Quantitative theory of hydrophobic effect as a driving force of protein structure

Nikolay Perunov and Jeremy L. England*

Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received 9 September 2013; Accepted 6 January 2014
DOI: 10.1002/pro.2420
Published online 9 January 2014 proteinscience.org

Abstract: Various studies suggest that the hydrophobic effect plays a major role in driving the folding of proteins. In the past, however, it has been challenging to translate this understanding into a predictive, quantitative theory of how the full pattern of sequence hydrophobicity in a protein shapes functionally important features of its tertiary structure. Here, we extend and apply such a phenomenological theory of the sequence-structure relationship in globular protein domains, which had previously been applied to the study of allosteric motion. In an effort to optimize parameters for the model, we first analyze the patterns of backbone burial found in single-domain crystal structures, and discover that classic hydrophobicity scales derived from bulk physicochemical properties of amino acids are already nearly optimal for prediction of burial using the model. Subsequently, we apply the model to studying structural fluctuations in proteins and establish a means of identifying ligand-binding and protein–protein interaction sites using this approach.

Keywords: hydrophobicity scale; protein structure; conformational fluctuations; ligand-binding sites; mutations

Introduction

Since the experiments of Anfinsen,[1] the field of structural biology has been motivated by the idea that the shape of a protein is completely determined by its sequence. Increasingly, however, it has been assumed that this mapping from sequence to structure is affected by such a diverse combination of physical interactions that a detailed simulation framework must be necessary to make accurate predictions about real proteins. Advances in hardware and simulation methods have led to various breakthroughs in the computer simulation of protein folding with all-atom resolution: the massive parallelization of trajectories for heavy sampling,[2] the optimization of supercomputing on the millisecond timescale,[3] and the improvement of algorithms for searching the energy landscapes of macromolecules have brought many structure-prediction and design goals within reach.[4]

However, even considering the success of such computational methods in shedding new light on macromolecular structure and function, their high-computational cost[2,3] and dependence on numerous modeling parameters raise the possibility that complementary insights might still be gained using a more theoretically and computationally simple approach. Such a method would potentially have at least two advantages: that on a fixed computational budget it could be applied to a much larger corpus of protein sequences or used to sample a wider diversity of low-energy structures; and, that the small number of modelling assumptions would make it easier to determine where the model is expected to succeed as well as where it might fail.

In the search for a simple physical principle to incorporate into the assumptions of such a model, the hydrophobic effect is a highly attractive choice. Various studies suggest that the hydrophobic effect plays a major role in the folding of proteins.[5–7] However, although the hydrophobic effect is well

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: J. L. England, Building 6C–307, 77 Massachusetts Avenue, Cambridge, MA 02139.
E-mail: jengland@mit.edu

understood at the level of individual amino acids—nonpolar amino acid residues tend to be buried in the core of the protein, and the polar residues are more likely to be on the surface—a quantitative theory of how the hydrophobic effect impacts structure as a whole in real globular proteins is difficult to construct. The lattice HP models, where a protein is represented as a sequence of nonpolar (H) and polar (P) residues with attractive interaction between H residues, quite often do not give unique native structures, so that the predictions of these models cannot be translated to real protein structures.[8,9] Hydrophobicity profiles, which are constructed by averaging sequence hydrophobicity, are known to correlate reasonably well with the burial of amino acid residues in globular proteins.[10,11] However, the methods that use hydrophobicity profiles to predict burial generally do not include nontrivial effects of the polypeptide chain and do not account for the limited space in the core of a protein domain, which limits application of these methods.

Previously, we introduced a model of protein folding, termed here the "burial mode model," that considers the hydrophobic effect, steric repulsion, and the polymeric constraints of the protein backbone to be the driving forces of protein structure.[12] Using only the amino acid sequence of a protein, this model allows one to compute not only the minimum energy conformational state of a protein but also an ensemble of low-energy excited states. Knowledge of these states has in turn been demonstrated to be useful for studying coupled motion of different parts of a protein in allosteric motion.

For a 100–300 residue protein, it takes less than a second to use the burial mode model to compute tertiary structural information on a single CPU. Thus, it might eventually be appealing to apply the model to studying the large collections of sequence homologs which became available with high-throughput genomic sequencing. However, before doing so one must clearly understand the model's domain of applicability, and which input parameters make it most successful in capturing the structural physics of protein domains.

In this study, we first examine whether our approach can be improved by choosing a better set of parameters. To accomplish this, we undertake to compute a new amino acid hydrophobicity scale from a large set of known protein structures, and compare this performance of this scale to those of known hydropathy scales. Having identified a suitable set of parameters, we then undertake to explore the confounding effects of interdomain interactions on the model's ability to predict burial in protein monomers. By doing so, we discover a new application for the model in the analysis of conformational fluctuations related to ligand-binding and mutation.

Results

Burial mode model

In the burial mode model, a globular protein domain is represented as a linear chain of N residues which are indexed by the number s and have position $\vec{r}(s) = [x(s), y(s), z(s)]$ relative to the center of mass of the globule (Fig. 1). The polymeric bonds and the hydrophobic effect are incorporated into the system energy

$$\mathcal{E} = \sum_{s=1}^{N-1} \kappa |\vec{r}(s+1) - \vec{r}(s)|^2 + \sum_{s=1}^{N} \varphi(s) |\vec{r}(s)|^2, \qquad (1)$$

The bond stiffness κ determines the strength of "harmonic spring-like" attraction between adjacent monomers along the chain, which sets the overall elastic extensibility of the polymer. The relative hydropathy $\varphi(s)$ reflects the tendency of each different amino acid in the chain to be exposed on the globule's surface or buried in its core, and is obtained by converting amino acid sequence into numbers using the standard Kyte-Doolittle (KD) hydrophobicity scale.[13] It should be noted that quadratic form of the hydrophobic contribution to the energy was chosen for two reasons: first, it allows the model to be analytically tractable; second, it has a physical intuition that force acting on the residue near the surface is larger than in the core because on the surface the amino acid is more likely to have larger area exposed to the solvent. The steric repulsion between different parts of a chain is taken into account as a global constraint on the ratio α of the gyration radius squared to the maximum distance to the center of mass squared R^2

$$R_g^2 = \frac{1}{N} \sum_{s=1}^{N} |\vec{r}(s)|^2 = \alpha R^2. \qquad (2)$$

The goal of this constraint is to prevent residues from collapsing into the center of the globule and, thus, to account for the limited space in the packed globular core.

To compute the lowest energy conformation of the protein one should minimize the system energy (1) subject to constraint (2). As shown in previous published work, this procedure can be reduced to an exactly solvable linear programming problem.[12] The optimized outcome of the linear program is given in the form of an energy-minimizing "burial trace," that is, the squared distance $|\vec{r}(s)|^2$ from each residue to the center of mass.

