Supporting Information

DeBartolo et al. 10.1073/pnas.0811363106

SI Methods

Secondary Structure Prediction. Table 1 in the main text demonstrates that the ItFix secondary (2°) structure prediction method can meet or exceed the percentage accuracy of the prediction from the programs PSIPRED and SSPro. The information conveyed by the percentage accuracy, however, is compromised because of disagreements between methods for the assignment of 2° structure. For example, the DSSP method, which we use to assign 2° structure, differs from DeepView (1) in specifying 2° structure for 1tif (Table 2). DeepView is more liberal when assigning β strands and designates residues 3–5 and 9 as β strand, whereas DSSP assigns this region as mostly coil. ItFix similarly favors assigning β strand over coil, implying that ItFix should achieve higher 2° structure accuracy for 1tif compared with the native DeepView assignment rather than the DSSP assignment. Nevertheless, we compare our prediction with DSSP assignments because DSSP is used to designate the 2° structure of the simulation models.

Another issue relating to the accuracy of 2° structure predictions is the varying assignment of 2° structures by different prediction methods. For example, some approaches consider an α helix and a 3–10 helix to belong to the same category, whereas we treat the 3–10 helix as a subtype of coil because the α helical hydrogen-bonding pattern requires at least four residues whereas the 3–10 helix only requires three. Notably, when Q3 level methods, such as PSIPRED and SSPro, predict a 3–10 helix as H, we consider them correct. Because our 2° structure sampling depends on DSSP, we adhere to its convention where a 3–10 helix is in the coil (C^G) class.

Secondary Structure-Fixing Protocol. The selection of thresholds has been made as an empirical compromise between prediction accuracy and the speed of specifying 2° structure. Some accuracy may be compromised to allow the largest number of positions to be fixed within a reasonable number of rounds. The protocol (see *Methods*) for fixing positions employs the following operations.

If the turn (C^T) probability exceeds 50% throughout a region with at least one 2° structure type removed in a previous region, that region is specified as coil. Also, when the overall 2° structure fixing is at an advanced state (>90% positions fixed) and a large stretch of positions are devoid of β strand, then β strand is removed from the library at those positions. If a position is fixed as β strand, at least three adjacent positions become fixed as strand when those positions subsequently attain a strand probability >50%. If the direction in which the fixing of strands is ambiguous, it proceeds away from the nearest segment of coil. This correction is added to make sure the maximum amount of 2° structure is fixed for a given target.

The number of rounds of folding required for convergence is reduced by using additional operations. If the strand option has been removed from the library at two positions that are separated by three or fewer residues without library restrictions, we remove strand from the intervening positions. Other operations for all types of library restrictions are used to refine small spaces between fixed regions, e.g., C–C and H–H are replaced by CCC and HHH, respectively.

The set of proteins studied typically requires 5–12 rounds. Convergence is slow for two proteins (1bm8, 1vqh), and those simulations are stopped after 12 rounds. After the final round for all proteins, any remaining unfixed positions have the 2° structure type determined by plurality. The DSSP 2° structure of each final structure in every round is calculated directly or from the original 2° structure for each residue of the final structure in the trimer library. β sheet and turn probabilities are taken from the origins in the library, whereas all other 2° structures are determined directly using DSSP. In a small minority of cases, 2° structure assignments disagree with those determined by DSSP. Incorrect assignments usually occur around the border between a helix and coil or β sheet and coil, and in most cases they tend to be at positions where 2° structure determination methods disagree. The most notable examples are 1sap, where ItFix assigns the fifth strand as coil, 1fwp, where the second helix is assigned as strand, and 1dcj, where the second helix and third strand are incorrectly specified. However, 1sap can folded accurately, implying that some errors do not affect the quality of the 3° structure prediction.

