

Materials and Methods

Tools for the generation and analysis of folded protein structures.

To generate the ensemble of compact homopolypeptide conformations, a protein model, protein force field and conformational search scheme are required. Then, given an ensemble of structures, we require clustering algorithms to select representatives from a diverse set of conformations. Finally, we need a sensitive algorithm to identify the structural analogs of the computer-generated conformations from a representative set of structures in the *PDB* and vice versa. In what follows, we describe the tools that address these issues.

Protein models, force fields and conformational search protocols.

Since there does not as yet exist a perfect force field capable of folding an arbitrary protein to its native structure, it is important to assess the sensitivity of the results to the particular force field and protein representation used. If the results are insensitive to protein representation (i.e. whether or not a continuous space, detailed atomic model or a C_α plus C_β , *CAS*, lattice based model is used), the particular force field, and the conformational search scheme, then this is suggestive that the resulting conclusions are robust. If, on the contrary, qualitatively different results emerge depending on the details of how the models are constructed, then one would have to be cautious in interpreting how well the simulations mimic the completeness of protein structural space. In what follows, we present an overview of the two models, their associated force fields and conformational search protocols.

Detailed atomic model, force field and conformational search scheme.

The first model employs a detailed atomic representation of the protein and accounts for all heavy atoms (*1, 2*). The backbone atoms are typed as peptide N, C_α , carbonyl C, and O. We consider a homopolypeptide where each side chain has a C_β atom. Atomic radii were calibrated to achieve realistic Ramachandran plots. Bond lengths, chain connectivity, and excluded volume, are always maintained. The potential contains two components: pairwise interactions and hydrogen bonding. Two-body interactions are represented by a square well, contact potential that does not require the knowledge of native structure (if any) of the protein to be simulated. Atoms A and B, with hard-sphere radii r_A and r_B , separated by a distance D are in contact if $0.75(r_A + r_B) < D < 1.8(r_A + r_B)$ and interact with a potential E_{AB} . In addition to the pair potential, a backbone hydrogen

bonding, E_{HB} , function is used to ensure proper secondary structure formation. The relative strength of hydrogen bonding and pairwise interaction is controlled by α , which balances polymer elongation and collapse.

As shown in Figure 5, a hydrogen bond is counted when the four atom pairs associated with donor nitrogen and hydrogen and acceptor carbonyl oxygen and carbonyl carbon are within a square well, eliminating the need for angle calculations and increasing computational efficiency. The indicated distances are as follows: d1 is the distance between the donor nitrogen and acceptor oxygen, d2 is the distance between the donor nitrogen and acceptor carbonyl carbon, d3 is the distance between the donor hydrogen and acceptor oxygen, and d4 is the distance between the donor hydrogen and acceptor carbonyl carbon. In this particular variant, interactions are only allowed between residues i and residues $i+2$, $i+3$ and $i+4$. Thus, hydrogen bonds using this potential are essentially limited to helical conformations. The hydrogen bond energy, E_{HB} , (in dimensionless units) for specific distance parameters are given in the accompanying hydrogen bond potential file tabulated in the format d1, d2, d3, d4, and E.

The total energy of a given conformation, E_{total} , is given by

$$E_{total} = \alpha E_{HB} + (1 - \alpha) E_{AB} \quad (1)$$

Based on previous work, we set $\alpha=0.9$. For a test set of 7 proteins, in the native state the ratio of E_{AB} to E_{HB} is roughly 3:1, even though the strength of an average atom-atom contact is only 1/10 that of the energy of a hydrogen bond.

The dihedral move set (1, 2) satisfies detailed balance (1, 3), with the amplitudes of moves drawn from a Gaussian distribution with zero mean and 2° variance for the backbone and 10° variance for the side-chain χ angles. Conformations are searched using replica exchange Monte Carlo (4).

CAS reduced protein model, force field, and conformational search scheme.

The conformation of a protein in the CAS model is described by its C_α atoms and the side chain centers of mass (SG), taken here to be a C_β (5). The force field employed in this study is a subset of the full *TASSER* force field (a protein structure prediction algorithm) (5) and consists of: (1) Uniform hydrophobic interactions between side chain residues for the purpose of generating compact conformations; (2). Hydrogen bonding described in further detail below; (3). Excluded volume; and (4). An energetic bias to a

pre-assigned secondary structure. For helices this is an energetic bias towards loosely defined helical conformations, whereas for β strands this is a weak bias towards extended conformations. The secondary structures regions are assigned in the following way: Each secondary structure fragment (helix or strands) is followed by a short loop. The sizes of the secondary structures and loops are randomly taken from a distribution derived according to *PDB* statistics (Figure 6). For $\alpha\beta$ proteins, the helices and strands are randomly ordered, each with 50% probability of assignment.

Hydrogen-bond interactions in the CAS model.

