PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

On the accuracy of inferring energetic coupling between distant sites in protein families from evolutionary imprints: Illustrations using lattice model

Zhenxing Liu,[1] Jie Chen,[1] and D. Thirumalai[1,2]*

[1] Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742
[2] Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742

ABSTRACT

It is suspected that correlated motions among a subset of spatially separated residues drive conformational dynamics not only in multidomain but also in single domain proteins. Sequence and structure-based methods have been proposed to determine covariation between two sites on a protein. The statistical coupling analysis (SCA) that compares the changes in probability at two sites in a multiple sequence alignment (MSA) and a subset of the MSA has been used to infer the network of residues that encodes allosteric signals in protein families. The structural perturbation method (SPM), that probes the response of a local perturbation at all other sites, has been used to probe the allostery wiring diagram in biological machines and enzymes. To assess the efficacy of the SCA, we used an exactly soluble two dimensional lattice model and performed double-mutant cycle (DMC) calculations to predict the extent of physical coupling between two sites. The predictions of the SCA and the DMC results show that only residues that are in contact in the native state are accurately identified. In addition, covariations among strongly interacting residues are most easily identified by the SCA. These conclusions are consistent with the DMC experiments on the PDZ family. Good correlation between the SCA and the DMC is only obtained by performing multiple experiments that vary the nature of amino acids at a given site. In contrast, the energetic coupling found in experiments for the PDZ domain are recovered using the SPM. We also predict, using the SPM, several residues that are coupled energetically.

Proteins 2009; 77:823–831.

Key words: energetic coupling; statistical coupling analysis; double mutant cycles; PDZ domain; lattice model; exact enumeration; structural perturbation method.

INTRODUCTION

It is suspected that in a given protein family, a network of residues is involved in signal transmission on ligand binding that drives the functionally relevant conformational dynamics.[1–4] The residues in the network, which are responsible for allosteric movements, seem to be strongly conserved.[5] Typically allosteric transitions are associated with multidomain proteins.[2] However, in recent years, it has been argued that functional dynamics in single domain proteins are also driven by coupling between multiple residues that are spatially separated.[6] Given the ubiquitous nature of allosteric control in a number of biological processes (movements of motors on polar tracks,[7] domain movements in molecular chaperones,[8] and enzyme catalysis[9]), it is important to determine the key residues that are involved in the signalling pathways. The standard method for experimentally ascertaining correlation between distant sites in a protein is through the use of double-mutant cycle (DMC).[10,11] Comparison of the free energy changes in the protein with simultaneous mutations at two sites m_1 and m_2 and two single site mutation at m_1 and m_2 done separately allows us to infer if sites m_1 and m_2 are correlated. Recently, we introduced a structure-based structural perturbation method (SPM) in which response to a local perturbation (or mutation) to residue m_1 at all other residues is monitored.[5,12] The network of residues with the large responses is energetically coupled to m_1.

It is desirable to devise sequence-based methods because a larger database of evolved protein with related function can be simultaneously analyzed.[13–16] Recently, a novel sequence-based method, the statistical coupling analysis (SCA), was introduced to obtain the network of energetically coupled residues in protein families.[17–19] All of the sequence-based methods, including the SCA, are heuristic without a firm theoretical basis or sound physical arguments. It should be stressed that sequence analysis is crucial in understanding

Grant sponsor: National Science Foundation; Grant number: CHE 05-14056; Grant sponsor: Air-Force office of Scientific Research
*Correspondence to: D. Thirumalai, Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742. E-mail: thirum@umd.edu.
Received 28 January 2009; Revised 28 April 2009; Accepted 25 May 2009
Published online 17 June 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22498

Table I
Interaction Matrix Elements $U_{i,j}$ [Eq. (1)] Between the Four Types of Residues

	H	P	A	B
H	-4ϵ	-2ϵ	$-\epsilon$	$-\epsilon$
P	-2ϵ	-3ϵ	-2ϵ	-2ϵ
A	$-\epsilon$	-2ϵ	0	-5ϵ
B	$-\epsilon$	-2ϵ	-5ϵ	0

the link between function of proteins and their evolution. However, it is unclear whether heuristic methods like SCA can rigorously be used to infer physical coupling between distant sites in a protein. Indeed, despite several studies that have given credence to the SCA,[16,17,20] others have cast doubt on its efficacy in obtaining physically meaningful covariation in residues.[14,21,22] A basic difficulty is that SCA attempts to glean meaningful coupling using evolutionary signals whereas structure-based methods, including experiments,[11] probe consequences of free energy changes using only a discrete set of mutations. Consequently, it is not easy to establish the extent to which the SCA provides insight into long-distance coupling in single domain proteins.

In this article, we use lattice models for which exact calculations can be performed to compare the predictions of the SCA and the DMC. We find, in accord with recent experiments on the PDZ family,[21] that there is essentially no correlation between the free energies obtained using the DMC and the energetic coupling ascertained from the SCA. In contrast, if the DMC free energies are averaged over multiple mutations at specific sites, then the resulting correlation improves significantly.

METHODS

Model

To provide a theoretical understanding of the SCA, we used a two dimensional lattice model representation of polypeptide chains with N, the number of residues, $N = 12$. Because our purpose is to investigate the efficacy of the SCA in obtaining the energetic coupling between residues that are nonbonded (i.e., linked by noncovalent interactions), we used a small value for N for which exact calculations can be performed. In the model, there are four types of resides, H (hydrophobic), P (polar), A (positively charged), and B (negatively charged). The energy of a given conformation, specified in terms of the coordinates $\boldsymbol{r_i}$ ($i = 1, 2, 3, \ldots, N$), is

$$E(\Gamma) = \sum_i \sum_{j>i+1} U_{i,j}\delta(r_{ij} - a) \qquad (1)$$

where $\delta(r_{ij} - a)$ is the Kronecker delta function, a is the lattice spacing, and the interaction matrix elements $U_{i,j}$ depend on residue type. The matrix elements $U_{i,j}$, which are taken from Refs. 23 and 24, are listed in Table I.