To quantify the performance of the model on a given protein, one may compute Pearson's correlation coefficient (PCC) between the burial trace computed from the sequence using the model and the burial trace generated from the known structure of

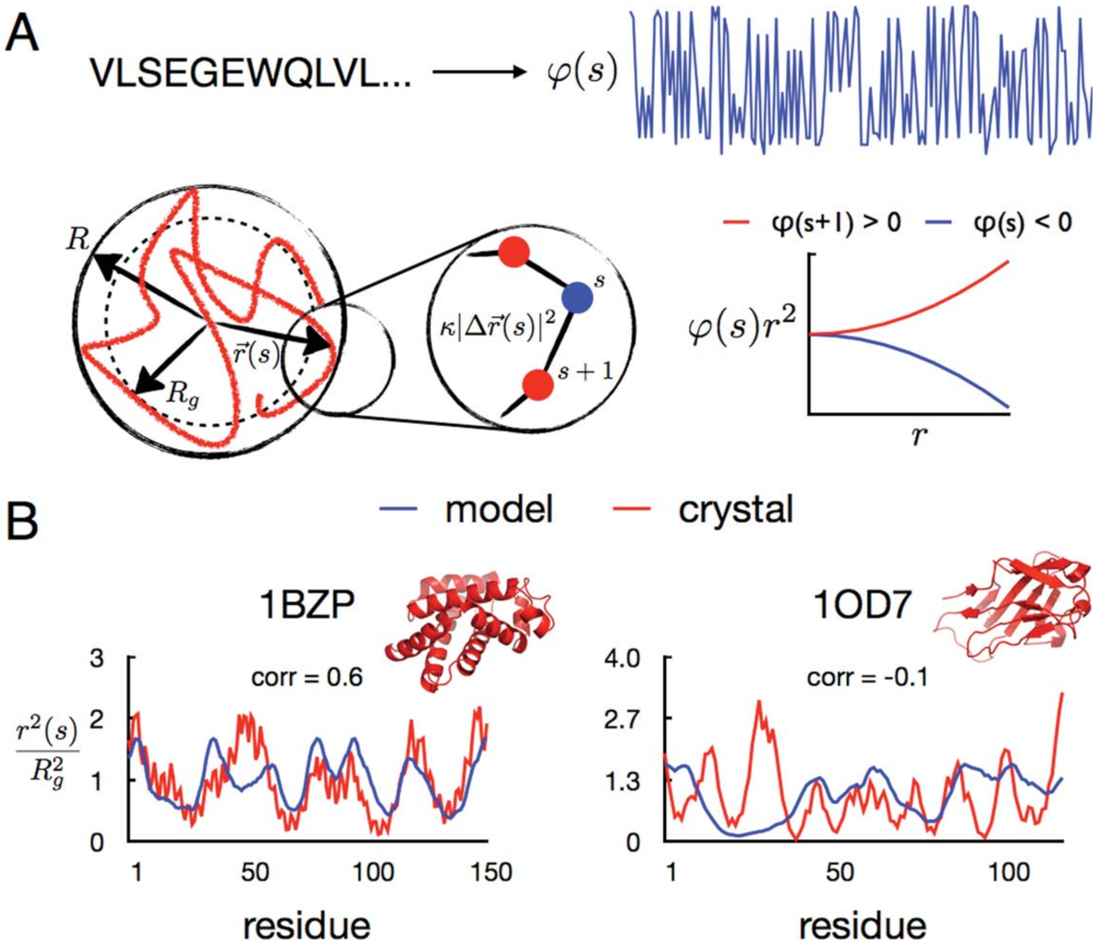

Figure 1. Basic assumptions of the burial mode model. (A) The protein backbone is represented as a linear chain (red solid line) with residues indexed by the number *s* and that have position $\vec{r}(s)$ relative to the center of the globule. The black solid line shows the maximum size of the globule, while the black dashed line shows the radius of gyration R_g. The hydropathy of each residue $\varphi(s)$ is determined by the type of the residue. Neighboring residues are connected by harmonic springs of stiffness κ. Blue and red residues represent hydrophilic and hydrophobic amino acids, respectively. The plot in the bottom right corner shows contribution of different residues to the system energy as a function of the distance to the center of the globule; and (B) Burial traces computed using the model (blue lines) and from the crystal structures (red lines) of sperm whale myoglobin (1BZP) and sialoadhesin (1OD7). The PCC between the model and the structure is 0.6 for myoglobin and -0.1 for sioloadhesin.

the protein using coordinates of C_α atoms. (Note: To compute burial traces one can also use coordinates of C_β atoms or side chain centroids but this does not change burial traces significantly.) Examples of proteins for which the model gives different PCC values are shown in Figure 1(B). As one can see from this figure, for the proteins with high PCC (>0.4) the resemblance between burial traces is striking, whereas for the proteins with low PCC (<0.1) the model correctly predicts only positions of a few local extrema of the burial trace.

In globular protein domains, burial traces show which parts of the protein are buried in the core and which parts are exposed to water. In this regard, burial traces are similar to hydrophobicity profiles or window-averaged sequence hydrophobicities $\varphi(s)$, which are widely used to find out information about the secondary and the tertiary structure of proteins from their sequences.[11,14] However, unlike hydrophobicity profiles, which do not contain any explicit information about conformational changes, the burial mode model allows one to compute the ensemble of burial traces for low-energy excited states of the chain and, thus, provides a framework for studying conformational fluctuations in proteins. Previously, this framework has been successfully used to explain allosteric motion in a panel of test proteins for which the PCC between the burial traces from the sequence and structure was greater than 0.4.[12]

The mapping of the sequence-structure relationship that is effected in the burial mode model simplifies, and, thus, accelerates the calculation so that it becomes an attractive tool for studying large collections of proteins. However, to use the model as a reliable method for analysis of conformational fluctuations one should identify the set of physical

Table I. *Comparison of the Model Performance with Different Hydrophobicity Scales for Different Classes of Proteins from the SCOP Database*

	Protein class			
Hydrophobicity scale	α	β	$\alpha+\beta$	α/β
Kyte-Doolittle	0.25±0.22	0.22±0.18	0.25±0.18	0.23±0.20
Wimley–White	0.24±0.23	0.21±0.19	0.21±0.18	0.21±0.19
Janin	0.22±0.23	0.18±0.19	0.23±0.18	0.20±0.19

Each column shows the mean and the standard deviation of the distributions of PCC between the burial traces computed from sequences using different hydrophobicity scales and the burial traces extracted from protein structures for a given SCOP class.

parameters that makes the model applicable to the broadest set of proteins. Thus, motivated by the previous successes of the model in explaining allosteric motion for the proteins with high PCC, we attempted to improve the model's power to predict burial traces by optimizing its input parameters.

Parameter optimization

There are 21 independent parameters in the burial mode model: the bond stiffness (κ), the ratio of the squared gyration radius to the squared maximum radius of the protein (α), and 19 relative hydrophobicities of amino acid residues. However, not all parameters can be changed given the model's assumptions. First of all, the bond stiffness κ fixes the unit of length, and must be chosen so that corresponding mean-square distance between neighboring C_α atoms is equal to one; the parameter α ranges from 0.4 to 0.6 in real proteins and is set to 3/5, which is the value that would hold for a globular protein that was spherical and had uniform density. The maximum radius of the protein, meanwhile, is estimated from the number of monomers in the chain, and is given by $R=(3N/4\pi\rho_0)^{1/3}$, where ρ_0 is the density of monomers estimated from the crystal structure of the TIM barrel fold (PDB ID 2VXN). Thus, it is the amino acid hydrophobicity scale that offers some remaining parametric flexibility and could perhaps be optimized to improve the model's burial trace prediction.