Since hydrogen bonding is essential to the results, we present the explicit details of the hydrogen bond scheme. The strength and occurrence of H-bonds in the CAS model are defined by the contact order (CO, residue distance along sequence) and relative orientation and geometry of donor and receptor residues. As shown in Figure 7, if we define $cc=\bar{c}_i \cdot \bar{c}_j$, $bb=\bar{b}_i \cdot \bar{b}_j$, $pp=\bar{p}_i \cdot \bar{p}_j$, $qq=\bar{q}_i \cdot \bar{q}_j$, $bri=|\varepsilon\bar{b}_i - \bar{r}|$, $brj=|\varepsilon\bar{b}_j - \bar{r}|$, and $r=|\bar{r}|$, the hydrogen bonding energy in the CAS model can be calculated by the following automated procedure:

For a hydrogen bond between residues that are located in an α -helix,

If $ss \neq \beta$ and $CO = 3$ and $bb > bb_\alpha$, then

If $cc > cc_\alpha$ and $r < r_\alpha$ and $pp > pp_\alpha$ and $qq > qq_\alpha$, then

$$E_{HB} = \lambda_\alpha (1 - |cc - cc_{\alpha 0}|) (1 - |bb - bb_{\alpha 0}|) / [(1 + |bri - br_{\alpha 0}|)(1 + |brj - br_{\alpha 0}|)]$$

For a hydrogen bond between residues that are located in antiparallel β strands,

If $ss \neq \alpha$ and $CO > 4$ and $bb < -bb_\beta$, then

If $cc > cc_\beta$ and $r < r_\beta$ and $pp < -pp_\beta$ and $qq < -qq_\beta$, then

$$E_{HB} = \lambda_\beta |bb| cc / [(1 + bri/2)(1 + brj/2)]. \quad (2)$$

For a hydrogen bond between residues that are located in parallel β strands,

If $ss \neq \alpha$ and $CO > 20$ and $bb > bb_\beta$, then

If $cc > cc_\beta$ and $r < r_\beta$ and $pp > pp_\beta$ and $qq > qq_\beta$, then

$$E_{HB} = \lambda_\beta bb * cc / [(1 + bri/2)(1 + brj/2)]$$

Here ' $ss \neq \alpha(\beta)$ ' means neither putative donor nor receptor residues are assigned as an α -helix (β -strand). $\lambda_{\alpha(\beta)} = 1$ if both donor and receptor residues are each assigned as an α -helix (β -strand); otherwise $\lambda_{\alpha(\beta)} = 0.5$. $\varepsilon = 5.0 \text{ \AA}$ for α -helix and 4.6 \AA for β -sheet. All other parameters are calculated from the statistics of 100 high resolution structures in *PDB* (50

in α -proteins and 50 in β -proteins according to DSSP assignments), and are summarized in Table 2.

This H-bond scheme is mainly designed for the backbone atoms inside α -helices and between β -strands. But rarely some backbone atoms in the loop or tail regions may also form a H-bond with other backbone atoms if their relative geometry satisfies any of the above conditions.

Starting conformations, move sets and sampling

The protein chain is confined to a high coordination number lattice (5), and the only input for the 150 chains is the secondary structure assignments. All conformations start from a random, extended coil. Unlike the full *TASSER* algorithm (5), no fragments are excised from the *PDB*, nor are idealized secondary structural elements used. Parallel Hyperbolic Monte Carlo sampling (6), an improved variant of Replica Exchange Monte Carlo(4), is used to explore conformational space. Conformational updates consist of 2-6 bond movements and multi-bond sequence shifts (5).

The *SPICKER* clustering algorithm.

To select representative structures from the trajectories of either the atomic or CAS protein models, we employ the structure clustering algorithm, *SPICKER* (7). *SPICKER* is a greedy algorithm where members of each cluster are selected as follows: For a given pairwise RMSD cutoff, R_{cut} , the first cluster contains the structure with the most neighbors (that comprise the “cluster center structure”), as well as the structures of all its neighbors. The second cluster contains the structure with the second largest number of neighbors, excluding all members of the first cluster, as well as the structures of all its neighbors, etc. As shown elsewhere, *SPICKER* (7) has been extensively benchmarked (5, 8-10) and found to show improvement over previous clustering algorithms (11) in selecting representative lowest free energy structures.

***TM-align*: A new structural alignment algorithm.**

Since our goal is to compare the computer-generated compact structures with protein structures found in the *PDB*, a tool to generate structural alignments between them is needed. Structural alignments assess the structural similarity between a pair of structures, where the set of equivalent residues required for the comparison is not *a priori* given. Therefore, an optimal alignment needs to be identified; this is in principle NP-hard (12). A variety of different heuristic approaches have been proposed to search for this

“best” structure alignment given a metric of structural similarity. These differ mainly in the metric used to assess the alignments and the search algorithm that identifies the putative best alignment. Representative approaches include *DALI* (13), *CE*(14), *STRUCTAL* (15) and *SAL*(16). For example, *STRUCTAL* (15) and *SAL* (16) use the *inter-structural* residue-residue distance based Levitt-Gerstein, *LG*, score matrix and maximize the cumulative *LG*-score (15) or relative root-mean-square-distance, RMSD, (17) by a heuristic iterative Needleman-Wunsch dynamic programming approach (18). During the iterations, both algorithms use a rotation matrix that is constructed to minimize the RMSD between a pair of structures. However, the average RMSD of randomly related proteins depends on the length of compared structures, which renders its absolute magnitude meaningless (17).

