PROTEINS: Structure, Function, and Bioinformatics 67:385–399 (2007)

Local Motions in a Benchmark of Allosteric Proteins

Michael D. Daily[1] **and Jeffrey J. Gray**[1,2]*
[1]*Program in Molecular and Computational Biophysics, Johns Hopkins University, Baltimore, Maryland 21218*
[2]*Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland 21218*

ABSTRACT **Allosteric proteins have been studied extensively in the last 40 years, but so far, no systematic analysis of conformational changes between allosteric structures has been carried out. Here, we compile a set of 51 pairs of known inactive and active allosteric protein structures from the Protein Data Bank. We calculate local conformational differences between the two structures of each protein using simple metrics, such as backbone and side-chain Cartesian displacement, and torsion angle change and rearrangement in residue–residue contacts. Thresholds for each metric arise from distributions of motions in two control sets of pairs of protein structures in the same biochemical state. Statistical analysis of motions in allosteric proteins quantifies the magnitude of allosteric effects and reveals simple structural principles about allostery. For example, allosteric proteins exhibit substantial conformational changes comprising about 20% of the residues. In addition, motions in allosteric proteins show strong bias toward weakly constrained regions such as loops and the protein surface. Correlation functions show that motions communicate through protein structures over distances averaging 10–20 residues in sequence space and 10–20 Å in Cartesian space. Comparison of motions in the allosteric set and a set of 21 nonallosteric ligand-binding proteins shows that nonallosteric proteins also exhibit bias of motion toward weakly constrained regions and local correlation of motion. However, allosteric proteins exhibit twice as much percent motion on average as nonallosteric proteins with ligand-induced motion. These observations may guide efforts to design flexibility and allostery into proteins. Proteins 2007;67:385–399.** © 2007 **Wiley-Liss, Inc.**

Key words: allosteric protein structure; conformational change; protein motion; protein flexibility; protein crystal structure database

INTRODUCTION

Allosteric regulation is a major mechanism of control in many biological processes, including cell signaling, gene regulation, and metabolic regulation.[1] Allostery presents a special problem in protein structure because contrary to the classical paradigm in which a protein folds into a single rigid structure based upon its sequence,[2,3] an allosteric protein adopts different structural and functional states depending upon the environmental conditions.[4] In this work, we systematically calculate and characterize local structural differences in known pairs of inactive and active allosteric protein crystal structures. Such an analysis reveals general trends about the structural basis of protein allostery that are not obvious from previous analyses of individual proteins. In addition, this detailed structural analysis is useful for potential medical and engineering applications such as rational drug design to target oncogenic allosteric proteins and creation of protein switches for nanotechnological applications.

In the past 50 years, much theoretical and experimental work has been directed toward illuminating the structural basis of protein allostery. Monod, Wyman, and Changeux (MWC)[5] and Koshland, Nemethy, and Filmer (KNF)[6] postulated that allosteric proteins are oligomeric and adopt two structural states called T ("tight," biochemically inactive) and R ("relaxed," active). In this work we refer to these two states as "I" and "A" rather than "T" and "R" to avoid implying that the active state is more flexible than the inactive state. In MWC's model, concerted changes in quaternary structure couple different subunits together, while in KNF's model, ligand-induced changes in tertiary structure propagate sequentially among subunits. Since then, allosteric crystal structures have revealed that most oligomeric allosteric proteins are complex systems with both tertiary and quaternary structure changes[7] and that allosteric conformational changes occur in monomeric proteins such as signaling proteins.[8,9] However, recent theoretical and experimental works have demonstrated that changes in dynamics as well as changes in average structure are important to allosteric switching.[10–12]

Nevertheless, comparisons of high-resolution I and A crystal structures of allosteric proteins (e.g. Refs. 13–16) often reveal both subtle local motions and obvious large-scale motions from which elegant, intuitively satisfying mechanistic models can be constructed. For example, Perutz found that for hemoglobin, oxygen binding or dis-

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: NIH; Grant number: K01-HG02316.

*Correspondence to: Jeffrey J. Gray, 3400 N. Charles Street, 221 Maryland Hall, Baltimore, MD 21218. E-mail: jgray@jhu.edu

Received 24 July 2006; Accepted 16 October 2006

Published online 12 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21300

sociation causes a change in the local structure of the individual α or β subunit which then propagates to and triggers a concerted rotation of the interface between the αβ dimers, facilitating corresponding tertiary structure changes in other monomers.[17] Recent NMR experiments have revealed functionally interesting dynamic differences between I and A states of some allosteric proteins, but at substantially lower resolution than that of crystallographic comparisons.[18–20] Thus, crystal structures are the most detailed currently available source of structural information about allosteric transitions in proteins. However, no systematic analysis of conformational changes in the many known pairs of I and A structures has yet been carried out.

Previous systematic characterizations of structural properties from databases of protein structures have revealed biophysically important principles about other protein phenomena. With a small set of proteins, Chothia[21] demonstrated that the surface area buried by a protein upon folding is a simple function of molecular weight and that protein residues are as closely packed as are crystal structures of amino acids. Rooman et al.[22] demonstrated from a set of 75 proteins that an unexpectedly high fraction of short sequences have a strong propensity to adopt a particular secondary structure. Systematic analyses of protein–protein complex structures[23,24] have shown that homodimers have more hydrophobic interface compositions than have heterodimers, that interfaces in heterodimers exhibit approximately the same polarity as surface residues in general, and that the interface area of a protein complex is higher for proteins with conformational changes upon complex formation than for proteins which associate rigidly. Of direct relevance to our study, databases of proteins with multiple crystal structures have shown that in most proteins, the extent of backbone and side-chain conformational change due to binding of small molecules[25] and other proteins[26] is small. Such databases have also been used to classify small- and large-scale motions in protein structures into regularly occurring conceptual categories such as hinge and shear.[27–30] The successes of these analyses suggest that a systematic analysis of local motions in allosteric proteins would yield significant biophysical insight into the problem of allostery.

In this work, we create a database of allosteric conformational changes in known pairs of I and A structures and characterize the structural properties of these conformational changes. We compile a set of 51 allosteric proteins with high-resolution crystal structures of both I and A states. We consider any protein allosteric that binds an effector molecule at one site with the result of a functional change at a second site. In each protein, we identify moving residues with a variety of types of local conformational difference metrics including backbone and side-chain motions in both Cartesian space and dihedral angle space and rearrangements in local contact structure. While some multidomain and oligomeric proteins in this set exhibit quaternary structure (domain and subunit) motions, we do not treat such large-scale motions in this work. Next, we calculate from the motions in the database structural statistics, including the extent of motion by each criterion, the propensities of motion in different local structural environments, and the correlations of motions at varying sequence and Cartesian space separations. These structural properties of motions quantify the magnitudes of allosteric effects in proteins and describe how allosteric conformational changes depend on simple properties of local structure. Since nonallosteric proteins can exhibit functionally significant motions, we also calculate and statistically analyze motions in a set of 21 ligand-binding nonallosteric proteins. We then compare structural properties of motions between this set and the allosteric set to discern which properties of allosteric motions are unique to allostery and which are general properties of protein motion.

RESULTS

Dataset of Allosteric Proteins

We compile a set of 51 ligand-induced allosteric proteins for which high-resolution structures of at least two allosteric states are available in the Protein Data Bank[31] through several online keyword searches. This set contains signaling proteins, DNA-binding proteins, and enzymes, which have diverse biochemical functions, structural topologies, and potential allosteric mechanisms. About one-third of the proteins in the set are also included in the protein motions part of Echols coworkers[32,33] Database of Molecular Movements. Only 42 of 51, which we refer to as the reduced allosteric set, are used in statistical analyses because the remaining nine are missing significant parts of the effector or substrate binding sites. The PDB codes for the two states of each protein, the proteins excluded from the reduced allosteric set, and the details of the selection criteria and literature search methods are given in the Methods section.

Calculating Local Motions

We measure local motion by six metrics (see Fig. 1). These include backbone torsional motion ($\max(|\Delta\varphi|, |\Delta\psi|)$), C_α displacement relative to the core of the protein (ΔC_α), angular movement of the $C_\alpha \rightarrow C_\beta$ bond vector ($\theta_{\alpha\beta}$), which captures backbone motions that reorient the side-chain, and side-chain motion by dihedral angle changes and root-mean squared displacement of side-chain atoms relative to the backbone (ΔSC). In addition to these classical metrics of conformational change, we measure fractional change in atom–atom contacts (set of atoms within 6.0 Å of a residue) to capture residues, which by moving affect other residues ($\max(f_I, f_A)$). $\theta_{\alpha\beta}$ is not defined for glycine, and and ΔSC are not defined for glycine, alanine, and proline. Residues with undefined motion by a metric are excluded from statistical analyses for that metric throughout this work. For determination of ΔC_α and $\theta_{\alpha\beta}$, we employ a method that detects the rigid core of the protein and includes only those residues in the superposition, since a standard least-squared superposition would produce a physically inappropriate result by including flexible residues in the calculation. The details of the superposition algorithm and the definition of are given in the methods.