We first investigated, how the burial mode model's performance changes when we use different standard hydrophobicity scales. Based on the methods by which they were developed, hydrophobicity scales can be divided into two groups: experimental scales, which are based on the measurements of the free energy of solvation of single amino acids or short peptides in water and ethanol[13,15,16] and numerical scales, which are derived from the partition of amino acid residues between the core and the surface in proteins with known three-dimensional (3D) structures.[5,17] In our previous study, the relative hydrophobicities of amino acid residues were taken from the KD scale and standardized so that the energy change associated with transfer of glutamine from surface to the core of the globule is equal to 0.5 k_BT. To compare the performance of the model with different hydrophobicity scales, we normalized all scales so that the difference between the maximum and the minimum hydrophobicities was the same as in the KD scale. Table I shows the mean and the variance of the distributions of PCC for different classes of proteins from the structural classification of proteins (SCOP) database.[18] Interestingly, despite the different origins of the hydrophobicity scales, none of the scales significantly altered the performance of the model on this large set of proteins (SCOP class).

Next, we did a brute-force search for a better hydrophobicity scale. For large groups of proteins (SCOP classes/folds), it is computationally costly to fit burial traces using a 20-letter amino acid alphabet, so we elected to use a reduced-size amino acid alphabet for these searches. We first split amino acids into four groups according to their hydrophobicity indices in the KD scale: (R, K, D, E, Q, N, H), (P, Y, W, S, T, G), (A, M, C, F), and (L, V, I). Because this is a somewhat arbitrary way to split amino acids into groups, as a control we also divided amino acids into random groups. Then, we generated a 4D rectangular grid with 10 nodes along each axis. The range of hydrophobicity indices was set between -9 and 9—twice the minimum and maximum values of KD scale, respectively. In the case, when amino acids were divided into groups at random, we found that the distributions of PCC for α-helical proteins were always broad (st. dev. ≈ 0.2) and their mean was never greater than 0.2 (the data are shown in Supporting Information); whereas when amino acids were grouped according to the KD scale, the mean of the distribution of PCC never exceeded 0.3 and the standard deviation was about 0.2. It should be noted that out of 10^4 different hydrophobicity scales, we examined only 2% had the mean of the distribution of PCC higher than 0.25, the mean PCC for the KD scale. Furthermore, the hydrophobicity scales that provided high values of the mean PCC were in good agreement with the KD scale (Supporting Information Fig. S2). Taking into account the data in Table I and the results of the exhaustive search, one can conclude that one cannot achieve a significantly better performance for the model on large groups of proteins using four-letter hydrophobicity scales.

To investigate if the model's power to predict burial traces can be improved with a 20-letter amino acid alphabet, we developed a method to derive a hydrophobicity scale from real protein structures, using physical assumptions in line with those of the model. In particular, we noted that any two amino acids of any two given types in adjacent positions on a protein chain are forced to "live" in nearly identical environments. Because of this, one might suppose that their relative position in space with respect to the center of the protein globule in a crystal structure could provide an all-things-equal comparison of the tendencies of each amino acid to be buried in the globular core. Put another way, a relatively greater tendency of one amino acid in such a pair to be buried might be indicative of a relatively greater hydrophobicity.

To pursue this idea, we treated a large collection of proteins with known 3D structures as an ensemble of amino acid pairs, in which the relative burial of neighboring amino acids is determined only by their relative hydrophobicity. We examined the distribution of amino acid positions inside globular protein domains with unique sequences and constructed a matrix M_{ij}, each element of which was defined to be the number of times that a residue of type i is further from the center of the globule than residue of type j, given that these residues are the nearest neighbors on a chain (Fig. 2). By positing that the probability of amino acid of type i being closer to the center of the globule than amino acid of type j is given by a Boltzmann weight, we find that the relative hydrophobicity $\Delta\varphi_{ij}$ of these amino acids is given by

$$\Delta\varphi_{ij}=\varphi_i-\varphi_j \propto \ln\frac{M_{ij}}{M_{ji}}.$$

Repeating this procedure for every pair of amino acids provides 190 relative hydrophobicities $\Delta\varphi_{ij}$. Thus, to compute 19 hydrophobicity indices φ_i of single amino acids we did a least squares optimization. Figure 2(B) shows the matrix of relative positions of amino acid residues M_{ij} and the hydrophobicity indices φ_i computed for a set of α-helical protein domains with unique sequences of length between 100 and 300 a.a. from the SCOP database (970, in total). To compute this matrix, we used only the residues that are far from the center of a domain ($|\vec{r}(s)|^2 > 0.5R^2$). Strikingly, this new hydrophobicity scale (called "α-rpm") that we computed from burial information in real crystal structures turned out to agree quite well with the both the KD scale and with the Wimley–White (WW) scale [Fig. 2(B)]. Thus, by devising a new procedure to quantify the empirical relative statistical force on adjacent amino acids on a protein chain, we seem to have somewhat surprisingly discovered that classic hydrophobicity scales determined decades ago from bulk physicochemical measurements on amino acids already constitute a nearly optimal model of how the hydrophobic effect drives burial trends of adjacent amino acids.

To confirm this, we tested how the model works with the new hydrophobicity scale. As one can see from Figure 2(C), the new parameters only slightly improve performance on a large set of proteins compared to the KD scale—roughly one quarter of all domains have PCC greater than 0.4. This finding, along with the results of our earlier searches of parameter space, suggests that there is no hydrophobicity scale that works significantly better than the KD scale, and there will always be many proteins whose structural physics cannot be captured by this simple model. Therefore, we sought next to understand better what other factors might limit the model's domain of applicability.

Sequence diversity in globins

In search of systematic blind-spots for the burial mode approach, we elected to look at a specific group of similar proteins for which the model's performance showed a wide range of outcomes. The rationale in taking this approach was to reduce the number of sequence and structural differences among the proteins being compared, so that it would be easier to correlate the remaining differences in these factors with resulting divergences in predicted burial trace.

An ideal group to consider for this purpose was the SCOP family of globins (SCOP ID a.1.1.2). The proteins in this family consist of eight α-helices forming a compact globule, which is appealing because the burial mode model does not account for nonlocal hydrogen bonding that is required for the formation of β-sheets. In light of the exceptionally good performance of the model in the case of myoglobin (PCC = 0.56), we at first expected that the calculation should work just as well for all globins. However, examining more closely the full distribution of PCC for nonredundant proteins in this family, we found that the mean PCC is only 0.40 and there are three separate peaks. Because the family of globins consists of two protein domains: myoglobin (a monomer) and hemoglobin (a heterotetramer), we decided to check if the peaks in the distribution of PCC corresponded to these proteins. As one can see from Figure 3(A), we, indeed, found that the model predicts burial traces significantly better for single domain myoglobins than for their multidomain hemoglobin cousins. For both chains of hemoglobin, Figure 3(B) shows that the model mistakenly predicts that the region 110–130, which corresponds to an interdomain interface in the tetramer, is buried.

