PROTEINS: Structure, Function, and Bioinformatics 68:48–56 (2007)

Benchmarking of TASSER in the Ab Initio Limit

Jose M. Borreguero and Jeffrey Skolnick*
Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30318

ABSTRACT **A significant number of protein sequences in a given proteome have no obvious evolutionarily related protein in the database of solved protein structures, the PDB. Under these conditions, ab initio or template-free modeling methods are the sole means of predicting protein structure. To assess its expected performance on proteomes, the TASSER structure prediction algorithm is benchmarked in the ab initio limit on a representative set of 1129 nonhomologous sequences ranging from 40 to 200 residues that cover the PDB at 30% sequence identity and which adopt α, α + β, and β secondary structures. For sequences in the 40–100 (100–200) residue range, as assessed by their root mean square deviation from native, RMSD, the best of the top five ranked models of TASSER has a global fold that is significantly close to the native structure for 25% (16%) of the sequences, and with a correct identification of the structure of the protein core for 59% (36%). In the absence of a native structure, the structural similarity among the top five ranked models is a moderately reliable predictor of folding accuracy. If we classify the sequences according to their secondary structure content, then 64% (36%) of α, 43% (24%) of α + β, and 20% (12%) of β sequences in the 40–100 (100–200) residue range have a significant TM-score (TM-score ≥0.4). TASSER performs best on helical proteins because there are less secondary structural elements to arrange in a helical protein than in a beta protein of equal length, since the average length of a helix is longer than that of a strand. In addition, helical proteins have shorter loops and dangling tails. If we exclude these flexible fragments, then TASSER has similar accuracy for sequences containing the same number of secondary structural elements, irrespective of whether they are helices and/or strands. Thus, it is the effective configurational entropy of the protein that dictates the average likelihood of correctly arranging the secondary structure elements. Proteins 2007;68:48–56.** © 2007 Wiley-Liss, Inc.

Key words: ab initio folding; protein folding; protein structure prediction

INTRODUCTION

For roughly 25% of the sequences in a given proteome, threading fails to identify a structural related template that can be used in subsequent modeling.[1] Since it will take many years until there is at least one deposited structure for every protein family,[2] ab initio or template-free modeling methods are the only tool available for the structure prediction of these hard cases. Among the various realizations of ab initio methods are those that employ either physics based[3–5] or knowledge-based potentials derived from a statistical analysis of protein structural databases.[6,7] While these approaches are in principle applicable to any sequence, in practice because no global template information is used, as evidenced by their recent performance in CASP6, their accuracy has been rather limited.[8] In the past, ab initio methods were validated on a relatively small number of proteins from which it is difficult to extract general trends, including the expected success rate. One trend which did emerge is that the ab initio folding of helical proteins was more successful than for proteins containing β sheets.[4–6,9] Often, this reflected a problem with the hydrogen bond term that did not work well for β sheet structures. Alternatively, for a given chain length, since the mean length of a helix is longer than that of a beta strand, the number of secondary structural elements is smaller in helical that in beta proteins.[10] This effect might have contributed to the success rate, but to establish this, a large, representative benchmark set is required.

Recently, we developed the TASSER (Threading/ASSEmbly/Refinement) algorithm, which is designed to span the comparative modeling to ab initio folding regimes. We reported the results for the large scale benchmarking in the limit of weakly homologous, single and multiple domain proteins,[11,12] where reasonable structural templates can be identified that may or may not be evolutionary related to the sequence of interest. We also explored its accuracy in the comparative modeling regime where there is a clear evolutionary relationship between the target and template structures.[13] As expected, the quality of the prediction deteriorates when the templates identified from threading are unrelated to the target structure of interest. We also applied TASSER to the comprehensive structure prediction of all human GPCRs below 500 residues,[14] as well as benchmarked

Grant sponsor: Division of General Medical Sciences, National Institutes of Health; Grant number: GM-37408.

*Correspondence to: Jeffrey Skolnick, Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, GA 30318. E-mail: skolnick@gatech.edu

Received 22 August 2006; Revised 10 November 2006; Accepted 2 January 2007

Published online 19 April 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21392

TASSER on all families of membrane proteins with solved crystal structures.[15] In all cases, the ab initio component of TASSER was applied to model the loops and tails regions lacking a template alignment. However, there has been no systematic examination of the performance of TASSER in the template free limit. Here, we address this issue for single domain proteins ranging from 40–200 residues in length.

METHODS AND MATERIALS

Construction of the Benchmark Sets

To generate the set of sequences below 100 residues, S_{100}, we retrieve a representative set of α, β, and α + β protein sequences from the PDB that are under 100 residues whose pairwise sequence identities are no higher than 30%.[1] The resulting set contains 131 α, 60 β, and 102 α + β proteins (293 total), according to SCOP.[16] For the set of sequences between 100 and 200 residues, S_{200}, we similarly retrieve α, β, and α + β sequences from the PDB with identical sequence identity cut-off (30%). The resulting set contains 230 α, 337 β, and 269 α + β proteins (836 in total). Predicted models and native structures are available on our website at http://cssb.biology.gatech.edu/skolnick/files/abinitio/

Overview of TASSER in the ab initio limit

The protein is described by a reduced protein model, where each residue is comprised of the Cα and the side-chain center of mass coordinates. Side-chain center of mass coordinates are determined with the Cα coordinates and a two-rotamer approximation. Initial Cα coordinates are generated by first projecting template coordinates of the Cα atoms onto a high coordinated cubic lattice, then connecting consecutive template fragments with an on-lattice random walk of Cα-Cα bond vectors.