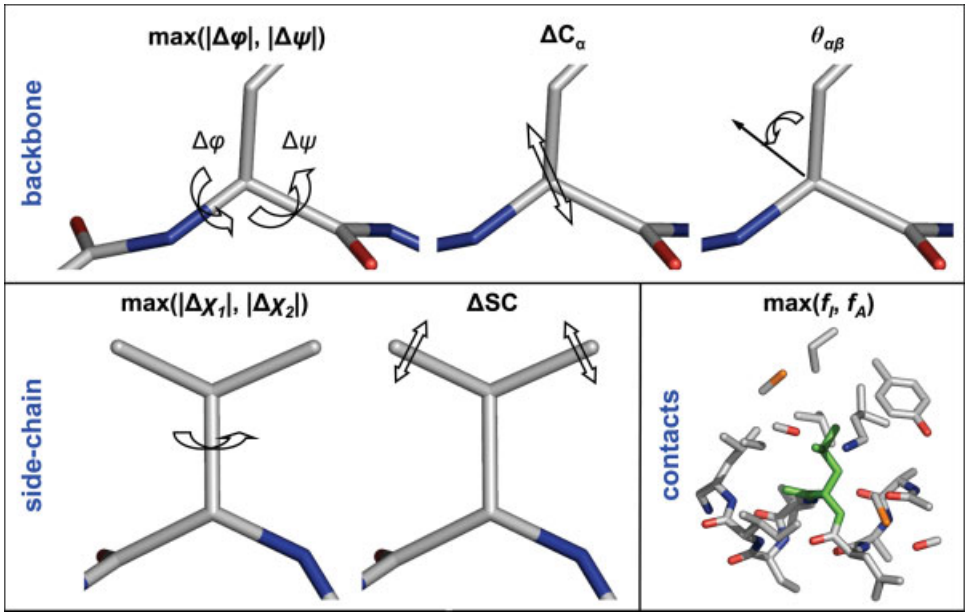

Fig. 1. Six ways of measuring local motion. max(|Δφ|, |Δψ|): maximum of the absolute values of the changes in φ and ψ backbone torsion angles, respectively. ΔC_α: Cartesian displacement of C_α between I and A conformations after superimposing A onto I. $\theta_{\alpha\beta}$: angle between the I and A conformations of the $C_\alpha \rightarrow C_\beta$ bond vectors after superimposing A onto I. max(|Δχ$_1$|, |Δχ$_2$|): maximum of the absolute values of the changes in χ_1 and χ_2 side-chain torsion angles, respectively. ΔSC: root-mean squared displacement of side-chain atoms beyond C_β after superimposing A coordinates of the backbone atoms (N, C_α, C) onto I coordinates. max(f_I, f_A): maximum of the fractions of atoms within 6.0 Å unique to I and A states, respectively.

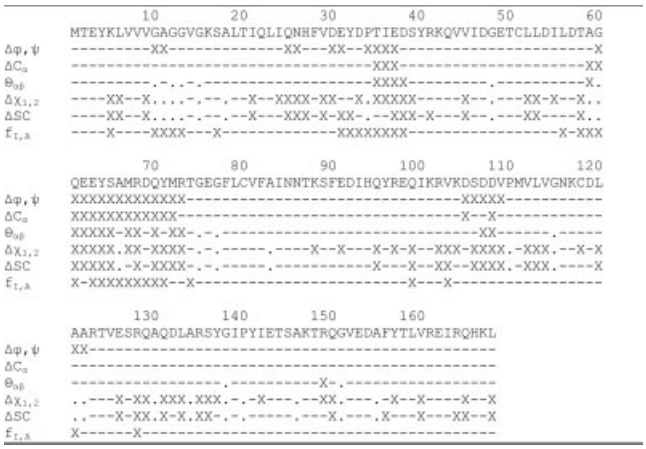

Fig. 2. Six motion metrics vs. sequence for ras. For each metric, 'X' denotes a residue which exceeds the threshold (methods) for that metric, '-' a residue below that threshold, and '.' a residue for which that metric is not defined. Motions were calculated from 4Q21 vs. 6Q21.

Because crystal structures vary slightly because of intrinsic flexibility, different crystallization conditions, and different refinement methods,[34] we identify thresholds for each of the metrics shown in Figure 1 to separate truly moving residues from residues that do not vary beyond this noise. As a reference for intrinsic flexibility in protein structures in general, we collect a dataset of five pairs of nonallosteric protein structures (Control 1). Since allosteric proteins might be more intrinsically flexible than nonallosteric proteins,[20] we also collect a set of nine pairs of allosteric protein structures in the same biochemical (I or A) state (Control 2). We set thresholds for allosteric and ligand-binding nonallosteric proteins to exclude approximately 99% of the background motion in the nonmoving control distributions. Rotameric shifts in torsion angles are not included as background motion. The resultant thresholds are 30° for $(\max(|\Delta\varphi|, |\Delta\psi|))$, 1.2 Å for ΔC_α, 28° for $\theta_{\alpha\beta}$, 46° for $(\max(|\Delta\chi_1|, |\Delta\chi_2|))$, 2.0 Å for ΔSC, and 0.20 for $\max(f_I, f_A)$. The PDB codes for the Control 1 and 2 sets and details of the threshold calculations are in the Methods section.

Ligand-Binding Nonallosteric Set

Because allosteric proteins might share some features of motion with proteins that bind ligands but are not allosteric, we collect a set of 21 proteins that bind ligands but are not known to be allosteric (Control 3) with two different liganded states in the PDB. Of these 21, 3 were selected manually and 18 were randomly selected from families in the PDB with two or more high-resolution structures. The details of this set are given in the Methods.

Motions in Protein Space

Figures 2 and 3 show moving residues mapped onto the sequences and three-dimensional structures, respectively, of three proteins from the allosteric benchmark and one ligand-binding nonallosteric protein. The sequence view of ras (see Fig. 2) reveals two major regions of motion at approximately residues 31–38 and 59–73 which in the three-dimensional structure (Fig. 3, top row) form a contiguous surface patch adjacent to the

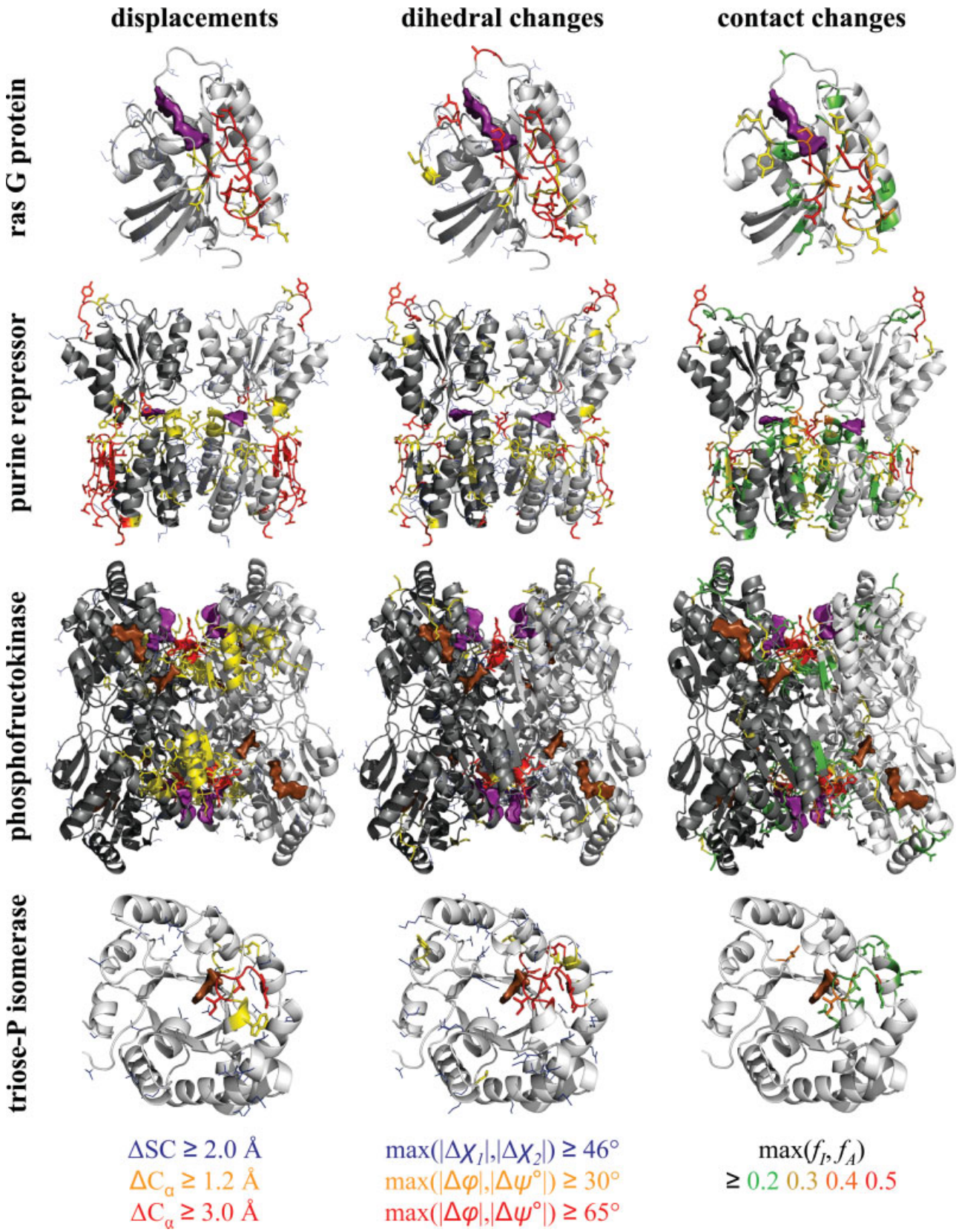

Fig. 3. Calculated local motions mapped onto the structures of three allosteric proteins and one ligand-binding non-allosteric protein. Top row: ras (4Q21 vs. 6Q21). Second: purine repressor (1DBQ vs. 1WET). Third: phosphofructokinase (6PFK vs. 4PFK). Bottom: triosephosphate isomerase (8TIM vs. 1TPH), a non-allosteric protein. The A state structure is displayed for allosteric proteins, and 1TPH for triosephosphate isomerase. Left column: backbone and side-chain Cartesian displacements. Center: backbone and side-chain dihedral angle changes. Right: contact changes. Colors corresponding to different motion thresholds noted below each column of panels. Purple surface: allosteric effector. Brown surface: substrate (if applicable). Figures made with PyMOL.[35]

