Coevolutionary signals across protein lineages help capture multiple protein conformations

Faruck Morcos[a,1], **Biman Jana**[a], **Terence Hwa**[b], **and José N. Onuchic**[a,1]

[a]Center for Theoretical Biological Physics and Departments of Physics and Astronomy, Chemistry, and Biochemistry and Cell Biology, Rice University, Houston, TX 77005; [b]Center for Theoretical Biological Physics and Department of Physics, Division of Physical Sciences Section of Molecular Biology, University of California San Diego, San Diego, La Jolla, CA 92093-0374

Edited by Arieh Warshel, University of Southern California, Los Angeles, CA, and approved November 8, 2013 (received for review August 21, 2013)

A long-standing problem in molecular biology is the determination of a complete functional conformational landscape of proteins. This includes not only proteins' native structures, but also all their respective functional states, including functionally important intermediates. Here, we reveal a signature of functionally important states in several protein families, using direct coupling analysis, which detects residue pair coevolution of protein sequence composition. This signature is exploited in a protein structure-based model to uncover conformational diversity, including hidden functional configurations. We uncovered, with high resolution (mean ∼1.9 Å rmsd for nonapo structures), different functional structural states for medium to large proteins (200–450 aa) belonging to several distinct families. The combination of direct coupling analysis and the structure-based model also predicts several intermediates or hidden states that are of functional importance. This enhanced sampling is broadly applicable and has direct implications in protein structure determination and the design of ligands or drugs to trap intermediate states.

conformational plasticity | covariation | statistical inference | molecular dynamics

As demonstrated by Anfinsen in 1973 (1) for small and intermediate-size proteins, amino acid sequences contain all of the necessary information to determine their native structure and function. In principle, a complete physical understanding of all molecular interactions should be sufficient to uncover not only the proteins' native structures, but also all their respective functional states, including functionally important intermediates. This landscape is required for a complete knowledge of functional mechanisms and therefore it has implications for drug discovery. Advances in computational approaches have been promising in sampling such conformational intermediates (2, 3). However, in general, computational methods are limited by uncertainties in protein models as well as insufficient computational resources to achieve proper sampling. Experimental techniques such as crystallography or NMR spectroscopy have been successful in identifying functional protein structures but only for a fraction of the complete set of known protein sequences (4, 5). Additionally, the determination of functionally important intermediate states using such methods has been challenging due to their transient nature. One idea to confront this challenge is to search for clues in genomic data (6–10). Functional states under conformational selection should leave a trace in the evolutionary history of proteins. Recent results inspired by this hypothesis have led to the development of the powerful "direct coupling analysis" (DCA), which was able to predict a large number of direct structural contacts between residues from sequences alone (11). Other useful methods have been developed to define coupling among residue pairs (12, 13). Others have also looked into correlated electrostatic mutations to study the evolution of protein topology toward minimized interaction frustration (14). Integrating the DCA-predicted contacts into coarse-grained physical models of proteins such as structure-based models (SBMs) (15–19) led to predictions on protein–protein interactions (20–22) as well as tools to aid the prediction of native structures (19, 23–25). The idea of using predicted contacts to estimate native structures was also explored by other methodologies (26, 27). Here, we show that DCA predicts important structural interactions related not only to the native state but also to distinct functional conformational states of a protein, including intermediates. We develop a hybrid computational method to recover such important conformations, which derives the SBM energy function from a single experimental structure and incorporates DCA residue contacts into the energy function (*Materials and Methods*). We show that this model samples well beyond a single native structure to reveal conformational diversity, including hidden functional configurations, in proteins. We refer to this methodology as SBM+DCA.

Results

Conformational Diversity Is Embedded in Evolutionary Information. In this study, we provide evidence that accurate information about conformational diversity can be extracted from evolutionarily related protein sequences. We focus on proteins that experience large conformational changes upon ligand binding (Table S1) (28, 29). Fig. 1 illustrates our first example, the L-leucine binding protein [Protein Data Bank (PDB) IDs 1usg and 1usi], which experiences large conformational changes upon binding to L-leucine (30). Fig. 1*C* displays a structural comparison between the ligand bound and unbound states. The residue–residue contact maps for open and closed conformations have very distinct signatures, as highlighted in Fig. 1*A*. This protein belongs to the domain family "periplasmic binding protein 6" (Pfam ID: PF13458). Applying DCA to 7,363 sequences in this domain, we obtained a large

Significance

We show that directly coupled evolutionary residue pairs provide a distinct footprint of conformational diversity in protein families. This is revealed as competing residue contacts unique to distinct configurations of proteins with multiple conformations. We demonstrate that combining this information with physical models of proteins is sufficient to uncover conformational diversity for several protein families. We discovered that such directly coupled residues not only allow us to accurately transition between apo and holo conformations but also help to uncover intermediate states that are not easily accessible to experimental or other computational methods. This enhanced sampling of the functional conformational space of proteins may have broad implications in protein structure determination and the design of ligands that can trap intermediate states.