Most of the energy potential terms in TASSER have been previously described.[17,18] Here, we outline its essential ingredients. The potential energy includes: (i) generic hydrogen bonding (ii) side chain contact energies between residues, (iii) short-range backbone correlations reflecting the propensity to adopt a particular secondary structure. Energy terms containing parameters that take into account the target protein's sequence are: (i) amino acid burial propensity; (ii) short-range backbone correlations and a bias in the hydrogen bond to adopt the PSIPRED[19] predicted secondary structure; and (iii) a contact potential derived from the alignment of pairs of small (11 residues) sequence fragments.[11,18] All templates whose global pairwise sequence identity is higher than 30% are a priori excluded from the calculations. Protein conformational space is searched with a variant of the replica-exchange Monte Carlo algorithm. For each target protein, 40 different simulations with a total of $8 \cdot 10^7$ Monte Carlo moves are attempted. Simulations are performed concurrently and in a broad range of temperatures (replicas). Protein conformations for replicas with similar temperatures are swapped at regular time intervals with a probability to accept the swap that is dependent on the energy difference between the two conformations. Within each replica simulation, Monte Carlo moves, comprising random selection plus coordinate change of a protein fragment ranging from two to six amino acids in size, are performed. Changes in the protein conformation are accepted or rejected based on an evaluation of the energy difference before and after the conformational change.[20]

Structural Similarity Measures

We use the root mean square deviation (RMSD),[21] the Z-score of the relative root mean square deviation (Z-rRMSD),[22] and the TM-score[23] as three metrics to assess the structural similarity of the models to the native structure. While RMSD is a more intuitive measure, the same RMSD value represents models of different quality for sequences of varying lengths. Z-rRMSD is independent of target sequence length, and from a practical point of view, we consider a protein as folded if the Z-rRMSD of the model is lower than −4.45 (P-value = 10^{-5}). In cases when only a fraction, albeit significant, of the residues fold close to native, the low RMSD signal from these residues is lost due to the high RMSD values of the other residues. The resulting RMSD and Z-rRMSD values don't differentiate these cases from a random structure to native. In contrast, the TM-score can report the subset of residues with coordinates close to native, and its value distinguishes these cases from a random global alignment. In addition, the TM-score is sequence-length independent. Again, for practical purposes we consider a protein as folded if it has a TM-score of 0.4 or higher. This value usually indicates that more than half of the residues have coordinates close to native. The average TM-score of a pair of randomly related structures is 0.17[24] and that of the best structural alignment of a pair of randomly related structures is 0.30, with a standard deviation of 0.01.[24]

Clustering Algorithm

We employ the SPICKER[25] algorithm to cluster the structures generated by TASSER, and obtain an average structure (model) for each of the top five clusters ranked by cluster density. The density of a cluster is the number of cluster members divided by the average RMSD of the members to the average structure. We report results for the average structures having the best RMSD, Z-rRMSD, and TM-score to native, termed the best model, and the average structure of the densest cluster, termed the first model.

RESULTS

S_{100} Set

The probability that the best model has an alignment to native better than some particular RMSD value [Fig. 1(a)] shows that models for α proteins are consistently more accurate than for α + β proteins, that in turn are more accurate than for β proteins. Other ab initio methods[6,9] also report that β proteins are the most difficult to fold.

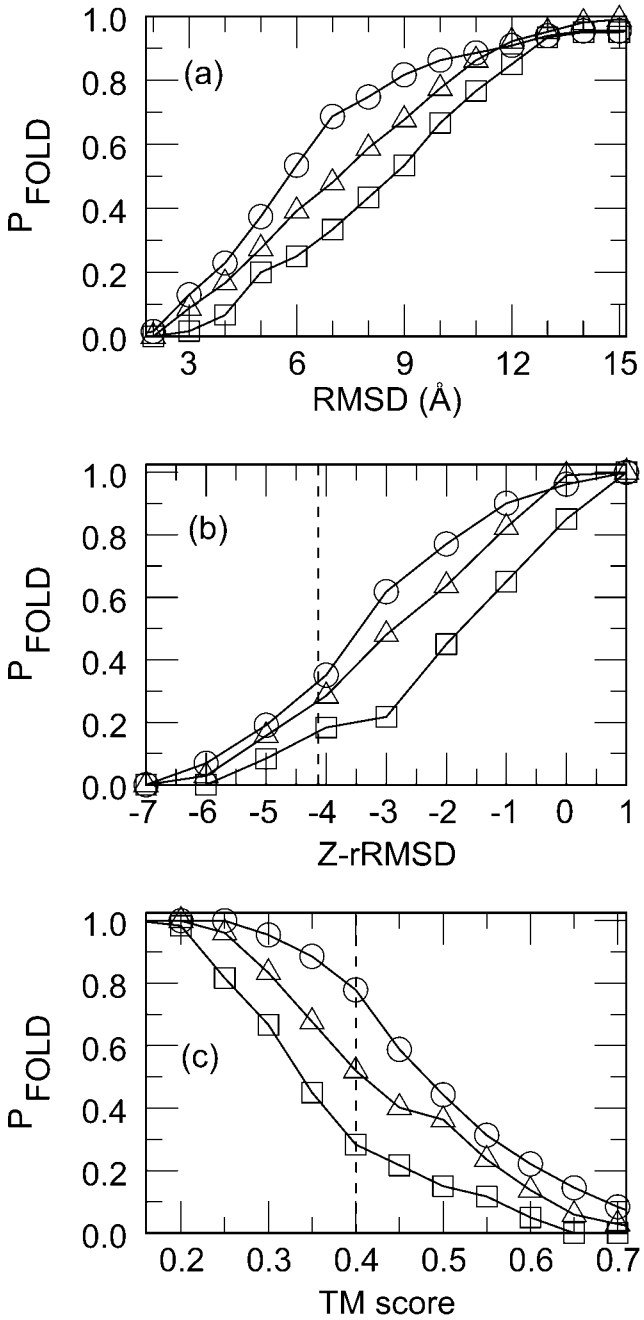

Fig. 1. (**a**) Probability of folding to native below a particular RMSD value for α (circle), α + β (triangle), and β (square) classes in the S_{100} set for the best model in the top five models. (**b**) Same as in (a), but using the Z-rRMSD measure. Dashed line indicates the Z-rRMSD = −4.25 threshold. (**c**) Probability of folding to native above a particular TM-score value.