GTP effector site. These two segments (switches I and II) and the patch they form have been previously identified by Milburn et al.,[36] and this patch was subsequently confirmed to be the interaction site with ras targets like raf.[37] In purine repressor (Fig. 3, second row), the calculations identify two clusters of motion, which have been previously identified by Schumacher et al.,[38] residues 69–75 near the guanine effector site and residues 144–160 and 311–315. In phosphofructokinase (PFK, Fig. 3, third row), the calculations identify a cluster of motion containing residues 154–162 and 211–216, which bridges the allosteric ADP site and the catalytic fructose-6-phosphate site. This cluster may form a pathway of motion between these two sites, in accordance with the PFK mechanism hypothesized by Schirmer and Evans.[39] Thus, the calculations reveal functionally interesting local motions for allosteric proteins that corroborate the most significant features of local motion previously identified manually by crystallographers. The three backbone motion metrics and identify broadly similar motions in sequence space for ras (see Fig. 2) and in three-dimensional space for all three allosteric proteins (see Fig. 3), capturing the underlying motions in different ways. However, these four metrics identify slightly different sets of residues, showing that no single metric captures all of the effects of the underlying motions. Side-chain motions overlap well with backbone and contact motions in the most flexible regions of the proteins, but they also show significant density in other regions, possibly reflecting dynamic side-chain fluctuations.

Furthermore, these calculated motions display patterns in sequence and structural space which suggest hypotheses pertaining to the structural basis of allostery. First, only small portions of the structures in Figure 3 experience backbone, side-chain, or contact motion, which argues that conservation of overall topology is important for the function of allosteric proteins. Second, backbone displacements and dihedral angle changes and contact changes cluster in both sequence and three-dimensional spaces, which implies that proteins change conformation in a structurally coordinated fashion. Third, allosteric proteins exhibit locally contiguous clusters of motion between allosteric and catalytic sites in some allosteric proteins (PFK and ras), which suggests that mechanical connectivity is important for linkage between sites in allosteric proteins.

The calculations also reveal a functionally relevant motion in the nonallosteric protein triosephosphate isomerase (Fig. 3, bottom row), a substrate-induced loop closure (residues 168–176) previously identified by Zhang et al.[40] Like the allosteric proteins, triosephosphate isomerase exhibits motion which is small relative to the size of the protein and tightly clustered.

Extent of Structural Rearrangement

Figures 2 and 3 show three cases where most of an allosteric protein maintains its local structure in the I to A transition. To assess the generality of this observation, we calculate f_p^m, the fraction of residues which move above the threshold for each metric m and each protein p in control and allosteric sets. Supplementary Table I shows f_p^m for each metric for each protein in the allosteric dataset. Table I shows that the average fraction of residues $\langle f_p^m \rangle$ moving by backbone and contact metrics is 3% or less in all three control sets, although the standard deviations are higher in Control 3 than in controls 1 and 2. Backbone and contact motion frequencies vary from 9% to 20% among metrics for the reduced allosteric set. All the allosteric proteins are clearly distinguishable from nonmoving and ligand-binding controls by at least one of these four metrics. Allosteric proteins with the largest extent of conformational rearrangement are chorismate mutase (47% by ΔC_α) and arf1 (47% by $\max(f_I, f_A)$). The average fraction of residues moving by side-chain metrics is 10–15% for control proteins and 25–30% for allosteric proteins. Eyal et al.[34] also observed that side-chain conformational changes are common even for crystal structures in the same space group. However, 15–20% more residues per protein experience side-chain motion in allosteric proteins than in nonmoving control proteins, which indicates that a substantial fraction of side-chain motions in allosteric proteins is functionally significant. The relative standard deviations of f_p^m for allosteric proteins are high (25–50% of the mean), which indicates significant variability in allosteric mechanisms among proteins. f_p^m varies among metrics depending on the physical meaning of the metric; for example, C_α displacements in allosteric proteins (20%) are more common than changes in local backbone conformation (9% by reorientation of the C_α–C_β vector and 15% by a change in the φ or ψ angle) because significant portions of a structure can move in Cartesian space with just a few changes in φ and ψ angles. For the three backbone metrics and $\max(f_I, f_A)$, there is significant variation in f_p^m, and thus in allosteric mechanisms among functional classes; however, no one class has systematically higher f_p^m than the others.

The average extent of motion in allosteric proteins is substantially higher than that in ligand-binding nonallosteric proteins. Most ligand-binding nonallosteric proteins show little or no motion; only 6 of the 21 Control 3 targets, including 4 of the 18 randomly selected targets, exhibit motion of 5% or more of residues by at least one of the backbone or contact metrics. The average motion frequency among these six varies from 5 to 10% among backbone and contact metrics; we refer to these as Control 3′. Only one Control 3 target, guanylate binding protein 1 (22% by ΔC_α and 16% by $\max(f_I, f_A)$, is in the middle range of allosteric motion frequencies.

Local Structural Environment Effects

Different sites in allosteric proteins may have different propensities for experiencing motion depending on the local structural environment. That is, allosteric proteins might use the details of local structure to regulate which residues move. Figure 4 compares the frequencies of motion among different categories of

TABLE I. Average Fractions of Residues Moving in Control and Allosteric Sets

	max(|Δφ|, |Δφ|)	ΔC_α	$\theta_{\alpha\beta}$	max(|$\Delta\chi_1$|, |$\Delta\chi_2$|)	ΔSC	max(f_I, f_A)
Control 1	0.03 ± 0.03	0.02 ± 0.03	0.01 ± 0.02	0.13 ± 0.09	0.11 ± 0.08	0.03 ± 0.02
Control 2	0.03 ± 0.02	0.01 ± 0.01	0.01 ± 0.01	0.16 ± 0.08	0.10 ± 0.05	0.02 ± 0.02
Control 3	0.03 ± 0.03	0.03 ± 0.06	0.02 ± 0.02	0.12 ± 0.08	0.08 ± 0.06	0.03 ± 0.04
Control 3′	0.08 ± 0.03	0.10 ± 0.07	0.04 ± 0.03	0.22 ± 0.06	0.14 ± 0.05	0.09 ± 0.04
Signaling	0.16 ± 0.06	0.22 ± 0.09	0.13 ± 0.05	0.33 ± 0.07	0.26 ± 0.05	0.23 ± 0.09
Transcription	0.20 ± 0.08	0.13 ± 0.08	0.07 ± 0.03	0.38 ± 0.10	0.27 ± 0.04	0.16 ± 0.05
Enzymes	0.12 ± 0.07	0.21 ± 0.11	0.06 ± 0.03	0.30 ± 0.09	0.21 ± 0.07	0.14 ± 0.06
All allosteric	0.15 ± 0.07	0.20 ± 0.10	0.09 ± 0.05	0.33 ± 0.08	0.24 ± 0.06	0.18 ± 0.09

The fractions of moving residues in control sets and in the reduced allosteric set and its classes are averages over all proteins in each respective set of $f_p^m = N_p^m/N_p$, where N_p is the total number of residues and N_p^m is the number of residues which moves by metric m, that is, exceeds the threshold for metric m, in all subunit types of protein p. The errors are standard deviations of f_p^m over all proteins in each of the respective sets. Control 3′ refers to six proteins from Control 3 for which each target moves by 5% or more by at least one backbone or contact metric.

sequence (primary structure), secondary structure, and solvent-exposure (tertiary structure) environments in proteins from the reduced allosteric set to quantify these possible effects. Polar residues are more likely to experience all types of motion, especially side-chain motion, than are apolar residues. Loop residues are substantially more likely to experience contact motion and all three types of backbone motion, especially a change in φ or ψ, than are helix or strand residues. These propensities likely result from the involvement of backbone atoms in hydrogen bonds to other backbone atoms in helices and strands. Side-chains, the atoms of which do not participate in such hydrogen bonds, are substantially less affected by the presence or absence of secondary structure. In addition, helix and sheet secondary structures exert slightly differing effects on motion propensities for some metrics. β-sheet residues, which hydrogen-bond to residues in adjacent strands, are more likely to experience φ or ψ changes than are α-helical residues, which hydrogen-bond within helices. However, α-helical residues are more likely to be displaced in Cartesian space than are β-sheet residues, probably because an α-helix can maintain its internal hydrogen bonds if it moves in a concerted fashion. Furthermore, exposed residues, which are constrained by fewer contacts than are their buried counterparts, are more likely to experience backbone, side-chain, and contact motion than are buried residues.

