Structure
Article

Allostery in Protein Domains Reflects a Balance of Steric and Hydrophobic Effects

Jeremy L. England[1,*]
[1]263 Icahn Laboratory, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
*Correspondence: jengland@princeton.edu
DOI 10.1016/j.str.2011.04.009

SUMMARY

Allosteric conformational change underlies biological function in many proteins. Allostery refers to a conformational event in which one region of a protein undergoes structural rearrangement in response to a stimulus applied to a different region of the same protein. Here, I show for a variety of proteins that a simple, phenomenological model of the dependence of protein conformation on hydrophobic burial energy allows one to compute low-energy conformational fluctuations for a given sequence by using linear programming to find optimized combinations of sequence-specific hydrophobic burial modes that satisfy steric constraints. From these fluctuations one may calculate allosteric couplings between different sites in a protein domain. Although the physical basis of protein structure is complex and multifactorial, a simplified description of conformational energy in terms of the hydrophobic effect alone is sufficient to give a mechanistic explanation for many biologically important allosteric events.

INTRODUCTION

Structural biology rests on the principle that each macromolecule reliably adopts a well-defined shape and that it is this shape that provides the basis for its function. What complicates this basic picture is that it is often the capacity to undergo conformational *change* in reaction to targeted stimuli that enables a given protein to fulfill its role in the biological context. Allosteric motion—i.e., the structural rearrangement of one part of a protein in response to a stimulus applied at some remote site on the same protein—plays a crucial role in many biochemical pathways, particularly those involved in regulation and signaling (Branden and Tooze, 1999). Proteins may redistribute themselves from one part of conformation space to another in ways that affect their functional interactions with other biomolecules, whether through a ligand-binding event, the hydrolysis of a substrate, or some covalent modification such as phosphorylation (Swain and Gierasch, 2006; Volkman et al., 2001).

In broad terms, the physical basis of allostery is clear: if a relatively small perturbation can bring about a large-scale conformational shift in a protein, it follows that there are at least two, structurally distinct ensembles of conformations with nearly the same free energy, such that a small amount of additional energy supplied by the right stimulus can shift the equilibrium from one basin to the other (Formaneck et al., 2006; Gunasekaran et al., 2004; Kern and Zuiderweg, 2003; Kumar et al., 2000; Swain and Gierasch, 2006). Making more precise claims about why a particular protein should exhibit the particular conformational multistability that it does proves to be much more difficult. Although normal mode analysis has helped elucidate the origins of certain functionally important coordinated motions in macromolecules (Levitt et al., 1985), such an approach is by definition limited to the domain of small deviations from a single local-energy minimum. Heroic efforts to extend normal modes beyond the linear regime have made it possible to describe the dynamics of barrier-crossing events that underlie some allosteric events (Miyashita et al., 2003), but not without relying on foreknowledge of initial and final states for the conformational transition in question (Daily and Gray, 2009; Hawkins and McLeish, 2004). More recently, some researchers have begun to circumvent this obstacle by using detailed, high-resolution, full-atom structure prediction or molecular dynamics-simulation algorithms to generate accurate predictions of alternative conformations for allosteric systems (Kidd et al., 2009), as well as physical estimates of correlations in motions of different parts of a fluctuating protein (Liu and Nussinov, 2008). Others, meanwhile, have had great success in approaching allostery from an evolutionary standpoint, uncovering potentially important groups of interacting residues by identifying rare sequence covariations in families of related proteins (Süel et al., 2003). Nevertheless, there remains a need for an analytically solvable, physical theory of allosteric motion that provides a general framework for explaining allostery mechanistically in terms of detailed features of protein sequence.

In this work such a model for globular protein domains is proposed, solved, and applied. Focusing on large-scale backbone arrangements at the expense of finer, angstrom-level details, I construct a phenomenological expression for hydrophobic burial energy whose global minimum may be computed exactly on constraints that account for the impact of intrachain steric repulsion. This approach enables rapid calculation of the energetically minimal backbone burial trace for many globular domains from genetic information alone. More significantly, it paves the way for a new understanding of allosteric motion as the outcome of a sterically constrained competition among different, sequence-specific collective modes of hydrophobic burial.

RESULTS

Model

The aim of this section is to develop a solvable physical model of how conformational fluctuations in the near-native ensemble of a globular protein depend on sequence, in order to provide a mechanistic explanation for the structural rearrangements that take place during allosteric motion. The guiding principle for the approach taken here is that large-scale motions in a protein chain are rough features of tertiary structure as a whole and, therefore, need not necessarily be described in terms of an atomistically exact account of conformational energetics. Rather, we may construct our phenomenological model so as to keep it simple and analytically tractable, while still capturing the essence of the physical effects at play in the phenomenon of interest.

The first step in determining which conformation or conformations a protein will prefer as a result of its sequence is to make some assumption about how a protein's energy varies with its shape. A panoply of forces (e.g., backbone-to-backbone hydrogen-bonding interactions, electrostatic attraction and repulsion between charged side chains, or sequence-specific propensities for particular backbone dihedral angles) do, in fact, affect the energy of a given conformation. However, in the interest of simplicity, it is worth noting that burial of hydrophobic amino acid side chains in a solvent-occluded core is a feature of tertiary structure common to nearly all globular proteins (Branden and Tooze, 1999; Camacho and Thirumalai, 1993). Indeed, various studies suggest that the hydrophobic effect (Chandler, 2005; Rose et al., 1985), and the drive it produces in a protein to bury hydrophobic amino acid side chains, may be the fundamental force that determines the native structure and stability of many polypeptides (Ghosh and Dill, 2009; Silverman, 2005). The most basic question to ask, then, is (Pereira De Araújo, 1999): Given a polypeptide chain of amino acids whose sequence gives rise to a certain pattern of hydrophobicity along its length, what is the energetically optimal way of burying the hydrophobic parts of the chain in a collapsed globule while obeying the constraints of polymeric bonds and steric repulsion?