In contrast, the accuracy of TASSER when global templates can be successfully identified (which is *not* the situation described here) is independent of secondary structure class.[11] The analogous probability distribution with the Z-rRMSD score [Fig. 1(b)] shows that TASSER predicts 33% of the best α models with a significant global alignment to native (Z-rRMSD ≤ -4.25, P-value $\leq 10^{-5}$), and the average and standard deviation of the RMSD to native of these models is 4.1 ± 0.5 Å. Corresponding success rates for α + β and β proteins are 27% and 16%, respectively. The rank distribution of the best model in these predictions (not shown) is not significantly different from a flat distribution ($\chi^2 \sim 0.5$ for the three secondary structure classes). However, we find that best and first models coincide much more often than randomly expected 32% for α, 48% for β, and 35% for α + β proteins. The average rank of the best model is independent of secondary structure class (α: 2.9 ± 0.7, α + β: 2.6 ± 0.7, β: 2.6 ± 0.7). The reason why the best and first models do not coincide even more often is that the five top models are structurally similar to each other, and the current ab initio TASSER potential may not discriminate in some cases from among a set of similar structures the one which is closest to the native state. The average and standard deviation of the TM-score among the top five models is 0.66 ± 0.16 for α proteins, 0.63 ± 0.16 for α + β proteins, and 0.50 ± 0.14 for β proteins.

Since there may be target sequences for which TASSER correctly predicts the coordinates of a significant fraction of the residues, we calculate the TM-score of the models to the native structure to detect such cases. The probability that the best model has an alignment to native better than some particular TM-score [Fig. 1(c)] shows that 78%, 42%, 28% of α, α + β, β sequences respectively have significant predictions (TM-score ≥0.4). These percentages are higher than those we obtain using the Z-rRMSD cut-off because the Z-rRMSD measure can detect only folds with overall global similarity to native. The average and standard deviation of the TM-score among the five models of the same target sequence is 0.70 ± 0.15 (α), 0.63 ± 0.16 (α + β), and 0.57 ± 0.15 (β), respectively. The structural similarity among the top five models as assessed by their average TM-score has a 0.5 correlation coefficient to the TM-score of the best model to native [Fig. 2(a)], and the correlation coefficient is independent of the secondary structure class. This structural similarity among the models arises when different initial conformations are driven via TASSER simulations towards conformations that are structurally close. This can happen if the parameterization of the potential energy reproduces some of the features of real proteins. Then, the different initial conformations converge to conformations that are structurally similar to the global minimum (native state), and therefore, are similar to each other.

S_{200} Set

Figure 3(a), for 100–200 residue proteins shows the probability of folding a target sequence below a certain RMSD threshold. Again, α proteins are the easiest secondary structure class to fold. The percentage of sequences with the best model having a significant global fold (Z-rRMSD ≤ -4.25, $P \leq 10^{-5}$) is 26%, 17%, 12% for α, α + β, β proteins [Fig. 3(b)], and the average and standard deviation of the RMSD to native of these models is 6.4 ±

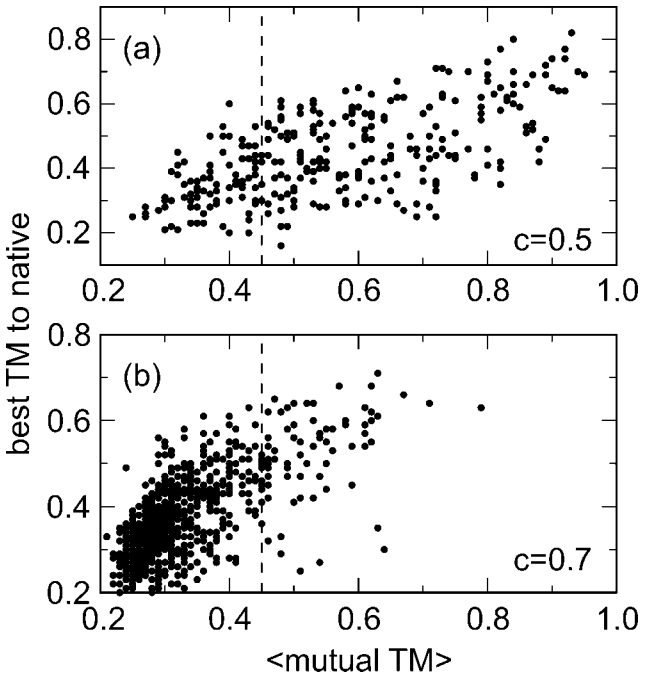

Fig. 2. (**a**) Scatter plot of the average TM-score between all possible pairs of the five top models versus TM-score-to-native of the best model for sequences in the S_{100} set. (**b**) Same as in (a), but for sequences in the S_{200} set.