These observations imply that the higher propensity of polar residues than of apolar residues for motion may simply reflect the higher propensity of polar residues than of apolar residues to be in motion-tolerant solvent-exposed environments. To assess this possibility, we compared conditional probabilities of motion between polar and apolar residues in 5% bins of all-atom, side-chain, and backbone solvent-accessible surface area (ASA) for proteins in the reduced allosteric set. Polar side chains move more frequently than do apolar side chains regardless of ASA. However, polar and apolar residue backbone and contact motion frequencies vary equivalently in bins of constant all-atom ASA (max(|Δφ|, |Δψ|)), side-chain ASA (max(|Δφ|, |Δψ|) and max(f_I, f_A)), or backbone ASA (ΔC_α and $\theta_{\alpha\beta}$). Therefore, these data most strongly

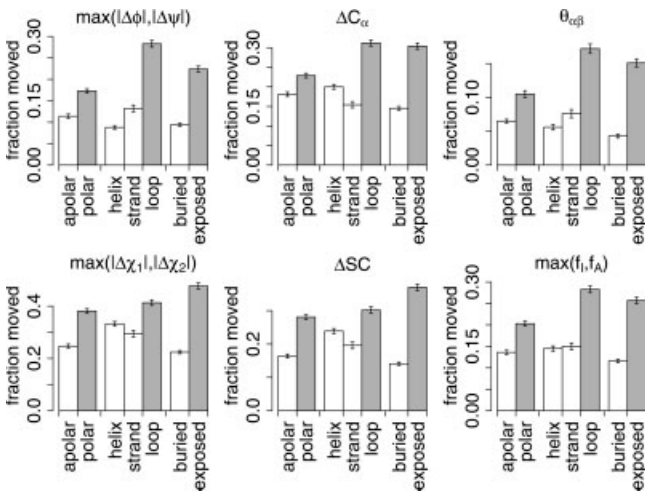

Fig. 4. Frequencies of motion (f_m) in different local structure environments. Frequencies of motion are among residues of the respective categories shown above in all subunit types of all proteins in the reduced allosteric set. Apolar residues include A, C, F, I, L, M, P, V, W, and Y, and polar residues include D, E, G, H, K, N, Q, R, S, and T. Secondary structure is assigned by DSSP,[41] buried residues are defined by ≤30% all-atom ASA by naccess,[42] and exposed residues are defined by >30% ASA. Error bars represent estimated sampling errors in the f_m values.

support the interpretation that polar side-chains are intrinsically more flexible than apolar side-chains but that side-chain polarity *per se* does not affect a residue's propensity to undergo backbone or contact motion. In summary, the local structural environment of a site in an allosteric protein strongly influences the propensity for motion at that site, and most notably, weakly constrained environments like loop backbones and solvent-exposed regions are most tolerant of motion.

The high propensity of motion in weakly constrained regions may be a general property of protein structure and not unique to allosteric proteins. Supplementary Figure 1 shows the frequencies of motion in different structural environments for six nonallosteric proteins with significant ligand-induced motion (Control 3′). The absolute motion frequencies are lower among nonallosteric proteins than among allosteric proteins, but the

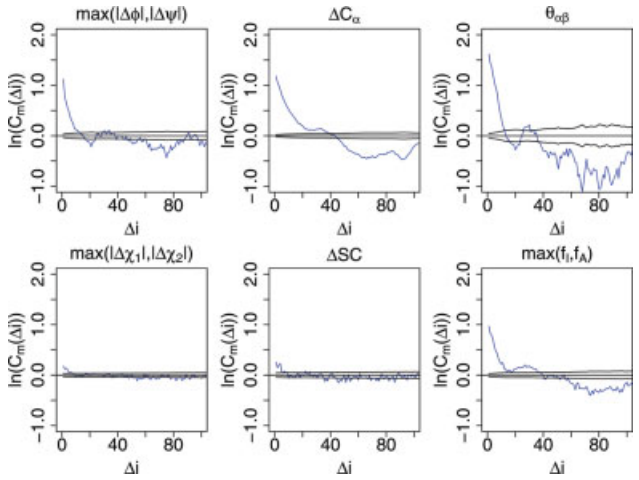

Fig. 5. Correlation of local motions vs. sequence separation. $C_m(\Delta i)$ is correlation and Δi is sequence separation between residues. Estimated sampling error in C_m is shown as dashed lines. C_m is calculated from all subunit types of all proteins in the reduced allosteric set as described in the methods. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

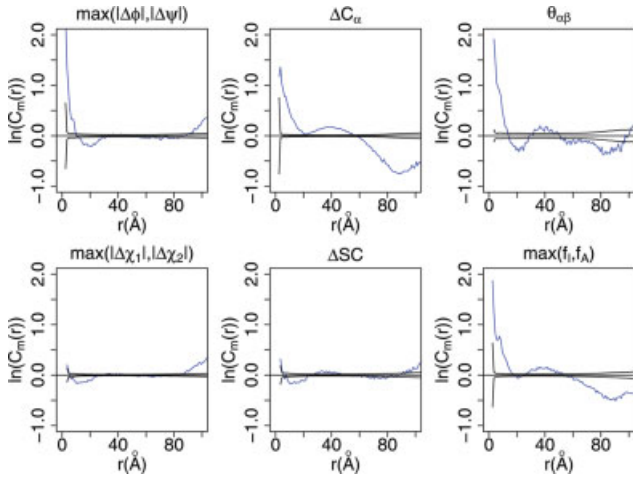

Fig. 6. Correlation of local motions vs. Cartesian distance separation. $C_m(r)$ is correlation and r is the integrally binned Cartesian distance separation between the all-atom centroids of residues. Estimated sampling error in C_m is shown as dashed lines. C_m is calculated from all subunit types of all proteins in the reduced allosteric set as described in the methods. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

propensities of motion in different protein structure environments are similar between these two sets.

Local Coupling of Motions

Figures 2 and 3 qualitatively show that local motions tend to cluster in sequence and three-dimensional space for three proteins; that is, moving residues are structurally linked with other moving residues. We compute correlation functions in sequence and Cartesian spaces for allosteric proteins which provide a statistical measure of the strength and significance of these cooperative effects and the distance ranges over which they occur. The sequence correlation function $C_m(\Delta i)$ (see Fig. 5) is the enrichment in the probability of finding a pair of moving residues at sequence separation Δi relative to that expected if moving residues are distributed randomly in sequence space. Similarly, the Cartesian space correlation function $C_m(r)$ (see Fig. 6) is the corresponding probability enrichment for pairs of moving residues at Cartesian distance separation r. The correlation plots use a logarithmic scale to equally offset positive and negative correlations of a given fractional magnitude. To minimize sampling error, we calculate correlation functions from the combined correlation data of all 42 proteins in the reduced allosteric set. However, the observed correlation effects may reflect general correlation features of local motions or structural anomalies of motions in one or a few proteins in each of these small samples. To distinguish these two kinds of correlation effects, we calculate correlation functions from 10 random subsets of 21 of the 42 allosteric proteins for each of the correlation functions and qualitatively compare correlation features among subsets.

In Figure 5, backbone motions of all three types and contact motions are positively correlated at sequence separations of less than 10–20 residues, followed by decreases in log correlation to near or below zero. In Figure 6, these same types of motions are positively correlated at Cartesian separations of less than 10–20 Å, with analogous drops in correlation thereafter. These effects are also present for both correlation functions in all subsets of the reduced allosteric set, and this indicates that they are general correlation features of allosteric proteins. These high local correlations may reflect sequentially or spatially contiguous clusters of motion, although motions can be correlated without being contiguous. For example, φ and ψ changes can compensate each other at nonadjacent sites which bracket a relatively rigid region such as a helix. Side-chain motions are only slightly correlated at short separations in both sequence space and Cartesian space, but these weak features are present for both correlation functions in all subsets of the reduced allosteric set. These weak positive correlations likely reflect weak coupling of moving side chains to nearby side chains. Side-chain motions are nearly uncorrelated at longer ranges in both sequence and Cartesian spaces. Backbone and contact motions appear to display a second effect; that is, correlations of these motions oscillate about zero at long-range sequence and Cartesian distance separations. However, the secondary positive correlations for ΔC_α and contact motions in sequence space are almost absent in 2 and 1 of 10 subsets of the reduced allosteric set, respectively, and the secondary peaks for ΔC_α, $\theta_{\alpha\beta}$, and contact motions in Cartesian space are absent in 1, 2, and 1 of 10 subsets, respectively. Thus, these apparent long-range correlations for some backbone and contact motions reflect peculiar features in several allosteric proteins rather than general correlation features of local motions. These peculiar features in some proteins are

probably between-cluster effects rather than direct correlation through sequence or three-dimensional space. For example, multiple sequentially distinct moving segments may cluster together in Cartesian space, and multiple symmetry-related spatially distinct clusters of motion may exist in some oligomers. In summary, backbone and contact motions, and to a lesser extent side-chain motions, are not randomly distributed in allosteric proteins, but they are more highly correlated than expected at short separations.