The most fundamental effect of a polymeric bond is to produce correlations in the spatial locations of pairs of monomers that are separated on the chain by relatively few bonds. The simplest and most mathematically tractable way of introducing these correlations into a model of a polymer with monomers indexed by s whose conformation is specified by the trajectory $\mathbf{r}(s) = [x(s), y(s), z(s)]$ is to have a term in the Hamiltonian that connects one monomer to the other with harmonic springs of stiffness κ. The partition function for this effective Hamiltonian by itself is simply the propagator for an unbiased random walk through space (Shakhnovich and Gutin, 1989). The parameter $\kappa(T)$ is a function of temperature and specifies a length scale for the typical separation in space between two adjacent monomers along the chain. For all calculations performed in this work, it is assumed that $\kappa = 3k_BT/2$, which corresponds to a random walk for which the mean-square distance between two adjacent monomers $\langle|\mathbf{r}(s+1) - \mathbf{r}(s)|^2\rangle$ is equal to unity. This choice effectively sets the units of length in the theory to be the typical distance between α carbons on a polypeptide chain.

To incorporate the hydrophobic effect into the model, it is necessary to make some choice about how the forces acting on the protein arise from its amino acid sequence. Positing a quadratic form as a rough approximation to the behavior of the hydrophobic force has the double appeal of its analytical tractability and its consistency with the physical intuition that the force on any given residue should be stronger in magnitude near the surface of the globule (where solvent is present) than it is near the core of the globule (where solvent is absent). In this case, one writes the full Hamiltonian as:

$$H = \int ds \left[\kappa \left| \frac{d\mathbf{r}(s)}{ds} \right|^2 + \varphi(s) |\mathbf{r}(s)|^2 \right].$$

Here, the scale of the relative hydropathy $\varphi(s)$ is fixed in terms of κ in units of k_BT by the expected free-energy change associated with moving a single amino acid from the hydrophobic core of the protein (often likened to an ethanol or octanol solution; Kyte and Doolittle [1982]) to the aqueous environment of the globule surface. The sequence-dependent Hamiltonian term above resembles that of a polymer in an external field (Grosberg, 1984), insofar as each amino acid is independently attracted toward or repelled from the center of the globule depending on whether it is hydrophobic or hydrophilic. However, in order for the Hamiltonian in the present discussion to make physical sense, the radial-squared distance $|\mathbf{r}(s)|^2$ must be taken from the globule's center of mass. It is distance from the center of the polymer, wherever it may be, and not distance from an arbitrary origin, that affects solvent-exposed surface area.

An exact treatment of steric repulsion is challenging, so much so that one is forced to resort to approximate methods even in the study of "simple" systems such as a fluid of hard spheres. This difficulty can be traced to the pairwise nature of the steric interaction: any atom in a protein chain should be able to occupy any location in space, in principle, unless that location is already occupied by another atom. It is arguable, though, that not all self-clashed conformations of a polymer that are disallowed by steric repulsion are equally forbidden. If a given conformation only is forbidden because a single pair of atoms overlap in space, then that conformation bears a great deal of structural similarity to a conformation that *is* permitted in the presence of steric repulsion. In contrast, conformations that pack hundreds of residues into a volume normally occupied by a single atom presumably must undergo dramatic structural rearrangement in order to come into line with steric constraints. This observation motivates the argument that the most essential structural constraint on a protein's conformation imposed by steric repulsion is to spread the polymer out over space enough that the latter category of "pathologically clashed" conformations that could never even resemble a protein are forbidden.

The simplest way to forbid pathological clashing is to analogize the polymeric globule to a sphere of maximum radius R. If the globule is assumed to have uniform mass density, so that the number of residues within any small subvolume of fixed size in the sphere is roughly the same, then it must be the case that $|\mathbf{r}|^2 = 3R^2/5$.

Thus, pathological clashing can be prevented by constraining the mean-square radius of the polymer's conformation to have a fixed ratio to its maximum squared radius of 3/5.

To proceed toward a solution to the problem as posed above, it is necessary to diagonalize the Hamiltonian. An orthonormal set of eigenfunctions ("burial modes") $\psi_k(s)$ may be defined for free-end boundary conditions and the center of mass constraint such that

$$-\kappa\frac{d^2\psi_k(s)}{ds^2}+\varphi(s)\psi_k(s)=\varepsilon_k\psi_k(s).$$

Any conformation of the polymer may be expressed in this basis as

$$\mathbf{r}(s)=[x(s),\ y(s),\ z(s)]=\left[\sum_k X_k\psi_k(s),\ \sum_k Y_k\psi_k(s),\ \sum_k Z_k\psi_k(s)\right],$$

in which case, defining $c_k=X_k^2+Y_k^2+Z_k^2$, one may write the Hamiltonian as

$$H=\sum_k c_k\varepsilon_k$$

and the steric constraint as

$$\sum_k c_k=\frac{3NR^2}{5}$$

with

$$R^2\geq|\mathbf{r}(s)|^2\approx\sum_{k=1}^{N}c_k\psi_k(s)^2,$$

which completes the picture (see Supplemental Experimental Procedures for derivation). Each "conformation" corresponds to a choice of the constants c_k, which allow one to compute a representative backbone trace $|\mathbf{r}(s)|^2$ that measures the relative burial of each part of the polymer with respect to the core of the globule. The energy, which is a linear function of the constants c_k, may be optimized on the steric constraints, which are simply linear inequalities. As a result, the search for a lowest-energy conformation reduces to an exactly solvable linear programming problem.

Data Analysis

The model considered here asserts that to each amino acid sequence, there corresponds a series of independent modes of hydrophobic burial that define a hierarchy of energetic favorability for global contortions of the protein. In the absence of steric repulsion, the optimal conformation would simply be the burial mode of lowest energy, but this would require that most of the polymer be crammed into a small subvolume of the globule at nonphysically high density. Introduction of the steric constraint forces the polymer to find an energetically optimal *combination* of the low-energy burial modes: one that unpacks the core of the globule most efficiently and thereby achieves a physically reasonable density without exposing too many hydrophobic residues to the surrounding solvent.