To classify families in the lattice model, we only considered, among the 4^{12} sequences, the ones (7, 316, 794 in all) with nondegenerate native (the lowest energy) states. Sequences that have identical native structures are further classified into 354 groups. The groups are analogous to protein families. A representative structure, shown in Figure 1, is the native state for 303,036 sequences. We further reduced the number of sequences to 48,014 based on the criterion that the sequences are stable at the simulation temperature (T_s).

For all sequences, we computed the stability of the native state, ΔG, using

$$\Delta G = -k_B T_s \ln\left(\frac{P_N}{1 - P_N}\right) \qquad (2)$$

where $k_B T_s = \epsilon$, k_B is the Boltzmann constant, $P_N = \frac{e^{-E_N/k_B T}}{Z}$ is the probability that the model protein is in the native state with energy E_N, and Z is the partition function. We calculated $Z = \sum_j e^{-E_j/k_B T_s}$ by summing over all the microstates of the $N = 12$ chain. The total number of microstates is 15,037. The free energies for all sequences are calculated using exact enumeration.

Computational DMC

Coupling between targeted pairs of sites is often inferred using the thermodynamic DMC analysis. For the lattice model, computational DMC can be used to deter-

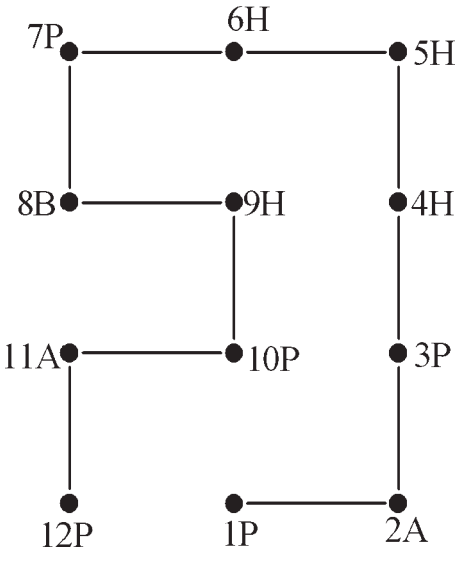

Figure 1
The native structure for the family of 48,014 well-evolved sequences that are stable at the simulation temperature $k_B T_s = \epsilon$. One of the sequences is explicitly shown.

mine the strength of coupling between all pairs of sites.[25] Let ΔG_{WT}, ΔG_{m_1}, and ΔG_{m_2} be the free energies of the folded states of the wild-type (WT) protein and proteins with mutations at m_1, m_2, respectively, and $\Delta G_{m_1,m_2}$ be the free energy for the double mutant. The coupling energy between sites m_1 and m_2[17] is,

$$\Delta\Delta\Delta G_{m_1,m_2} = \Delta G_{m_1,m_2} + \Delta G_{WT} - \Delta G_{m_1} - \Delta G_{m_2}. \quad (3)$$

If ΔG_{WT} is taken as the reference free energy, the coupling energy can be rewritten using,

$$\Delta\Delta\Delta G_{m_1,m_2} = \Delta\Delta G_{m_1,m_2} - \Delta\Delta G_{m_1} - \Delta\Delta G_{m_2} \quad (4)$$

where $\Delta\Delta G_{m_1,m_2} = \Delta G_{m_1,m_2} - \Delta G_{WT}$, $\Delta\Delta G_{m_1} = \Delta G_{m_1} - \Delta G_{WT}$, $\Delta\Delta G_{m_2} = \Delta G_{m_2} - \Delta G_{WT}$. Such a form is identical to the one used in the DMC experiments.[11] Coupling energies $\Delta\Delta\Delta G_{m_1,m_2}$ could be positive or negative or zero (no correlation between sites m_1 and m_2). Here, we use absolute values of the coupling energies $|\Delta\Delta\Delta G_{m_1,m_2}|$ to compare with the predictions from the SCA since coupling energy in the SCA formalism is always positive [see Eqs. (5) and (6) later]. We performed the DMC analysis on the family shown in Figure 1. As there are only four types of residues in the model and only mutations which do not alter the native state structure are accepted, the number of acceptable mutations for a given sequence is relatively small. Therefore, to obtain statistically meaningful results, we use randomly picked 1000 sequences from the family with 48,014 stable and well-evolved sequences. The model chain has 12 monomers with $\left(\frac{12\times11}{2} - 11 = 55\right)$ pairs of sites (the neighbor sites along the chain are excluded), and the DMC is implemented for each of the 55 pairs of sites.