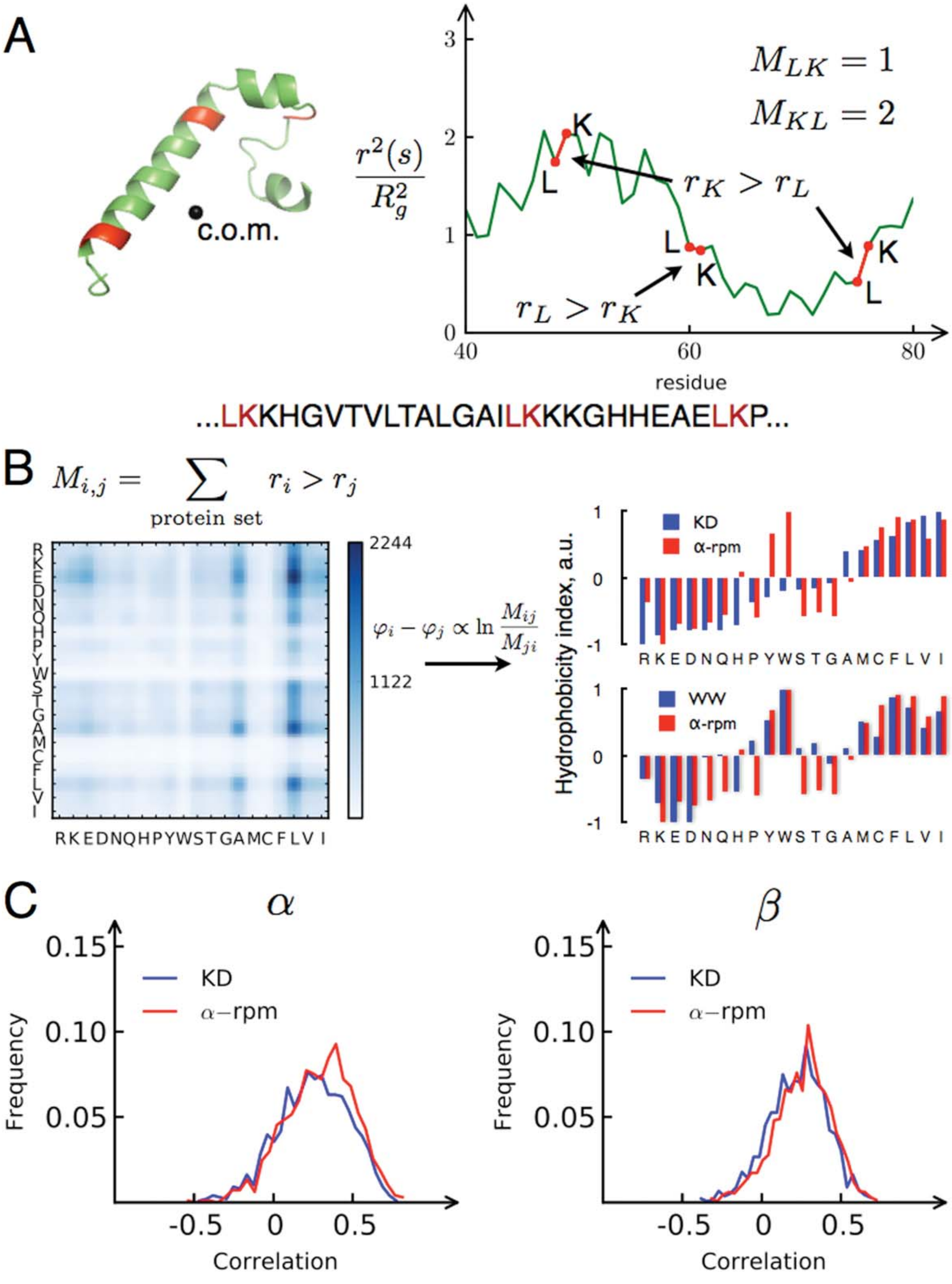

Figure 2. Extracting a model hydrophobicity scale from a set of proteins with known structures. (A) For a given protein one can compute the burial trace (right panel) corresponding to its 3D structure (left panel). Then, one can count how many times a residue of type *i* [leucine (L) in the figure] is closer to the center of the globule than residue of type *j* [lysine (K) in the figure] given that they are the nearest neighbors on the chain, (B) Repeating the procedure described earlier for all proteins from the set, one can compute the matrix of relative positions M_{ij} (left panel). On the right, comparison of the hydrophobicity scale (α-rpm) calculated from the matrix of relative positions M_{ij} with KD and WW hydrophobicity scales. The matrix M_{ij} was constructed using α domains with unique sequences of length between 100 and 300 a.a. from the SCOP database (970, in total). To compute, this matrix we used only the residues that are far from the center of a domain ($|\vec{r}(s)|^2 > 0.5R^2$), and (C) Distribution of PCC between the burial traces predicted by the model using KD and α-rpm scales and the burial traces computed from the crystal structures for α-helical and β-stranded proteins from SCOP.