Figure 1 demonstrates the application of this procedure to the sequence of sperm whale myoglobin. Myoglobin is an attractive test case for the model because it is a single, α-globular domain that does not reside in membrane (transmembrane proteins, which experience a nonuniform solvent environment, sit at the other end of the spectrum [Branden and Tooze, 1999]: their tertiary structure clearly will be dominated by effects that the model ignores). The standard Kyte-Doolittle hydropathy scale (Kyte and Doolittle, 1982) provides a means to convert the amino acid sequence of the protein into a string of numbers from which the independent hydrophobic burial modes may be computed (Figure 1B). Linear programming allows energetically optimal construction of the protein's backbone burial trace (Pereira de Araújo et al., 2008) $|\mathbf{r}(s)|^2$ from the sequence's low-energy modes, and the result is compared in Figure 1C to the same burial trace calculated from the positions of atoms in the crystal structure of myoglobin. The resemblance between the two traces is unmistakable (correlation = 0.58), and substantially more pronounced than the equivalent result obtained through a simple local averaging of sequence hydrophobicity within windows along the chain, where window width is selected so as to best fit the crystal structure (correlation = −0.41, width = 6 residues) (see Figure S1 available online). The same burial mode analysis carried out for a variety of other proteins yields comparable agreement, with the locations (as opposed to the exact heights) of peaks and troughs in the burial trace tending to match best with the crystal structure (see Figures S1B and S1C).

A large-scale study of protein structure space provides further evidence that the model-predicted energetically optimal trace computed from sequence alone does a good job of roughly capturing the burial patterns of many proteins. As a basis for comparison, a control method of burial calculation was also implemented, in which each sequence hydrophobicity pattern was averaged within windows of fixed size, and where window size was chosen to optimize the correlation between the window-averaged hydrophobicity trace and the burial trace computed from the crystal structure. Thus, the model-predicted burial traces, which are derived from sequence alone without any fitting parameters, were compared to a fitted-window control method in which information about the crystal structure was used to produce the best correlation possible.

In Figures 2A–2D, the distributions of Pearson correlations between burial traces from sequence and corresponding crystal structure are plotted for the four relevant classes of single protein domains from the Structural Classification of Proteins (SCOP) (Murzin et al., 1995). In all classes the distribution of traces derived from burial modes (blue solid curves) had positive mean (ranging from 0.17 to 0.25, Figure 2E), and more significantly, roughly 20% of all sequences produced a correlation with the crystal structure of 0.4 or better (Figure 2F), indicating that the model can successfully predict the qualitative shape of a protein's burial trace from sequence alone in thousands of cases. The success of this performance is underlined by a comparison with the control distributions computed for random permutations of each sequence: whereas the fitted-window averaging method, which makes use of information from the crystal structure, produces a positive correlation on average even for random control sequences (dashed orange curves), the use of burial modes involves no fitting to the structure and, therefore, exhibits far less correlation for randomly permuted sequences (dashed cyan curves). Thus, the computation of energetically optimal burial traces from the burial modes of primary sequences successfully extracts accurate tertiary

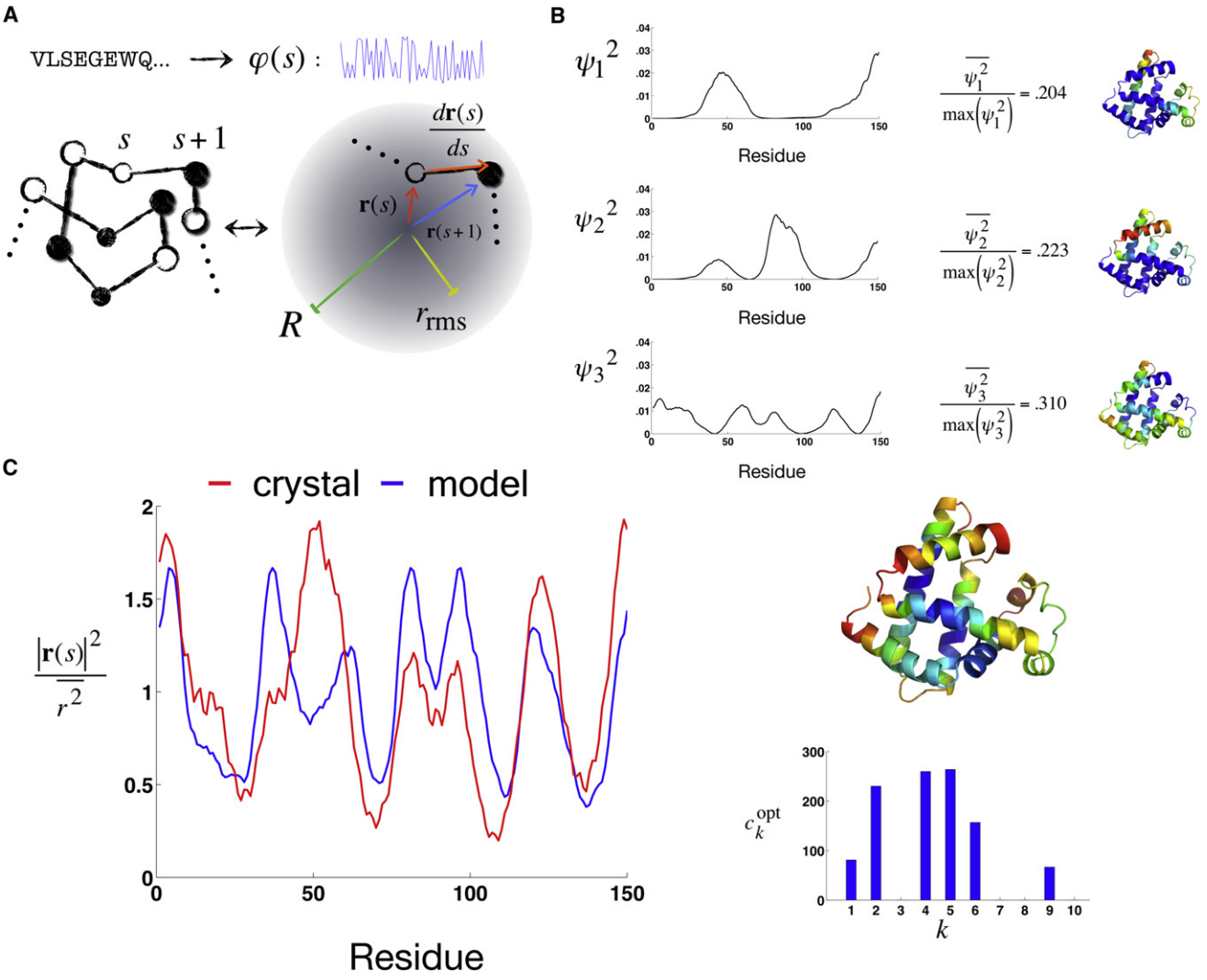

Figure 1. Computation of Burial Traces through Linear Optimization of Burial Modes