Two versions of the SCA

The sequence-based SCA, introduced by Lockless and Ranganathan,[17] is used to measure the statistical interactions between amino acid positions. We use the statistical coupling energies to ascertain if they correlate well with the physical coupling energies from DMC [Eq. (3)]. The energetic couplings within the SCA are computed using the equation introduced in Ref. 17,

$$\Delta\Delta G^{LR}_{i,j} = k_B T \sqrt{\sum_{x=1}^{4}\left(\ln\frac{P^x_{i|\delta j}}{P^x_{MSA|\delta j}} - ln\frac{P^x_i}{P^x_{MSA}}\right)^2}, \quad (5)$$

and the one introduced by Dima and Thirumalai,[26]

$$\Delta\Delta G^{DT}_{i,j} = k_B T \sqrt{\frac{1}{c_i}\sum_{x=1}^{4}\left(P^x_{i|\delta j}\ln\frac{P^x_{i|\delta j}}{P^x_{MSA}} - P^x_i\ln\frac{P^x_i}{P^x_{MSA}}\right)^2}, \quad (6)$$

where k_BT is an arbitrary energy unit, c_i is the number of types of amino acids that appear at position i, P^x_{MSA} is the mean frequency of amino acid type x in the multiple sequence alignment (MSA). In Eqs. (5) and (6), $P^x_i = n^x_i/N_i$, where n^x_i is the number of times amino acid type x that appears at i in the MSA, and $N_i = \sum_{x=1}^{4} n^x_i$, $P^x_{i|\delta j} = n^x_{i|\delta j}/N_{i|\delta j}$, $n^x_{i|\delta j}$ is the number of sequences in the subalignment in which x appears in the ith position, and $N_{i|\delta j} = \sum_{x=1}^{4} n^x_{i|\delta j}$. Note that the computation of P^x_i differs from the procedure used in Ref. 17.

Let $f = P/N_{MSA}$, where P is the number of sequences in the subalignment and N_{MSA} is the total number of sequences in the MSA. We choose $f(=0.25$ for the 48,014 lattice family) to satisfy the central limit theorem[26] so that the statistical properties from the subalignments coincide with the full MSA. Using $f = 0.25$, we calculated the matrix elements $\Delta\Delta G^{LR}_{i,j}$ and $\Delta\Delta G^{DT}_{i,j}$ that estimate the response of position i in the MSA to all allowed perturbations at j (where a given amino acid is fully conserved).

Letting the number of allowed perturbation types at position j be n_j, we performed an average over all perturbations to obtain $\langle\Delta\Delta G_{i,j}\rangle_\tau = \frac{\sum_{t=1}^{n_j}\Delta\Delta G^t_{i,j}}{n_j}$. The matrix $\langle\Delta\Delta G_{i,j}\rangle_\tau$ is asymmetric while DMC results yield symmetrical values. To compare the DMC and the SCA results, we used $\langle\Delta\Delta G_{i,j}\rangle_{SCA} = \frac{\langle\Delta\Delta G_{i,j}\rangle_\tau + \langle\Delta\Delta G_{j,i}\rangle_\tau}{2}$, where τ=LR or DT.

Structural perturbation method (SPM)

To rationalize the DMC experiments on the PDZ domain, we use a structure-based method to determine the network of residues that are energetically coupled. We introduced the SPM to map out the allostery wiring diagram for biological machines. The SPM, which has been remarkably successful in predicting the key residues that transmit allosteric signals in a number of systems,[5,12,27,28] is based on probing the propagation of response of a local perturbation at a given site to all other sites in a given structure.[5,12,27] To implement the SPM for the PDZ domain, we represent the native state using the C_α-side chain elastic network model (ENM) in which each amino acid (except Gly) is represented using the coordinates of C_α atoms and the center of mass of the side chain atoms.[27] For Gly, only the C_α atom is used. In the spirit of the ENM, we use a harmonic potential between all interaction sites (C_α and side chains) that are within a cutoff radius R_c in the folded structure. For the PDZ domain (PDB code 2BE9) R_c = 8 Å. In the C_α-*SC* ENM the potential energy is[27]

$$E = \frac{1}{2}\sum_{i,j;d^o_{ij}<R_c}\kappa_{ij}(d_{ij} - d^o_{ij})^2 \quad (7)$$

where d_{ij} is the distance between interaction sites i and j, d^o_{ij} is the corresponding structure in the native state, and