The recently proposed *TM-score* (19) overcame this issue, where the *TM-score* is defined as:

$$\text{TM-score} = \text{Max} \left[\frac{1}{L_{\text{Target}}} \sum_i^{L_{\text{ali}}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\text{Target}})} \right)^2} \right]. \quad (3)$$

Here, L_{Target} is the length of target protein that the other structure is aligned to; L_{ali} is the number of aligned residues; d_i is the distance between the i^{th} pair of aligned residues.

$d_0(L_{\text{Target}}) = 1.24 \sqrt[3]{L_{\text{Target}} - 15} - 1.8$ is a distance parameter that normalizes the distance so that the average *TM-score* is independent of protein size for a random structure pair. The *TM-score* (whose range is (0,1]) has an average value of 0.17 for a pair of randomly related structures (19) and a value of 1.0 for two identical protein structures.

Our recently developed structural alignment algorithm *TM-align* (20) exploits these insights and extends the approaches of *STRUCTAL* (15) and *SAL* (16), by using the *TM-score* rotation matrix to speed up the identification of the best structure alignments. When the *best structural alignments* between a pair of randomly related structures are considered, the average *TM-score* is 0.30 and the standard deviation is 0.01. We examine this issue further in Figure 8, where we show the histogram of the *TM-score* of the best structural alignment of 158 distinct, compact conformations of 200 residue freely jointed

chains, FJC, to the representative *PDB* template library of 6,967 proteins that cover the *PDB* at a 50% sequence identity cutoff. Interestingly, the average *TM-score* of 0.30 is independent of the particular set of unrelated structures that are compared. It is the same average value for structural alignments of FJC to *PDB* structures or for structural alignments of unrelated single domain proteins in the *PDB*. Of course, we have to remove the set of related protein structures in the *PDB* to calculate the standard deviation for randomly related structure pairs; this is why the FJC are used to obtain this value.

The *TM-align* algorithm is about 4 times faster than *CE* (14) and 20 times faster than *DALI* (13) and *SAL*. On average, the resulting structure alignments have higher accuracy and coverage than those provided by these most often-used methods. Here, this approach is used to identify the optimal structural alignments between the computer-generated conformations and *PDB* structures. Beside the RMSD and the alignment coverage, *TM-align* also reports the *TM-score*. The *TM-align* program can be downloaded from our website at <http://www.bioinformatics.buffalo.edu/TM-align/>.

TASSER modeling starting from compact homopolypeptide templates.

We selected the 10 proteins in the *PDB150* set that have the worst structural matches to our 15,000 member library of compact, homopolypeptide models, based on their *TM-score* that may be found at http://www.bioinformatics.buffalo.edu/sticky_alignments/150_pdb/summary15000 (see Table 1). We then employed *TASSER* to build full-length models for these 10 proteins starting from the *TM-align* structural alignments, where the spatial contact and distance restraints are taken from the selected compact, homopolypeptide templates to guide the *TASSER* simulation. The results are summarized in Table 1. Although the *TM-score* of the structural alignments is modest (~0.37), the alignments provide the correct topology for around 2/3 of the core-region residues. Among the reasons for the modest *TM-score* is the presence of long tails in a number of the templates. One of the tasks of *TASSER* is to connect the continuous fragments by building appropriate loops. The average global RMSD from the corresponding *PDB* structure of the first *TASSER* model is 5.11 Å. Most targets have a RMSD below 6.5 Å, except for 1fjgl that has a long, unfolded tail in the N-terminus; see Figure 9. If we cut the tail (from PRO1 to SER18), then the global RMSD is 5.8 Å.

TASSER also considerably improves the topology of the structurally aligned regions. The average RMSD to native of the compact homopolypeptide templates and the

refined *TASSER* models is 4.79 Å and 4.15 Å, respectively, from the *PDB* structure for the same aligned residues; see Table 1. Figure 10 shows a representative example, 1at0_, which demonstrates a significant improvement due to *TASSER*'s ability to readjust the protein's core. There is only one target, 1nkws, where the RMSD of the *TASSER* model is higher than that of the TM-align structural alignments. In this example, see Figure 11, *TASSER* places the N-terminal in the wrong direction, a known problem of *TASSER* in modeling the orientation of long tails as well as mutual orientation of protein domains.

Three-dimensional active site template library.

For the detection of substructures whose geometry resembles enzyme active sites, we use an updated version of our library of Automated Functional Templates or *AFTs* (21). The *AFTs* are based on the three-dimensional arrangement of residues important for defining the molecular function of a given enzyme. The procedure for building an *AFT* consists of three steps: (1). Retrieval of functionally important substructures from all *PDB* structures associated with a specific EC number, (2). Generation of tentative distance-based templates describing the active site, and, (3). A specificity assessment of the *AFTs*. Previously, we defined an *AFT* as the spatial arrangement of k functional building blocks ($3 \leq k \leq 5$), each composed of the C_α atom of a functional residue, the two adjacent C_α atoms, and (for non-glycine residues) one pseudo-atom corresponding to the side chain center of mass (SG). To better suit the present analysis, since we focus on sticky homopolypeptides (whose only side chain heavy atom is a C_β), we use the C_β rather than the SG pseudo-atom. Also, to speed up the calculations, we set $k=3$. We now base the *AFTs* on functionally important substructures where all involved residues are annotated with the ACT_SITE key name in the Swiss-Prot (22) database (indicating a direct contribution to the enzyme's activity). Finally, we use a stricter definition for the restrictive cutoff to establish the significance of a match: the maximum distance RMSD, dRMSD, (the average root mean square difference between corresponding distances in the compared substructures), observed between a true positive hit and the corresponding *AFT*. The permissive cutoff is defined so that the expected number of false positive matches is less than 0.005 per true negative structure. Following this procedure, a library of 150 *AFTs* associated with 118 different EC numbers is obtained.

