Supporting Information

Weinkam et al. 10.1073/pnas.1116274109

SI Materials and Methods

Constructing the Allostery Model. For a given protein, the allostery model defines several effector bound and unbound landscapes that differ by the size of the allosteric site. The energy function that defines each landscape is a sum of nonbonded distance terms that control the attractive interactions between atoms and bonded terms that maintain proper stereochemistry. The nonbonded distance terms determine the efficient sampling of the allosteric transition and vary with the size of the allosteric site. Interactions involving atoms in the allosteric site, defined as residues within a radius of the effector ligand (r^{AS}), are given a single energy minima corresponding to distances in either the bound or unbound crystal structure (Fig. 2*A*). The remaining interactions between atoms have two energetically equivalent minima corresponding to distances from both crystal structures (Fig. S2). Changing r^{AS} modulates the strength of the allosteric signal. An order parameter for allostery is obtained by changing r^{AS} while restraining the allosteric site first to the unbound and then to the bound structure. In other words, changing r^{AS} allows interpolation between the effector bound and unbound landscapes (Fig. 1). The r^{AS} varies between 4 and 20 Å.

In our allostery model, a landscape is given by a potential energy function that is a sum of bonded and nonbonded terms implemented using MODELLER (1), following CHARMM (2): $\mathbf{E}_i = E_{\mathrm{bonded}} + E_{\mathrm{nonbonded}}$. Correct stereochemistry is achieved by the same terms MODELLER uses for standard comparative modeling: $E_{\mathrm{bonded}} = E_{\mathrm{bond}} + E_{\mathrm{angle}} + E_{\mathrm{dihedral}} + E_{\mathrm{improper\ dihedral}}$. To induce allostery, we add a truncated Gaussian distance term to the soft-sphere atom overlap term, to obtain total nonbonded energy: $E_{\mathrm{nonbonded}} = E_{\mathrm{soft\ sphere}} + E_{\mathrm{distance}}$. This distance term results in efficient sampling of the allosteric transition and is given by a sum over all heavy atom pairs more than two residues apart in sequence and less than 11-Å apart in distance. The energy function for a single atom pair has one or two minima, depending on the distance to the effector (Fig. S2). For an interaction involving atoms in the allosteric site (within a cutoff distance to the effector, r^{AS}), the function has one energetic minimum corresponding to the distance in either the effector bound or unbound structure. For all other pairwise interactions, the function has two minima corresponding to the distances in the bound and unbound structures. The energy and width of the distance interaction was parameterized to reproduce experimental folding temperatures. Varying r^{AS}, an order parameter for allostery, changes how the distance energy is distributed across the structure, thereby driving the simulation to sample different regions of the conformational space relevant to the allosteric transition.

The nonbonded distance energy is a sum of pairwise distance terms $\epsilon(r_{ij})$ applied to all atoms in amino acids that are separated in sequence by at least two residues and are in contact in any of the crystal structures:

$$E_{\mathrm{distance}} = \sum_{i_{\mathrm{index}}^{\mathrm{residue}} > j_{\mathrm{index}}^{\mathrm{residue}}+2} \epsilon(r_{ij})\delta(r_{ij}^{t})$$

in which $\delta({r_{ij}}^{t}) = 1$ if the distance between the side-chain centers of mass is less than 11 Å and $\delta({r_{ij}}^{t}) = 0$ otherwise. The pairwise distance term is found by taking the negative logarithm of a probability density function:

$$\epsilon(r_{ij}) = -RT\log\left(\sum_{t}^{N_{ij}} \mathrm{P}_t^{trG}(r_{ij})\right),$$

in which $\mathrm{P}_t^{trG}(r_{ij})$ is a truncated Gaussian. The probability density function is a sum of truncated Gaussians, each Gaussian pertaining to a maximum at the distance $({r_{ij}}^{t})$ between atoms i and j taken from N_{ij} templates (Fig. S2). Each truncated Gaussian is given by

$$P_t^{trG}(r_{ij}) = \begin{cases} \text{if } r_{ij} \le r_{ij}^t: & [1-0.5(1+\tanh(mr_{ij}+m(\delta r - r_{ij}^t)))]\cdot g_{\min} + [0.5(1+\tanh(mr_{ij}+m(\delta r - r_{ij}^t)))]\cdot P_t^G(r_{ij}) \\ \text{if } r_{ij} > r_{ij}^t: & [1-0.5(1+\tanh(mr_{ij}-m(\delta + r_{ij}^t)))]\cdot P_t^G(r_{ij}) + [0.5(+\tanh(mr_{ij}-m(\delta r + r_{ij}^t)))]\cdot g_{\min} \end{cases}$$