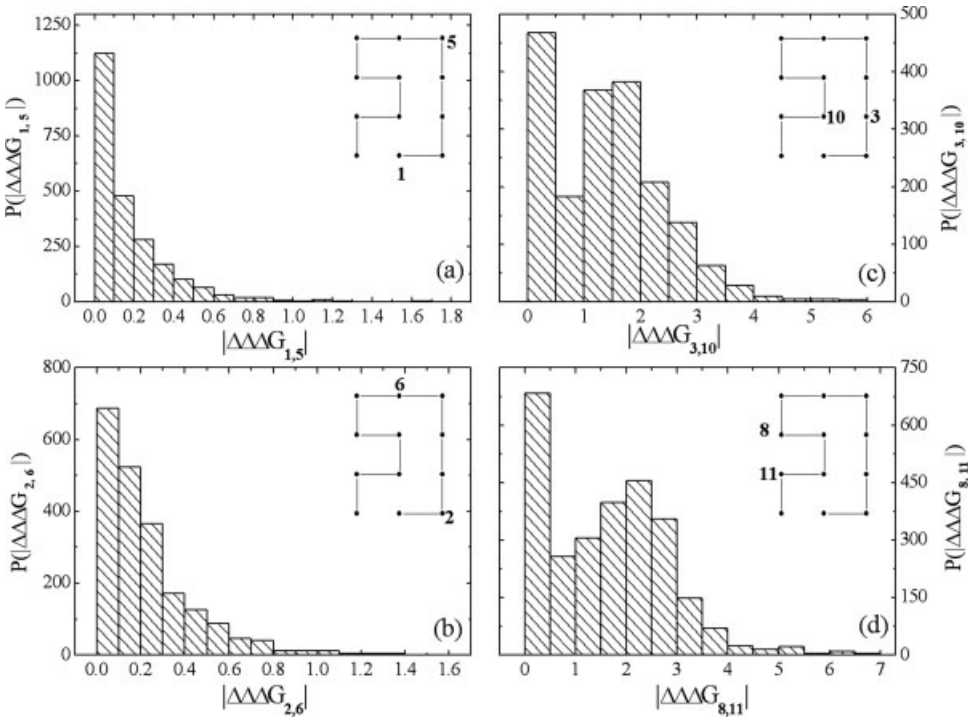

Figure 2

Distribution $P(|\Delta\Delta\Delta G_{i,j}|)$ of $|\Delta\Delta\Delta G_{i,j}|$ calculated using Eq. (4). We calculated $P(|\Delta\Delta\Delta G_{i,j}|)$ using 1000 sequences belonging to the family with the native state shown in Figure 1. (**a**) and (**b**) are obtained by mutations of sites that are not in contact, whereas $P(|\Delta\Delta\Delta G_{i,j}|)$ for sites in contact are shown in (**c**) and (**d**). The mutation sites are shown in the insets. For each sequence, there are nine possible mutations. Only those that do not alter the native structure are considered. The number $n_{i,j}$ of allowed mutations, which depends on the location i and j, corresponds to a range of $|\Delta\Delta\Delta G_{i,j}|$ is shown on the vertical scale. For example, $n_{i,j}$ with $|\Delta\Delta\Delta G_{i,j}|$ between 2 and 2.5 in (c) is $\approx$ 200.

κ_{ij} is the ij -dependent spring constant. The values of κ_{ij} are chosen based on the physical and chemical properties of the residues, and are given by $\kappa_{ij} = 4\epsilon_{ij}/(\sigma_i + \sigma_j)^2$, where ϵ_{ij} is the strength of interaction between i and j as specified in the Betancourt–Thirumalai statistical potential,[29] σ_i is the van der Waals diameter of the ith residue. To implement the SPM, we first perform a normal mode analysis for the energy function in Eq. (7). For the PDZ domain, we identify the modes that best describe the structural changes on peptide binding. The modes that best describe the structural transition are assessed using the overlap function[30]

$$I_M = \frac{|\sum_{i=1}^{3N} \boldsymbol{v}_{iM} \Delta \boldsymbol{r}_i|}{|\sum_{i=1}^{3N} \boldsymbol{v}_{iM}^2 \sum_{i=1}^{3N} \Delta \boldsymbol{r}_i^2|^{1/2}} \quad (8)$$

where $\Delta \boldsymbol{r}_i = \boldsymbol{r}_i^B - \boldsymbol{r}_i^U$ with $\boldsymbol{r}_i^B$ ($\boldsymbol{r}_i^U$) being the positions of site i in the peptide bound (peptide unbound) states, and v_{iM} is the component eigenvector associated with mode M.

After identifying the modes with largest values of I_M ($0 \leq I_M \leq 1$), we perturb the values of κ_{ij} for a residue i. Such a perturbation approximately mimics the effect of mutation. The response to such a mutation at various sites is calculated using

$$\delta\omega_{iM} = \frac{1}{2} \sum_{j;d_{ij}^0 < R_c} \delta\kappa_{ij} (d_{ij,M} - d_{ij}^0)^2 \quad (9)$$

where $\delta\kappa_{ij}$ is the strength of perturbation, and $(d_{ij} - d_{ij}^0)$ is the change in the distance between sites i and j in mode M. In practice, we used the criterion that significant response corresponds to $\delta\omega_{iM} > 2\langle\delta\omega_M\rangle$, where

$$\langle\delta\omega_M\rangle = \frac{1}{N} \sum_{i=1}^{N} \delta\omega_{iM} \quad (10)$$

with N being the total number of residues.

RESULTS AND DISCUSSIONS

Coupling energies are distance-dependent

The coupling energies obtained using the DMC [Eq. (4)] show that there is a large response to mutations at sites that are in contact and relatively weak changes at sites that are not in contact in the native conformation. The distribution, $P(|\Delta\Delta\Delta G_{i,j}|)$, calculated from mutations of the 1000 randomly chosen sequences, at four represen-

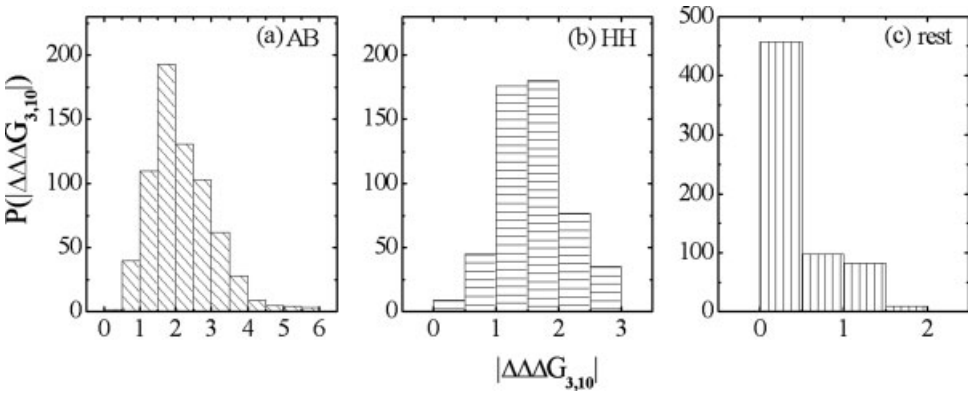

Figure 3

Dependence of $|\Delta\Delta\Delta G_{3,10}|$ [see Fig. 2(c)], calculated using Eq. (4), on the nature of mutations. The histogram with oblique lines corresponds to substitution of the oppositively charged residues (A and B) at (3,10) sites in either the WT or in one of three mutants (the single site mutants and the double mutant). The histogram with horizontal lines shows $|\Delta\Delta\Delta G_{3,10}|$ for the HH residues. The bars with vertical lines represent other possible mutations (e.g., HA, BB) at sites (3,10).