Sampling Library. We obtain our trimer library from a set of proteins culled from the PDB using PICSES (2) and a resolution cutoff of 2.5 Å. As 2° structure is progressively restricted during the simulations, the total number of trimers available for a given sequence becomes smaller and less diverse. Consequently, diversity is enhanced by including trimers with amino acid substitutions within the following four groups of correlated amino acids: (FVI), (LM), (KRQH), and (WYF) (e.g., the three trimers XFY, XVY, XIY, are considered equivalent). We add 5° noise to each angle pulled from the library. Bond lengths and angles are all set to ideal values.

1. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* 18:2714–2723.

2. Wang G, Dunbrack RL, Jr (2003) PISCES: A protein sequence culling server. *Bioinformatics* 19:1589–1591.

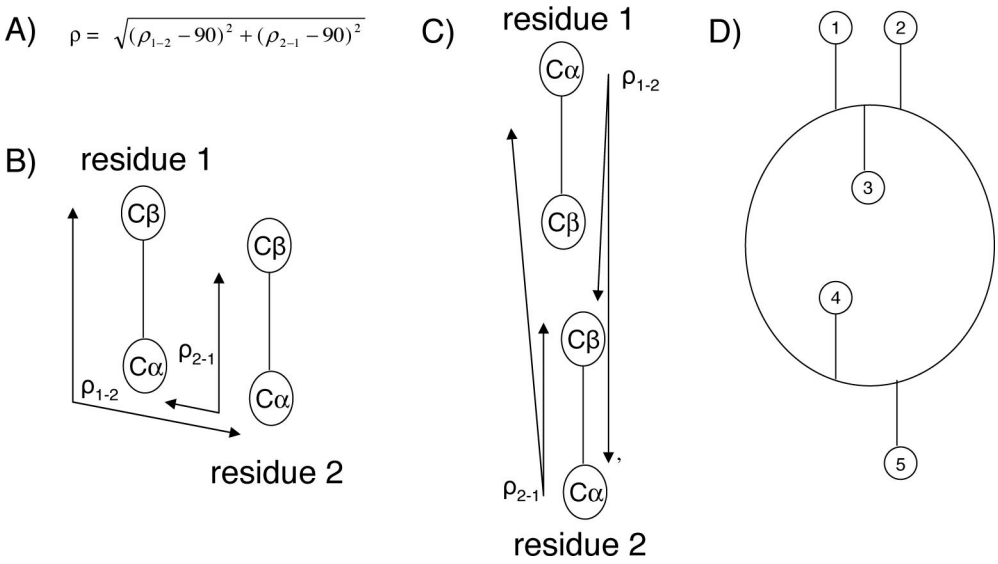

Fig. S1. Orientation dependence of statistical potential. Each interacting residue pair has two angles. One angle, ρ_{1-2}, is the angle between the C^{α}–C^{β} vector of residue 1 and the C^{α}–C^{α} vector from residue 1 to residue 2, whereas the other angle, ρ_{2-1}, is the angle between the C^{α}–C^{β} vector of residue 2 and the C^{α}–C^{α} vector from residue 2 to residue 1. (*A*) Relative orientation of the side chains is quantified as ρ. (*B*) Two residues have angles ρ_{1-2} and ρ_{2-1} close to 90°, yielding a small ρ value. (*C*) A residue pair with large ρ has angles ρ_{1-2} and ρ_{2-1} that are both far from 90°. (*D*) Illustration of protein with residue pair orientations having small ρ (1–2, 2–3, 1–3, 4–5) and large ρ (1–4, 2–4, 3–4, 1–5, 2–5, 3–5).