Scanning of native structures and sticky homopolypeptides with the *AFT* library.

We used our *AFT* library to scan, in a sequence-independent manner, three sets of structures: (1). The top 5 clustered structures generated by the simulation for each of the one hundred and fifty 200-residue homopolypeptides (750 structures), (2). The same number of native structures from the *PDB*, and (3). The representative set of 3500 compact homopolypeptide structures used to assess the completeness of the compact fold library with respect to the *PDB*. The 750 native structures are nonredundant (at the level of 40% sequence identity), with lengths ranging from 163 to 230 residues and an average length of 199.7 residues. To eliminate direct effects due to evolution, before scanning the set of native structures with a given *AFT*, we remove those that correspond to enzymes whose EC numbers share the first two components of the EC number of the *AFT* under analysis.

By way of illustration, we show in Figure 11, the relative frequency distributions of substructures of the top 10 *AFTs* that have the best match to one of the 750 sticky homopolypeptide structures. Given that the sticky homopolypeptide structures are generated at random with no knowledge of the *AFT* geometry, the results suggest active site geometries at the level of $C_{\alpha}s$ and $C_{\beta}s$ arise from the packing of compact secondary structural elements and at the level of substructure geometry are not special. We find that there is no relationship with the set of enzymes functions that have the best match and ancient enzymes.

Structural Alignment Library

In the library of aligned structures (see http://www.bioinformatics.buffalo.edu/sticky_alignments), there are 913 representative *PDB* structures and three sets of computer-generated models for the compact sticky, homopolypeptide models: (1) 100-residue chains by atomic off-lattice modeling; (2) 100-residue chains by reduced on-lattice modeling; (3) 200-residue chains by reduced on-lattice modeling. The files are in the following directories:

- **100_255:** 255 models of 100 residue proteins (all are α -proteins) from the atomic, off lattice model.
- **100_150x14:** 100 residues, 150 chains each with 14 clusters (i.e. Cluster-1, 2, 3, 4, 5, 10, 25, 50, 75, 100, 125, 150, 175, 200) from the reduced, on-lattice model. For the file name, “a” stands for α -proteins; “b” for β -proteins; “ab” for $\alpha\beta$ -proteins.

- **200_150x14:** 200 residues, 150 chains each with 14 clusters (i.e. Cluster-1, 2, 3, 4, 5, 10, 25, 50, 75, 100, 125, 150, 175, 200).
- **200_15000:** 200 residues, 150 chains each with the top-100 clusters.
- **200_7000:** 200 residues, 7000 models from clustering of the 200_15000 set.
- **200_3500:** 200 residues, 3500 models from clustering of the 200_7000 set.
- **150_PDB:** 913 compact *PDB* structures whose length is between 41 and 150 residues with a pairwise sequence identity <30%.
 - **X.PDB:** *PDB* structures (“X” stands for *PDB* IDs).
 - **X.ali_15000:** *TM-align* structure alignment of X to the closest model from the 200_15000 set.
 - **X.ali_7000:** *TM-align* structure alignment of X to the closest model from the 200_7000 set.
 - **X.ali_3500:** *TM-align* structure alignment of X to the closest model from the 200_3500 set.
 - **Summary15000:** Summary of *TM-align* results of the *PDB150* set to the 200_15000 set.
 - **Summary7000:** Summary of *TM-align* results of the *PDB150* set to the 200_7000 set.
 - **Summary3500:** Summary of *TM-align* results of the *PDB150* set to the 200_3500 set.
 - **10worst:** Structural alignments of the 10 worst *PDB150* proteins to the closest homopolypeptide model in the 15,000 compact, homopolypeptide library.