tative pairs of sites (1,5), (2,6), (3,10), and (8,11) (Fig. 2) show that the values of $|\Delta\Delta\Delta G_{i,j}|$ for residues that are not in contact in the native state are small [Fig. 2(a, b)], which implies that the coupling between sites that are physically distant is weak. In contrast, from $P(|\Delta\Delta\Delta G_{i,j}|)$ for sites that form nonbonded contacts [Fig. 2(c, d)] we infer that $|\Delta\Delta\Delta G_{i,j}|$ for (3,10) and (8,11) pairs are large. These observations, based on precise computations, accord well with DMC experiments on the PDZ domain by Chi *et al.*,[21] who noted that the energetic coupling is strongest between residues that are in proximity in the folded state.

It is interesting that among the mutations involving sites that are close in space [(3,10) for example], the ones involving AB ("charged" residues) have a stonger coupling than the HH mutations. The distribution $P(|\Delta\Delta\Delta G_{3,10}|)$ shows that the DMC coupling free energies are large for sites containing AB, which is consistent with experimental results.[11] Figure 3 also shows that the extent of energetic coupling depends on the nature of mutations that supports the observation that different mutant at the same sites can give different coupling.[21] Thus, in evaluating the extent of correlation between two sites both spatial distance and the nature of mutation have to be considered.

Comparison of energetic coupling obtained from DMC and SCA

The SCA formalism is statistical in nature while the coupling free energies inferred from DMC is based on measurements of free energy changes done to mutations at specific sites. The observations using calculations (Fig. 3) and experiments[21] that energetic coupling based on the DMC depends on the nature of mutations implies that a proper comparison with the prediction of the SCA requires averaging the DMC results over as large a dataset of mutants as possible. While such a procedure is cumbersome to carry out in vitro experiments, it can be executed using the lattice model. To compare the DMC and the SCA results, we define

$$\lambda_{i,j} = \frac{\sum |\Delta\Delta\Delta G_{i,j}|}{N_{i,j}}, \qquad (11)$$

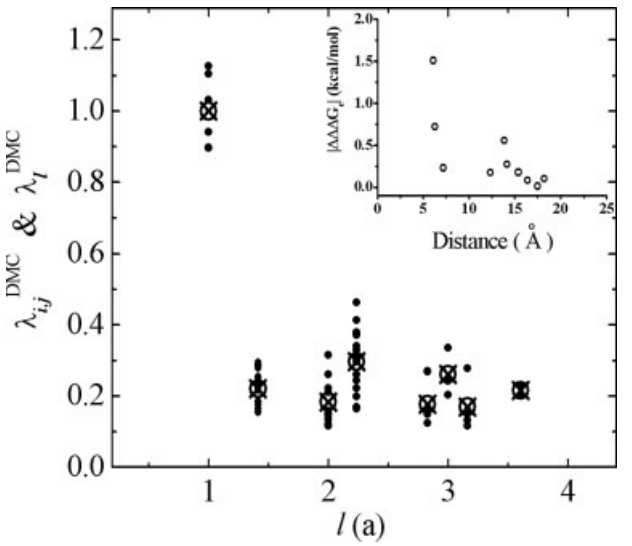

Figure 4

Dependence of $\lambda_{i,j}^{DMC}$ and λ_l^{DMC} on the spatial separation l (in unit of a). Black dots represent $\lambda_{i,j}^{DMC}$ calculated using Eq. (13), and the open circles with crosses correspond to λ_l^{DMC} [Eq. (14)]. For comparison, experimental DMC results on PDZ domain $|\Delta\Delta\Delta G_c|$ taken from Ref. 21 are shown in the inset. The sharp decrease in $\lambda_{i,j}^{DMC}$ and λ_l^{DMC} as l increases are also evident in $\lambda_{i,j}^{SCA1}$ and λ_l^{SCA1} as well as in $\lambda_{i,j}^{SCA2}$ and λ_l^{SCA2} (data not shown).

$$\lambda_l = \frac{\sum \lambda_{i,j}\delta(r_{i,j} - la)}{N_l}. \quad (12)$$

The sum in Eq. (11) is over all accepted mutations that do not alter the native state structure at sites (i,j) and $N_{i,j}$ is number of such mutations. As before for all mutations the physical free energies [Eq. (3)] are computed from exact partition function. The sum in Eq. (12) is over all the residue pairs that are separated by a distance la, and N_l is the number of such pairs. For example, for $l = 1$, $N_l = 6$ (Fig. 1). The quantities,

$$\lambda_{i,j}^{\mathrm{DMC}} = \lambda_{i,j}/\lambda_1 \quad (13)$$

$$\lambda_l^{\mathrm{DMC}} = \lambda_l/\lambda_1, \quad (14)$$

as a function of l are plotted in Figure 4. It follows from Figure 4 that the physical coupling between sites are the strongest when la is small and decreases substantially as la increases. For comparison, we also show the experimental data on PDZ family in the inset of Figure 4. There is a striking resemblance between the experimental and computational results despite the severe limitations of the small lattice size used here.