Author contributions: F.M., B.J., T.H., and J.N.O. designed research; F.M. and B.J. performed research; F.M., B.J., T.H., and J.N.O. analyzed data; and F.M., B.J., T.H., and J.N.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1] To whom correspondence may be addressed. E-mail: faruckm@rice.edu or jonuchic@rice.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1315625110/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1315625110

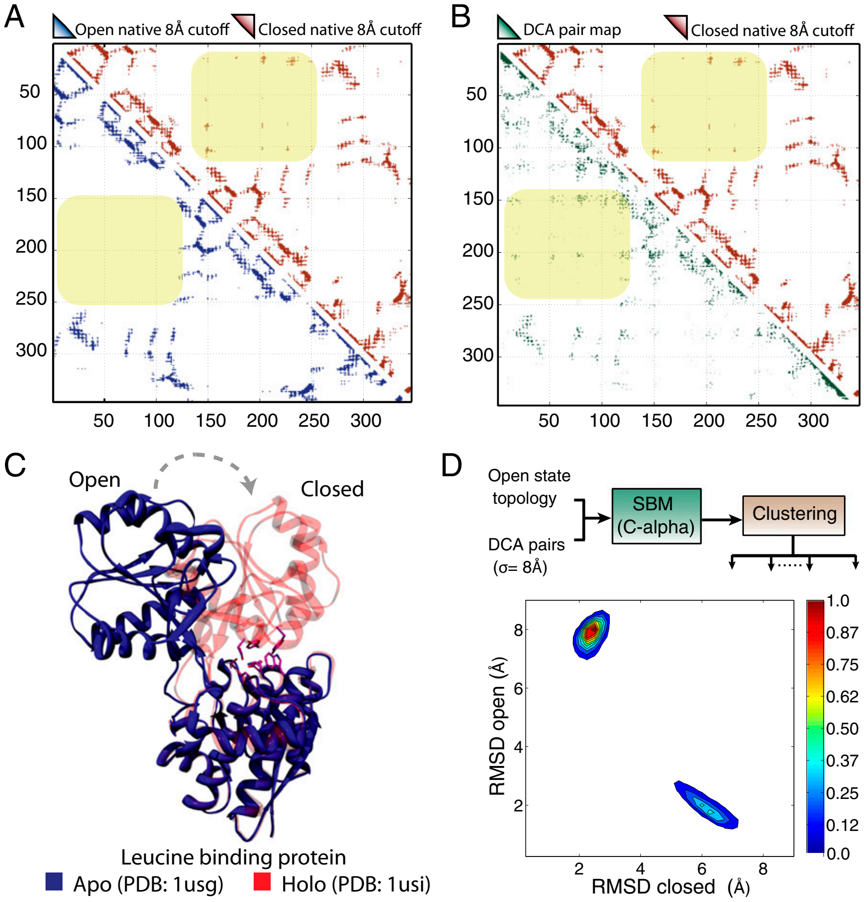
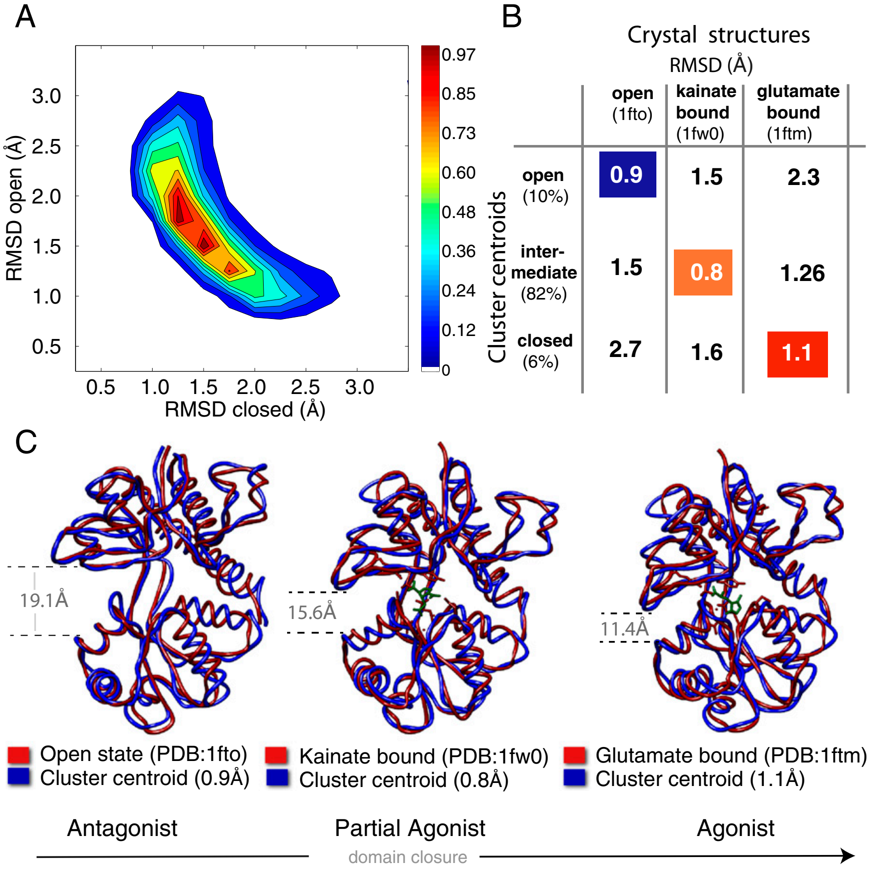