References

1. Shimada, J., Kussell, E. L. & Shakhnovich, E. I. (2001) *J Mol Biol* **308**, 79-95.
2. Hubner, I. A., Edmonds, K. A. & Shakhnovich, E. I. (2005) *J Mol Biol* **349**, 424-34.
3. Shimada, J. & Shakhnovich, E. I. (2002) *Proc Natl Acad Sci U S A* **99**, 11175-80.
4. Swendsen, R. H. & Wang, J. S. (1986) *Physical Review Letters* **57**, 2607-2609.
5. Zhang, Y. & Skolnick, J. (2004) *Proc Natl Acad Sci U S A* **101**, 7594-9.
6. Zhang, Y., Kihara, D. & Skolnick, J. (2002) *Proteins* **48**, 192-201.
7. Zhang, Y. & Skolnick, J. (2004) *J Comput Chem* **25**, 865-71.
8. Skolnick, J., Kihara, D. & Zhang, Y. (2004) *Proteins* **56**, 502-18.
9. Zhang, Y. & Skolnick, J. (2005) *Proc Natl Acad Sci U S A* **102**, 1029-34.
10. Zhang, Y., Arakaki, A. & Skolnick, J. (2005) *Proteins*, submitted.
11. Betancourt, M. R. & Skolnick, J. (2001) *J. Comp. Chem.* **22**, 339-353.
12. Lathrop, R. H. (1994) *Protein Eng* **7**, 1059-68.
13. Holm, L. & Sander, C. (1993) *J Mol Biol* **233**, 123-38.
14. Shindyalov, I. N. & Bourne, P. E. (1998) *Protein Eng* **11**, 739-47.
15. Levitt, M. & Gerstein, M. (1998) *Proc Natl Acad Sci U S A* **95**, 5913-20.
16. Kihara, D. & Skolnick, J. (2003) *J Mol Biol* **334**, 793-802.
17. Betancourt, M. R. & Skolnick, J. (2001) *Biopolymers* **59**, 305-9.
18. Needleman, S. B. & Wunsch, C. D. (1970) *J Mol Biol* **48**, 443-53.
19. Zhang, Y. & Skolnick, J. (2004) *Proteins* **57**, 702-710.
20. Zhang, Y. & Skolnick, J. (2005) *Nucleic Acids Res* **33**, 2302-9.
21. Arakaki, A. K., Zhang, Y. & Skolnick, J. (2004) *Bioinformatics* **20**, 1087-1096.
22. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003) *Nucleic Acids Res* **31**, 365-70.

Table 1. *TASSER* modeling of the 10 proteins in the *PDB150* set with the worst matches to structures in the 15,000 member, compact homopolyptide library.

ID ^a	Homopolyptide templates selected by <i>TM-align</i>					Final models by <i>TASSER</i>		
	Model_A ^b	Lch ^c	Coverage ^d	<i>TM-score</i> _A ^e	R_A_ali ^f	<i>TM-score</i> ^g	R_all ^h	R_ali ⁱ
1at0_	b10R_70	142	0.62	0.371	4.46	0.874	2.05	2.13
1fjgl	b49R_85	125	0.70	0.371	5.01	0.502	7.33	5.61
1gqva	ab23R_30	135	0.64	0.369	5.00	0.763	4.0	2.65
1ilja	b50R_19	106	0.67	0.364	4.68	0.518	5.96	3.92
1khia	ab9R_6	147	0.63	0.362	5.07	0.591	6.13	5.03
1knma	b14R_77	129	0.66	0.371	4.64	0.512	5.11	4.69
1mi8A	b11R_42	141	0.62	0.365	4.43	0.605	5.22	3.62
1nkws	a6R_77	113	0.60	0.363	4.37	0.480	6.14	6.27
1urk_	a46R_1	130	0.64	0.372	4.95	0.538	5.73	4.79
2ila_	ab33R_71	145	0.66	0.357	5.25	0.775	3.49	2.80
<...>		131	0.64	0.366	4.79	0.616	5.11	4.15

^a10 *PDB* proteins in the *PDB150* set which have the worst match (lowest *TM-score*) to the homopolyptide model.

^bTemplates found by *TM-align* in the 15,000 member compact, homopolyptide structure library.

^cSize of the *PDB* target.

^dCoverage of the *TM-align* structural alignments of the homopolyptide model to the *PDB* target structure.

^e*TM-score* of *TM-align* of the homopolyptide model to the *PDB* target structure.

^fRMSD of the structurally aligned regions of the homopolyptide model compared to the target *PDB* structure.

^g*TM-score* of the first models built by *TASSER* (ranked by cluster density) started from homopolyptide templates.

^hRMSD to native of the *TASSER* modes over all residues of the target protein.

ⁱRMSD to native of the *TASSER* models over the same structure aligned region as the compact, homopolyptide template.

Table 2. H-bond parameters calculated from 100 high resolution PDB structures.

bb_{α} 0.45	cc_{α} 0.1	r_{α} 6.03 Å	pp_{α} 0	qq_{α} 0	$cc_{\alpha 0}$ 0.4	$bb_{\alpha 0}$ 0.815	$br_{\alpha 0}$ 1.56 Å
bb_{β} 0.25	cc_{β} 0.4	r_{β} 6.15 Å	pp_{β} 0.35	qq_{β} 0.35			

Figure Legends

Figure 5. Schematic representation of the hydrogen bond potential, E_{HB} . A hydrogen bond is counted when the four atom pairs are within a square well, eliminating the need for angle calculations and increasing computational efficiency. The indicated distances are as follows: d1 is between the donor nitrogen and acceptor oxygen, d2 is between the donor nitrogen and acceptor carbonyl carbon, d3 is between the donor hydrogen and acceptor oxygen, and d4 v is between the donor hydrogen and acceptor carbonyl carbon. The interaction energies for specific distance parameters are given in the accompanying file, tabulated in the format d1 d2 d3 d4 E.

Figure 6. Histogram of occurrence of loops, helices, beta strands, and coils at the tails (N and C-termini) found in a representative set of 1489 proteins in the *PDB*.