(A) A protein in a collapsed globular conformation may be represented as a chain with residues indexed by the number *s* that have position **r**(s) (red and blue vectors) relative to the center of mass of the globule. The root-mean-square distance r_{rms} (yellow vector) from the globule center averaged over the whole polymer is necessarily less than the maximum radius *R* (green vector). Each position *s* has an associated hydropathy $\varphi(s)$ determined by the type of amino acid at that point along the chain.

(B) The three lowest-energy burial modes for the sequence of sperm whale myoglobin are plotted and colored on the myoglobin crystal structure (PDB ID 1BZP), with blue corresponding to most buried and red to least buried. Each individual mode has a ratio of mean-square to max-square radius far below the value of 0.6 for a sphere of uniform mass density and, therefore, fails to satisfy the steric constraint.

Thus, in (C) the optimal solution (which is both colored on the crystal structure on the right-hand side as in B, and also plotted on the left-hand side in blue against the same trace computed from the crystal structure in red) must be constructed from multiple burial modes, with the heaviest weights c_k not corresponding to the modes of lowest energy. The Pearson correlation between model and crystal structure is 0.58.

Other representative burial traces for various sequences can be found in Figure S1.

structural information for thousands of proteins in the space of all observed folds without the use of any fitted parameters. It is moreover quite encouraging that the burial mode approach is noticeably more accurate in extracting this information from α-helical proteins than from ones dominated by β structure (Figure 2F), as this result is consistent with the underlying assumptions of the model: the long-range intrachain contacts necessary for the formation of β sheets are not represented in the model, and therefore, one would expect β-rich domains to be more challenging for the model to describe.

The data in Figure 2 demonstrate that the burial mode approach succeeds in roughly predicting the pattern of burial for some, but not all, protein domains. Although a high correlation between the optimal burial pattern predicted from sequence and that observed in the crystal structure does not establish with certainty that the burial mode picture is able to describe the conformational energetics of a given protein, a low correlation is a good indication that the approach has failed. Thus, in order to use burial modes to study the conformational fluctuations that underlie allosteric motion in a given protein domain, it is clearly