that is an interpolation between a Gaussian function (P_t^G) and a constant value given by $g_{\min}$. These terms are given by

$$P_t^G(r_{ij}) = \frac{1}{N_{ij}\sigma_{ij}\sqrt{2\pi}}\exp[(-0.5(r_{ij}-r_{ij}^t)/\sigma_{ij})^2]$$

$$g_{\min} = \frac{1}{N_{ij}}\exp\left[-\delta E/RT + \log\left(\sum_{\tilde{t}\neq t}^{N_{ij}} P_{\tilde{t}}^{trG}(r_{ij}^{\max})\right)\right],$$

in which σ_{ij} is the standard deviation and $r_{ij}^{\max}$ is the distance between atoms i and j that yields the maximum probability. The truncated Gaussian function limits information taken from any template, which is equivalent to setting the energy of a contact between two atoms. This contact energy, given by δE, was parameterized empirically (along with the distance cutoff) by comparing the results in the current study to experimental data from folding studies (3–6) and studies on the proteins' functional behavior in solution (7, 8). The truncated Gaussian form allows the protein to interconvert between allosteric states. By setting the appropriate pairwise contact energy, the unfolding temperatures of the three proteins are approximately correct (Fig. S3):

$$\delta E = 3.6(N_{\mathrm{res}}/N_{\mathrm{contacts}}),$$

in which N_{res} is the number of residues in the target sequence and N_{contacts} is the number of atom–atom nonbonded contacts. The equation for δE ensures an average energy per residue that is 2 to 3 times the energy required to rotate a backbone dihedral angle. Similar energetic ratios for balancing backbone rigidity to interresidue interactions have been used to successfully predict protein folding routes in previous models (9–12). The standard deviation of the Gaussian function is small to strongly restrain atoms in the backbone, but is given systematically larger values for interactions involving side chains and for interactions involving residues that are unstructured in one or more of the allosteric states. The standard deviation is given by

$$\sigma_{ij} = 2.0(N_{\mathrm{tot}}/N_{ij})^2\theta_{ij}^{\mathrm{SC/BB}},$$

in which N_{tot} is the total number of allosteric states used to define the landscape and N_{ij} is the number of templates that are used to define the interaction between atoms i and j. Interactions are scaled using $\theta_{ij}^{\mathrm{SC/BB}}$ so contacts between backbone atoms have a value of 1.0, side-chain–backbone contacts is 1.5, and side-

chain–side-chain contacts is 1.5^2. The factor 1.5 arises due to the observation that side-chain atoms are approximately 50% more mobile than backbone atoms in molecular dynamics trajectories as well as in ensembles generated from NMR data (13, 14).

We varied a number of parameters within a wide range, without affecting our conclusions based on the simulations; the absolute rates of motions within the simulation change but the relative rates of motions remain similar (Fig. 3 and Fig. S3). Monitoring the variability of results as a function of r^{AS}, which provides an order parameter for allostery, allows an estimate for how well each landscape is sampled.

Simulations. The simulation protocol in MODELLER is set up to most efficiently sample regions of the energy landscape that are important for allostery by initializing structures along relevant regions of the energy landscape, similar to variational calculations in protein folding (15). The initial structure is generated by first aligning the two allosteric structures; second by interpolating the positions of each atom between the two allosteric states; and third, randomizing each atom by 2 Å. The structures are first relaxed with conjugate gradient steps using only the bonded energy term. Conjugate gradient relaxation is then performed in successive steps of increasing strength in absence of the soft-sphere energy. Molecular dynamics at 300 K is used to optimize the structure as the strength of the soft-sphere energy term is gradually increased. Further molecular dynamics at gradually increasing temperatures equilibrates the structure until the desired sampling temperature is reached, which is 300 K for the allostery model. The bulk of computational time is spent sampling the landscape using constant temperature molecular dynamics with 3 fs time steps and velocity rescaling every 200 steps. Sampling for each landscape involves 30 simulations that are first equilibrated and then followed by a 6-ns run. The total sampling for each protein is more than 1.08 ms and over 2 million structures.