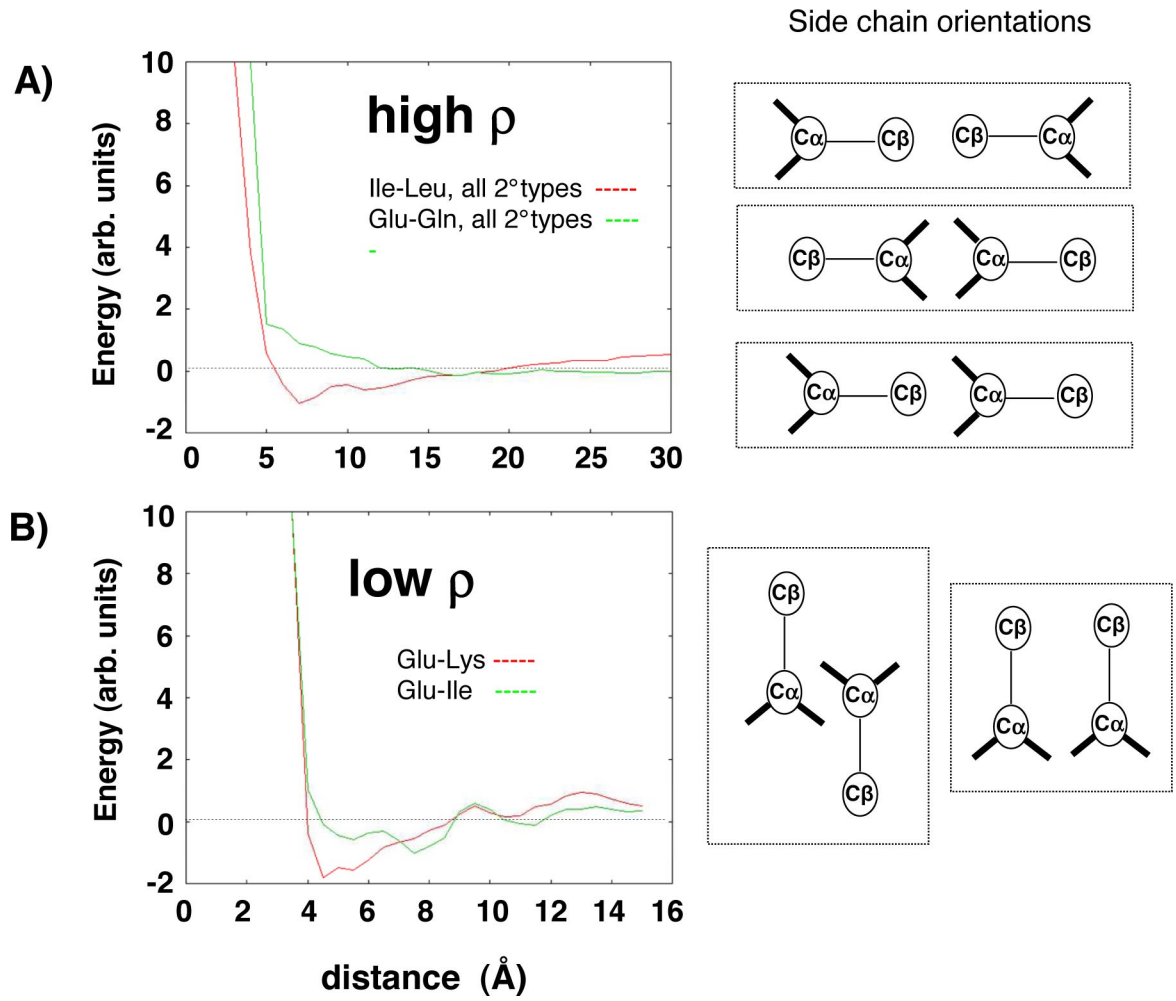

Fig. S2. Statistical potential energy profiles illustrating orientation dependence. This dependence reflects the basic protein structural principles of hydrophobic burial, hydrophilic exposure, and 2° structure conformation. (*A*) Interatomic potential for two C^{β} atoms with three different relative orientations that produce a high ρ value. In such cases, hydrophobic amino acids are favored to be situated at closer distances, corresponding to buried residues pointing at each other in the core of the protein. The opposite applies for hydrophilic amino acids, which prefer larger separations corresponding to surface-exposed residues on opposite sides of the protein. (*B*) Potential for two C^{β} atoms with two different orientations for a pair of residues on strands of β sheets and small ρ values. Shorter C^{β}–C^{β} distances are preferred for residues on the same side of the sheet, and larger ones are favored for those on opposite sides of the sheet.