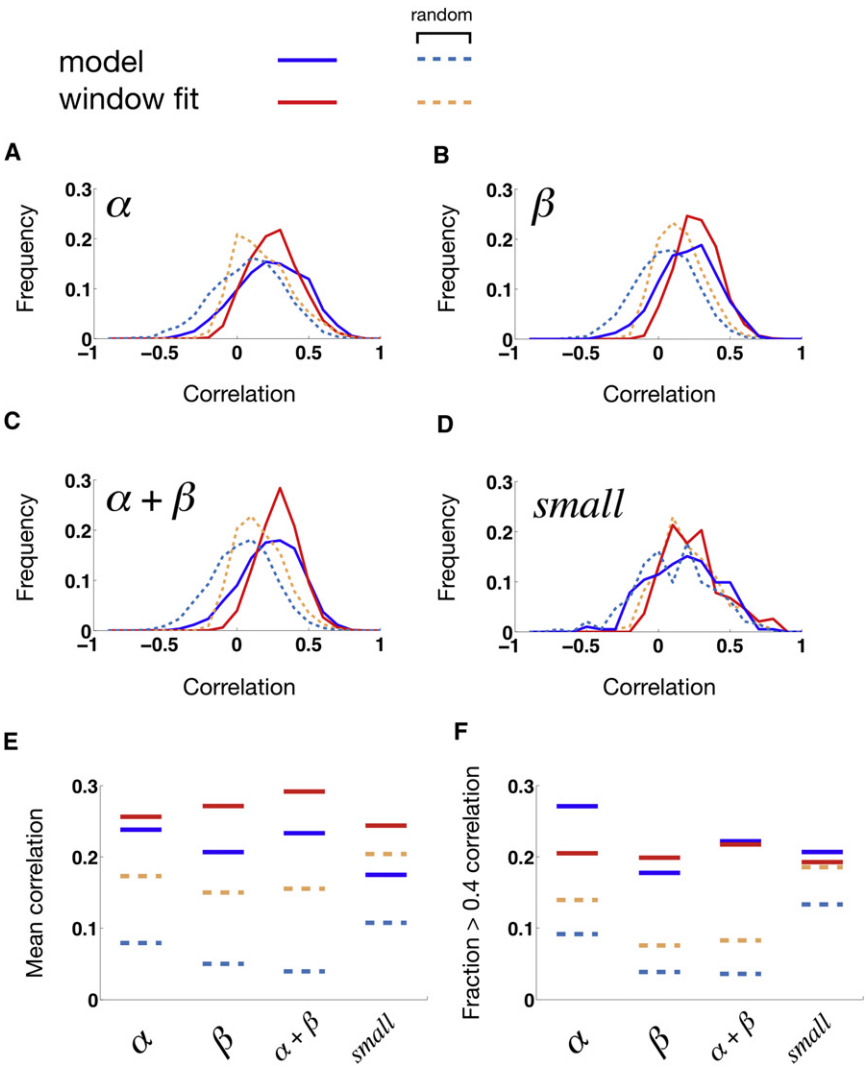

Figure 2. Application of Burial Mode Analysis to the Space of Protein Folds

Histograms of Pearson correlation between burial traces predicted from sequence and extracted from structure are plotted for (A) α-helical (1985 sequences), (B) β stranded (2197 sequences), (C) mixed α-β (5318 sequences), and (D) small non-α-β domains (619 sequences) in the SCOP space. (E) and (F) respectively report the mean correlation and fraction of domains above correlation 0.4 for each distribution. For each sequence structure pair, the burial trace was first computed from the crystal structure, and also predicted from the sequence. The distributions of correlations between these pairs of traces are drawn in each panel as the blue solid curve. The dashed cyan curves show the distributions of correlations for the same comparison between burial mode prediction and crystal structure, but for control sets of random permutations of the sequences. Because the solid blue distributions in all cases have a mean and mode substantially greater than zero, whereas the dashed cyan distributions do not, it is clear that the burial mode method is extracting accurate tertiary structural information from the real sequences, but not from the random control sequences. The solid red curves were generated by averaging sequence hydrophobicity within windows of fixed width along the chain, and finding for each sequence the width that optimized the correlation between the window-averaged hydrophobicity trace and the crystal structure burial trace. The dashed orange curves apply the same optimally fitted-window method to random control permutations of sequence for each structure. Because the window-fitting method uses information from the crystal structure, its distribution has a positive mean even for randomly permuted sequences (dashed orange curve), and yet applying the same method to real sequences (solid red curve) cannot outperform the distribution of burial mode-based predictions (solid blue curve), which are derived from sequence alone.

necessary, if not sufficient, that the predicted optimal burial trace correlate well with the known structure of the domain. Thus, burial trace correlation becomes a useful litmus test for selecting allosteric systems for study.

A previous study (Kidd et al., 2009) collected a set of eight experimentally characterized single-domain allosteric systems from the structural biology literature and analyzed them computationally. In the present work five out of eight of these systems demonstrated predicted-to-measured burial trace correlations of 0.4 or higher and were, therefore, selected for further analysis. The cutoff of 0.4 was chosen because it is approximately one standard deviation above the mean for the correlation distributions plotted in Figure 2, and because it is roughly the point at which the similarities between the predicted and measured burial traces start to be qualitative and obvious from visual inspection. It should be noted that approximately one in five sequences in the space of all SCOP domains would satisfy this correlation criterion (Figure 2F).

To assay whether analysis of burial modes aids the identification of allosteric couplings between sites in a polypeptide with a given amino acid sequence, one need only analyze the correlated motions in that polypeptide's ensemble of low-energy conformations (Gunasekaran et al., 2004; Kidd et al., 2009; Kumar et al., 2000; Levitt et al., 1985; Süel et al., 2003; Swain and Gierasch, 2006). Figure 3A shows a heat map of the pairwise covariances (Liu and Nussinov, 2008) in squared radial position between different sites along the length of the lymphocyte function-associated antigen-1 (LFA-1). LFA-1 binds to intracellular adhesion molecule (ICAM)-1, which is involved in activation of a downstream immune response (Last-Barney et al., 2001; Zhang et al., 2009). Allosteric inhibitors developed to block the LFA-1 interaction with ICAM have been found to bind LFA-1 at a site distant from points on the protein known to interface directly with ICAM. The absolute value of the sum of the columns of the burial covariance matrix that correspond to residues on the protein that contact the allosteric inhibitor (Zhang et al., 2009) estimates the magnitude of the conformational response at each point along the protein to the binding of the inhibitor. As one would expect, the binding site of the inhibitor is the region of the protein most strongly affected by the binding event

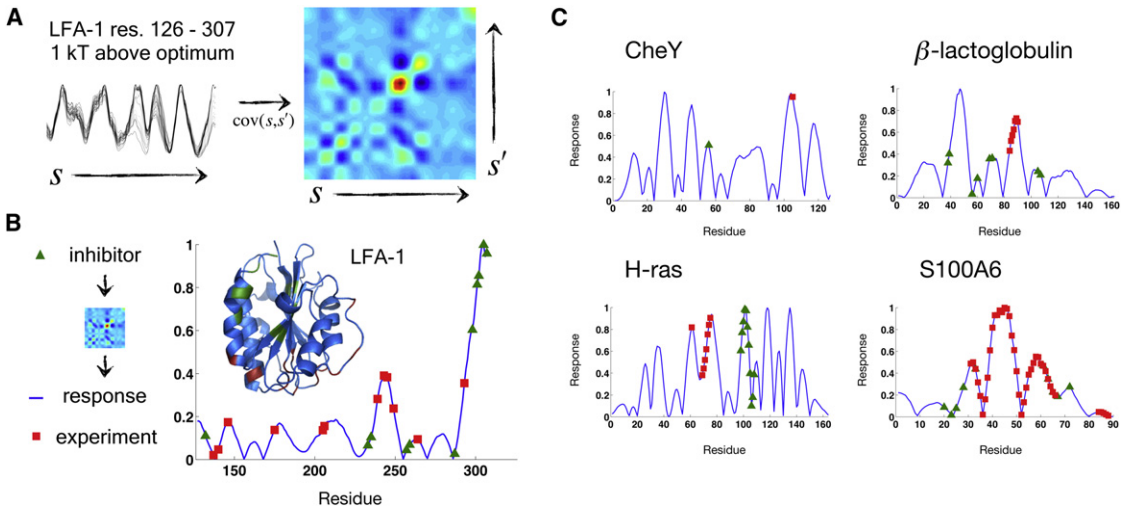

Figure 3. Allosteric Motion Predicted from Conformational Fluctuations

(A) Using the burial traces of LFA-1 conformations 1 k_BT above the energy minimum, a burial covariance matrix may be constructed between pairs of points s and s' along the chain, where $\mathrm{cov}(s,s') = \langle r^2(s)\ r^2(s')\rangle - \langle r^2(s)\rangle\langle r^2(s')\rangle$, and the brackets denote an average over all burial traces in the 1 k_BT ensemble. In the color map shown here, blue denotes negative covariance and red denotes positive.