0.8 Å. An analogous analysis with the TM-score shows that 55%, 38%, 21% of α, α + β, β sequences have their best model with a significant portion of the structure acceptably predicted, viz. with a TM-score ≥0.4 [Figure 3(c)]. The fraction of amino acids with coordinates close to native, the coverage, shows a strong linear correlation with TM score (coverage $\sim -0.01 + 1.50 \cdot$TM, $r = 0.92$). For instance, a model with a TM-score = 0.4 has 48% to 61% of its residues with a RMSD to native typically between 3.0 and 4.5 Å. The RMSD of this region shows also a linear correlation with TM score (RMSD $\sim$ 5.2 Å–3.8 Å·TM, $r = -0.66$). Thus, while only a few percent of targets have a global RMSD below 5 Å [see Fig. 3(a)], there are many more targets with at least half of their residues below this RMSD value, usually located in the protein's core. Using the TM-score measure, the rank distribution of the best models is not significantly different from a flat distribution ($\chi^2 \sim 0.3$ for all three secondary structure classes), but as with the previous S_{100} set, the best and first models coincide more often than the randomly expected 20% (α: 40%, α + β: 28%, β: 46%). The average rank for the best model is 2.5 ± 0.7 (α), 2.6 ± 0.7 (α + β), and 2.3 ± 0.7 (β) respectively. Finally, we find a correlation coefficient of 0.7 between the TM-score of the best model to native and the average TM-score among the top five models [Figure 2(b)], so that the average TM-score among the models is a moderately reliable indicator of a successful prediction in the 100–200 residue range.

We show in Figure 4 some representative target examples, with lengths in between 64 and 141 residues, where

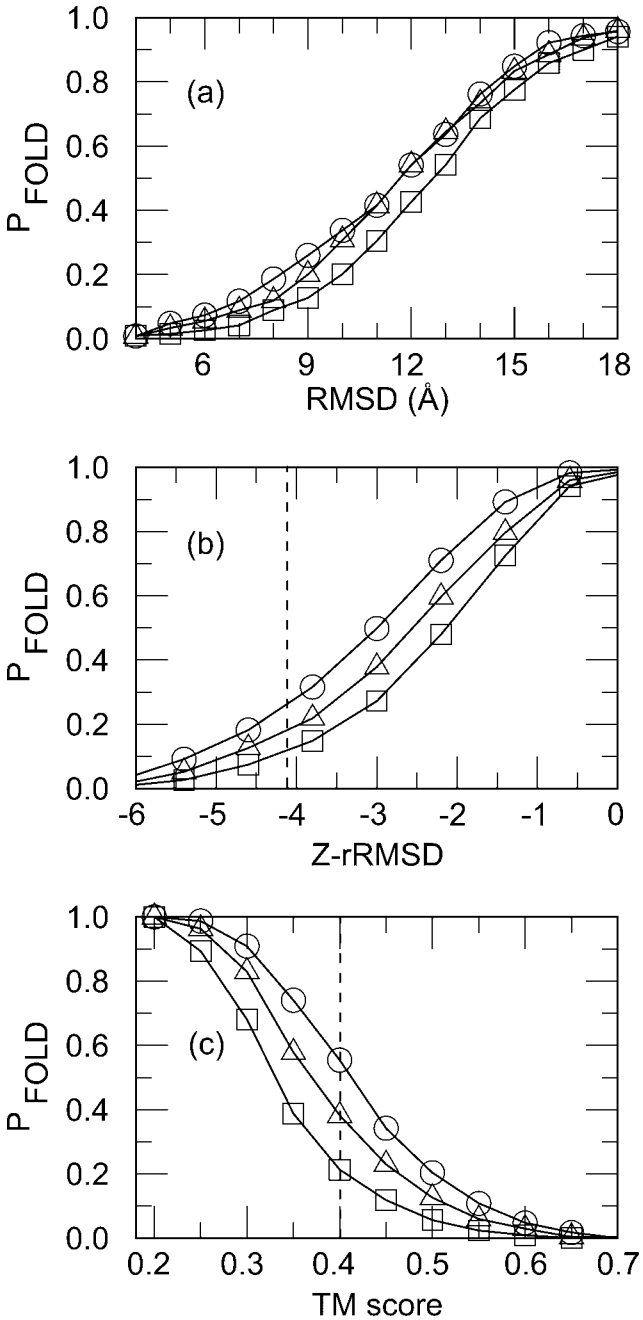

Fig. 3. (**a**) Probability of folding to native below a particular RMSD value for α (circle), α + β (triangle), and β (square) classes in the S_{200} set. (**b**) Same as in (a), but using Z-rRMSD. The dashed line indicates the Z-rRMSD = −4.25 threshold. (**c**) Probability of folding to native above a particular TM-score value.

TASSER provides a significant prediction. Figure 4(a) shows the best of the top five models (rank = 2) for Granulysin from human cytolytic T lymphocytes (PDB code 1l9lA, 74 residues), which adopts the Saposin-like fold (an orthogonal bundle of four helices). For this target, TASSER correctly predicts the positions of all Cα atoms, with a global RMSD of only 1.64 Å. Figure 4(b) shows the best of the top five models (rank = 1) for one monomer of