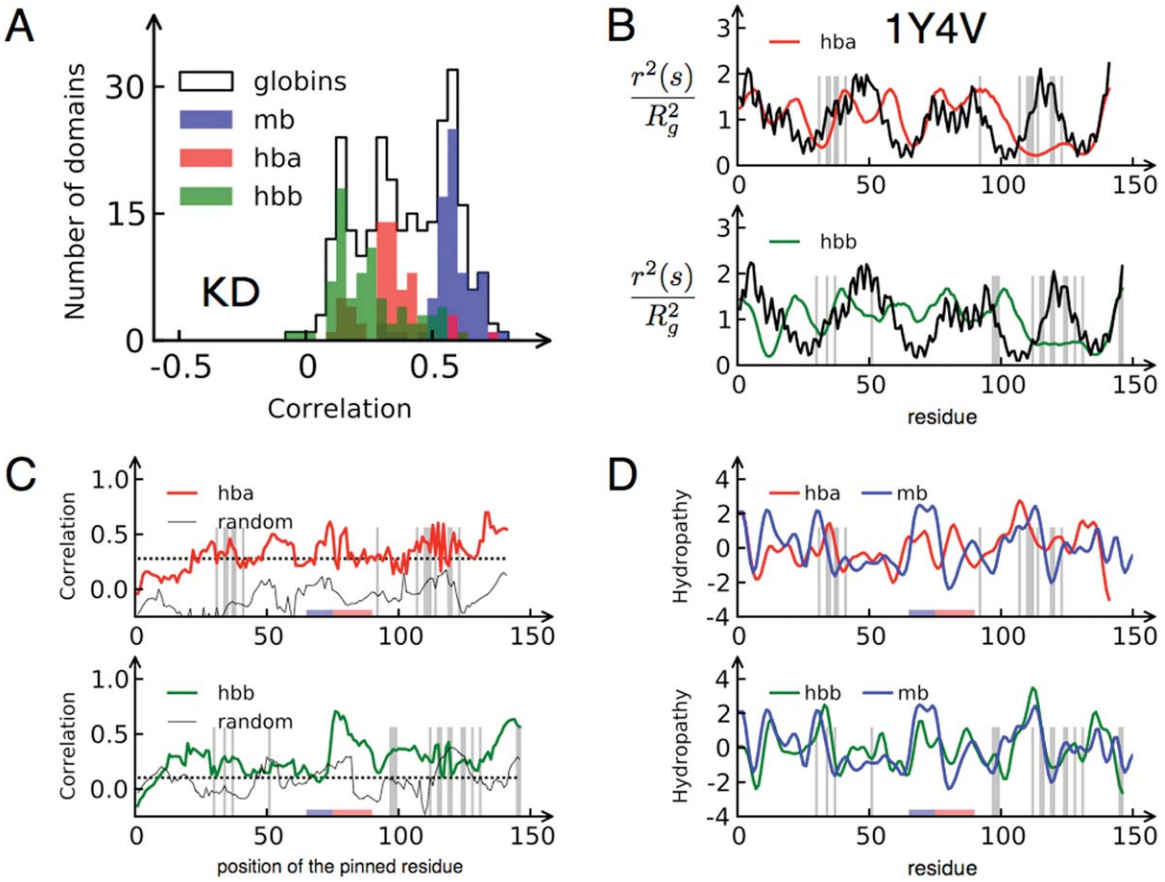

Figure 3. Interdomain interaction in hemoglobin. (A) Distribution of PCC between the burial traces predicted from the sequence using the KD scale and the burial traces computed from the crystal structures for the family of globins (SCOP ID a.1.1.2), (B) Burial traces of α and β chains of hemoglobin (1Y4V) computed from crystal structures (black lines) and using the model (red and green lines). Gray bars correspond to interdomain contacts, which were determined by the distance between C_α atoms with the threshold 6.5 Å, (C) PCC between the burial traces extracted from crystal structures of α and β chains of hemoglobin (1Y4V) and the burial traces computed using the model when one of the residues is pinned to the surface of the globule. Black solid thin lines correspond to the same procedure for the random sequence. The dashed horizontal lines correspond to PCC without pinning (0.28 for α-chain and 0.10 for β-chain), whereas black solid lines correspond to the random sequence, and (D) Hydrophobicity profiles of myoglobin (blue line) and hemoglobin (red and green lines) calculated using a sliding window of 10 residues.

These results suggested to us that interdomain interaction, which is not included in the model, might change the amino acid propensity to burial by allowing hydrophobic residues to be a part of interdomain interfaces on the surfaces of single domains.

To account for interdomain interactions in hemoglobin we introduced a perturbation to the original burial mode model. In particular, we generated ensembles of burial traces where each residue of the chain was successively pinned to the surface of the globule by setting its hydrophobicity index to a large negative number. The PCC between these burial traces and the burial traces computed from the structures of α and β chains of hemoglobin as a function of pinning position is shown in Figure 3(C). The idea behind this approach was that pinning the hydrophobic residues that are parts of interdomain interfaces to the surface would push a protein into the correct shape by changing the amount of room in the protein core, and as one can see from Figure 3(C), the model indeed predicted the burial traces better when regions corresponding to interdomain interfaces (residues 35–40, 110–130, and C-terminus) were forced to be on the surface. However, the highest PCC was achieved when residues 75–85 were pinned to the surface.

To understand why pinning this region, which is not a part of interdomain interface, improves the performance of the model, we compared the hydrophobicity profiles of myoglobin and hemoglobin [Fig. 3(D)]. As one can see from the hydrophobicity profiles, the regions of hemoglobin corresponding to interdomain interfaces are more hydrophobic than the same regions in myoglobin, but the largest differences in hydrophobicity occur in regions 62–72 and 75–85. The first region is more hydrophobic in myoglobin and is in close contact with a heme molecule,[19] whereas the second region contains more hydrophobic residues in hemoglobin and can bind to 2,3-bisphosphoglyceric acid in the deoxy state of hemoglobin.[20,21] Because of these differences in hydrophobicities, burying region 62–72 and exposing region 75–85 of hemoglobin is energetically less favorable in the framework of the original burial mode model. Therefore, by pinning residues 75–85 to the surface we just restored the propensity of this region to exposure. To summarize, from the family of globins, we have learned that the tendency of amino acid residues to be buried or exposed might be determined not only by their hydrophobicity and the available space in the core but also by whether the residues are potential sites of interaction.

Binding and mutation as triggers of conformational change

The realization that regions involved in interactions have marginal propensities to be buried gave us the idea to look at conformational fluctuations, which we would expect the burial model to predict in regions least able to "decide" whether to be buried or exposed. Continuing to study the family of globins, we generated an ensemble of burial traces with energy $\Delta E = 1 - 5k_{\mathrm{B}}T$ above the ground state energy for the sequence of sperm whale myoglobin (PDB ID 1BZP),[12] and then from these burial traces we computed the variance of squared radial distance $\mathrm{var}\,[r^2(s)]$ as a function of residue position along the chain. This function indicates the ability of each part of the chain to change its shape. Figure 4 shows the structural variability $\mathrm{var}\,[r^2(s)]$ and the 3D structure of the myoglobin colored according to this function. Strikingly, the most variable region of myoglobin corresponds the location of histidine 93, which chelates the protein's heme cofactor.[19] This result is consistent with our initial idea that the regions which can freely shift from core to surface are located close to interaction sites.

We decided to look at other proteins and to check if our method of fluctuation analysis can be used to provide analogous insight into function in a broader range of cases. We selected two proteins in which the relation between function and conformational motion is understood and for which the model succeeds in predicting ground state burial traces: H-Ras protein (3K8Y, PCC = 0.42) and chymotrypsinogen (1PYT, D chain, PCC = 0.49). H-Ras is an intracellular protein which is involved in cell division regulation, while chymotrypsinogen is a secreted protein which possesses serine protease activity. H-Ras acts as a switch in a signal transduction from membrane to the cell nucleus. In its active state H-Ras binds to GTP and converts it to GDP by cleaving the phosphate group. Figure 4(C) shows the 3D structure of H-Ras bound to GTP and the structural variability of H-Ras computed using burial mode analysis method. As one can see from this figure, the GTP binding sites of the H-ras protein (10–17, 57–61, 116–119) are located in highly fluctuating/variable regions.[22]

Figure 4(C) shows the results of similar analysis performed for chymotrypsinogen and chymotrypsin (the active form of chymotrypsinogen). The conversion of chymotrypsinogen into its active form occurs in several steps: first, chymotrypsinogen is secreted and the signal peptide (residues 1–16) is cut; then, the activation peptide (residue 17–29) is removed by trypsin. The active form of chymotrypsin (residues 30–268) has catalytic activity.[23] As one can see from Figure 4(D), both the activation peptide and the catalytic sites of chymotrypsin have high structural variability. These findings increase our confidence that the model correctly explains structural rearrangements in proteins, where the burial trace prediction matches well to the known structure.

Structural variability may, indeed, be an important physical mechanism for biological function in many proteins, however, there are also situations where one would not expect to see a signature of conformational change in this metric. It is possible that a protein's native fold might be well-structured but that it could exhibit strong sensitivity to small changes in its sequence. For example, in a recent study, Alexander et al.[24] demonstrated that it is possible to design a version of the streptococcal protein G such that a single point mutation (L45Y) leads to switching from 3α to $4\beta + \alpha$ fold. Furthermore, they obtained high-resolution NMR structures of two proteins (2KDL, 2KDM) different by three mutations (L20A, I30F, L45Y). These structures and the corresponding burial traces are shown at the top panel of Figure 5(A). While the L20A and I30F mutants do not lead to a conformational rearrangement in the protein, the L45Y mutation does, and it is clear that the map of structural variability does not reflect the corresponding pattern of mutational sensitivity.