Locally correlated motions may exist in proteins with nonallosteric motions as well as in allosteric proteins. Supplementary Figures 2 and 3 show that for nonallosteric proteins with ligand-induced motion, motions are correlated in approximately the same ranges as in allosteric proteins, up to 10–20 residues sequence separation and up to 10–20 Å Cartesian distance separation, respectively. Nonallosteric proteins with ligand-induced motion also exhibit weak correlation of side-chain motions in sequence space, but the positive correlation is within sampling error in Cartesian space.

DISCUSSION

Local Motion Calculations

We measure and analyze three types of local conformational changes which represent different frames of reference important to protein structure. Dihedral angles define the local conformation; thus, dihedral angle changes reflect the bond rotations that underlie allosteric transitions. However, Cartesian coordinates describe the positions of atoms in space, and atoms can move through space without changing local torsion angles due to lever-arm effects of distant torsional changes. Finally, residue–residue contacts represent the network of noncovalent interactions which define the energetics of protein structure; thus, contact changes correspond both to communication among residues and to changes in the local energetics. Each of these metrics is important in describing conformational change, and a conformational change must simultaneously obey constraints in local conformational space, Cartesian space, and noncovalent interaction space.

The automated calculations in this work reveal detailed, functionally interesting pictures of conformational changes in allosteric proteins. The calculated motions corroborate many of the most significant local motion features previously observed by crystallographers, and they might reveal some statistically significant motions missed in analysis by eye. However, the calculations may miss other functionally important features that would only be obvious by manual inspection. For example, the calculations only identify motions which exceed the background motion level estimated from nonallosteric proteins, but smaller motions may be important for function in some proteins, especially for enzymes that are thought to require precise placement of atoms in the active site.

The calculated local motions for 51 allosteric proteins are a key step toward the goal of mechanistic descriptions of allosteric transitions in these proteins. At least two additional kinds of calculations are needed to complete these mechanistic descriptions. First, measurements of large-scale (domain and subunit) motions in multidomain and oligomeric proteins would complete the description of differences between I and A state crystal structures in these proteins. Second, analysis of spatial relationships among tertiary and quaternary structural changes in each protein would produce a network-scale representation of the conformational differences which might sufficiently describe communication between the allosteric and functional sites in that protein.

Toward the Structural Basis of Protein Allostery from Calculated Motions

On average in an allosteric protein, about 15% of residues differ in backbone dihedral conformation between the I and A states while 20 and 18% of residues each differ by backbone position in space and contact structure, respectively. Since not all protein residues which move relative to the core or experience a change in contact structure need to change their φ or ψ angles, these data mean that about 20% of an allosteric protein changes local structure in the allosteric transition. This is substantially higher than the average extent of motion (about 10%) among nonallosteric proteins with significant (5% or more residues) motion upon ligand binding. These results, taken together with the observation that most nonallosteric proteins experience little or no motion upon ligand binding, argue that protein structures have evolved to be robust to changes in environmental conditions and to undergo large local motion only for some specific functional purpose. In addition, these results show that transmitting a ligand-binding signal to a second site requires substantially more rearrangement in a protein structure than does a functional change in the immediate vicinity of a single isolated ligand-binding site. The extent of motion varies significantly among allosteric proteins (by 25–50% of the mean depending on the metric), which provides evidence for significant structural heterogeneity in allosteric mechanisms. Furthermore, while the results clearly demonstrate that allosteric mechanisms in many proteins involve significant changes in average structure, Cooper and Dryden[43] have demonstrated that allostery without a change in average structure is thermodynamically possible, and recent experiments have shown that some allosteric proteins exhibit functionally significant changes in dynamics.[12,20] Changes in dynamics are beyond the scope of our observations and could add a significant additional dimension for these proteins. However, the extent and generality of the observed differences between I and A structures argues that many allosteric proteins may be adequately described by simpler static and mechanical mechanisms.

Highly constrained local structural environments like helices, strands, and buried regions suppress motion relative to weakly constrained environments like loops and solvent-exposed regions in allosteric proteins. This rather intuitive result implies that protein structures use constraints to control the location of flexible regions. It is also reasonable to hypothesize that proteins arrange these constraints to direct signal communication between allosteric and functional sites in a biochemically productive manner. Like allosteric proteins, non-allosteric proteins display biases of motion toward weakly constrained regions, which indicates that the inverse relationship between motion and density of local constraints is a general protein structure phenomenon rather than a unique phenomenon of allosteric proteins. Previously, elastic network models have predicted temperature factors with high accuracy based on native state packing geometry,[44,45] and a constraints-based model has qualitatively predicted flexible regions based on native-state hydrogen bonds, covalent bonds, and contacts.[46] These successful flexibility predictions have offered indirect evidence that structural constraints control flexibility in proteins, and the data provided here about structural environment effects on flexibility provide support for the hypotheses underlying these prediction algorithms.

Backbone and contact motions are positively correlated at distances of up to 10–20 residues in sequence space and of up to 10–20 Å in Cartesian space in allosteric proteins. That is, protein motions can communicate through the primary and tertiary structures, respectively, of proteins over approximately these distances. This Cartesian communication distance is several times the atom-atom contact distance (3–4 Å); that is, it is potentially large enough to bridge two spatially distinct sites over several residues, and it is possible that some proteins communicate perturbations through tertiary structure over longer distances. The kinetic mechanisms by which allosteric proteins communicate between allosteric and functional sites may be complex. Koshland and coworkers[47,48] have suggested that this communication may occur by a series of conformational distortions upon ligand binding. However, recent experiments have shown that dynamic fluctuations along the pathway of allosteric motion can exist in a single state,[18,19] and a recent simulation has suggested nonlinear propagation of the conformational change between sites in CheY.[49] Regardless of the kinetic mechanism, these correlations suggest the hypothesis that mechanically coupled motion is a common contributor to allosteric mechanisms. Correlated motions were also observed in six nonallosteric proteins with 5% or more motion, which implies that correlated motions are a general protein phenomenon. Previous works have offered indirect evidence that protein motions are correlated. Correlated backbone torsional rotations of residues one or two positions apart have been observed in low-resolution Monte Carlo simulations of T4 lysozyme,[50] and hydrogen exchange protection factors have been accurately predicted in several proteins with a cooperative local unfolding model.[51] Spatially contiguous clusters of coevolving residues have been observed in some allosteric proteins,[52,53] and a putative intramolecular signaling pathway was suggested by modeling anisotropic thermal diffusion.[54]

Finally, these analyses show that allosteric proteins and nonallosteric proteins with ligand-induced motion are related but different. Their similarities in distribution of motion among local structural environments and local correlation of motion imply that they are controlled by the same principles of protein structure. However, allosteric proteins clearly constitute a distinct subset of nonallosteric proteins with ligand-induced motion because they average twice as much local motion as the whole set.

Possible Further Applications of the Allosteric Benchmark

The calculated local motions for 51 proteins have possible uses beyond the statistical analyses presented here. For example, just as previous benchmark sets have provided excellent broad tests for protein folding and docking structure predictions,[55–57] this benchmark potentially provides such a resource for protein flexibility prediction methods such as COREX,[51] elastic network models,[44,45] FIRST,[46] and statistical coupling analysis.[52] The six metrics of motion reveal different features of protein motions, which can be compared with the various results of these different flexibility prediction algorithms. In addition, the local motions identified from I and A state crystal structures in these 51 proteins provide candidates for biochemical analyses of allosteric mechanisms like the recent identification of thermodynamically important residues for allosteric inhibition and activation in phosphofructokinase.[58] Finally, the calculated movements between I and A state structures provide possible new targets for structure prediction methods, which have traditionally targeted rigid protein structures[59] and complexes.[60] Such predictions of effector-induced backbone and/or rigid-body motions most closely resemble homology modeling and refinement problems, and recent progress in these two problems[61,62] is encouraging for predicting allosteric transitions.

CONCLUSIONS

Forty years ago, MWC[5] and KNF[6] revolutionized the allostery field by developing gross models of how proteins respond allosterically to ligand binding based on early low-resolution allosteric structures. Here, we undertook a structural analysis of allosteric transitions by calculating several different types of local motion at the residue scale from 51 pairs of high-resolution I and A allosteric structures which have been solved since the MWC and KNF models were published. Statistical analyses of the calculated motions quantified the magnitude of local allosteric effects in proteins and revealed principles of motion consistently exhibited by allosteric proteins. For example, allosteric proteins exhibit local conformational changes comprising approximately 20% of

residues upon introduction or removal of the effector. In addition, the propensity for motion at a site in a protein is inversely related to the density of constraints such as contacts and hydrogen bonds at that site. Motions in proteins are mechanically coupled at distances of tens of angstroms, enough to bridge allosteric and functional sites in many proteins. Furthermore, a comparison of properties of motions in allosteric and nonallosteric proteins with ligand-induced motions revealed that allosteric proteins are controlled by similar principles of protein structure as nonallosteric proteins but that on average, allosteric proteins have evolved to have a significantly higher extent of motion than their nonallosteric counterparts. These properties of allostery may be useful to guide rational manipulation and/or design of allosteric proteins for therapeutic and engineering applications.