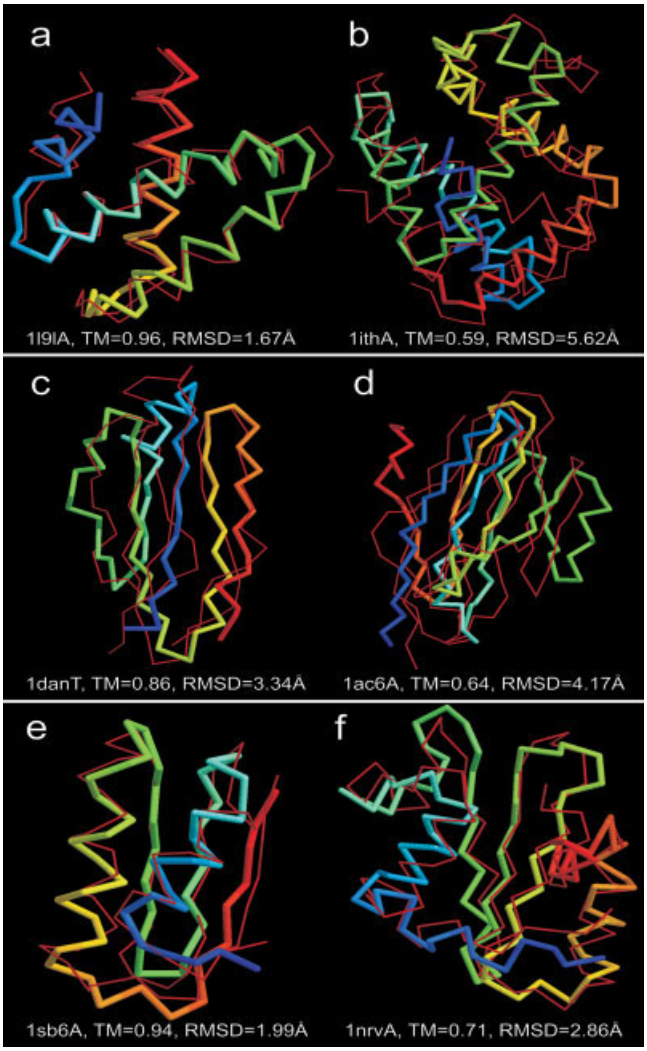

Fig. 4. Two illustrative examples of successful global superposition for each of the α (**a,b**), β (**c,d**), and α + β (**e,f**) classes. We superimpose the model (colored backbone, from red in the N-terminal to blue in the C-terminal) onto the native structure (thin red backbone). Every case shows its PDB code, TM-score, and global RMSD to native.

the homo-tetrameric hemoglobin of *Urechis caupo* (1ithA, 141 residues). TASSER predicts a structure with an RMSD of 5.6 Å and a TM-score of 0.59, with the major errors due to one long loop and the C-terminus. Two reasons, both extrinsic to the protein chain, converge in this target to account for the misoriented residues. First, the absence of an explicit representation of the Heme group in our model forces the C-terminal to occupy part of the volume left by the absent Heme group in the protein core. Second, extensive interactions with the other monomers of the biological unit produce a tight geometry in the long loop that otherwise may not be the most stable in the monomeric state. Figure 4(c,d) show two representatives of the Immunoglobulin-like sandwich fold of β proteins. Figure 4(c) shows the best of the top five models (rank = 2) for the human soluble tissue factor (1danT, 75 residues). TASSER yields a structure with a global RMSD of 3.34 Å and a TM-score of 0.86. Figure 4(d) shows the best of the top five models (rank = 1) for a mutant T cell receptor (TCR) V alpha domain (1ac6A, 110 residues) with a RMSD of 4.17 Å. This protein contains 12 strands arranged in two sheets. Figures 4(e,f) show two representative results for α + β proteins. Figure 4(e) shows the best of the top five models model (rank = 1) for the cyanobacterial copper metallochaperone, ScAtx1, (1sb6A, 64 residues) with a global RMSD to native of only 1.99 Å. Finally, Figure 4(f) shows the best of the top five models model (rank = 1) for one of the protein chains of the Grb10 Src homology 2 domain, a natural dimer (1nrvA, 100 residues). The global RMSD is 2.86 Å.

In addition to these previous examples, we show in Figure 5 two targets for which TASSER predicts a reasonably good substructure but with a high global RMSD. Figure 5(a) shows the predicted the best of top five models (rank = 2) for the human interferon β (PDB code 1au1A, 166 residues), which adopts the 4-helical cytokine fold. For this target, TASSER correctly predicts 66% of the structure with a RMSD = 3.09 Å, corresponding to three helices of the four-helix bundle plus the extra helix characteristic of the cytokine fold. The remaining residues are located in the extra helix at the N-terminal and two long, connecting loops. The model has a global RMSD of 15.2 Å, and a TM-score of 0.51. From these examples, we observe that unaligned residues tend to be located in the termini and long loops, resulting from incorrect assignment of secondary structure and/or the inherent disorder of the tails. The presence of other protein chains, prosthetic groups, metals and binding molecules/peptides in the native state may also force the protein chain to adopt some local geometry that our monomer potential ignores. Figure 5(b) shows the predicted best of the top five models (rank = 1) of the mannose 6-phosphate receptor (1c39A, 152 residues). TASSER correctly predicts 64% of the structure with a RMSD = 3.08 Å, corresponding to seven of the nine strands. An incorrect assignment of secondary structure in the first 51 residues by the PSIPRED program results in TASSER generating a helix in place of the first strand, forcing the misalignment of the N-terminal and a global RMSD to native of 14.7 Å. A more dramatic example of incorrect secondary structure assignment occurred for target protein 1a30A. PSIPRED assigned helices to four of the nine native strands, resulting in a model with different fold than native (TM = 0.27, RMSD = 11.7 Å). On the other hand, the JPRED secondary structure predictor server[28] correctly assigned eight of the nine strands. Thus, the use of a secondary structure meta-predictor could aid in improving the accuracy of secondary structure assignments. One example of correct secondary structure assignment but incorrect assembly into the native fold is target 1m4oA (TM = 0.23, RMSD = 10.1 Å), composed of three helixes and eight strands. Both the native structure and the best predicted model have a very similar radius of gyration, but the native structure contains almost double number of long range