Ensemble Analysis. In the allostery model, simulation trajectories sampling each landscape ($\mathbf{E}_i$) are combined for analysis. For most results, trajectories representing a single landscape are combined ($N_{ASi} = 1$), but for pseudocorrelation maps, data from all trajectories are combined ($N_{ASi} = 6$). We sample related landscapes that differ by the size of the allosteric site (r^{AS}) and whether the allosteric site is in the bound or unbound configuration. The probability for a given structure is

$$P(i) = \frac{\exp[-\mathbf{E}_i/\sigma_{AS_i}]}{N_{AS_i} Z_{AS_i}},$$

where N_{ASi} is the number of different landscapes used in the analysis. Structures are weighted using the energy for each sampled landscape ($\mathbf{E}_i$) and the standard deviation of the energy for each landscape (σ_{ASi}). There is likewise a separate partition function for each landscape:

$$Z_{AS_i} = \sum_i \exp[-\mathbf{E}_i/\sigma_{AS_i}].$$

Structural Analysis. We compare structures from simulations to crystal structures using pairwise distance similarity scores (11, 12, 15). For a given structure, an overall fold similarity to any other structure t is given by Q^t, reflecting the fraction of similar contacts:

$$Q^t = \frac{1}{N} \sum_{i<j+1}^{N} \exp[-(r_{ij} - r_{ij}^t)^2/2(\sigma_{ij})^2]$$

where r_{ij} is the distance between the centers of mass of side chains i and j, $\sigma_{ij} = 2.0$, and the sum is over all pairs of atoms within 11 Å of each other for which $|i - j| > 1$. To determine if a simulated structure is more similar to the effector bound ($t+$) or the effector unbound ($t-$) crystal structures, we calculate

$$Q_{diff} = \frac{Q^{t+} - Q^{t-}}{(1 - \Delta Q)}$$

where ΔQ is the structural similarity (Q^t) between the two allosteric crystal structures. Restricting the calculation to a subset of contacts, such as $Q_{diff}(X)$, results in a score for region X. Also, $QI_{diff}(X)$ refers to a score of the interface between X and the remaining protein and $QI_{diff}(X \text{ to } Y)$ refers to a score of the interface between X and Y.

1. Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815.
2. Vanommeslaeghe K, et al. (2010) CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* 31:671–690.
3. Novokhatny V, Ingham K (1997) Thermodynamics of maltose binding protein unfolding. *Protein Sci* 6:141–146.
4. Nagy A, Malnasi-Csizmadia A, Somogyi B, Lorinczy D (2004) Thermal stability of chemically denatured green fluorescent protein (GFP)—a preliminary study. *Thermochim Acta* 410:161–163.
5. Andrews BT, Gosavi S, Finke JM, Onuchic JN, Jennings PA (2008) The dual-basin landscape in GFP folding. *Proc Natl Acad Sci USA* 105:12283–12288.
6. Johnson SE, Ilagan MXG, Kopan R, Barrick D (2010) Thermodynamic analysis of the csl-notch interaction distribution of binding energy of the notch ram region to the csl beta-trefoil domain and the mode of competition with the viral transactivator ebna2. *J Biol Chem* 285:6681–6692.
7. Millet O, Hudson RP, Kay LE (2003) The energetic cost of domain reorientation in maltose-binding protein as studied by NMR and fluorescence spectroscopy. *Proc Natl Acad Sci USA* 100:12700–12705.
8. Friedmann DR, Wilson JJ, Kovall RA (2008) RAM-induced allostery facilitates assembly of a Notch pathway active transcription complex. *J Biol Chem* 283:14781–14791.
9. Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG (2001) Evaluating protein structure-prediction schemes using energy landscape theory. *IBM J Res Dev* 45:475–497.
10. Whitford PC, et al. (2009) An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins Struct Funct Bioinf* 75:430–441.
11. Weinkam P, Romesberg FE, Wolynes PG (2009) Chemical frustration in the protein folding landscape: Grand canonical ensemble simulations of cytochrome c. *Biochemistry* 48:2394–2402.
12. Weinkam P, Zong CH, Wolynes PG (2005) A funneled energy landscape for cytochrome c directly predicts the sequential folding route inferred from hydrogen exchange experiments. *Proc Natl Acad Sci USA* 102:12401–12406.
13. Zhou YQ, Vitkup D, Karplus M (1999) Native proteins are surface-molten solids: Application of the Lindemann criterion for the solid versus liquid state. *J Mol Biol* 285:1371–1375.
14. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. *Nature* 433:128–132.
15. Portman JJ, Takada S, Wolynes PG (2001) Microscopic theory of protein folding rates. I. Fine structure of the free energy profile and folding routes from a variational approach. *J Chem Phys* 114:5069–5081.