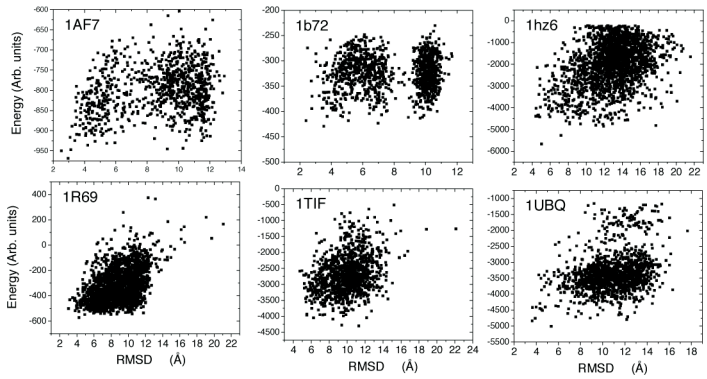

Fig. S3. Illustrative scatter plots of rmsd vs. energy for selected targets. ItFix is often capable of identifying a low rmsd model from energy alone (see main text Tables).

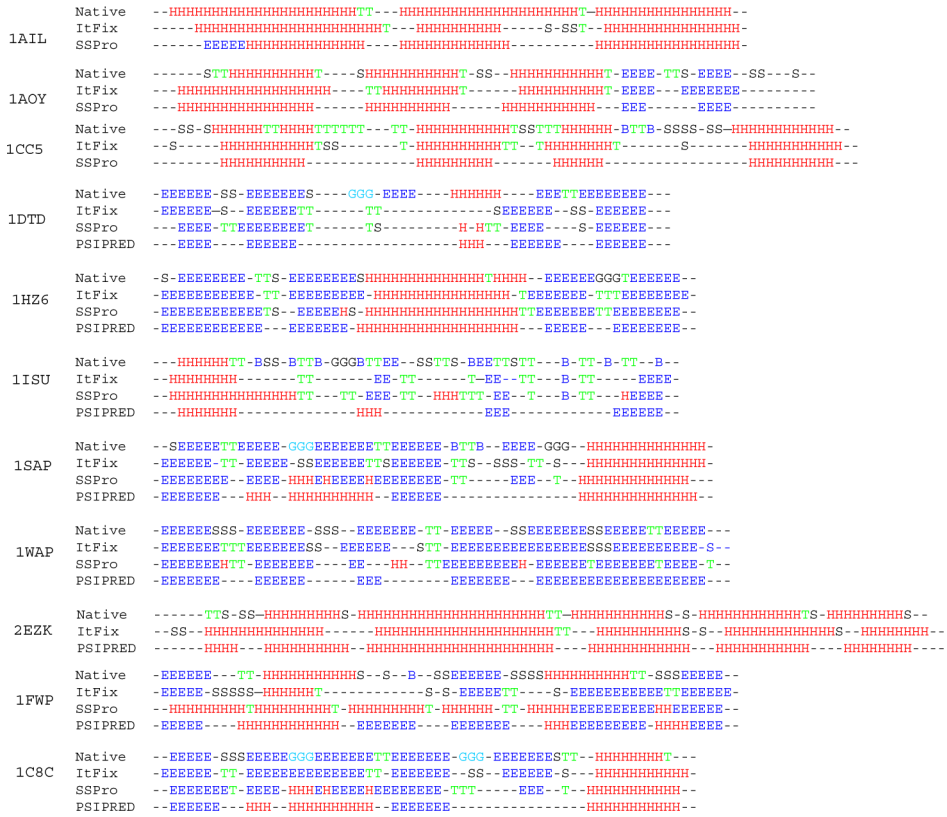

Fig. S4. Secondary structure prediction of targets from Table 1. These targets have rather few homologs in the sequence database and are more difficult targets for tertiary (3°) prediction. ItFix predicts 2° structure at the Q8 level, defined by DSSP as helix (H), β sheet (E), turn (C^T), bend (C^S), 3–10 helix (C^G), π helix (C^I), β bridge (C^B), and coil (C^N).