(B) Summing together the covariance matrix columns corresponding to residues in LFA-1 that contact the allosteric isoflurane inhibitor (green triangles) and computing the absolute value of the result generates a measure of the amplitude of structural response to drug binding (blue line). The most strongly responding regions of LFA-1, aside from the site of drug binding itself, are those that are part of the ICAM-LFA-1 protein-protein interface (red squares).

(C) The same method is applied to analysis of allosteric motion in the proteins CheY (top left), β-lactoglobulin (top right), H-ras (bottom left), and S100A6 (bottom right).

Tests of statistical significance are reported in Figure S2.

(Figure 3B). However, it is quite notable that the residues that interact with ICAM nevertheless are clustered at other sites along the chain that undergo especially large structural rearrangements. Put another way, the summed covariance response curve correctly identifies the ICAM-LFA-1 protein-protein interface (Last-Barney et al., 2001) as a strong allosteric responder to inhibitor binding. An identical analysis performed for the proteins H-ras (Buhrman et al., 2010), β-lactoglobulin (Wu et al., 1999), S100A6 (Otterbein et al., 2002), and CheY (Formaneck et al., 2006) yields comparable results: in all cases the expected response to the stimulus localizes well with the region known from experiment to undergo an induced conformational change (Figure 3C).

To test the significance of this result, a metric for the overlap between the predicted and known allosteric response was generated for each protein, where the blue curves in Figure 3 were normalized to their maximum height outside the region of the stimulus (green triangles) and summed over the region of the response (red squares). This number was compared in each case to a control distribution generated from random sequence permutations whose predicted burial traces correlated with coefficient 0.4 or better with the known structure (Figure S2A). The p values generated from this procedure (LFA-1, 0.12; CheY, 0.02; β-lactoglobulin, 0.19; H-ras, 0.04; S100A6, 0.40) indicate overwhelming significance for the set as a whole, although the value for S100A6 on its own is marginal due to the large size of the allosterically responsive region. Similar results were obtained when the response per residue for the allosteric systems in Figure 3 was compared to the distribution of normalized pairwise residue-to-residue burial covariances for each wild-type sequence (LFA-1, 0.14; CheY, 0.03; β-lactoglobulin, 0.09; H-ras, 0.17; S100A6, 0.17) (Figure S2B). However, it should further be noted that the metric for significance employed here does not take into account other features of apparent agreement between the predicted and measured response, such as the clustering of the relevant residues near local maxima in the predicted response, and the absence of any predicted response peaks that are dramatically higher than the response expected in the experimentally predicted region. These additional features should further increase our confidence that the burial mode model employed here allows one to extract a significant amount of physical information about allostery from sequence alone.

Perhaps most striking of all, the basis for the allosteric motion in each case becomes clear upon examination of the specific burial modes that contribute to each native ensemble. For the proteins LFA-1 (top), H-ras (middle), and S100A6 (bottom), Figure 4 identifies specific pairs of low-energy modes whose competition within the native ensemble gives rise to allostery. Strikingly, for each protein there is one mode whose peaks line up well with the residues associated with a ligand-binding event, and another mode whose peaks line up well with the residues known to exhibit an induced conformational rearrangement as a result of ligand binding (see Figure S3 for tests of statistical significance). As the scatter plots on the right-hand side of Figure 4 show, the weights on each pair of modes display a significant nonzero correlation (LFA-1, −0.30 to −0.42; H-ras, −0.42 to −0.51; S100A6, −0.85 to −0.97) over a range of low-energy excitations above the ground state (1–5 k_BT). Thus, a low-energy structural rearrangement that changes the weight of one mode

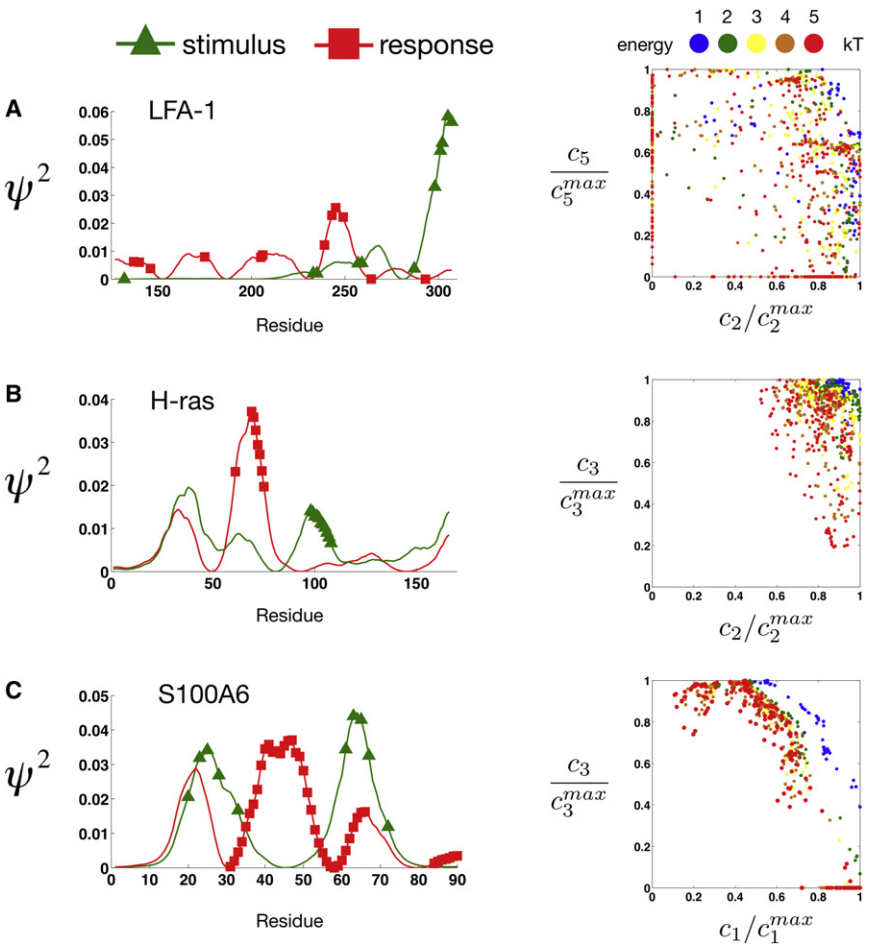

Figure 4. Allostery from Switching between Specific Pairs of Burial Modes

For the proteins LFA-1 (A), H-ras (B), and S100A6 (C), the expected allosteric motion is revealed to be the result of a trade-off between two different low-energy burial modes present in the low-energy, native ensemble. In each case the residues corresponding to the experimentally known stimulus to the protein line up well with the peaks of one mode (number 2 for LFA-1, number 2 for H-ras, and number 1 for S100A6), whereas the residues known from experiment to rearrange themselves in response to the stimulus localize well with the peaks of another mode (number 5 for LFA-1, number 3 for H-ras, and number 3 for S100A6). The significant nonzero correlation in the weights given to each pair of modes at low energy is preserved over a range of energies above the ground state in the native ensemble (right-side panels). Tests of statistical significance are reported in Figure S3.