Figure 7. Illustration of geometry of H-bond donor and receptor residues in CAS model. \bar{p} and \bar{q} are unit C_{α} - C_{α} vectors. $\bar{c} = \frac{\bar{p} - \bar{q}}{|\bar{p} - \bar{q}|}$ is the unit bisector vector and $\bar{b} = \frac{\bar{c} \times \bar{q}}{|\bar{c} \times \bar{q}|}$ is the unit normal vector. \bar{r} is a vector pointing from donor C_{α} atom to receptor C_{α} atom.

Figure 8. Histogram of the *TM-score* of the best structural alignment to the PDB template library of 158 distinct compact FJC conformations of 200 residue chains. The mean and median are 0.2963 and 0.2956 respectively.

Figure 9. Superposition of the *TASSER* model (thick backbone) with 1fjgl (thin backbone); this is the only case where *TASSER* generates models with $RMSD > 6.5 \text{ \AA}$ because of the misorientation of the N-terminal tail.

Figure 10. Representative example, 1at0_, of *TASSER* modeling improvement. **Left:** Superposition of the homopolyptide structure template (thick backbone) and the native (thin backbone). **Right:** Superposition of the *TASSER* model (thick backbone) and the native (thin backbone).

Figure 11. An example, 1nkws, where *TASSER* generates a model having a higher RMSD to native than the initial structural alignment because of misorientation of the dangling tail. The *TASSER* model and the target are indicated by thick and thin backbones respectively.

Figure 12. Detection of substructures resembling enzyme active sites. Relative frequency distributions of the dRMSD of selected *AFTs* (identified by their associated EC numbers, left) to the best sequence-independent hit detected in 750 sticky homopolypeptides (magenta boxes) or 750 native structures functionally unrelated to the specified *AFT* (blue boxes). The statistics represented in the box-and-whisker plots are: 10th percentile (whisker, left), 25th percentile (box, left), median (thick line), 75th percentile (box, right), 90th percentile (whisker, right) and outliers (circles). Only the top ten *AFTs* are shown. They are ranked by increasing median dRMSD of the best hit in sticky homopolypeptides. Restrictive (pointed red line) and permissive *AFT* cutoffs (pointed green line) are plotted as references to assess the significance of a sequence-dependent match to an *AFT*.

