSER (49,50), one can use the models from a first round of chunk-TASSER modeling in a second round of chunk-TASSER. This possibility of improvement is currently under investigation. Since chunk-TASSER, like TASSER, models a protein only in terms of its Cα atoms and side-chain centers of mass, its accuracy is limited by the force-field resolution (~1.5 Å). Therefore, development of methods to refine chunk-TASSER models and to rebuild the finer structural details using full atomic potentials is also needed.

Recently, our laboratory has developed the TASSER-Lite algorithm (51) that implements a limited time TASSER simulation. The algorithm was used in MetaTASSER (unpublished) that participated in CASP7 and was among one of the top performing servers. We are currently investigating a similar strategy for chunk-TASSER that will facilitate its public use.

We thank Dr. Adrian Arakaki for help in preparation of the figures.

This research was supported in part by grant Nos. GM-37408 and GM-48835 of the Division of General Medical Sciences of the National Institutes of Health.

REFERENCES

1. Skolnick, J., J. Fetrow, and A. Kolinski. 2000. Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.* 18: 283–287.
2. Baker, D., and A. Sali. 2001. Protein structure prediction and structural genomics. *Science*. 294:93–96.
3. Pieper, U., N. Eswar, H. Braberg, M. S. Madhusudhan, F. P. Davis, A. C. Stuart, N. Mirkovic, A. Rossi, M. A. Marti-Renom, A. Fiser, B. Webb, D. Greenblatt, C. C. Huang, T. E. Ferrin, and A. Sali. 2004. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* 32:D217–D222.
4. Murzin, A. G. 2001. Progress in protein structure prediction. *Nat. Struct. Biol.* 8:110–112.
5. Altschul, S. F., W. Gish, W. Miller, E. Myers, and D. Lipman. 1990. Basic local alignment tool. *J. Mol. Biol.* 215:403–410.
6. Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. A new approach to protein fold recognition. *Nature*. 358:86–89.
7. Vingron, M., and M. S. Waterman. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.* 235:1–12.
8. Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.-J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
9. Karplus, K., C. Barrett, and R. Hughey. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. 14:846–856.
10. Jones, D. T. 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287: 797–815.
11. David, R., M. J. Korenberg, and I. W. Hunter. 2000. 3D–1D threading methods for protein fold recognition. *Pharmacogenomics*. 1:445–455.
12. Lundström, J., L. Rychlewski, J. Bunnicki, and A. Elofsson. 2001. PCONS: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.* 10:2354–2362.
13. Zhang, Y., and J. Skolnick. 2004. Automated structure prediction of weakly homologous proteins on genomic scale. *Proc. Natl. Acad. Sci. USA*. 101:7594–7599.
14. Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. B. Jr, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein DataBank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.
15. Simons, K. T., C. Strauss, and D. Baker. 2001. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* 306:1191–1199.
16. Bonneau, R., J. Tsai, I. Ruczinski, and D. Baker. 2001. Functional inferences from blind ab initio protein structure predictions. *J. Struct. Biol.* 134:186–190.
17. Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*. 23:740–744.
18. Simons, K. T., C. Kooperberg, E. Huang, and D. Baker. 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.
19. Eyrich, V. A., D. M. Standley, A. K. Flets, and R. A. Friesner. 1999. Protein tertiary structure prediction using a branch and bound algorithm. *Proteins*. 35:41–57.
20. Lazaridis, T., and M. Karplus. 2000. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* 10:139–145.
21. Petrey, D., and B. Honig. 2000. Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Sci.* 9:2181–2191.
22. Jaroszewski, L., L. Rychlewski, W. Li, and A. Godzik. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* 9:232–241.
23. Jaroszewski, L., L. Rychlewski, and A. Godzik. 2000. Improving the quality of twilight-zone alignment. *Protein Sci.* 9:1487–1496.
24. Xia, Y., E. S. Huang, M. Levitt, and R. Samudrala. 2000. Ab initio construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* 300:171–185.