Using $\Delta\Delta G_{i,j}$ obtained from the SCA, we computed averages analogous to Eqs. (13) and (14). The linear correlation coefficients between $\lambda_{i,j}^{\mathrm{DMC}}$ and $\lambda_{i,j}^{\mathrm{SCA1}}$, $\lambda_{i,j}^{\mathrm{SCA2}}$ [the numbers 1, 2 refers to calculations using Eqs. (5) and (6)] are 0.57 and 0.67, respectively. This finding is in stark contrast with very weak correlation between the SCA and the DMC predictions found in Ref. 21 where, due to understandable experimental constraints, $\Delta\Delta\Delta G_{i,j}$ could not be averaged using a number of different types of mutations. Indeed, there is also a strong correlation between λ_l^{DMC} and $\lambda_l^{\mathrm{SCA1}}$, $\lambda_l^{\mathrm{SCA2}}$, which again illustrates the need for performing multiple DMC experiments using as diverse a set of mutations as possible without destablizing the native states. It would be stressed that for a number of reasons it may not be possible to perform multiple DMC experiments by varying the nature of mutations at sites m_1 and m_2. Many non-conservative mutations can render the folded state completely unstable.

In experiments energetic coupling between two sites is often inferred using conservative deletion mutations.[10] Such mutations typically have negligible influence on the native state of the protein. To mimic the effect of deletion mutations we introduced a fifth residue D whose

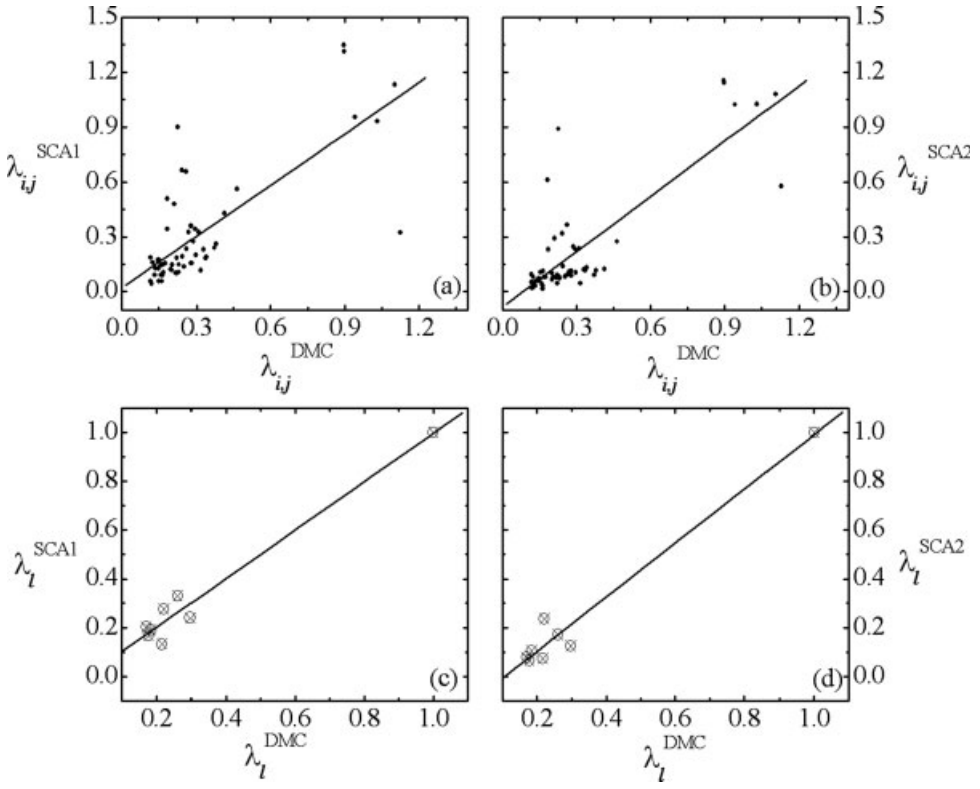

Figure 5

(**a**) Plots of $\lambda_{i,j}^{\mathrm{SCA1}}$ [calculated using Eq. (5) and the averaging procedure similar to that described for Eq. (13)] and $\lambda_{i,j}^{\mathrm{DMC}}$ [Eq. (13)]. The correlation coefficient is 0.57; (**b**) same as (a) except $\lambda_{i,j}^{\mathrm{SCA2}}$ [calculated using Eq. (6)] are used. The correlation coefficient is 0.67; (**c**) correlation between $\lambda_l^{\mathrm{SCA1}}$ and λ_l^{DMC}. The correlation coefficient is 0.96. In the absence of the datapoint corresponding to the largest values of $\lambda_l^{\mathrm{SCA1}}$ and λ_l^{DMC}, the correlation coefficient reduces to 0.31; and (**d**) correlation between $\lambda_l^{\mathrm{SCA2}}$ and λ_l^{DMC}. The correlation coefficient is 0.96. In the absence of the datapoint corresponding to the largest values of $\lambda_l^{\mathrm{SCA2}}$ and λ_l^{DMC}, the correlation coefficient becomes 0.19.