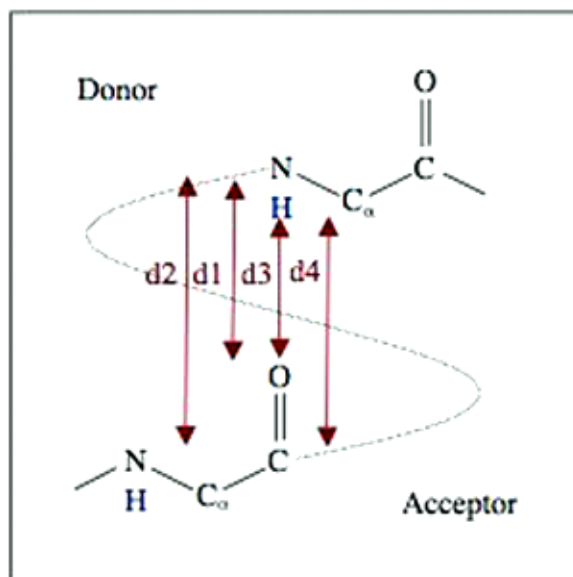


Figure 5

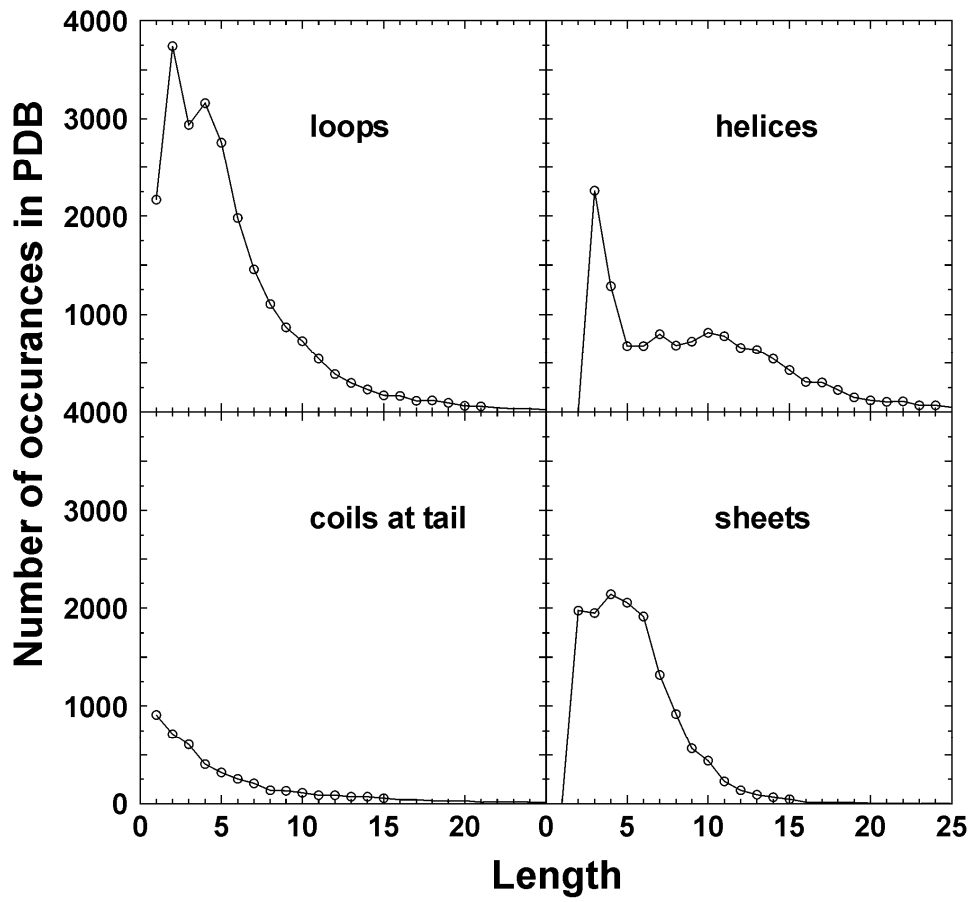


Figure 6

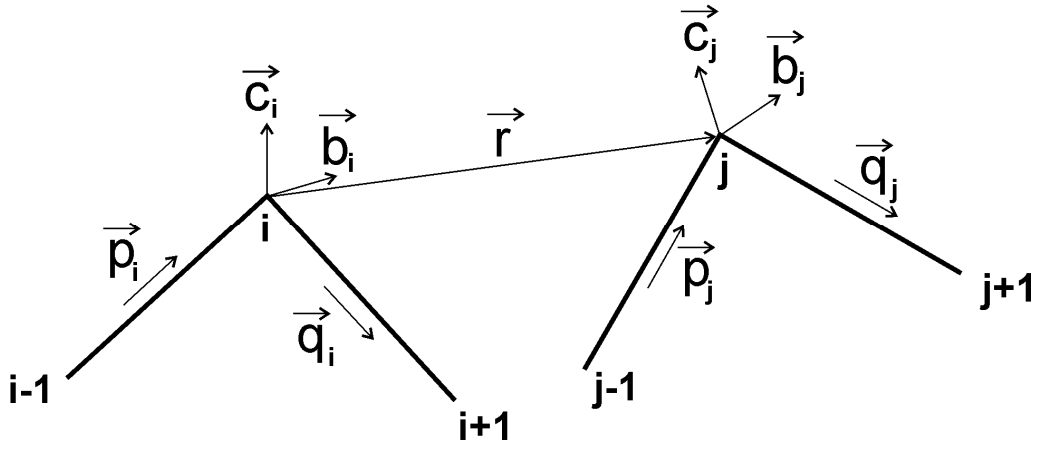


Figure 7

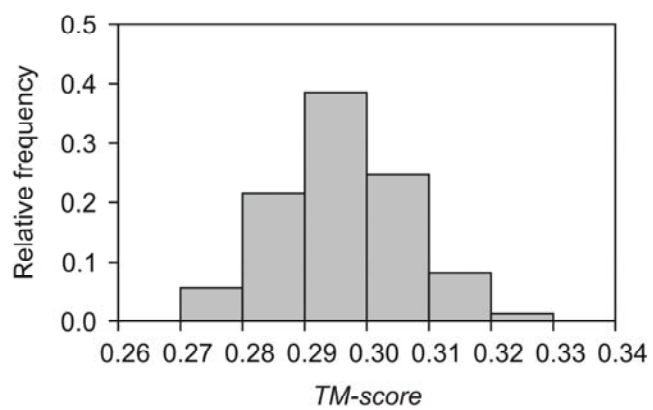


Figure 8