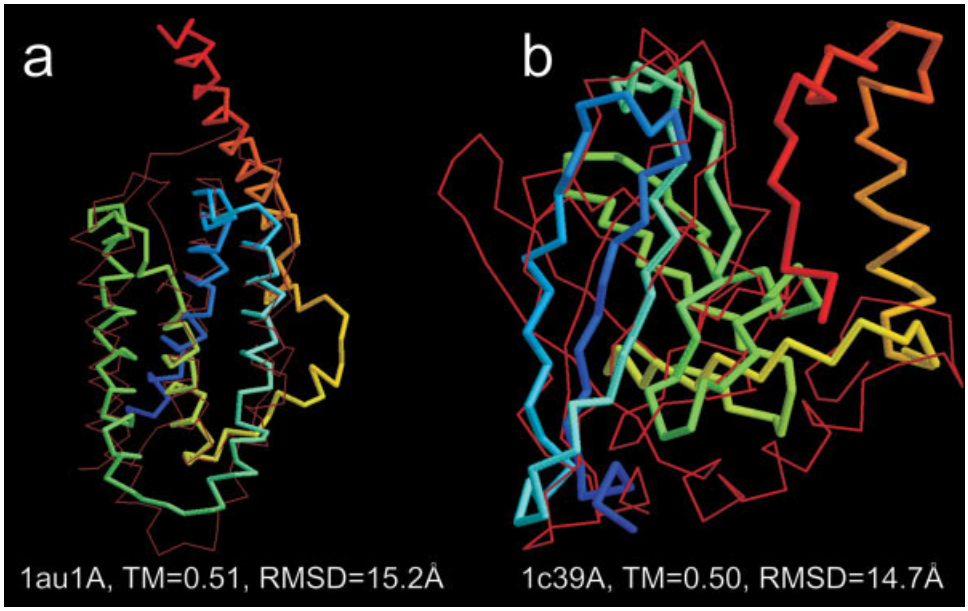

Fig. 5. Two examples of significant substructure predictions with a high global RMSD. We superimpose the model (colored backbone, from red in the N-terminal to blue in the C-terminal) onto the native structure (thin red backbone). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

contacts than the best model. Thus structures of lower contact order are predicted. The ratio (ρ) of model long-range contacts to native long-range contacts decreases from a $\rho = 0.9$ ratio for a contact between two residues separated by 30 residues to a $\rho = 0.6$ ratio for a contact separated by 160 residues. This scenario may be typical of a target protein with a pair contact potential that is not specific enough to the target. Finally, other failed predictions include target proteins with very open structures (1mhlA), or two-domain proteins for which TASSER fails to reproduce the correct domain orientation (1bcpB). We will address these more complex cases after we fine-tune ab initio TASSER to produce low RMSD models for globular, single-domain proteins.

Dangling Termini

The prediction of protein termini is of special difficulty, as it is often observed that termini do not adopt a particular secondary structure or lack interactions with the rest of the protein. There are several scenarios for which the termini may not be correctly predicted: (i) incorrect packing against the protein core; (ii) interactions with another protein; and (iii) the termini point away from the protein core in the native structure for no apparent reason. One can argue that there is information missing in the last two scenarios that may prevent TASSER from predicting the correct orientation of the termini.

We examine the disorder present in the termini of targets of the S_{200} set by counting the number of dangling residues. We define residue at position i as dangling if it has no contacts with other residues, excluding the $[i - 4, i + 4]$ range of local neighbors. In addition, we define two residues in contact if the center of mass of their respective chains is below some cut-off distance, taken from an analysis of PDB structures. By adding all the dangling residues found in the native structures of the 836 targets, we find 2682 dangling residues, which means that on average, there are four dangling residues per target. Figure 6(a) shows the probability that a target in the S_{200} set has less than some particular number of dangling residues, either in the N, C or both termini. Only a relatively few number of targets have dangling tails of considerable length. Thus, if we trim the dangling residues off all the targets and recalculate the percentage of targets in S_{200} having the best model with significant TM-score, then the percentage of acceptable predictions shows a gain of 2.1% (TM-score >0.4) with respect to the calculation including the dangling residues. This percentage gain is higher if we select the targets having dangling tails of considerable length, instead of all targets in S_{200}. Figure 6(b) shows the percentage gain in TM-score if we select targets having a number of dangling residues above certain cut-off value, trim these residues off, and then recalculate the percentage of targets with a significant TM-score (>0.4). We find that the gain is exponential with the cut-off value, as shown in the fit of Figure 6(b). This indicates the relevance of dangling tails in the prediction of the native structure of some proteins.

Since TASSER has difficulty predicting the native coordinates of dangling tails (if indeed there are any), we examine the ability of TASSER to predict whether a residue will be dangling in the native state. Models generated by TASSER correctly identify 33% of these residues as dangling. The remaining 77% of the dangling

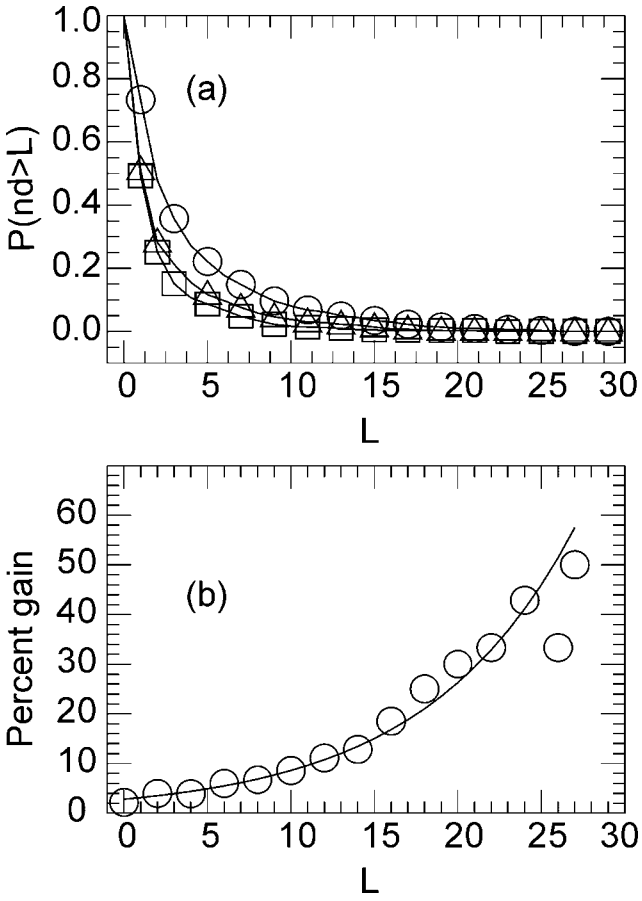

Fig. 6. (**a**) Probability that the number of dangling residues, nd, is bigger than some value L for N terminal (squares), C terminal (triangles) and both termini (circles). (**b**) For a subset of targets in the S_{200} set having L dangling residues in both termini, we show the percentage gain in the fraction of these targets having best centroid with TM-score (0.4 after we trim the dangling residues. We show only half of the circles for clarity of presentation. The curve shows and exponential fit with $r = 0.99$.