Fig. 1. A hybrid SBM+DCA model of the L-leucine binding protein is able to uncover its two-state (apo/holo) conformational landscape. *A* compares the native open and closed contact maps and *B* compares a DCA contact map with the native closed state. In *A*, comparing the native contact map of the open conformation (PDB ID 1usg; lower triangular map) and the closed conformation (PDB ID 1usi; upper triangular map) shows a clear set of contacts (shaded box) that are exclusive to the closed state. In *B* a predicted contact map using highly ranked DCA residue pairs (lower triangular map) shows a very accurate reconstruction of the complete map that includes the extra contacts in the closed conformation (upper triangular map). (*C*) Structural comparison between the apo and the holo states of the L-leucine binding protein, showing domain closure. (*D*) Integrating a SBM of the open-state topology with DCA contacts produces a distinct bimodal landscape, as opposed to the single-basin distribution observed when we use the same number of extra contacts but randomly distributed.

number of high-ranking pairings predicted to be physical contacts (69% true positives in the top 500 predictions). Interestingly, the predicted contacts belong to both open and closed conformations; compare highlighted regions in Fig. 1*B*. We use such predicted couplings in combination with a structure-based model (SBM+DCA) to study the conformational dynamics of the L-leucine binding protein. Using only structure parameters taken from the open-state topology (residue contacts and distances and dihedral angles and bond distances), the DCA-predicted contacts led the model to identify an ensemble of closed conformations in addition to the open state (Fig. 1*D*). We used the gromos clustering algorithm (31) to obtain representative structures (cluster centroids) in the ensemble. Both centroids are within 2-Å rmsd accuracy to the experimental structures (Movie S1). For comparison, replacing those DCA contacts by the same number of random contacts led to fluctuations only around the open conformation (Fig. S1). A dual-basin SBM (32) is also able to recover both the open and the closed conformations. However, it requires, for both bound and unbound conformations, the complete knowledge of the contacts and their experimentally determined distances (Fig. S2).

Multiple States with Functionally Important Intermediates Are Found by SBM+DCA Methodology. We next studied the glutamate receptor (GluR2), which belongs to the family of "bacterial extracellular solute-binding proteins" (PF00497). Armstrong and Gouaux (33) provided structural evidence suggesting that GluR2 uses an agonist-induced domain closure mechanism to gate the transmembrane channel and that its activation is dependent on the degree of domain closure.

We analyzed the 20,059 protein sequences in this family using DCA, and the predicted contacts are again used to study the dynamics of the structure. The 2D rmsd frequency distribution in Fig. 2*A* illustrates how DCA-predicted couplings can be used to sample the conformational space of the open and closed states to an accuracy of 1 Å. However, the conformational space is more complex in this case. Fig. 2*B* shows the rmsd measures of the top three clusters from the SBM trajectory. One cluster centroid (6%) has an rmsd of 1.1 Å with respect to the closed glutamate-bound state (PDB ID: 1ftm); a second cluster centroid (10%) has an accuracy of 0.9 Å with respect to the apo state (PDB ID: 1fto). Interestingly, the centroid of the most populated cluster (82%) is far away from both the open and the closed state. However, this centroid structure is only 0.8 Å apart from a Kainate-bound structure of the receptor (PDB ID 1fw0). Kainate was shown to be a partial agonist that induced an intermediate semi-closed state in the glutamate receptor (33). Fig. 2*C* compares the cluster centroids with the experimentally determined structures and shows the sequential closure of the domains. As a control, we



Crystal structures

RMSD (Å)

Cluster centroids	open (1fto)	kainate bound (1fw0)	glutamate bound (1ftm)
open (10%)	0.9	1.5	2.3
inter-mediate (82%)	1.5	0.8	1.26
closed (6%)	2.7	1.6	1.1