25. Yue, K., and K. A. Dill. 2000. Constraint-based assembly of tertiary protein structures from secondary structure elements. *Protein Sci.* 9: 1935–1946.
26. Lee, M. R., J. Tsai, D. Baker, and P. A. Kollman. 2001. Molecular dynamics in the endgame of protein structure prediction. *J. Mol. Biol.* 313:417–430.
27. Pillardy, J., C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kamierkiewicz, S. Odziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y.-J. Ye, and H. A. Scheraga. 2001. Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA*. 98:2329–2333.
28. Zhang, C., J. Hou, and S.-H. Kim. 2002. Fold prediction of helical proteins using torsion angle dynamics and predicted restraints. *Proc. Natl. Acad. Sci. USA*. 99:3581–3585.
29. Zhang, Y., A. Kolinski, and J. Skolnick. 2003. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.* 85: 1145–1164.
30. Chinchio, M., C. Czaplewski, S. Odziej, and H. A. Scheraga. 2006. A hierarchical multiscale approach to protein structure prediction: production of low-resolution packing arrangements of helices and refinement of the best models with a united-residue force field. *Multiscale Model. Sim.* 5:1175–1195.
31. Odziej, S., C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth, R. Kamierkiewicz, D. R. Ripoll, J. Pillardy, J. A. Saunders, Y. K. Kang, K. D. Gibson, and H. A. Scheraga. 2005. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. *Proc. Natl. Acad. Sci. USA*. 102:7547–7552.
32. Crivelli, S., T. Phillips, R. Byrd, E. Eskow, R. Schnabel, R. Yu, and T. Head-Gordon. 2000. A global optimization strategy for predicting protein tertiary structure: α-helical proteins. *Comput. Chem.* 24: 489–497.
33. Zhang, Y., A. Arakaki, and J. Skolnick. 2005. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins*. 61:91–98.
34. Borreguero, J., and J. Skolnick. 2007. Benchmarking of TASSER in the ab initio limit. *Proteins*. In press.
35. Zhou, H., and Y. Zhou. 2005. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*. 58:321–328.
36. Zhou, H., and Y. Zhou. 2005. SPARKS 2 and SP3 servers in CASP 6. *Proteins*. 7:S152–S156.
37. Domingues, F. S., P. Lackner, A. Andreeva, and M. J. Sippl. 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.* 297:1003–1013.
38. Zhou, H., and Y. Zhou. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* 11:2714–2726.
39. Zhang, C., S. Liu, H. Zhou, and Y. Zhou. 2004. An accurate residue-level pair potential of mean force for folding and binding based on the distance-scaled ideal-gas reference state. *Protein Sci.* 13: 400–411.
40. Ramachandran, G. N., and V. Sasisekharan. 1968. Conformation of polypeptides and proteins. *Adv. Protein Chem.* 23:283–438.
41. Zhang, Y., D. Kihara, and J. Skolnick. 2002. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins*. 48:192–201.
42. Zhang, Y., and J. Skolnick. 2004. SPICKER: a clustering approach to identify near-native protein fold. *J. Comput. Chem.* 25:865–871.

43. Zhang, Y., and J. Skolnick. 2004. A scoring function for the automated assessment of protein structure template quality. *Proteins*. 57:702–710.
44. Zhang, Y., I. Hubner, A. Arakaki, E. Shakhnovich, and J. Skolnick. 2006. On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. USA*. 103:2605–2610.
45. Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1989. Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge, UK.
46. Ginalski, K., A. Elofsson, D. Fischer, and L. Rychlewski. 2003. 3D-jury: a simple approach to improve protein structure predictions. *Bioinformatics*. 19:1015–1018.
47. Reference deleted in proof.
48. Marti-Renom, M. A., M. S. Madhusudhan, A. Fiser, B. Rost, and A. Sali. 2002. Reliability of assessment of protein structure prediction methods. *Structure*. 10:435–440.
49. Lee, S., and J. Skolnick. 2007. Development and benchmarking of TASSERiter for the iterative improvement of protein structure predictions. *Proteins*. In press.
50. Wu, S., J. Skolnick, and Y. Zhang. 2007. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology* 5:17.
51. Pandit, S. B., Y. Zhang, and J. Skolnick. 2006. TASSER-Lite: an automated tool for protein comparative modeling. *Biophys. J*. 91:4180–4190.