residues make some contact(s) in the TASSER generated models. Conversely, these results change only slightly if we focus in the N terminus (31%) or in the C terminal (34%). TASSER generated models also predict dangling residues which are not dangling in the native state. About 33% of the TASSER dangling residues are correctly identified. TASSER predicts a larger excess of dangling residues in the N-terminus (18% of them are correctly identified) than in the C terminus (45% identified), due to the fact that there are less dangling residues in the N terminus of native states than in the C terminus. These results suggest that inclusion of a bias for predicted intrinsic disordered regions[26] in the TASSER potential energy may increase the ability of TASSER to discriminate a dangling tail from a terminal that interacts with the rest of the protein.

Number of Secondary Structure Elements

Independent of the structural similarity measure that we use (either Z-rRMSD or TM-score), the results consistently show that the prediction accuracy decreases with increasing number of strands in the protein. Thus, α + β proteins are harder to predict than α proteins, and β proteins are harder to predict than α + β proteins. Can the observed lower folding probability of β sequences be attributable to the relative higher number of secondary structure elements (NSS), when compared to α sequences of same length? As NSS increases, we expect that the potential energy loses its ability to discriminate the unique native structure among the different ways in which the elements of the secondary structure can arrange and produce different topologies. We estimate of the order of $e^{N\cdot\ln(z)}$ arrangements, where N is the number of independent elements here taken to be the secondary structure elements, and z is the partition function of the internal degrees of freedom of a typical secondary structure element. If no energy function is used, then all arrangements are equally likely and the probability of finding the unique native structure among the set of arrangements would be at best

$$P_{\mathrm{F}} \sim 1/e^{\mathrm{NSS}\cdot\ln(z)} = e^{-\mathrm{NSS}\cdot\ln(z)} \text{ or } \ln(P_{\mathrm{F}}) \sim \mathrm{NSS}.$$

In addition to the number of secondary structure elements, structural similarity measure between the model and native structures will be adversely affected by the presence of residues that are not part of the secondary structure elements, that is, loops and dangling tails. These coil-residues may be flexible and therefore their alignment to native is of increased difficulty to predict.

To eliminate the effect of the coil-residues from our resulting TM-score values, we will only take into account those residues predicted to adopt a helix or strand conformation when calculating the TM-score. The resulting scores will assess the significance of the model topology to the native topology. Figure 7(a) shows the logarithm of the percentile probability that a sequence of given length and secondary structure class will have a significant TM-score, $\log(P_{\mathrm{F}}(\mathrm{TM} > 0.4|L,\mathrm{class}))$. The probabilities are relatively high for sequences below 150 residues, a direct consequence of removing coil-residues from the calculation of the TM-score. P_{F} has a monotonic decrease with increasing sequence length, which becomes more acute for sequences above 150 residues. We observe that the probabilities for α proteins are consistently higher than those of β proteins, with an average difference of 18% over the whole range of sequence lengths. Figure 7(b) shows the logarithm of the percentile probability that a sequence of given number of secondary structure elements will have a significant TM-score, $\log(P_{\mathrm{F}}(\mathrm{TM} > 0.4|\mathrm{NSS},\mathrm{class}))$. We observe again a monotonic decrease of the probability with increasing NSS, except for a flattening in the curve corresponding to α + β proteins in the small NSS range (NSS <5). The reason for this exception may be an insufficient number of target proteins, since one additional pseudo count for each NSS in this range, having significant TM-score would give the monotonic decrease in this NSS range. α Proteins have a slightly higher probability than β proteins in the

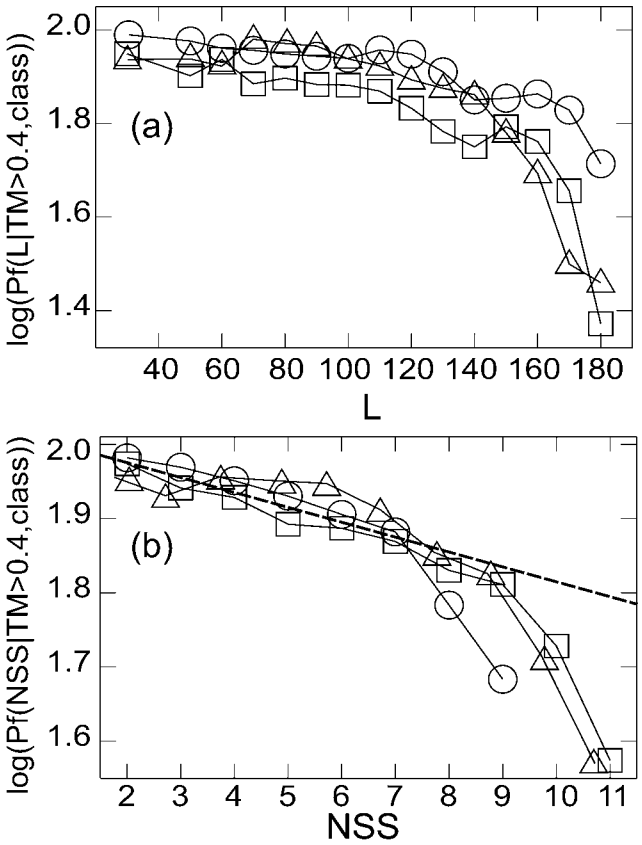

Fig. 7. (**a**) Logarithm of the percentile probability of obtaining a model with significant TM-score for α (circle), α + β (triangle), and β (square) secondary structure classes versus sequence length and (b) versus number of secondary structure elements. The dashed line represents the pure exponential decay $10^{2-0.045\ NSS}$.