Fig. 2. The glutamate receptor has a ligand-dependent domain closure. (*A*) The conformational landscape observed after combining an open-state topology with coevolutionary restraints obtained from the family of bacterial extracellular solute-binding proteins (PF00497). The landscape includes conformations with an rmsd of less than 2 Å from the crystal structures of open and closed states. An intermediate state is also present that is between the closed and the open conformations. (*B*) After using the gromos clustering algorithm (31) for the molecular dynamics trajectory, the top three cluster centroids cover 98% of the conformations. The centroid of the most populated cluster is in fact structurally very similar to the kainite-bound structure of the glutamate receptor (rmsd 0.8 Å). Kainate is a partial agonist that brings the protein to a semiclosed state. (*C*) Structural comparison between the centroids and experimental structures of antagonist (open; PDB ID 1fto), partial agonist (semiclosed; PDB ID 1fw0), and agonist (closed; PDB ID 1ftm) states. The predictions for these three states have an ∼1-Å rmsd with respect to the crystal structures.

show that a dual-basin SBM can sample a similar conformational space but it requires the complete knowledge of open and closed structures (Fig. S3). These results suggest that coevolutionary information can sample multiple functionally relevant states with high accuracy. We analyzed another protein of the same family, glutamine binding protein (GBP) (PDB IDs 1ggg and 1wdn) (34). We predict an intermediate state (Fig. S4), for which there is only indirect experimental evidence of its existence (35). We propose that this intermediate state for the GBP can be trapped experimentally by designing an appropriate ligand, in the same manner that kainate was used to crystallize the intermediate for GluR2.

We next examined sugar-binding proteins that also experience large conformational changes. The D-Ribose binding protein is a protein of the "family of periplasmic protein binding domains" (PF13407) with more than 8,000 members. Fig. 3*A* shows a comparison between the open (PDB ID: 1urp) and closed (PDB ID: 2dri) residue contact maps (upper triangular map) as well as the contacts obtained via DCA (lower triangular map). We identified contacts common for both open and closed structures but also some unique to the closed state (red dashed boxes). We also found a series of contacts (black dashed box) that belong to neither the open nor the closed state. These couplings gave rise to a conformational landscape that includes a very distinct third state (Fig. 3*B*). For comparison, a dual-basin model of the open and closed conformations combined with the same number of random extra contacts did not yield the intermediate-state basin (Fig. S5), but instead a broadening is observed compared with the dual-basin landscape (Fig. S6). This finding suggests that this third state is a feature uniquely captured by the extra DCA couplings. Fig. 3*B* shows a closed-state ensemble with a tighter domain closure with respect to the experimental closed state but conserving the same topological features (rmsd open >5 Å). We attribute this to the difference between the ligand-bound and the ligand-free closed states (3). We model the presence of the ligand by using the exact experimental contact distances for only a few contacts (landscape in Fig. 3*C*). Now we observe the ligand-bound closed ensemble instead of the ligand-free state present in Fig. 3*B*, while still accessing the intermediate state. Fig. 3*D* shows the contact maps of the cluster centroids from the distributions observed in Fig. 3*B*. The open-state cluster centroid has an rmsd of 2.2 Å whereas the closed-state cluster has an rmsd of 2.6 Å. The intermediate-state cluster centroid has a twisted semiclosed state that is equidistant to the open and closed conformations and has unique contacts that are also present in the DCA estimated map (compare black dashed boxes in Fig. 3 *A* and *D*). A comparison between the three cluster centroids is available in Movie S2. Using umbrella sampling molecular dynamics simulations, Ravindranathan et al. provided computational evidence supporting the existence of such a twisted state (3). Such an intermediate state