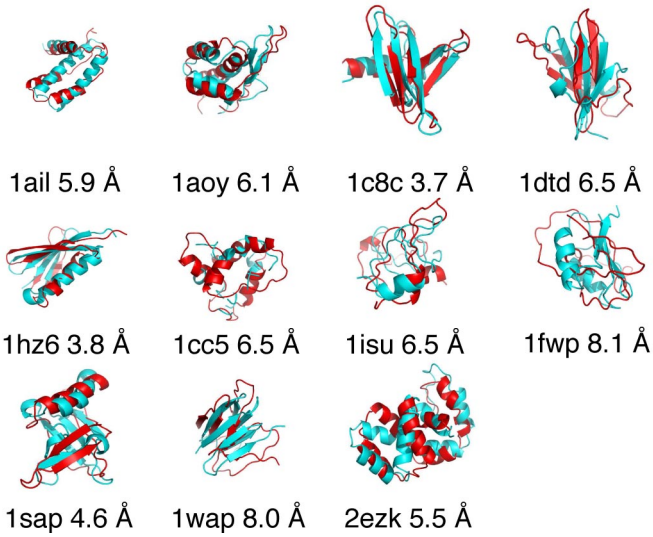

Fig. S5. Tertiary structure prediction for low-homology set. Alignments of the ItFix lowest rmsd 3° structure (red) with the native structure (blue) using PyMol visualization software are shown. C^{α} rmsd values of ItFix model and native structure (Table 1, 10th column) are listed below each target.