through the binding of a ligand necessarily must produce a reaction along the other mode in the statistical ensemble of conformations because of the correlation between the weights for the modes at low energy. In retrospect, it should not be unexpected to observe this alignment of an allosterically coupled site with a single mode: the way to get a coherent, large-scale structural rearrangement out of a small structural perturbation is to concentrate that perturbation's impact along a single conformational degree of freedom whose associated energy scale for deviation from equilibrium is small. The simplicity of this explanation suggests that analysis of burial modes might be a quite generally applicable tool for characterization of allosteric motions in proteins, and may open the door to new strategies for selection of target sites for drug design.

DISCUSSION

Allostery is challenging to describe in analytical terms because it is, on the one hand, a collective phenomenon that arises from the convergence of many weak interactions among a large number of degrees of freedom, yet, on the other hand, it often can be triggered by a small perturbation that acts on only a few of those degrees of freedom. This inherent sensitivity rules out a straightforward linear response theory, and has spurred the innovation of various augmented elastic models (Daily and Gray, 2009; Hawkins and McLeish, 2004; Miyashita et al., 2003), which aim to capture dynamics by making use of predetermined information about the beginning and endpoint of the expected conformational change. The model presented here is, in its own way, elastic, insofar as it associates an energy scale with each independent mode out of which a given conformation is constructed. What distinguishes the burial mode approach is that it sacrifices atomistic detail in favor of a highly approximate steric constraint, and by doing so succeeds in introducing much-needed nonlinearity into the model without rendering things intractable. Thus, the description of allosteric motion is broken into two steps: first, one solves a linear problem to get the burial modes specific to a given sequence, and then the constrained competition among these modes in the presence of sterics can give rise to the multistability needed for allostery (Kumar et al., 2000). A particularly intriguing outcome from this line of inquiry is that sites involved in allostery in a protein tend to line up strongly along single burial modes of the sequence. In this light, burial modes may be seen as the conformational pressure points of sequence that have been selected by evolution.

It should be noted in closing that, although the method of burial mode analysis presented here was applied to the study allostery, it arguably has the potential to motivate other new lines of inquiry into how function emerges from primary sequence in proteins. Armed with a model of how small changes in a pattern of sequence hydrophobicity can give rise to global rearrangements in a polypeptide chain, researchers will have the opportunity to develop a fuller understanding of how various mutations lead to temperature sensitivity, structural instability (Liu and Nussinov, 2008), and aggregation. It will also be worthwhile to investigate whether the mapping of burial optimization to

a linear programming problem might be co-opted into a more sophisticated, full-atom structure prediction algorithm (Das and Baker, 2008). The burial mode approach also suggests a new lens through which to examine folding kinetics: low-energy collective modes of the protein chain may provide a natural coordinate system for charting folding pathways in terms of a small number of highly relevant degrees of freedom. Finally, the approximate, yet informative means for mapping sequence to structure described in this work has the distinct advantage of being extremely fast; the search for the global energy minimum of the myoglobin sequence takes less than 1 s on a 3.06 GHz Intel Core 2 Duo Processor. In this respect a qualitatively new kind of tool may now be available to drug designers, protein engineers, and evolutionary theorists alike in their efforts to decode principles of protein architecture from the wealth of genomic data produced by recent and future breakthroughs in sequencing technology.

EXPERIMENTAL PROCEDURES

Proteins

In all cases the amino acid sequences used were taken from the FASTA sequence of a structure in the Protein Data Bank (PDB) at http://www.rcsb.org. The structures and sequences used were: myoglobin/1BZP, CheY/1F4V, H-ras/3K8Y, LFA-1/3F74, S100A6/1K9K, and β-lactoglobulin/1BEB. Burial traces were generated from crystal structures by computing the center of mass of all polypeptide atoms in the PDB file and then measuring the distance of each α carbon from that center point. The resulting squared distance from the center was averaged within windows four residues in width all along the chain to remove high-frequency noise from local α-helical oscillations in position.

Optimization

The bond stiffness κ was chosen so that the corresponding free random walk would have a mean-square intermonomer distance of unity, fixing the units of length in the model to be the typical distance between α carbons in a protein chain. Density of monomers in a collapsed spherical globule was estimated from the TIM barrel structure (PDB ID 2VXN), idealized as a sphere of radius 4. This density was used to calculate the maximum radius for a globule of N residues from

$$R^2 = \left(\frac{3N}{4\pi\rho_0}\right)^{2/3}.$$

The hydropathies corresponding to each amino acid were taken from the standard Kyte-Doolittle scale but rescaled by a constant factor to ensure that, regardless of the number of residues in the chain, the energy change associated with motion from the surface to the center of the globule corresponded to 0.5 kT for glutamate. This fixed the energy scale of the hydrophobic effect at the correct order of magnitude for known transfer free energies of amino acids from water to solvents such as octanol or ethanol.

For a protein of N residues, an N by N energy matrix was constructed from the sequence and diagonalized, yielding a matrix of independent eigenmodes. The elements of this matrix were squared to yield a burial mode matrix. The MATLAB function linprog() was then used to find the lowest-energy combination of burial modes satisfying the linear constraints (see Supplemental Experimental Procedures for MATLAB code).