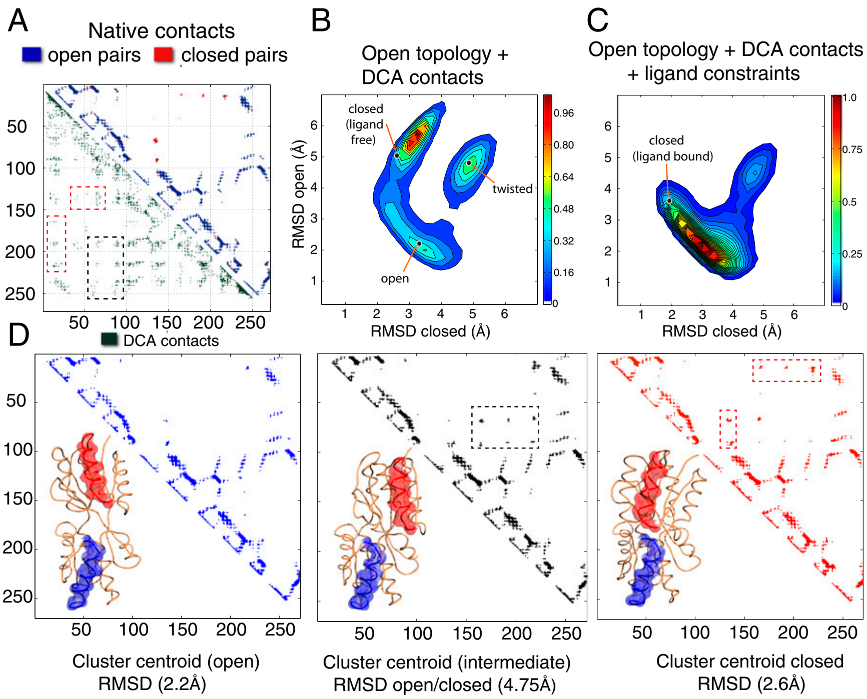


Fig. 3. D-Ribose binding protein (PDB IDs 1urp and 2dri) goes through large conformational changes upon ribose binding. (*A*) Comparison between the open and the closed native contact maps. Blue marks illustrate open-state contacts and red contacts are uniquely found in the closed state. The lower triangular part of the map shows the contacts estimated using DCA for which we can identify the global structure as well as the closed-state contacts (red dashed box). We can also see in the map an additional set of contacts (black dashed box) that may lead to a third state that has not been observed experimentally. (*B*) The conformational landscape shows a very distinct third state, which is not observed in the dual-basin control simulations, even if random extra contacts are added to increase the sampling of the conformational space (control in Fig. S5). (*C*) If we represent ligand constraints with few contacts with experimental distance parameters, then the landscape shows the ligand-bound closed state while the population of the intermediate state is still present. (*D*) The contact maps and structures for the cluster centroids are shown for the open, intermediate, and closed states. There is a distinct set of competing contacts that induce the observed landscape only when we integrate our model with DCA information. The comparison of the centroid cluster structures shows how two helices from the two domains are aligned for the open and closed states, as in their native structures, whereas a twisted alignment was found for the intermediate state. Ravindranathan et al. have provided evidence for this intermediate state (3). Similar evidence exists for other sugar-bound proteins like D-Glucose and Maltose binding protein (36, 38, 39).

seems to be functionally relevant because it has been hypothesized to facilitate ribose transfer in the permease complex—a partially closed conformation with a more weakly bound ribose might help in providing an easier release of ribose into the membrane-bound permease (3).

A similar twisted intermediate state was suggested for the D-Glucose binding protein, another member of the same family, based on experimental studies using disulfite-trapping and fluorescence spectroscopy (36, 37). Very recently, accelerated molecular dynamics were used to suggest the existence of a semiclosed state for the Maltose binding protein, a member of a closely related family (38, 39). Our SBM+DCA-based model predicts another twisted intermediate for the D-Allose binding protein (PDB IDs 1gud and 1rpj) of the same family (Fig. S7). The existence of such a semiclosed twisted state seems to be a general feature of the sugar-bound periplasmic proteins.

We have investigated additional systems using this methodology with consistent results. For example, in the case of the 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase (PDB IDs 1rf5 and 1rf4), we have also found a hidden intermediate state (Fig. S8). Experimental evidence supporting this claim exists in the form of an ortholog (PDB ID 3roi) with similar topological features. The structure of this ortholog suggests that the state we found may be of functional relevance for EPSP. As a matter of negative control, we have also studied the case of the D class β-lactamase (PDB ID 1h8y) that is a member of a largely populated family of transpeptidases (PF00905). This protein is known to have only one state. When adding the same amount of DCA pairs to the open-state topology, we did not observe any spurious additional state (Fig. S9).

Discussion

In this study, we combined physical models of proteins, for instance SBM, with coevolutionary constraints obtained using direct coupling analysis. Our results provide support that information about conformational plasticity can be retrieved from a collection of evolutionary related protein sequences. This is a consequence of the fact that diverse states (intermediates/closed), which are of functional importance, are selected by evolution, because DCA captures evolutionarily significant residue–residue correlations, regardless of whether the interaction stabilizes the final or intermediate state. To clearly observe these types of transitions, the conformational states need to have unique subsets of contacts. Subtle conformational changes where only contact distances are changed, e.g., conformation differences between ATP and ADP binding in the active site, are harder to capture by SBM+DCA. Nonetheless, this enhanced sampling of the functional conformational space of proteins might have broader implications in protein structure determination as well as in the design of ligands that can trap intermediate states. Such ligands could be used to crystallize states that were previously difficult to access and also be used in the process of rational drug design. Our observations and theoretical framework are general enough to be applied to many protein families with enough sequence information, in principle even for those families without any experimental structural state available.