METHODS

Culling the Dataset of Allosteric Proteins

We seek proteins with 3.0 Å or better resolution structures in different allosteric states in the Protein Data Bank.[31] We limit the set to proteins induced by small molecules rather than by proteins, nucleic acids, or solution conditions (pH, temperature, salt, etc.). We consider biologically relevant small molecules, peptides of five or fewer residues, and small covalent modifications which are not constitutively part of the protein, such as phosphorylations, to be valid small molecule ligands. Like Hu et al.,[63] we exclude from consideration buffers and other ligands present in the structure only as a result of the crystallization process. In addition, we only consider a protein allosteric if it is known to exhibit a change in function a second (substrate) site distinct from the effector site. That is, the molecules which bind at substrate and effector sites must not interact directly with one another or be involved in the same chemical reaction or other biochemical function. We require that the effector and substrate-binding sites be known and that the effector site be occupied in at least one of the two structures. Finally, the two or more structures of a protein in the set must be unambiguously in different biochemical (I or A) states according to literature articles pertaining to one or both structures.

We first conducted several online keyword searches of the literature for publications pertaining to crystal structures of allosteric and signaling proteins in general and to specific classes of allosteric and signaling proteins such as enzymes, DNA-binding proteins (repressors and activators), G proteins, protein kinases, and bacterial response regulators. Given the large amount of human effort needed in assessing the validity of potential targets for the allosteric set, we did not perform an exhaustive analysis of proteins with more than one structure in the PDB. Initial keyword searches on PubMed resulted in several thousand abstracts, from which we identify about 70 putative allosteric structure pairs. After carefully screening these targets against the above criteria, we reduced the set to 51 proteins (Table II). We excluded nine of these from statistical analysis because they are missing substantial portions of the effector or substrate sites; we refer to the remaining 42 as the reduced allosteric set. Supplementary Table II gives the resolution and identifies the relevant allosteric effectors and substrates for each structure.

Motion Calculation Methods

Superposition algorithm

Calculation of ΔC_α and $\theta_{\alpha\beta}$ requires a superposition of the I and A state coordinates, which often differ substantially in structure. Therefore, we develop a heuristic algorithm to automatically find a core set of residues for a flexible protein from which a physically reasonable superposition can be performed.

We first align the sequences of the two structures with clustalw[64] to identify matching fragments for superposition. We carry out all superpositions using the SVDSuperimposer package of BioPython (www.biopython.org), which calculates the optimal translation vector and rotation matrix of a set of points using a singular value decomposition algorithm.

We define the core fragment of two conformations P_A and P_B of a domain as the subset of residues S_f for which the C_α displacement between P_A and P_B is less than some threshold f. However, there is no analytical solution to find this set unless one knows beforehand the optimal rotation matrix R_{BA} and translation vector T_{BA} for superimposing P_B onto P_A. Previous algorithms which have targeted this problem, including those of Wriggers and Schulten[65] and Damm and Carlson,[66] have used iterative procedures. Thus, we employ a two-stage heuristic algorithm that will find S_f for most protein structure pairs using $f = 1.0$ Å. The first stage selects an initial set of locally rigid residues $S_{f,0}$, that is, those residues which superimpose locally with an RMSD of 0.5 Å or better in nonoverlapping segments of seven residues. The second stage iteratively optimizes S_f as the convergence of the sequence $S_{f,i}$ starting from $S_{f,0}$. Each iteration i finds $R_{BA,i}$ and $T_{BA,i}$, the optimal rotation and translation, respectively, for superimposing the residues in set $S_{f,i-1}$. It then transforms P_B by $R_{BA,i}$ and $T_{BA,i}$ and assigns those residues for which the C_α displacement between P_A and P_B is less than f to a new core set $S_{f,i}$. The algorithm repeats this process, updating $R_{BA,i}$, $T_{BA,i}$, and $S_{f,i}$ at each step until $S_{f,i} = S_{f,i-1}$. This algorithm resembles the Gaussian-weighted superposition algorithm of Damm and Carlson,[66] except we exclude flexible residues from the superposition while Damm and Carlson down-weight such residues. The algorithm consistently and accurately superimposes high-resolution structures of flexible proteins using a substantial portion of the residues in the protein. For the 66 domains of the 51 allosteric proteins in the benchmark, the algorithm determines the core after an average of 4.3 iterations, includes an average of 78% of the residues in the core, and produces an average core RMSD of 0.47 Å and a maximum core RMSD of 0.62 Å.

TABLE II. A Benchmark of 51 Allosteric Proteins

Protein	Inactive	Active	Protein	Inactive	Active
A: Signaling proteins (25)			B: DNA-binding proteins (8)		
A1: G proteins			AraC[b]	2ARA	2ARC
arf1	1HUR	NA	arg repressor	1XXC	1XXA
arf6[a]	1E0S	1HFV	biotin repressor[b]	1BIA	1HXD
cdc42	1AN0	1NF3(A)	lac repressor[a]	1TLF	1EFA
EF-Tu[a,c]	1TUI	1EFT	met repressor	1CMB	1CMA
rab11	1OIV	1OIW	OxyR[a,b]	1I69	1I6A
rab7[b]	1VG1	1VG8	PurR	1DBQ	1WET
rac1	1HH4(A)	1MH1	tet repressor[a]	2TRT	1QPI
ran[a]	NA	1IBR(A)			
rap2a	1KAO	2RAP	C: Enzymes (18)		
ras[a]	4Q21	6Q21	anthranilate synthase	1I7S	1I7Q
rheb	1XTQ	1XTS	ATCase[a]	1RAC	1D09
rhoA	1FTN	1A2B	ATP sulfurylase[a]	1M8P	1I2D
sec4	1G16	1G17	ATP-PRT	1NH8	1NH7
ypt7p[b]	1KY3	1KY2	caspase[b]	1SHJ	1F1J
YsxC[b]	1SVI	1SVW	chorismate mutase	2CSM	1CSM
$G_{i\alpha 1}$[a,b]	1GDD	1GIA	DAHP synthase	1KFL	1N8F
$G_{t\alpha}$	1TAG	1TND	FBPase-1[a]	1EYJ	1EYI
A2: Protein kinases			glcN-6-P deaminase	1CD5	1HOT
ERK2	1ERK	2ERK	glycogen phosphorylase[a]	1GPB	7GPB
IGF-1R	1P4O	1K3A	GTP cyclohydrolase I	1WPL	1IS7
IRK	1IRK	1IR3	hemoglobin[a]	4HHB	1HHO
PKB[b]	1GZK	1O6K	lactate DH[a]	1LTH(T)	1LTH(R)
A3: response regulators			NAD-malic enzyme	1QR6	1PJ2
CheY[a]	3CHY	1FQW	phosphofructokinase[a]	6PFK	4PFK
DctD	1L5Z	1L5Y	phosphoglycerate DH	1PSD	1YBA
fixJ[a]	1DBW	1D5W	PTP1B	1T48	1PTY
SpoIIAA	1H4Y	1H4X	uracil PRT	1XTU	1XTT

NA: Not available in PDB; coordinates obtained directly from authors.
[a]Protein is also included in the Database of Molecular Movements. In some cases, we use different structure pairs than the Database of Molecular Movements.
[b]One or both structures is missing significant functional regions. Local motion data has been calculated for these nine proteins, but that data is not included in the statistical analyses of motions.
[c]We analyze only the GTP binding domain (residues 1–211).

Definition of f_I and f_A

If $I_{6,i}$ is the set of atoms within 6.0 Å of a residue i in the I state and $A_{6,i}$ is the corresponding set for residue i in the A state (see Fig. 1 for an example), we calculate the fractional change in each residue's environment, f_I and f_A, by

$$f_\mathrm{I} = \frac{N_\mathrm{I} - M_\mathrm{IA}}{N_\mathrm{I}} \quad \text{and} \quad f_\mathrm{A} = \frac{N_\mathrm{A} - M_\mathrm{IA}}{N_\mathrm{A}},$$

where N_I and N_A are the numbers of atoms in $I_{6,i}$ and $A_{6,i}$, respectively, and M_IA is the number of atoms in both $I_{6,i}$ and $A_{6,i}$. Protein atoms but not ligand atoms are included in $I_{6,i}$ and $A_{6,i}$; that is, f_I and f_A capture internal rearrangements in the protein structure but not immediate effects of the introduction or removal of the effector.

Multidomain and multimeric proteins

For multimeric proteins, we perform calculations on the biological unit coordinates rather than on the asymmetric unit coordinates. For proteins with multiple rigid domains which move relative to one another, we assign domain boundaries visually based on a superposition of the entire protein (Supplementary Table III), and we superimpose rigid domains separately for the determination of ΔC_α and $\theta_{\alpha\beta}$. For multimeric proteins, we determine and report the values of $\max(|\Delta\varphi|, |\Delta\psi|)$, ΔC_α, $\theta_\alpha\beta$, $\max(|\Delta\chi_1|, |\Delta\chi_2|)$, and ΔSC for the residues of the first copy of each subunit type. We calculate $\max(f_\mathrm{I}, f_\mathrm{A})$ based on all atoms in the oligomer, but we report f_I and f_A only for the first copy of each subunit type.