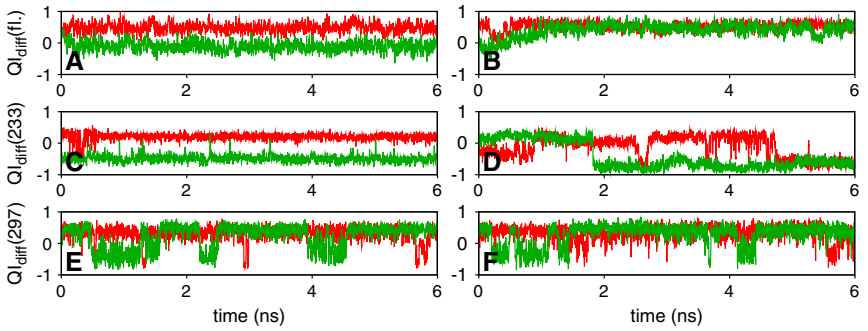

Fig. S1. Structural similarity metrics for the regulated sites are shown for representative simulation trajectories: (*A* and *B*) CaGFP, (*C* and *D*) maltose binding protein, and (*E* and *F*) CSL transcription factor. Red curves are from simulations in the effector bound state and green curves are from simulations in the effector unbound state. The plots in the left column are from simulations with a large r^{AS} (roughly half the distance between the allosteric and regulated sites) and represent the effector bound/unbound landscapes most consistent with experiment. The plots in the right column are from simulations with a small r^{AS} (roughly 5 Å) and represent an interpolation between the landscapes represented on the left. Some trajectories involve interconversions between substates, including a partial folding transition for CaGFP.

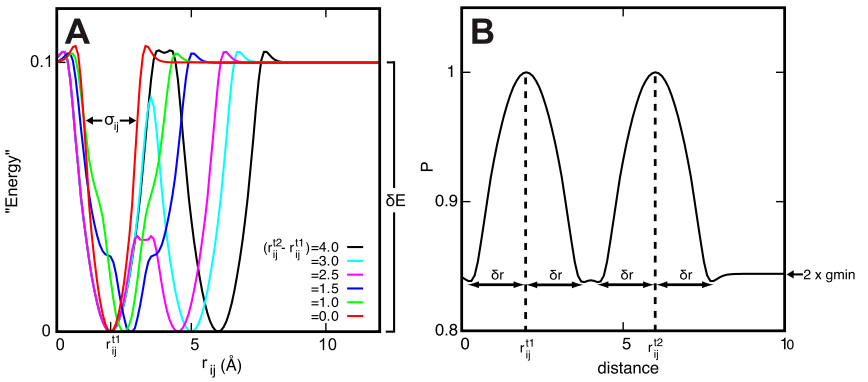

Fig. S2. (*A*) Plot of $E_{distance}$ for several contacts with two minima. The value r_{ij}^{t1} is the distance between atoms *i* and *j* in template *t*1 and σ_{ij} corresponds to the width of the Gaussian for that contact. (*B*) A sum of two truncated Gaussian probability density functions that correspond to the energy plot in *A*. Several parameters in the truncated Gaussian probability density function are depicted.

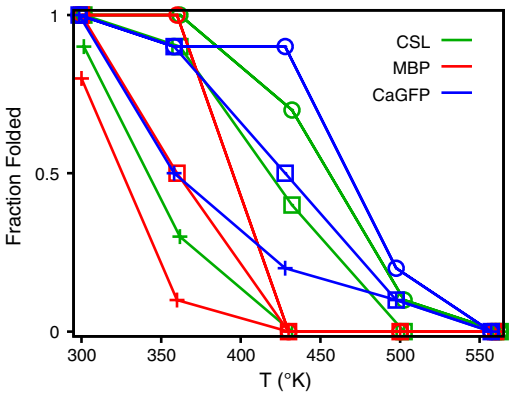

Fig. S3. The unfolding temperatures of CSL transcription factor (CSL), maltose binding protein (MBP), and the GFP domain of CaGFP are accurately predicted. Each point represents the fraction of folded proteins after 10 6-ns simulations in which different distance cutoffs are used: 9 (+), 11 (□), and 15 Å (○). Because unfolding likely occurs much more slowly than 6 ns, these curves represent an approximate upper bound for unfolding within the model. The experimental unfolding temperatures for MBP and GFP are 345 (3) and 356 K (4) respectively. Guanidine unfolding experiments also seem to place the stability of CSL in between MBP and GFP (5, 6). A structure is defined as folded if all of the domains have a Q^t with respect to the native crystal structure above 0.5.