Materials and Methods

Directly Coupled Residue Pairs. We use DCA to estimate directly coupled coevolving residue–residue physical contacts. DCA models the joint probability distribution of amino acid sequences with an exponential function that depends on single-site amino acid frequencies and pairwise interactions. An approximation of pairwise energies is calculated by inverting the connected correlation matrix computed from multiple sequence alignments. These pairwise terms are used to compute probabilities of "direct couplings" among amino acid pairs. When applying DCA to a set of sequences of a given family [e.g., Pfam domains (40)], then residue pairs that show the largest amount of direct coupling or direct information (DI) tend to be a proxy of residue–residue contacts in the 3D fold of a protein that is part of such family. For more details about DCA and an evaluation of its performance using a mean field formulation, refer to Morcos et al. (11). For each of the families analyzed, we used the top ranked pairs based on the DI metric. We used a pseudocount value of $\lambda = M_{eff}$. The value of M_{eff} in turn was computed using a correction of sampling bias for proteins with sequence identity of 80%. The number of M_{eff} is shown as "effective sequences" in Table S1. The number of DCA contacts used was proportional to the total number of native contacts in the open state of a protein, using shadow contact maps. For all of the systems studied, the number of DCA contacts used was 1.75 times the number of native contacts. The results are robust for a range of 1.5–2 times the number native shadow contacts. We used a cutoff value for shadow contacts of 6 Å + 1 Å atom "shadowing" radius; this is the standard value used in the smog web server (16).

Single-Basin SBM. We built our single-basin SBM from a single native structure (open state of the binding protein) by placing a single bead of unit mass for each amino acid at the location of the Cα atom (10, 11). The energy function used for the SBM is given as

$$\mathrm{H}_{\mathrm{SBM}}\left(\left\{\vec{r}_i\right\}\right) = H_b^O + H_{nb}^O. \qquad \textbf{[1]}$$

Here, the superscript O refers to the open state and H_b^O represents the local bonded component of the Hamiltonian,

$$H_b^O = \sum_{i=1}^{N-1} \frac{K_r}{2}\left(r_{i,i+1} - r_{i,i+1}^{0(O)}\right)^2 + \sum_{i=1}^{N-2} \frac{K_\theta}{2}\left(\theta_i - \theta_i^{0(O)}\right)^2 + \sum_{i=1}^{N-3}\sum_{n=1,3} K_\phi^{(n)}\left(1 - \cos\left[n\left(\phi_i - \phi_i^{0(O)}\right]\right)\right). \qquad \textbf{[2]}$$

The first term in H_b^O ensures that the bond distance $r_{i,i+1}$ between the neighboring residues i and $i+1$ is constrained harmonically with respect to its native bond distance $r_{i,i+1}^{0(O)}$ by a spring constant $K_r = 20(\mathrm{kJ/mol.\mathring{A}}^2)$. The second term constrains the angle θ_i among the residues i, $i+1$, and $i+2$ with respect to its native value $\theta_i^{0(O)}$ by a harmonic spring constant $K_\theta = 20(kJ\,/\,mol.rad^2)$. The third term represents the dihedral angle potential with $K_\phi^{(1)} = 2K_\phi^{(3)}$ that describes the rotation of the backbone involving successive residues from i to $i+3$. The native values $r_{i,i+1}^{0(O)}$, $\theta_i^{0(O)}$, and $\phi_i^{0(O)}$ are taken from the open conformation crystal structure. The value of $K_\phi^{(1)}$ is chosen carefully to ensure better sampling of the conformational space while ensuring sufficient stabilization of the open state. The nonlocal part of the Hamiltonian, H_{nb}^O is given by

$$H_{nb}^O = \sum_{i=1}^{N-4}\sum_{j=i+4}^{N}\left[\varepsilon^O\left(\left(\frac{r_{ij}^{0(O)}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{0(O)}}{r_{ij}}\right)^{6}\right)\Delta_{ij}^O + \varepsilon_r\left(\frac{\sigma}{r_{ij}}\right)^{12}\left(1-\Delta_{ij}^O\right)\right]. \qquad \textbf{[3]}$$

The 6–12 Lennard–Jones (LJ) potential is used in H_{nb}^O to describe the interactions that stabilize the nonbonded native contacts. Native contact pairs (i and j) are obtained using the shadow contact map that is implemented in SMOG (16). If i and j residues are in contact in the native state, $\Delta_{ij}^O = 1$; otherwise $\Delta_{ij}^O = 0$. Native contact pair distance $r_{ij}^{0(O)}$ is obtained from open-state structure. Nonnative pairs with $\Delta_{ij}^O = 0$ are under repulsive potential with a distance parameter $\sigma = 4$ Å. The strength of repulsive potential ε_r is 1 kJ/mol. However, the value of ε^O is chosen carefully to ensure better sampling of the conformational space and sufficient stabilization of the open state. These parameters are sampled until we observe the existence of new basins with a sufficiently large population.