1<NSS<8 range, with an average difference of 5%. For higher NSS (7 < NSS <10), β proteins have a 7% higher probability than α proteins. The overall difference between α- and β proteins over the whole NSS range is 1.3% in favor of α proteins, much smaller than the 18% when we plot the probabilities against sequence length. We show the exponential fit (dashed line) of all three secondary classes in the NSS<9 range, with a 0.91 correlation coefficient. The fit suggest that the specificity of the TASSER potential decreases exponentially with the number of secondary structures, due to the increasing number of arrangements with a potential energy similar to that of the native structure. The trend is independent of the type of secondary structure so that in the ab initio limit, TASSER has a similar accuracy for sequences of different secondary structure classes and the same number of secondary structures. Above NSS = 8, TASSER more frequently assigns an energy to the native structure that is significantly higher than the structure having the minimum energy. Thus, the probability of predicting the native structure in this scenario is lower than the case when no energy function is used and all structures are equally accessible. Hence, the sharp decrease of the folding probability in the high NSS range.

CONCLUSIONS

We have assessed the ability of TASSER in the template free limit to predict the global fold of a comprehensive set of nonhomologous α, α + β, and β sequences below 200 residues. For representative sequences below 100 residues, in the top five ranked models TASSER predicts structures whose global fold bears a statistically significant similarity to the native structure (Z-rRMSD ≤ −4.25) in 43% of α, 33% of α + β, and 19% of β proteins. For sequences in the 100–200 residue range, the corresponding success rates are 33% for α, 24% for α + β, and 15% for β proteins. Even when the entire fold is not correctly predicted, TASSER can in some cases predict the correct structure of the protein core. For sequences below 100 residues, it can generate models in the top five ranked models with significant TM-score (TM-score ≥0.4) in 64% of α, 43% of α + β, and 20% of β sequences. For sequences 100 to 200 residues in length, these percentages are 36% for α, 24% for α + β, and 12% for β proteins with the poorly predicted regions located in loops and at the N- and C-termini. Furthermore, structural similarity among the top five clusters is a moderately reliable predictor of folding success. Finally, for all sequences below 200 residues, the ability of TASSER to predict the structure of the protein core, represented by its secondary structure elements, is very similar for α, α + β, and β sequences with the same number of elements. Thus, the success of TASSER is strongly dictated by the size of the conformational space which must be searched, which is a function of the number of secondary structural elements.

While the results of this comprehensive benchmark are encouraging, clearly improvements in the TASSER force field are required. One possibility is to reparameterize the TASSER force field specifically for the template free limit (at present, the relative weights of the terms in the potential are the same regardless of whether template information is used or not[18]). Alternatively, additional terms might need to be added to the potential to enhance the sequence-structure specificity. These might include a bias towards intrinsic-disordered residues,[26] distance-dependent pair potentials,[7] as well as three-body terms.[27] Finally, the use of a secondary structure meta-predictor could improve on the current secondary structure assignments. These issues as well as others will be examined in detail in the near future.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Dr. A. Arakaki for his careful reading of the manuscript and the production of Figures 1–3.

REFERENCES

1. Skolnick J, Kihara ,D, Zhang ,Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 2004;563:502–518.
2. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–603.

3. Snow CD, Sorin EJ, Rhee YM, Pande VS. How well can simulation predict protein folding kinetics and thermodynamics? Annu Rev Biophys Biomol Struct 2005;34:43–69.
4. Simmerling C, Strockbine B, Roitberg AE. All-atom structure prediction and folding simulations of a stable protein. J Am Chem Soc 2002;124:11258–11259.
5. Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Amautova YA, Jagielska A, Makowski M, Schafroth HD, Kazmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, Scheraga HA. 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 2005;102:7547–7552.
6. Kussell E, Shimada J, Shakhnovich EI. A structure-based method for derivation of all-atom potentials for protein folding. Proc Natl Acad Sci USA 2002;99:5343–5348.
7. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726.
8. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins 2005;61 (Suppl 7):67–83.
9. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
10. Zhu ZY, Blundell TL. The use of amino acid patterns of classified helices and strands in secondary structure prediction. J Mol Biol 1996;260:261–276.
11. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
12. Zhang Y, Skolnick J. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. Biophys J 2004;87:2647–2655.
13. Pandit S, Skolnick J. TASSER-Lite: an automated tool for protein comparative modeling. Biophysical J, in press.
14. Zhang Y, Devries ME, Skolnick J. Structure modeling of all identified G protein-coupled receptors in the human genome. PLoS Comput Biol 2006;2:e13.
15. Zhang Y, Skolnick J. Tertiary structure predictions on a comprehensive benchmark of medium and large size proteins. Biophysical J, in press.
16. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothla C. SCOP: a Structural Classification of Proteins database. Nucleic Acids Res 1999;27:254–256.
17. Zhang Y, Arakaki A, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins, in press.
18. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164.
19. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16:404–405.
20. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002;48:192–201.
21. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vecotrs. Acta Cryst A 1978;34:827–828.
22. Betancourt MR, Skolnick J. Universal similarity measure for comparing protein structures. Biopolymers 2001;59:305–309.
23. Zhang Y, Skolnick J. A scoring function for the automated assessment of protein structure template quality. Proteins 2004;57:702–710.
24. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci USA 2006;103:2605–2610.
25. Zhang Y., Skolnick J., SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004, 25(6):865–71.
26. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006;7:208.
27. Li X, Liang J. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. Proteins 2005;60:46–65.
28. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics 1998;14:892–893.