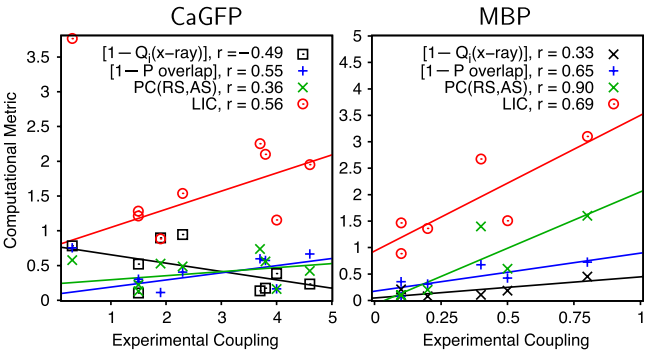

Fig. S4. Correlations of several computational metrics with the observed experimental coupling for CaGFP and maltose binding protein (MBP). Computational metrics are presented such that positive correlation implies accuracy. Ligand-induced cooperativity (LIC) shows the best overall correlation. PC (RS, AS), which refers to pseudocorrelation (Fig. 4) between the allosteric site (AS) and regulated site (RS), shows good correlation for MBP but poor correlation for CaGFP. *P* overlap, which refers to the overlap of QI_{diff} (Fig. 3), also shows good correlation because a small *P* overlap implies large degrees of coupling. Q_i(X-ray), a residue-specific structural similarity measurement applied between the effector bound and unbound crystal structures (i.e., ΔQ_i), fails to be well correlated with experimental coupling. Experimental coupling for CaGFP is defined as the average absolute deviation of fluorescence (Table S1). Experimental coupling for MBP is defined as $|\log(K_d^{wt}/K_d^{mut})|$. Correlations for CaGFP are shown without the data point for residue 377 because this residue is contained in the allosteric site in the simulations and is therefore predicted to have arbitrarily large coupling to effector binding.

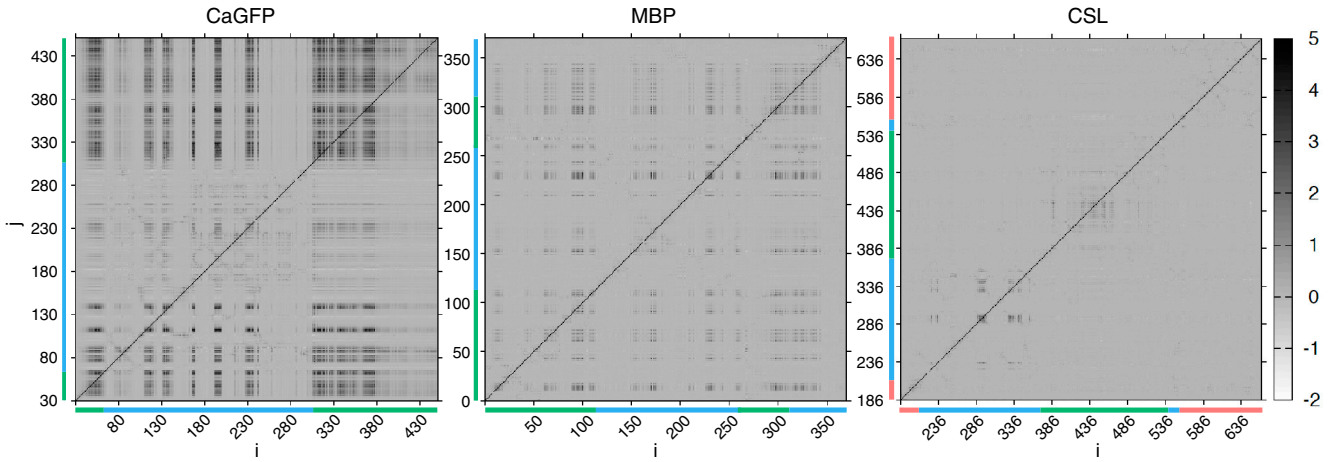

Fig. S5. Pseudocorrelation maps ($PC^{t-}(j, i)$) are obtained by assigning all residues into the effector bound or effector unbound substate using QI_{diff}. Colors along the *x* and *y* axes correspond to domains, which are not contiguous in these proteins. For CaGFP, green is the calmodulin domain and blue is the GFP domain. For maltose binding protein (MBP), green and blue represent the two domains on either side of the effector binding site. For CSL transcription factor, green is the β-trefoil domain, blue is the Ig-like domain containing the regulated site, and red is the Ig-like domain that does not participate in the allostery.

Other Supporting Information Files

Table S1 (DOCX)
Table S2 (DOCX)