To compute the low-lying conformations close to the ground state in energy, an additional linear constraint was added to the linear programming procedure, fixing the energy to remain below the chosen ceiling. The objective function optimized was then the dot product of the vector of burial mode weights with a vector of N elements independently taken from a normal distribution using the MATLAB function randn(). For all proteins discussed in this work, 500 random samples were generated in this way for analysis.

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.str.2011.04.009.

ACKNOWLEDGMENTS

The author thanks E. Shakhnovich, V. Pande, E. Kussell, D. and M. Kaganovich, and M. Levitt for helpful discussions. The author also thanks the Lewis-Sigler Institute for Integrative Genomics for financial support. J.E. and Princeton University have submitted the provisional patent application (#61/323,175) as filed in the United States Patent and Trademark Office on the invention entitled "Hydrophobic Burial-based Method for Computation of Conformational Energy Spectrum of a Protein from its Amino Acid Sequence."

Received: January 3, 2011
Revised: April 5, 2011
Accepted: April 6, 2011
Published: July 12, 2011

REFERENCES

Branden, C., and Tooze, J. (1999). Introduction to Protein Structure (New York: Garland Publishing, Inc.).

Buhrman, G., Holzapfel, G., Fetics, S., and Mattos, C. (2010). Allosteric modulation of Ras positions Q61 for a direct role in catalysis. Proc. Natl. Acad. Sci. USA *107*, 4931–4936.

Camacho, C.J., and Thirumalai, D. (1993). Minimum energy compact structures of random sequences of heteropolymers. Phys. Rev. Lett. *71*, 2505–2508.

Chandler, D. (2005). Interfaces and the driving force of hydrophobic assembly. Nature *437*, 640–647.

Daily, M.D., and Gray, J.J. (2009). Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. PLoS Comput. Biol. *5*, e1000293.

Das, R., and Baker, D. (2008). Macromolecular modeling with Rosetta. Annu. Rev. Biochem. *77*, 363–382.

Formaneck, M.S., Ma, L., and Cui, Q. (2006). Reconciling the "old" and "new" views of protein allostery: a molecular simulation study of chemotaxis Y protein (CheY). Proteins *63*, 846–867.

Ghosh, K., and Dill, K.A. (2009). Computing protein stabilities from their chain lengths. Proc. Natl. Acad. Sci. USA *106*, 10649–10654.

Grosberg, A.Y. (1984). On the theory of the condensed states of heteropolymers. J. Stat. Phys. *38*, 149–160.

Gunasekaran, K., Ma, B., and Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? Proteins *57*, 433–443.

Hawkins, R.J., and McLeish, T.C. (2004). Coarse-grained model of entropic allostery. Phys. Rev. Lett. *93*, 098104.

Kern, D., and Zuiderweg, E.R. (2003). The role of dynamics in allosteric regulation. Curr. Opin. Struct. Biol. *13*, 748–757.

Kidd, B.A., Baker, D., and Thomas, W.E. (2009). Computation of conformational coupling in allosteric proteins. PLoS Comput. Biol. *5*, e1000484.

Kumar, S., Ma, B., Tsai, C.J., Sinha, N., and Nussinov, R. (2000). Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci. *9*, 10–19.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Last-Barney, K., Davidson, W., Cardozo, M., Frye, L.L., Grygon, C.A., Hopkins, J.L., Jeanfavre, D.D., Pav, S., Qian, C., Stevenson, J.M., et al. (2001). Binding site elucidation of hydantoin-based antagonists of LFA-1 using multidisciplinary technologies: evidence for the allosteric inhibition of a protein-protein interaction. J. Am. Chem. Soc. *123*, 5643–5650.

Levitt, M., Sander, C., and Stern, P. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. *181*, 423–447.

Liu, J., and Nussinov, R. (2008). Allosteric effects in the marginally stable von Hippel-Lindau tumor suppressor protein and allostery-based rescue mutant design. Proc. Natl. Acad. Sci. USA *105*, 901–906.

Miyashita, O., Onuchic, J.N., and Wolynes, P.G. (2003). Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. Proc. Natl. Acad. Sci. USA *100*, 12570–12575.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Otterbein, L.R., Kordowska, J., Witte-Hoffmann, C., Wang, C.L., and Dominguez, R. (2002). Crystal structures of S100A6 in the Ca(2+)-free and Ca(2+)-bound states: the calcium sensor mechanism of S100 proteins revealed at atomic resolution. Structure *10*, 557–567.

Pereira De Araújo, A.F. (1999). Folding protein models with a simple hydrophobic energy function: the fundamental importance of monomer inside/outside segregation. Proc. Natl. Acad. Sci. USA *96*, 12482–12487.

Pereira de Araújo, A.F., Gomes, A.L., Bursztyn, A.A., and Shakhnovich, E.I. (2008). Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. Proteins *70*, 971–983.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. (1985). Hydrophobicity of amino acid residues in globular proteins. Science *229*, 834–838.

Shakhnovich, E.I., and Gutin, A.M. (1989). Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. Biophys. Chem. *34*, 187–199.

Silverman, B.D. (2005). Underlying hydrophobic sequence periodicity of protein tertiary structure. J. Biolmol. Struct. Dyn. *22*, 411–423.

Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat. Struct. Biol. *10*, 59–69.

Swain, J.F., and Gierasch, L.M. (2006). The changing landscape of protein allostery. Curr. Opin. Struct. Biol. *16*, 102–108.

Volkman, B.F., Lipson, D., Wemmer, D.E., and Kern, D. (2001). Two-state allosteric behavior in a single-domain signaling protein. Science *291*, 2429–2433.

Wu, S.Y., Perez, M.D., Puyol, P., and Sawyer, L. (1999). beta-lactoglobulin binds palmitate within its central cavity. J. Biol. Chem. *274*, 170–174.

Zhang, H., Astrof, N.S., Liu, J.H., Wang, J.H., and Shimaoka, M. (2009). Crystal structure of isoflurane bound to integrin LFA-1 supports a unified mechanism of volatile anesthetic action in the immune and central nervous systems. FASEB J. *23*, 2735–2740.