However, we also analyzed the sensitivity of both structures to changes in sequence hydrophobicity pattern. Using the burial mode model, we constructed the response matrix

$$\chi_{s,s'} = \frac{\delta r^2(s)}{\delta \varphi(s')},$$

where $\delta r^2(s)$ is the change in predicted optimal burial trace at position s following a small change in hydrophobicity $\delta\varphi(s')$ at position s' along the chain. The rows of this matrix show how sensitive the optimal structure of the protein is to mutations. The bottom panel of Figure 5(B) depicts the response matrices computed from the sequences of 2KDL and 2KDM proteins. It should be noted that for both proteins, small changes in hydrophobicity in the region 43–47 lead to large changes in predicted burial trace. This

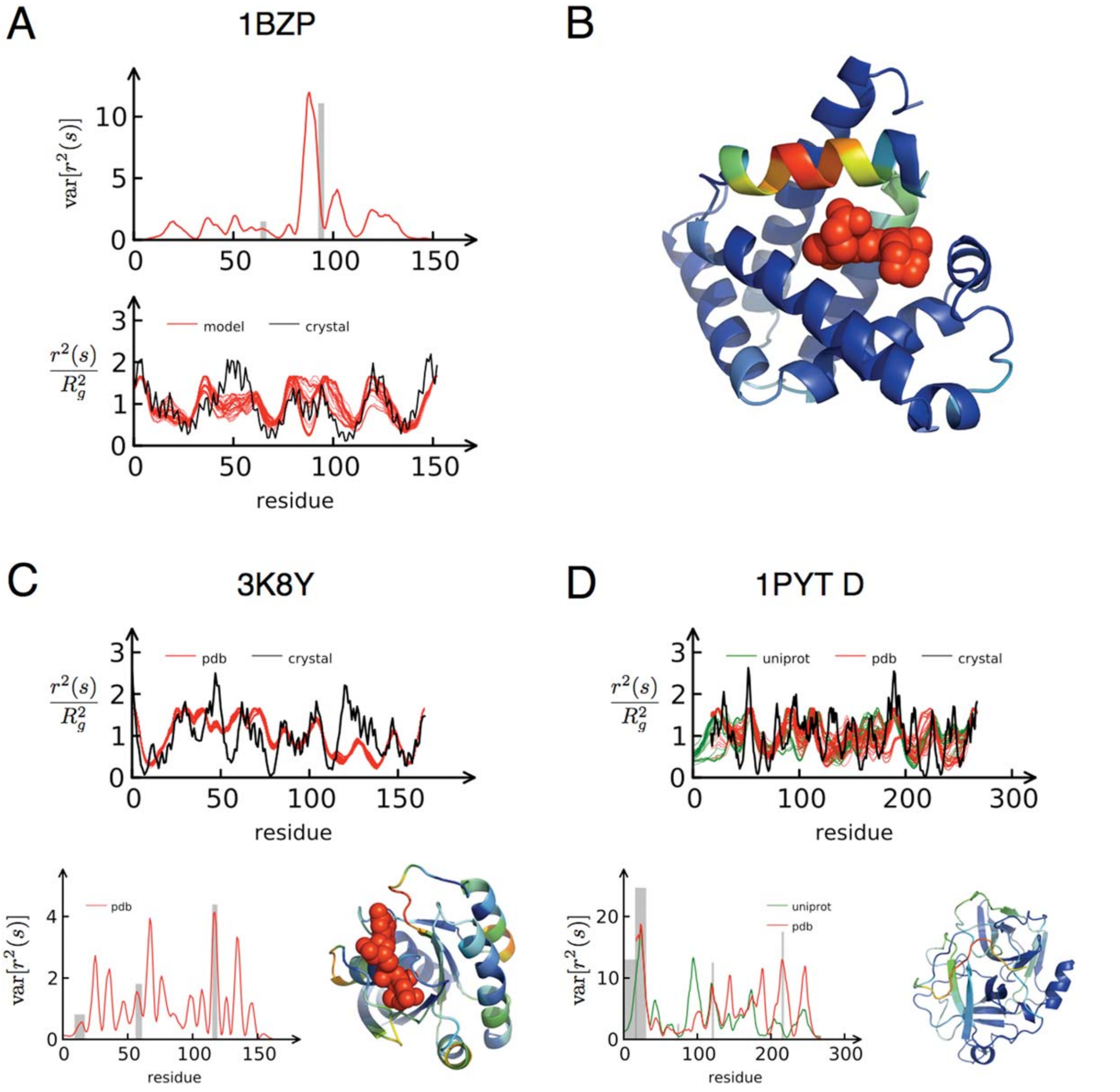

Figure 4. Conformational changes in sperm whale myoglobin (1BZP), H-Ras (3K8Y), and chymotrypsinogen (1PYT, D). (A) On the bottom panel, the solid black line corresponds to the burial trace of myoglobin computed from the crystal structure, while red lines correspond the burial traces of low-energy excited states ($\Delta E=4\ k_BT$). On the top panel, structural variability var $[r^2(s)]$ is computed from these burial traces. The gray bars on both subplots correspond to heme binding sites (residues 65 and 94), (B) The crystal structure of myoglobin is colored according to the structural variability var $[r^2(s)]$. A heme molecule is shown in red, (C) Conformational changes in H-Ras (3K8Y). On the top, burial traces of low-energy excited states of H-Ras are depicted. On the bottom, the structural variability is both plotted and colored on the crystal structure for H-Ras, as computed for burial traces of $\Delta E=4\ k_BT$. GTP binding sites are shown as gray bars, while GTP is shown in red, and (D) Structural variability of chymotrypsinogen (1PYT, D). Here, green lines correspond to the burial traces and structural variability computed for the uniprot sequence (before the signal peptide of chymotrypsinogen is cut), while red lines were computed for chymotrypsinogen sequence take from the PDB file (before the activation peptide is cleaved). Catalytic sites (H74, D121, and S216), signal and activation peptides are shown in gray. On all subplots, the structural variability var $[r(s)]$ is shown in arbitrary unit.

result is strikingly consistent with the experimental fact that mutation L45Y triggers a complete change of fold in the protein. Thus, the physical model of conformational energetics provided by the burial mode picture enables a diverse set of approaches to analyzing structural phenomena in globular protein domains.

Discussion

The problem of protein structure prediction from amino acid sequence has a long history. The most reliable approach to this problem so far—all-atom simulation—is computationally costly because it explicitly keeps track of the multitude of