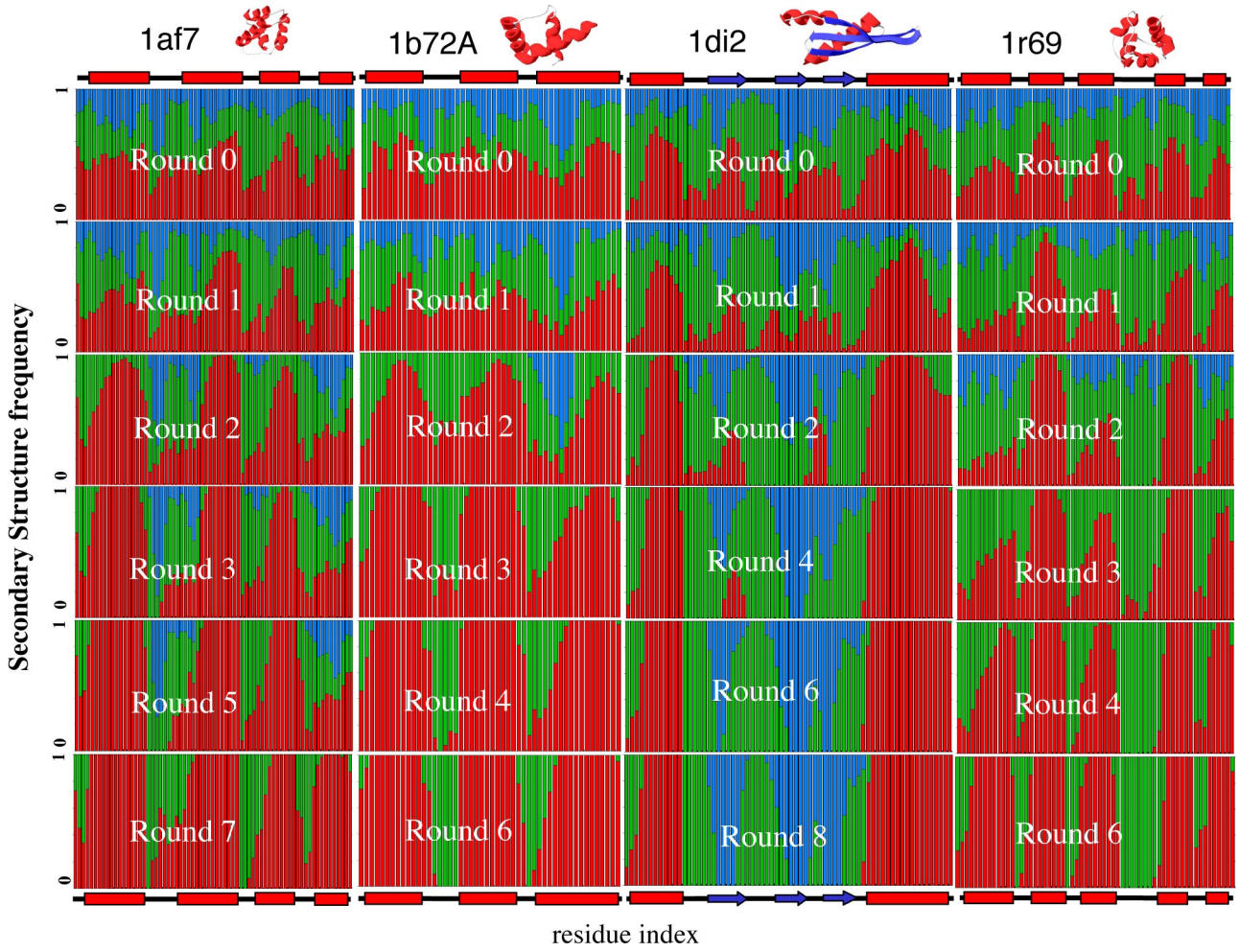

Fig. S6. Progression of fixing structure for 1af7, 1b72A, 1di2, and 1r69. The position dependence of the 2° structure frequencies at the end of each round, E (blue), H (red), and C (green) is shown. A single color bar represents a residue assigned to a particular 2° structure type. The native 2° structure is shown with red boxes (helices) and blue arrow (strands) at the top and bottom. The order of fixing of structural elements may recapitulate major features of the authentic pathway. Round 0 frequencies are the average 2° structure assignments obtained from the initial trimer library that is contingent only on the sequence. As the rounds progress, uncertainties in 2° structure diminish.

Table S1. Relative weights of the components of the statistical potential

Energy component	Contiguous helix	Contiguous strand	Contiguous coil	Antiparallel β sheet (small ρ)	Parallel β sheet (small ρ)	Non-β sheet (small ρ)	Non-β sheet (medium ρ)	Non-β sheet (large ρ)
Min/max distance, Å	0.0/15.0	0.0/15.0	0.0/15.0	0.0/15.0	0.0/15.0	0.0/30.0	0.0/30.0	0.0/30.0
Bin size, Å	0.5	0.5	0.5	0.5	0.5	0.5	1.0	1.0
Attractive coefficient	0.0	0.0	0.0	5.0	10.0	1.0	1.0	1.0

The coefficient of the attractive component for pairs in contiguous units of 2° structure is set to zero to eliminate bias toward specific 2° structure types (first 3 columns). The noncontiguous strand–strand term is only counted when two consecutive hydrogen bonds are formed, defined according to the amide nitrogen of one residue being within 3.5 Å of the carbonyl oxygen of its partner on the other strand. This condition forces the highly weighted strand–strand term to enter only for pairs that are oriented in a proper β sheet geometry. β sheets terms are included with a 5- to 10-fold higher weight factor than the other terms. In addition, favorable interactions are included in the strand–strand interaction term between two residues i and j that have not been fixed to E after the prior round only for $|i–j| > 4$ to avoid unwanted kinking and excess turn formation. When the residues i and j are specified as E, favorable interactions are allowed for $|i–j| > 2$. Favorable interactions for other than strand–strand interactions only are allowed when $|i–j| > 6$, thereby applying only to two residues that are within different units of 2° structure. Relative weight factors for the different terms have been derived semiempirically.