Determination of Motion Thresholds

Supplementary Table IV shows the PDB codes for structures in the Control 1 and 2 sets, which are references for intrinsic flexibility in proteins in general and dynamic fluctuations in allosteric proteins, respectively. We apply the calculations shown in Figure 1 to all residues in all subunit types of all proteins in the Control 1 and 2 sets and the reduced allosteric set (Table II), respectively. Figure 7 shows the resultant distributions of residues with various motion values. In Control 1 and 2 sets and the reduced allosteric set, most residues have small motion values for each metric corresponding to

TABLE III. Thresholds for Local Motions

		99th percentile			
Metric	Statistic	Control 1	Control 2	Average	Threshold
max(|Δφ|, |Δψ|)	|Δφ| (°)[a]	27.1	32.7	29.9	30°
	|Δψ| (°)[a]	30.5	30.3	30.4	
ΔC_α	ΔC_α (Å)	1.38	1.01	1.20	1.2 Å
$\theta_{\alpha\beta}$	$\theta_{\alpha\beta}$ (°)[a,b]	23.9	31.8	27.8	28°
max(|$\Delta\chi_1$|, |$\Delta\chi_2$|)	|$\Delta\chi_1$| (°)[a,c]	42.8	49.5	46.1	46°
ΔSC	ΔSC (Å)[d]	4.12	3.69	3.91	2.0 Å
max(f_I, f_A)	f_I	0.213	0.190	0.202	0.20
	f_A	0.200	0.190	0.195	

The 99th percentiles of each metric are calculated from all residues of all subunit types of all proteins in the control 1 and 2 datasets, respectively.
[a]For dihedral angle motion parameters (\|Δφ\|, \|Δψ\|, $\theta_{\alpha\beta}$, and \|$\Delta\chi_1$\|), residues which change by more than ±60° are excluded from the percentile calculation.
[b]For $\theta_{\alpha\beta}$, gly residues are excluded from the percentile calculation.
[c]For \|$\Delta\chi_1$\|, gly, ala, and pro residues are excluded from the percentile calculation.
[d]For ΔSC, we set the threshold at 2.0 Å because side-chain rotameric shifts cannot be readily excluded.

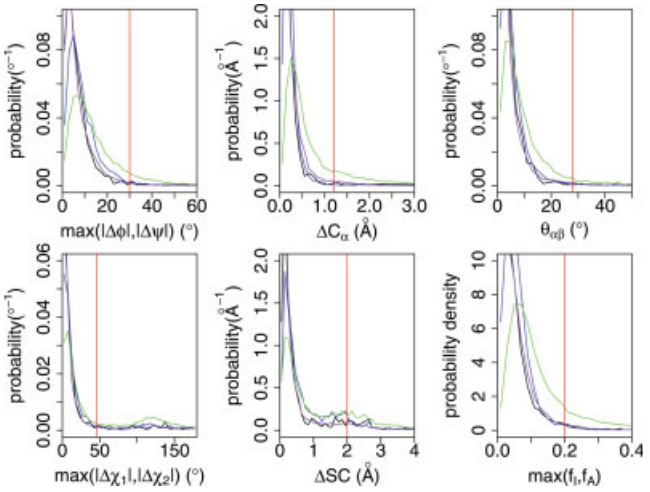

Fig. 7. Motion value distributions for local motion metrics. Histograms of motion values include all residues in all subunit types of all proteins in four respective datasets: Control 1 (non-allosteric, black), Control 2 (allosteric, blue), Control 3 (ligand-binding non-allosteric, purple) and the reduced allosteric dataset (green), with the threshold (Table III) shown as a red line. Residues for which motion is undefined by a given metric (e.g. glycine for $\theta_{\alpha\beta}$) are excluded from the distributions for that metric.

intrinsic flexibility and crystallographic error, and the probability density decays in bins of higher motion values. However, for backbone and contact motions, the reduced allosteric set distributions are shifted notably toward higher motion values relative to the corresponding distributions for the Control 1 and 2 sets. The side-chain motion value distributions are similar between the allosteric and control sets, with the exception that allosteric proteins exhibit more rotameric changes (near 120° in the max(\|$\Delta\chi_1$\|, \|$\Delta\chi_2$\|) graph) than control proteins.

Using Control 1 and 2 distributions as a guide, we identify a threshold for each metric (Table III) which excludes all but the most statistically significant motions, even though smaller motions may be functionally significant in some allosteric proteins. For max(\|Δφ\|, \|Δψ\|), and max(f_I, f_A), we calculate the threshold for each parameter individually (e.g. \|Δφ\| and \|Δψ\|) and then average those thresholds. For max(\|$\Delta\chi_1$\|, \|$\Delta\chi_2$\|), we use the threshold for \|$\Delta\chi_1$\| for both χ angles. We define the threshold for a metric as the average of the 99th percentiles of the background motion distributions for the two respective control sets. For ΔC_α and max(f_I, f_A), we consider the entire motion value distributions of the control sets to be background motion, and the 99th percentiles yield thresholds of 1.2 and 0.2 Å, respectively. The ΔC_α threshold is consistent with the crystallographic uncertainty identified by Eyal et al.[34] (<1 Å RMSD for independently solved crystal structures). For dihedral angle measurements max(\|Δφ\|, \|Δψ\|), $\theta_{\alpha\beta}$, and max(\|$\Delta\chi_1$\|, \|$\Delta\chi_2$\|), we only consider changes of less than ±60° to be background motion because larger changes represent rotameric shifts, which results in thresholds of 30°, 28°, and 46° for max(\|Δφ\|, \|Δψ\|), $\theta_{\alpha\beta}$, and max(\|$\Delta\chi_1$\|, \|$\Delta\chi_2$\|), respectively. The χ_1 cutoff is on the same order as the 90% χ_1 confidence levels calculated by Zhao et al.[67] (7–74° depending on the side-chain). However, it is difficult to compare these numbers because of the mathematical differences between our measure and theirs and because Zhao et al. use a lower resolution cutoff (2.2 Å) for their dataset than we do for our datasets (3.0 Å). For ΔSC, background motion and rotameric shifts are difficult to separate because the data include many types of side chains with different lengths and numbers of rotatable bonds. However, because Figure 7 reveals that at about 2.0 Å, the probability density for the allosteric distribution begins to significantly exceed the density for the control distributions, we set the ΔSC threshold to 2.0 Å.

Ligand-Binding Nonallosteric Set

As a reference for motion in nonallosteric proteins, we collect a set of 21 pairs of structures in different liganded states (Supplementary Table V). We collect three from keyword searches for structures of proteins known to bind only a single ligand. We cull the other

18 targets directly from a set of 14,000 95% sequence identical (by BLAST) families of protein chains from the PDB as of June 8, 2006. From 5900 of these families containing chains from at least two different structures, we remove 1100 which were reduced to less than two members after eliminating non-X-ray structures and X-ray structures of poorer than 3.0 Å resolution. We then remove 1400 more families for which all structures possess identical sets of ligands, excluding ligands known to be crystallization reagents. From the remaining 3400, we randomly select 40 putative ligand-binding nonallosteric targets and then narrow this set to 18 families with biochemically meaningful differences among liganded states according to the literature articles for one or both structures. Motion distributions of the Control 3 set are shown in Figure 7 for reference but are not used to set thresholds for motion metrics.

Correlation Functions

Correlation in sequence space

The correlation function captures the enrichment in the probability of finding a pair of moving residues at sequence separation Δi over that expected if moving residues are distributed uniformly in sequence space. It is calculated by the number of moving pairs at separation Δi for all polypeptide chains in the reduced allosteric set divided by the total number of moving pairs expected at separation Δi:

$$C_m(\Delta i) = \frac{\sum_c \sum_{i=1}^{L_c-\Delta i} \delta_i^m \delta_{i+\Delta i}^m}{\sum_c (L_c - \Delta i) f_{m,c}^2},$$

where $C_m(\Delta i)$ is the correlation for metric m, the first sum in the numerator is over all subunit types (unique chains) c in the dataset, the second sum in the numerator is over all $L_c - \Delta i$ pairs of residues separated by Δi in chain c, L_c is the number of residues in polypeptide chain c, δ_j^m is 1 if residue j moves by metric m and 0 otherwise, and $f_{m,c}$ is the fraction of moving residues for metric m in chain c.

For each Δi bin, the sampling error $\varepsilon(\Delta i)$ can be estimated as for random sampling from a heterogeneous set, $\varepsilon = \sqrt{n}$, where n is the total number of pairs of moving residues sampled, in this case $\sum_c \sum_{i=1}^{L_c-\Delta i} \delta_i^m \delta_{i+\Delta i}^m$.

Correlation in three-dimensional space

The Cartesian correlation function captures the enrichment in the probability of finding a pair of moving residues at a separation of Cartesian distance r over that expected if moving residues are distributed randomly in Cartesian space within the protein. To smooth these functions over the finite data set, the metrics are evaluated within discrete distance bins defined by $r \in \left(k + \frac{1}{2}\right)$, where k is a nonnegative integer and b is a bin size, for which we use 1.0 Å. The function δ_{ij}^r is 1 if $r - \frac{b}{2} \le d_{ij} < r + \frac{b}{2}$ and 0 otherwise, where d_{ij} is distance between the all-atom centroids of residues i and j, as an indicator function of whether a pair of residues are within a particular distance bin. Since d_{ij} may differ between I and A states, each pair of residues contributes one-half count to the bin corresponding to d_{ij}^{I} and one-half count to the bin corresponding to d_{ij}^{A}. The correlation $C_m(r)$ is then calculated by the number of moving pairs separated by r in all proteins in the dataset divided by the total number of pairs expected to be moving and separated by r,

$$C_m(r) = \frac{\sum_p N_p^m(r)}{\sum_p N_p(r) f_p^m},$$

where the sums are over all proteins p, $N_p(r)$ is the total number of pairs separated by $r - \frac{b}{2}$ in protein p,

$$N_p(r) = \frac{1}{\omega_p} \sum_{j>i} \delta_{ij}^r,$$

and $N_p^m(r)$ is the corresponding number of pairs of residues for which both move by metric m,

$$N_p^m(r) = \frac{1}{\omega_p} \sum_{j>i} \delta_{ij}^r \delta_i^m \delta_j^m.$$

The sums are over all pairs of residues i and j in the protein, ω_p is the number of asymmetric units in the protein, which is the total number of chains in the protein (N_p^c) divided by the number of subunit types in the protein (N_p^{uc}), $\omega_p = N_p^c / N_p^{uc}$. The normalization of $N_p(r)$ and $N_p^m(r)$ by ω_p ensures that symmetry-related pairs of residues in an oligomer are only counted once. f_p^m is the fraction of all pairs of residues in protein p for which both residues move by metric m,

$$f_p^m = \frac{\sum_r N_p^m(r)}{\sum_r N_p(r)}.$$

As in the sequence correlation calculation, sampling error can be estimated from the number of pairs of moving residues counted n as $\varepsilon(r) = \sqrt{n} = \left(\sum_p N_p^m(r)\right)^{1/2}$.