Dual-Basin Structure-Based Model. As the name suggests, the dual-basin structure-based model (dSBM) is built using two experimental structures of a single protein, namely open and ligand-bound closed states. The energy function used for this model is given as

$$\mathrm{H}_{\mathrm{dSBM}}\left(\left\{\vec{r}_i\right\}\right) = H_b^O + H_{nb}^{\mathrm{dSBM}}. \qquad \textbf{[4]}$$

Here, H_b^O represents the local part of the Hamiltonian as in the single SBM. H_{nb}^{dSBM} describes the nonlocal part of the dual-SBM Hamiltonian and has the form

$$H_{nb}^{\mathrm{dSBM}} = \sum_{i=1}^{N-4}\sum_{j=i+4}^{N}\left[\varepsilon^O\left(\left(\frac{r_{ij}^{0(O)}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{0(O)}}{r_{ij}}\right)^{6}\right)\Delta_{ij}^O + \varepsilon^C\left(\left(\frac{r_{ij}^{0(C)}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{0(C)}}{r_{ij}}\right)^{6}\right)\Delta_{ij}^{C_u} + \varepsilon_r\left(\frac{\sigma}{r_{ij}}\right)^{12}\left(1-\left(\Delta_{ij}^O \vee \Delta_{ij}^{C_u}\right)\right)\right]. \qquad \textbf{[5]}$$

The first term in the summation is same as the first term in H_{nb}^O and derived from open structure. The second term is derived from the closed-state structure and $\Delta_{ij}^{C_u} = 1$ if residues i and j are in contact in closed state but not in open state. For these unique closed-state contacts, we also use the native contact distances $r_{ij}^{0(C)}$ derived from the closed-state crystal structure. The last term accounts for the repulsion between the nonbonded pairs that are not in contact either in open state or in closed state and the logic or operator ∨ is used for that purpose. Here, the values for ε^O and ε^C are chosen carefully to sample both the states.

SBM+DCA Model. For our hybrid SBM+DCA model, we combine the Hamiltonian of the open state and supplement with DCA-predicted contact pairs. The energy function for this hybrid method is given by

$$\mathrm{H}_{\mathrm{hybrid}}\left(\left\{\vec{r}_i\right\}\right) = H_b^O + H_{nb}^{\mathrm{hybrid}}. \qquad \textbf{[6]}$$

Here, H_b^O represents the local part of the Hamiltonian as in the single SBM and derived from open-state crystal structure. H_{nb}^{hybrid} has the form

$$H_{nb}^{\mathrm{hybrid}} = \sum_{i=1}^{N-4}\sum_{j=i+4}^{N}\left[\varepsilon^O\left(\left(\frac{r_{ij}^{0(O)}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{0(O)}}{r_{ij}}\right)^{6}\right)\Delta_{ij}^O + \varepsilon^{DCA}\left(\left(\frac{\sigma^{DCA}}{r_{ij}}\right)^{12} - 2\left(\frac{\sigma^{DCA}}{r_{ij}}\right)^{6}\right)\Delta_{ij}^{DCA_u} + \varepsilon_r\left(\frac{\sigma}{r_{ij}}\right)^{12}\left(1-\left(\Delta_{ij}^O \vee \Delta_{ij}^{DCA_u}\right)\right)\right]. \qquad \textbf{[7]}$$

The first term in the summation is the same as the first term in H_{nb}^O and derived from open structure. The second term is derived from the DCA pair and $\Delta_{ij}^{DCA_u} = 1$ if the residue pair i and j appear as a top-ranked DCA contact but not in open state. For these unique DCA contacts, we use the native contact distance $\sigma^{\mathrm{DCA}} = 8$ Å and LJ well-depth $\varepsilon^{\mathrm{DCA}} = 0.7\ \mathrm{kJ \cdot mol^{-1}}$. The last term accounts for the repulsion between the nonbonded pairs that are not in contact either in open state or in DCA pairs and the logic or operator ∨ is used for that purpose. Here, the value of ε^O is chosen carefully to sample the phase space efficiently. The values for ε^O and $K_\phi^{(1)}$ range between 0.4 and 0.6.

ACKNOWLEDGMENTS. This work was supported by the Center for Theoretical Biological Physics sponsored by the Welch Foundation Grant C-1792, by the National Science Foundation (NSF) (Grant PHY-1308264), and by Grant NSF-MCB-1214457. J.N.O. is a CPRIT Scholar in Cancer Research sponsored by the Cancer Prevention and Research Institute of Texas.