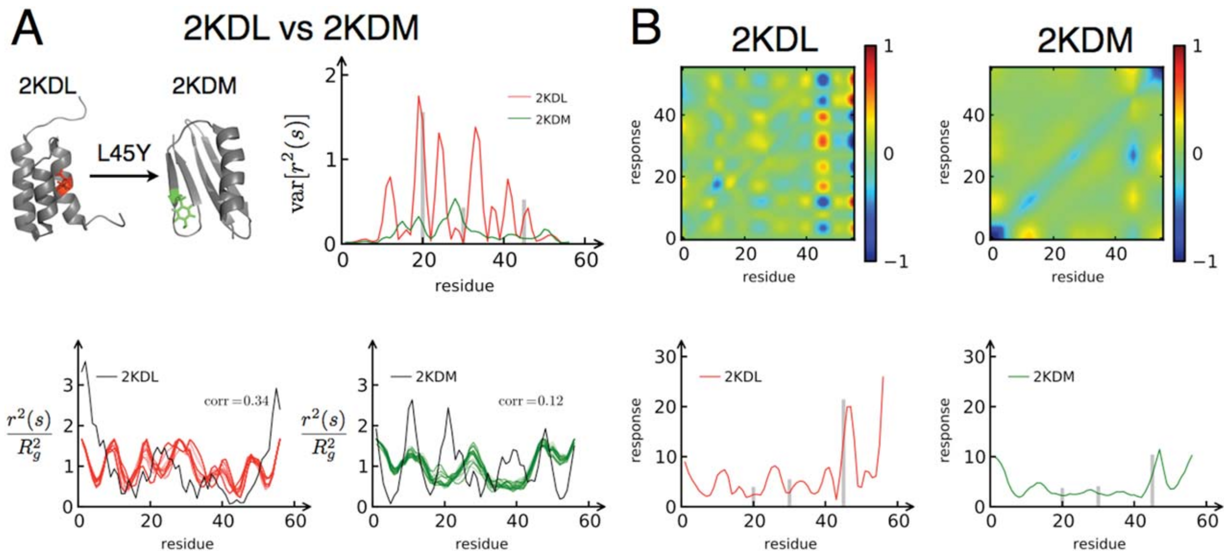

Figure 5. Conformational change triggered by mutation. (A) 3D structures of the 2KDM and 2KDL proteins show that mutation L45Y leads to the transformation of a 3-α fold into a $4\beta+\alpha$ fold. Structural variability, plotted in red and green, was computed from the burial traces of the low-energy excited states ($\Delta E=4\ k_BT$). The positions of other mutations are shown as the gray bars on the plot; and (B) Response matrices $\delta r^2(s)/\delta\varphi(s')$ of the 2KDM and 2KDL proteins. The plots on the bottom were obtained by taking the sum of the absolute values along the rows of the response matrices. In both proteins, the residues near the termini and residues 43–47 are the most sensitive to changes in amino acid hydrophobicity.

interactions among all atoms inside a protein. In this study, we set out to characterize a model of protein folding which sacrifices atomic details and which considers only backbone stretching, steric repulsion, and the hydrophobic effect to explain conformational preference in proteins. The advantages of this approach to studying the sequence-structure relationship are its high speed and the simplicity of interpreting results. However, a stumbling block preventing us from using the model to study large collections of proteins was a lack of clear understanding of the model's limitations.

The parameter space of the burial mode model is defined by the hydrophobicity scale by which the amino acid sequence is mapped into a quantified string of relative burial tendencies. Thus, to improve the predictive power of the model, we searched for a better hydrophobicity scale. Having not found another standard hydrophobicity scale that works significantly better than KD scale, we did a brute force search for a new hydrophobicity scale with a reduced amino acid alphabet. Because this approach was not more effective than using KD scale, we devised a method to infer relative hydrophobicities of amino acid residues from analysis of known protein structures. This method is based on the idea that two amino acid residues that are the nearest neighbors on the chain are essentially in the same environment, and their tendency to burial is determined only by their relative hydrophobicity. It should be noted that using statistics of amino acid contacts and distances to infer amino acid interactions has been widely used before.[25,26] However, our method is fundamentally different from Miyazawa, Jernigan, and Sippl's statistical potentials as it considers only local interactions affecting relative burial of adjacent residues along the chain and focuses on the relative positions of amino acid residues with respect to the center of mass of the protein rather than pairwise distances.

Strikingly, the hydrophobicity scale computed with our method was in good agreement with the experimentally measured scales. This fact supports the idea that a large collection of proteins can be treated as a statistical ensemble of sequences, and that our model of folding is based on sound physical assumptions about the forces driving native structure. Testing the model with the new scale, we found that performance on a large set of proteins was not improved; apparently, the model has limitations which may come from neglecting other intrachain and/or interdomain interactions that may be important to protein structure in any given case. Indeed, it is not surprising that the hydrophobic effect is not sufficient to explain the tertiary structures of globular proteins in all cases. Long-range hydrogen bonding interactions (such as in beta sheets), disulfide linkages, salt bridges, and dihedral angle constraints all are forces not included in the burial mode model that might play a definitive role in selecting a particular native structure in the case of a given protein. In this light, it is easy to understand why the alpha-rich globins proved such a fertile testing ground for the model.

Nonetheless, it should also be noted that the matrix of relative positions M_{ij} that we used to compute our new hydrophobicity scale contains more information about amino acid residues than a simple hydrophobicity scale, because it treats each pair of letters as having a unique local interaction. Thus, there are 190 parameters in this matrix that correspond to relative burial tendencies of different pairs, and an exciting future avenue of research will be to develop a model similar to the burial trace model that exploits all of the information in this statistical potential to predict the conformational physics of proteins. For example, it may eventually be possible using this information to develop better criteria for distinguishing between sequence trends that promote burial in the globular core and sequence trends that facilitate surface interaction with a hydrophobic ligand or protein–protein interface. While both such trends might correspond to elevated hydrophobicity on the KD scale, one type of sequence composition could well be distinguishable from the other with a more detailed model of the nontransitive relative burial tendency in each amino acid pair.

Having found that the burial mode model could not be substantially improved simply through parametric optimization, we set out to explore the origins of the model's limitations. In particular, we looked at the family of globins, where the model performs exceptionally well with myoglobin and does not succeed with hemoglobin. From the comparison of these two proteins, we learned that the propensity of amino acid residues to burial might depend not only on their hydrophobicity but also on the interactions with molecules external to the monomeric protein chain, which are not included in the model. This realization gave us the idea to study conformational fluctuations in order to identify potential sites of interactions. For various proteins with good burial trace agreement (myoglobin, H-Ras protein, and chymotrypsinogen) we demonstrated that ligand-binding and catalytic sites are located in the regions of high structural variability.

This finding is consistent with the "conformational selection" paradigm that has been suggested previously in the study of binding events[27]—regions of proteins that have to accommodate ligands, whether small molecules or other proteins, benefit from being structurally variable because the free energy of interaction is improved when the protein can optimize its shape to accommodate the moieties of the ligand. This process is accompanied by large structural rearrangements if there is an energy exchange between protein regions with "discrete breathers" (localized excitations).[28–31] The conformational selection paradigm implies that "discrete breathers" should be located close to ligand-binding sites. Although at first sight, the conformational selection paradigm and the approach that we used in this study look different, the similarity between them becomes clear if we make an analogy between "discrete breathers" and the eigenmodes of the burial mode model energy function[12]—in both descriptions, ligand-binding suppresses one mode and stimulates another, coupling large scale motions to the transduction of small forces. Furthermore, it should be noted in passing that, unlike methods which use the normal mode analysis to compute structural variability and mechanical response,[32–35] burial mode analysis relies only on sequence information and is not limited to small perturbations about a local energy minimum in a particular conformational state. Thus, burial mode analysis may yet prove useful as a general tool for prediction of catalytic and ligand-binding sites from primary sequence information.

To conclude, we presented a simplified model of protein folding which allows one to compute information about protein structure directly from its sequence. In our attempt to optimize the input parameters, we discovered that the KD hydrophobicity scale provides nearly optimal performance and the limitations of the model come in part from the interactions with external molecules that are not considered in the model. To predict potential sites of ligand interaction, we exploited the idea of conformational selection and demonstrated that the burial mode model captures functionally relevant conformational changes in several cases of good burial trace agreement. Finally, we showed that sometimes the requirement for good burial trace agreement can be relaxed and the model can also be used to predict regions most sensitive to mutations. This information can potentially be used in drug design to identify target sites and in SNP genotyping to distinguish neutral and disease-causing mutations. The model can also provide auxilliary information for MD simulations that use burial traces to generate initial protein configurations.[36] In addition, because of the high speed, the model can be used as a tool to study large collections of homologous sequences, which became available with high-throughput genomic sequencing and to access structural information about different mutants that are not yet crystallized.