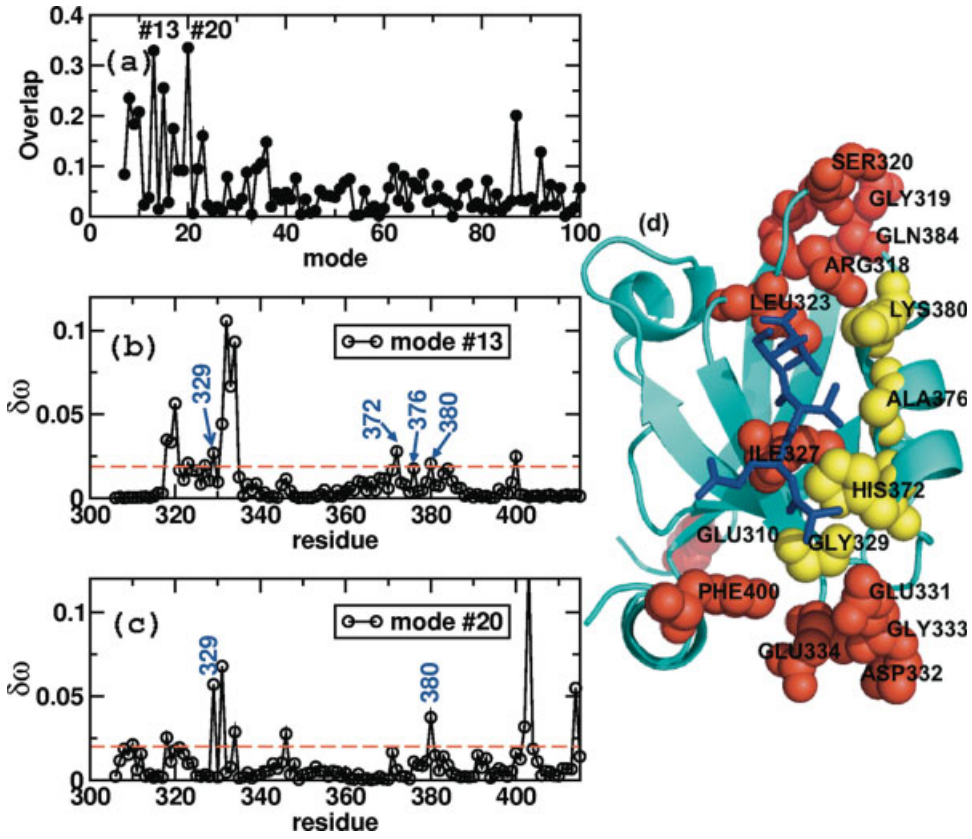

Figure 6

Structural perturbation method for the PDZ domain. (**a**) Overlap [Eq. (8)] of the 100 lowest frequency modes that describe the structural changes in the PDZ domain on peptide binding. Modes 13 and 20, with significant overlaps, are highlighted. (**b**) The response, $\delta\omega_i$ [Eq. (9)], of PDZ on perturbation at residue *i* for mode 13. The cutoff value, 2 $\langle\delta\omega\rangle$ [Eq. (10)], is indicated by the red line. The allosteric coupled residues identified in experiments are labeled. (**c**) Same as in (b), except $\delta\omega_i$ corresponds to mode 20. (**d**) Predicted hotspot residues are mapped onto the structure with red spheres. The hotspot residues that are also identified in experiments are shown in yellow spheres. The figure was drawn using PYMOL.

interactions with all other residues is set to -0.7ϵ. The relatively small value ensured that the native state is unaffected. With this choice of deletion mutation in the lattice model we performed the DMC analysis as described in the Methods section. The coupling energies averaged over all the sequences correlate well with the SCA [Eqs. (5) and (6)] with correlation coefficients that are similar to that shown in Figure 5(a,b). We conclude that it may be possible to perform a number of conservative deletion mutation experiments to assess the success of the SCA for a particular protein family.

Energetically coupled residues from SPM for the PDZ domain

The overlap of the 100 lowest frequency modes [see Eq. (8)] for the bound to unbound transition in the PDZ domain shows that this transition can be accurately described using the 13th and 20th modes [Fig. 6(a)]. The SPM analysis for the two modes [Fig. 6(b,c)] clearly identify a set of residues that have the largest response to local perturbation. Among the key residues, coupling involving Gly329, His372, and Ala376 have been deemed to be important in experimental studies.[17,21] In addition, our method also predicts Leu323, Ile327, and Gln384 that are near neighbours of Gly322, Phe325 and Val386, which are in the network of energetically coupled residues. The comparison shows that the SPM is remarkably successful in predicting the key residues that are involved in dynamics of even single domain proteins. The complete list of residues predicted by SPM are mapped onto the structure of PDZ domain (see the right side of Fig. 6).

CONCLUSIONS

Methods for detecting covariation in residues that are separated along the sequence using evolutionary imprints are based on a set of assumptions. The utility of such methods can only be discerned by making exhaustive

comparisons with experiments. Indeed, making such a comparison is not always easy because sequence-based methods are statistical in nature whereas the double mutant cycle experiments (used to extract coupling between distant sites) use limited data sets of mutations. We have used lattice models to self-consistently examine the reliability of the sequence-based SCA to predict the physical coupling between various sites. Based on this study several conclusions and inferences may be drawn.