ACKNOWLEDGMENTS

MDD was supported by an Achievement Rewards for College Scientists (ARCS) Foundation fellowship. We thank Robert Schleif and Maria Zavodszky for reading the manuscript and for giving helpful advice regarding the structure of this study and J. Goldberg and A. Wittinghofer for sharing protein structures not available in the PDB.

REFERENCES

1. Berg JM, Tymoczko JL, Stryer L. Biochemistry. New York: W. H. Freeman; 2002.

2. Mirsky AE, Pauling L. On the structure of native, denatured, and coagulated proteins. Proc Natl Acad Sci USA 1936;22:439–447.
3. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
4. Monod J, Changeux JP, Jacob F. Allosteric proteins and cellular control systems. J Mol Biol 1963;6:306–329.
5. Monod J, Wyman J, Changeux JP. On the nature of allosteric transitions: a plausible model. J Mol Biol 1965;12:88–118.
6. Koshland DE, Jr, Nemethy G, Filmer D. Comparison of experimental binding data and theoretical models in proteins containing subunits. Biochemistry 1966;5:365–385.
7. Jardetzky O. Protein dynamics and conformational transitions in allosteric proteins. Prog Biophys Mol Biol 1996;65:171–219.
8. Vetter IR, Wittinghofer A. The guanine nucleotide-binding switch in three dimensions. Science 2001;294:1299–1304.
9. Huse M, Kuriyan J. The conformational plasticity of protein kinases. Cell 2002;109:275–282.
10. Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R. Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci 2000;9:10–19.
11. Luque I, Leavitt SA, Freire E. The linkage between protein folding and functional cooperativity: two sides of the same coin? Annu Rev Biophys Biomol Struct 2002;31:235–256.
12. Gunasekaran K, Ma B, Nussinov R. Is allostery an intrinsic property of all dynamic proteins? Proteins 2004;57:433–443.
13. Lambright DG, Noel JP, Hamm HE, Sigler PB. Structural determinants for activation of the α-subunit of a heterotrimeric G protein. Nature 1994;369:621–628.
14. Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brennan RG, Lu P. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science 1996;271:1247–1254.
15. Lipscomb WN. Aspartate transcarbamylase from *Escherichia coli*: activity and regulation. Adv Enzymol Relat Areas Mol Biol 1994;68:67–151.
16. Barford D, Hu SH, Johnson LN. Structural mechanism for glycogen phosphorylase control by phosphorylation and AMP. J Mol Biol 1991;218:233–260.
17. Perutz MF. Stereochemistry of cooperative effects in haemoglobin. Nature 1970;228:726–739.
18. Volkman BF, Lipson D, Wemmer DE, Kern D. Two-state allosteric behavior in a single-domain signaling protein. Science 2001; 291:2429–2433.
19. Lukin JA, Kontaxis G, Simplaceanu V, Yuan Y, Bax A, Ho C. Quaternary structure of hemoglobin in solution. Proc Natl Acad Sci USA 2003;100:517–520.
20. Kern D, Zuiderweg ER. The role of dynamics in allosteric regulation. Curr Opin Struct Biol 2003;13:748–757.
21. Chothia C. Structural invariants in protein folding. Nature 1975; 254: 304–308.
22. Rooman MJ, Rodriguez J, Wodak SJ. Relations between protein sequence and structure and their significance. J Mol Biol 1990; 213:337–350.
23. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
24. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
25. Najmanovich R, Kuttner J, Sobolev V, Edelman M. Side-chain flexibility in proteins upon ligand binding. Proteins 2000;39: 261–268.
26. Betts MJ, Sternberg MJ. An analysis of conformational changes on protein–protein association: implications for predictive docking. Protein Eng 1999;12:271–283.
27. Gerstein M, Lesk AM, Chothia C. Structural mechanisms for domain movements in proteins. Biochemistry 1994;33:6739–6749.
28. Boutonnet NS, Rooman MJ, Wodak SJ. Automatic analysis of protein conformational changes by multiple linkage clustering. J Mol Biol 1995;253:633–647.
29. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res 1998;26:4280–4290.
30. Gerstein M, Echols N. Exploring the range of protein flexibility, from a structural proteomics perspective. Curr Opin Chem Biol 2004;8:14–19.
31. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The protein data bank. Acta Crystallogr D Biol Crystallogr 2002;58(Pt. 6):899–907.
32. Echols N, Milburn D, Gerstein M. MolMovDB: analysis and visualization of conformational change and structural flexibility. Nucleic Acids Res 2003;31:478–482.
33. Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, Wells S, Yu EZ, Thorpe M, Gerstein M. The database of macromolecular motions: new features added at the decade mark. Nucleic Acids Res 2006;34(Database issue):D296–D301.
34. Eyal E, Gerzon S, Potapov V, Edelman M, Sobolev V. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J Mol Biol 2005;351:431–442.
35. Delano WL. The PyMOL Molecular Graphics System (www.pymol.org). 2002.
36. Milburn MV, Tong L, deVos AM, Brunger A, Yamaizumi Z, Nishimura S, Kim SH. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. Science 1990;247:939–945.
37. Nassar N, Horn G, Herrmann C, Scherer A, McCormick F, Wittinghofer A. The 2.2 Å crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue. Nature 1995;375:554–560.
38. Schumacher MA, Choi KY, Lu F, Zalkin H, Brennan RG. Mechanism of corepressor-mediated specific DNA binding by the purine repressor. Cell 1995;83:147–155.
39. Schirmer T, Evans PR. Structural basis of the allosteric behaviour of phosphofructokinase. Nature 1990;343:140–145.
40. Zhang Z, Sugio S, Komives EA, Liu KD, Knowles JR, Petsko GA, Ringe D. Crystal structure of recombinant chicken triosephosphate isomerase-phosphoglycolohydroxamate complex at 1.8-Å resolution. Biochemistry 1994;33:2830–2837.
41. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
42. Hubbard SJ, Thornton JM. NACCESS. London: Department of Biochemistry and Molecular Biology, University College; 1993.
43. Cooper A, Dryden DT. Allostery without conformational change. A plausible model. Eur Biophys J 1984;11:103–109.
44. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 1997;2:173–181.
45. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 2001;80:505–515.
46. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. Protein flexibility predictions using graph theory. Proteins 2001;44:150–165.
47. Koshland DE. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 1958;44:98–104.
48. Yu EW, Koshland DE, Jr. Propagating conformational changes over long (and short) distances in proteins. Proc Natl Acad Sci USA 2001;98:9517–9520.
49. Formaneck MS, Ma L, Cui Q. Reconciling the "old" and "new" views of protein allostery: a molecular simulation study of chemotaxis Y protein (CheY). Proteins 2006;63:846–867.
50. Bahar I, Erman B, Haliloglu T, Jernigan RL. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. Biochemistry 1997;36:13512–13523.
51. Hilser VJ, Freire E. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. J Mol Biol 1996;262:756–772.
52. Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 2003;10:59–69.
53. Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R. Allosteric determinants in guanine nucleotide-binding proteins. Proc Natl Acad Sci USA 2003;100:14445–14450.
54. Ota N, Agard DA. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. J Mol Biol 2005;351:345–354.
55. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of a comprehensive benchmark. Pac Symp Biocomput 1996:300–318.

56. Chen R, Mintseris J, Janin J, Weng Z. A protein–protein docking benchmark. Proteins 2003;52:88–91.
57. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. Protein–protein docking benchmark 2.0: an update. Proteins 2005;60:214–216.
58. Fenton AW, Paricharttanakul NM, Reinhart GD. Identification of substrate contact residues important for the allosteric regulation of phosphofructokinase from *Eschericia coli*. Biochemistry 2003;42:6453–6459.
59. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
60. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331:281–299.
61. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2006; 103:5361–5366.
62. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
63. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (mother of all databases). Proteins 2005;60:333–340.
64. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the clustal series of programs. Nucleic Acids Res 2003;31:3497–3500.
65. Wriggers W, Schulten K. Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. Proteins 1997;29:1–14.
66. Damm KL, Carlson HA. Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. Biophys J 2006;90(12):4558–4573.
67. Zhao S, Goodsell DS, Olson AJ. Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation. Proteins 2001;43:271–279.