1. Anfinsen CB (1973) Principles that govern folding of protein chains. *Science* 18(4096): 223–230.
2. Kim MK, Chirikjian GS, Jernigan RL (2002) Elastic models of conformational transitions in macromolecules. *J Mol Graph Model* 21(2):151–160.
3. Ravindranathan KP, Gallicchio E, Levy RM (2005) Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *J Mol Biol* 353(1):196–210.
4. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242.

5. UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 40(Database issue):D71–D75.
6. de Juan D, Pazos F, Valencia A (2013) Emerging methods in protein co-evolution. *Nat Rev Genet* 14(4):249–261.
7. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18(4):309–317.
8. Park S, Kono H, Wang W, Boder ET, Saven JG (2005) Progress in the development and application of computational methods for probabilistic protein design. *Comput Chem Eng* 29(3):407–421.
9. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295–299.
10. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. *Cell* 138(4):774–786.
11. Morcos F, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108(49):E1293–E1301.
12. Taylor WR, Sadowski MI (2011) Structural constraints on the covariance matrix derived from multiple aligned protein sequences. *PLoS ONE* 6(12):e28265.
13. Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190.
14. Haq O, Andrec M, Morozov AV, Levy RM (2012) Correlated electrostatic mutations provide a reservoir of stability in HIV protease. *PLoS Comput Biol* 8(9):e1002675.
15. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* 298(5):937–953.
16. Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN (2010) SMOG@ctbp: Simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res* 38(Web Server issue):W657-W661.
17. Oklejas V, Zong C, Papoian GA, Wolynes PG (2010) Protein structure prediction: Do hydrogen bonding and water-mediated interactions suffice? *Methods* 52(1):84–90.
18. Davtyan A, et al. (2012) AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J Phys Chem B* 116(29):8494–8503.
19. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci USA* 109(26):10340–10345.
20. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* 106(1):67–72.
21. Schug A, et al. (2010) Computational modeling of phosphotransfer complexes in two-component signaling. *Methods Enzymol* 471:43–58.
22. Dago AE, et al. (2012) Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proc Natl Acad Sci USA* 109(26):E1733–E1742.
23. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12):e28766.
24. Hopf TA, et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149(7):1607–1621.
25. Dill KA, MacCallum JL (2012) The protein-folding problem, 50 years on. *Science* 338(6110):1042–1046.
26. Nugent T, Jones DT (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci USA* 109(24):E1540–E1547.
27. Taylor WR, Jones DT, Sadowski MI (2012) Protein topology from predicted residue contacts. *Protein Sci* 21(2):299–305.
28. Brylinski M, Skolnick J (2008) What is the relationship between the global structures of apo and holo proteins? *Proteins* 70(2):363–377.
29. Seeliger D, de Groot BL (2010) Conformational transitions upon ligand binding: Holo-structure prediction from apo conformations. *PLoS Comput Biol* 6(1):e1000634.
30. Magnusson U, Salopek-Sondi B, Luck LA, Mowbray SL (2004) X-ray structures of the leucine-binding protein illustrate conformational changes and the basis of ligand specificity. *J Biol Chem* 279(10):8747–8752.
31. Daura X, et al. (1999) Peptide folding: When simulation meets experiment. *Angew Chem Int Ed* 38(1):236–240.
32. Okazaki K, Koga N, Takada S, Onuchic JN, Wolynes PG (2006) Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc Natl Acad Sci USA* 103(32):11844–11849.
33. Armstrong N, Gouaux E (2000) Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: Crystal structures of the GluR2 ligand binding core. *Neuron* 28(1):165–181.
34. Hsiao CD, Sun YJ, Rose J, Wang BC (1996) The crystal structure of glutamine-binding protein from Escherichia coli. *J Mol Biol* 262(2):225–242.
35. Hsiao CD, et al. (1994) Crystals of glutamine-binding protein in various conformational states. *J Mol Biol* 240(1):87–91.
36. Careaga CL, Sutherland J, Sabeti J, Falke JJ (1995) Large amplitude twisting motions of an interdomain hinge: A disulfide trapping study of the galactose-glucose binding protein. *Biochemistry* 34(9):3048–3055.
37. Messina TC, Talaga DS (2007) Protein free energy landscapes remodeled by ligand binding. *Biophys J* 93(2):579–585.
38. Bucher D, Grant BJ, Markwick PR, McCammon JA (2011) Accessing a hidden conformation of the maltose binding protein using accelerated molecular dynamics. *PLoS Comput Biol* 7(4):e1002034.
39. Bucher D, Grant BJ, McCammon JA (2011) Induced fit or conformational selection? The role of the semi-closed state in the maltose binding protein. *Biochemistry* 50(48): 10530–10539.
40. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38(Database issue):D211–D222.