- Free energetic changes on mutations of residues that are in contact in the native structure are much greater than those that are not. In other words, energetic coupling between distinct residues fall off (almost exponentially) with distance between them (Fig. 4). This finding, obtained using precise computations based on lattice model representation of polypeptide chains, confirms earlier experimental observations.[11,21]
- We also find that strongly interacting residues have the largest response to mutations and hence are highly correlated in their evolution. This finding is also in accord with inferences drawn from bioinformatic analysis[26] and double mutant cycle experiments on barstar-barnase complex.[11] Taken together these results affirm the observations that strongly interacting residues that are spatially adjacent are highly correlated. In a statistical sense the SCA can predict the covariation among these residues reliably.
- The SCA method, as currently formulated can only infer covariation using perturbation at a single site. It would be interesting to obtain sequence variation by considering perturbation at two sites j and k, which can be implemented by considering subalignments of the MSA in which the sequence entropies at sites j and k are zero. Such a formulation can be used to infer multisite covariations, which are automatically recovered using structure-based methods.[5,12]
- It should be pointed out that in some instances the SCA is useful in deciphering energetically coupled network of residues that transmit allosteric signals in proteins. It was shown by Chen *et al.*[31] that the SCA prediction of the sparse of network of residues in dihydrfolate reductase (DHFR) also were intimately related in the kinetics of the conformational transitions between the open to the occluded states. The link between the sequence-based approach and structure-based method that determines the kinetics of transition between two allosteric states lends credence to the utility of the SCA. More recently, SCA has been used to engineer new allosteric sites by fusing two proteins (LOVE domain and DHFR) to initiate hydride transfer reaction in DHFR by photolysis of the LOVE domain.[20] The lack of firm theoretical basis for the SCA (and other evolutionary approaches) requires that only by applications to a variety of systems can the overall efficacy of sequence-based methods be assessed.
- Our results show that the SPM method, which depends on the response of local perturbation on the the residues that are spatially far apart (allosteric effect), yields an allosteric wiring diagram for PDZ. All of the residues that have been deemed to have significant free energetic coupling based on experiments were identified using the SPM using the C_α-SC representation elastic network representation of the PDZ domain. In addition we also predict few other relevant set of residues (Fig. 6) that are potential candidates for DMC experiments. The physically motivated SPM method, which has been carried out for a number of systems by us and others, is an alternative way to infer the AWD, which reflects the potential energetic coupling between distinct sites in a protein.

ACKNOWLEDGMENTS

The authors thank Dr. Riina Tehver for providing a program to compute the results shown in Figure 6. They also thank Prof. Ruxandra I. Dima for useful discussions.

REFERENCES

1. Horovitz A, Fridmann Y, Kafri G, Yifrach O. Review: allostery in chaperonins. J Struct Biol 2001;135:104–114.
2. Changeux JP, Edelstein SJ. Allosteric mechanisms of signal transduction. Science 2005;308:1424–1428.
3. Eaton WA, Henry ER, Hofrichter J, Mozzarelli A. Is cooperative oxygen binding by hemoglobin really understood? Nat Struct Biol 1999;6:351–358.
4. Hyeon C, Lorimer GH, Thirumalai D. Dynamics of allosteric transitions in GroEL. Proc Natl Acad Sci USA 2006;103:18939–18944.
5. Zheng WJ, Brooks BR, Doniach S, Thirumalai D. Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. Structure 2005;13:565–577.
6. Kern D, Zuiderweg ER. The role of dynamics in allosteric regulation. Curr Opin Struct Biol 2003;13:748–757.
7. Vale RD, Milligan RA. The way things move: looking under the hood of molecular motor proteins. Science 2000;288:88–95.
8. Thirumalai D, Lorimer GH. Chaperonin-mediated protein folding. Annu Rev Biophys Biomol Struct 2001;30:245–269.
9. Hammes-Schiffer S, Benkovic SJ. Relating protein motion to catalysis. Annu Rev Biochem 2006;75:519–541.
10. Horovitz A, Fersht AR. Strategy for analyzing the cooperativity of intramolecular interactions in peptides and proteins. J Mol Biol 1990;214:613–617.
11. Schreiber G, Fersht AR. Energetics of protein-protein interactions — analysis of the Barnase-Barstar interface by single mutations and double mutant cycles. J Mol Biol 1995;248:478–486.
12. Zheng WJ, Brooks BR, Thirumalai D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci USA 2006;103:7664–7669.
13. Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins 2002;48:611–617.
14. Fodor AA, Aldrich RW. On evolutionary conservation of thermodynamic coupling in proteins. J Biol Chem 2004;279:19046–19050.
15. Neher E. How frequent are correlated changes in families of protein sequences. Proc Natl Acad Sci USA 1994;91:98–102.
16. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol 2002;12:368–373.

17. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–299.
18. Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 2003;10:59–69.
19. 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.
20. Lee J, Natarajan M, Nashine VC, Socolich M, Vo T, Russ WP, Benkovic SJ, Ranganathan R. Surface sites for engineering allosteric control in proteins. Science 2008;322:438–442.
21. Chi CN, Elfström L, Shi Y, Snäll T, Engström Å, Jemth P. Reassessing a sparse energetic network within a single protein domain. Proc Natl Acad Sci USA 2008;105:4679–4684.
22. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 2004;56:211–221.
23. Harrison PM, Chan HS, Prusiner SB, Cohen FE. Conformational propagation with prion-like characteristics in a simple model of protein folding. Protein Sci 2001;10:819–835.
24. Dima RI, Thirumalai D. Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. Protein Sci 2002;11:1036–1049.
25. Noivirt-Brik O, Unger R, Horovitz A. Analysing the origin of long-range interactions in proteins using lattice models. BMC Struct Biol 2009;9:4–13.
26. Dima RI, Thirumalai D. Determination of network of residues that regulate allostery in protein families using sequence analysis. Protein Sci 2006;15:258–268.
27. Tehver R, Chen J, Thirumalai D. Allostery wiring diagrams in the transitions that drive the GroEL reaction cycle. J Mol Biol 2009;387:390–406.
28. Taly A, Corringer PJ, Grutter T, de Carvalho LP, Karplus M, Changeux JP. Implications of the quaternary twist allosteric model for the physiology and pathology of nicotinic acetylcholine receptors. Proc Natl Acad Sci USA 2006;103:16965–16970.
29. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361–369.
30. Zheng WJ, Doniach S. A comparative study of motor-protein motions by using a simple elastic-network model. Proc Natl Acad Sci USA 2003;100:13253–13258.
31. Chen J, Dima RI, Thirumalai D. Allosteric communication in dihydrofolate reductase: signaling network and pathways for closed to occluded transition and back. J Mol Biol 2007;374:250–266.