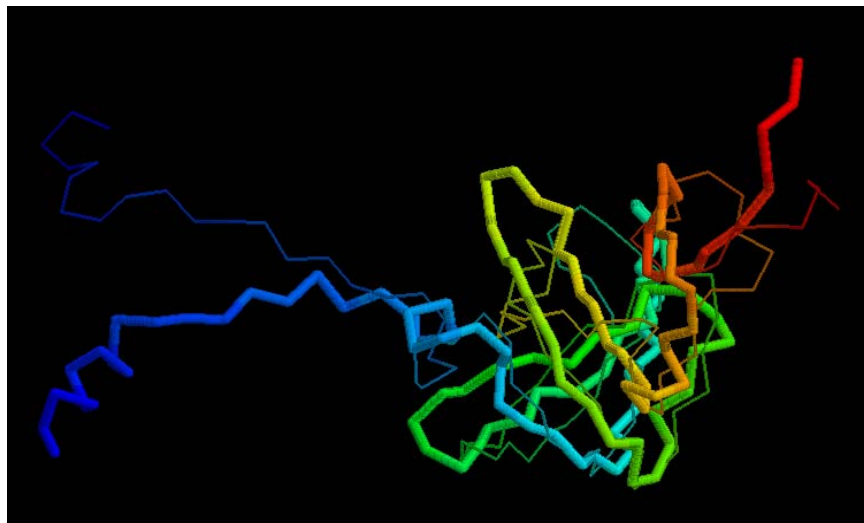


Figure 9

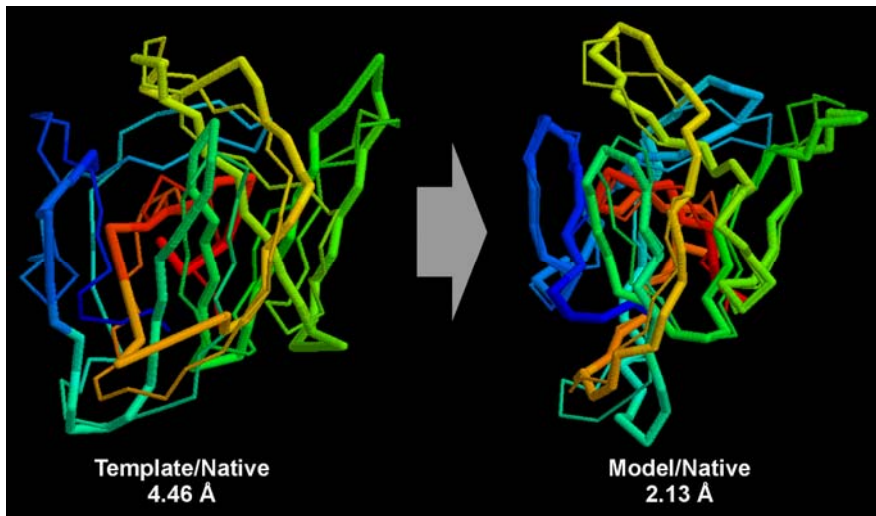


Figure 10

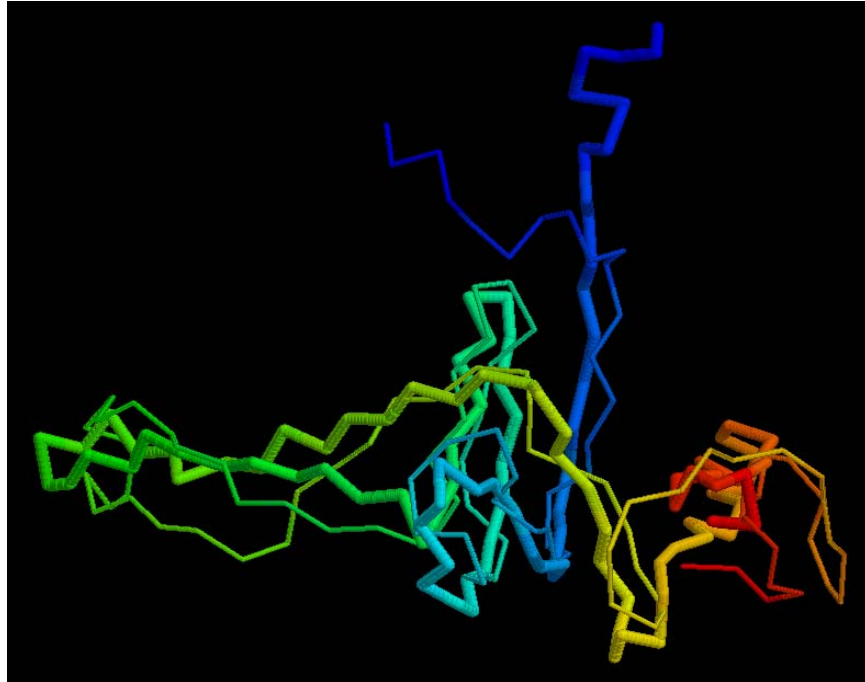


Figure 11

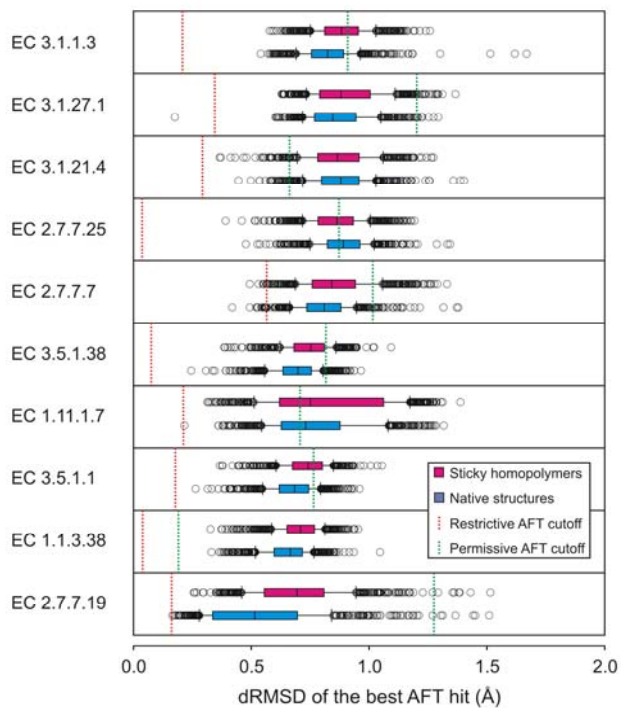


Figure 12