Materials and Methods

Calculation of hydrophobicity scale from the matrix of relative positions

To calculate hydrophobicity scale of n-letter amino acid alphabet from the matrix of relative positions, we first constructed two matrices $A_{n(n-1)/2\times n}$ and $B_{n(n-1)/2\times 1}$ elements of which were computed as follows:

$$A_{mi}=1,\ A_{mj}=-1, \tag{3}$$

$$A_{mk}=0,\ \text{for}\ k\neq i,j, \tag{4}$$

$$B_{m1} = -\ln M_{ij}/M_{ji} = \Delta\varphi_{ij}, \tag{5}$$

where $i=[1, n-1]$, $j=[i+1, n]$, and $m=(i-1)\times(2n-i)/2+j-i$. Then, we used the method of least squares to find approximate solution for overdetermined system of linear equations $A\cdot\varphi=B$, where φ is n-letter hydrophobicity scale.

Generation of the burial traces of near-native states

The burial traces in the model can be written in terms of the eigenmodes $\psi_k(s)$ of energy function (1) and coefficients c_k: $r^2(s)=\sum_k c_k\psi_k^2(s)$. Thus, to compute the burial trace of the lowest energy state, one should minimize

$$E=\sum_k c_k\varepsilon_k, \tag{6}$$

where ε_k are the eigenvalues of the model energy function (1), subject to the steric constraints:

$$\sum_k c_k = \alpha N R^2, \tag{7a}$$

$$0 \le \sum_k c_k\psi_k^2(s) \le R^2, \text{ for } s \in [1, N], \tag{7b}$$

$$c_k \ge 0, \text{ for all } k. \tag{7c}$$

These equations set an exactly solvable linear programming problem with variables c_k, objective function (6), and linear constraints (7). The solution of this problem provides the energy of the lowest energy state $E_{\min}$ and optimal coefficients c_k^{opt}. To find the burial traces of excited states with energy $E_{\min}+\Delta E$, we generated a set of coefficients c_k which are the solution of another linear programming problem with constraints (7) and $\sum_k c_k\varepsilon_k = E_{\min}+\Delta E$, and objective function $\sum_k c_k r_k$, where r_k are random numbers.

To compute the structural variability $\operatorname{var}[r^2(s)]$, we first computed $n=100$ burial traces of near-native states $r_i^2(s)$ $(i=1, 2, ..., n)$, and then for every position s we calculated the variance of $r^2(s)$:

$$\operatorname{var}[r^2(s)] = \frac{1}{n}\sum_i (r_i^2(s) - \operatorname{mean}[r^2(s)])^2, \quad \text{where} \tag{8}$$

$$\operatorname{mean}[r^2(s)] = \frac{1}{n}\sum_i r_i^2(s). \tag{9}$$

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230.
2. Jayachandran G, Vishal V, Pande VS (2006) Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. J Chem Phys 124:164902.
3. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330:341–346.
4. Das R, Baker D (2008) Macromolecular modeling with Rosetta. Annu Rev Biochem 77:363–382.
5. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino-acid residues in globular-proteins. Science 229:834–838.
6. Baldwin RL (2007) Energetics of protein folding. J Mol Biol 371:283–301.
7. Chandler D (2005) Interfaces and the driving force of hydrophobic assembly. Nature 437:640–647.
8. Lau KF, Dill KA (1989) A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. Macromolecules 22:3986–3997.
9. Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA (1995) A test of lattice protein-folding algorithms. Proc Natl Acad Sci USA 92:325–329.
10. Chothia C (1976) Nature of accesible and buried surfaces in proteins. J Mol Biol 105:1–14.
11. Rose GD, Roy S (1980) Hydrophobic basis of packing in globular-proteins. Proc Natl Acad Sci USA 77:4643–4647.
12. England JL (2011) Allostery in protein domains reflects a balance of steric and hydrophobic effects. Structure 19:967–975.
13. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132.
14. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino-acid-sequences. Proc Natl Acad Sci USA 78:3824–3828.
15. Wimley WC, Creamer TP, White SH (1996) Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. Biochemistry 35: 5109–5124.
16. Nozaki Y, Tanford C (1971) Solubility of amino acids and 2 glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. J Biol Chem 246:2211–2217.
17. Janin J (1979) Surface and inside volumes in globular proteins. Nature 277:491–492.
18. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540.
19. Takano T (1977) Structure of myoglobin refined at 2–0 Å resolution: I. crystallographic refinement of metmyoglobin from sperm whale. J Mol Biol 110:537–568.
20. Benesch R, Benesch RE (1967) The effect of organic phosphates from the human erythrocyte on the allosteric properties of hemoglobin. Biochem Biophys Res Commun 26:162–167.
21. Arnone A (1972) X-Ray-diffraction study of binding of 2,3-diphosphoglyrecate to human deoxyhemoglobin. Nature 237:146–149.
22. McCormick F, Clark BF, la Cour TF, Kjeldgaard M, Norskov-Lauritsen L, Nyborg J (1985) A model for the tertiary structure of p21, the product of the ras oncogene. Science 230:78–82.
23. Branden C, Tooze J (1991) Introduction to protein structure, Vol. 2. New York: Garland Science.
24. Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci USA 106: 21149–21154.

25. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal-structures-quasi-chemical approximation. Macromolecules 18:534–552.
26. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213:859–883.
27. Csermely P, Palotai R, Nussinov R (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci 35:539–546.
28. Piazza F, Sanejouand YH (2008) Discrete breathers in protein structures. Phys Biol 5:026001.
29. Piazza F, Sanejouand YH (2009) Long-range energy transfer in proteins. Phys Biol 6:046014.
30. Kopidakis G, Aubry S (1999) Intraband discrete breathers in disordered nonlinear systems. I. Delocalization. Phys D 130:155–186.
31. Kopidakis G, Aubry S, Tsironis GP (2001) Targeted energy transfer through discrete breathers in nonlinear systems. Phys Rev Lett 87:165501.
32. Bathe M (2008) A finite element framework for computation of protein normal modes and mechanical response. Proteins 70:1595–1609.
33. Kim D-N, Sedeh RS, Nguyen CT, Bathe M Finite element framework for mechanics and dynamics of supramolecular protein assemblies. In: Proceedings of the ASME First Global Congress on Nanoengineering for Medicine and Biology (NEMB2010). (2010), New York: American Society of Mechanical Engineers, pp 315–316.
34. Hawkins RJ, McLeish TC (2004) Coarse-grained model of entropic allostery. Phys Rev Lett 93:098104.
35. Levitt M, Sander C, Stern PS (1985) Protein normal-mode dynamics: trypsin-inhibitor, crambin, ribonuclease and lysozyme. J Mol Biol 181:423–447.
36. Rocha JR, van der Linden MG, Ferreira DC, Azevedo PH, Pereira de Araujo AF (2012) Information-theoretic analysis and prediction of protein atomic burials: on the search for an informational intermediate between sequence and structure. Bioinformatics, 28: 2755–2762.