Interactions in Native Binding Sites Cause a Large Change in Protein Dynamics

Dengming Ming[1] and Michael E. Wall[1,2*]

[1]Computer and Computational Sciences Division and [2]Bioscience Division, Los Alamos National Laboratory, Los Alamos NM 87545

*E-mail address of the corresponding author: mewall@lanl.gov

Keywords: allosteric regulation; protein dynamics; protein evolution; ligand binding; functional sites; protein function prediction

Abbreviations used: DPA, dynamics perturbation analysis; ENM, elastic network model; tri-NAG, tri-N-acetyl-D-glucosamine

Los Alamos National Laboratory LA-UR-05-8671.

To appear in *Journal of Molecular Biology.*

Subject area: Protein and nucleic acid structure, function and interactions

Abstract

Cellular functions are regulated by molecules that interact with proteins and alter their activities. To enable such control, protein activity, and therefore protein conformational distributions, must be susceptible to alteration by molecular interactions at functional sites. Here we investigate whether interactions at functional sites cause a large change in the protein conformational distribution. We apply a computational method, called dynamics perturbation analysis (DPA), to identify sites at which interactions have a large allosteric potential $D_{\mathbf{x}}$, which is the Kullback-Leibler divergence between protein conformational distributions with and without an interaction. In DPA, a protein is decorated with surface points that interact with neighboring protein atoms, and $D_{\mathbf{x}}$ is calculated for each of the points in a coarse-grained model of protein vibrations. We use DPA to examine hundreds of protein structures from a standard small-molecule docking test set, and find that ligand-binding sites have elevated values of $D_{\mathbf{x}}$: for 95% of proteins, the probability of randomly obtaining values as high as those in the binding site is 10^{-3} or smaller. We then use DPA to develop a computational method to predict functional sites in proteins, and find that the method accurately predicts ligand-binding-site residues for proteins in the test set. The performance of this method compares favorably with that of a cleft analysis method. The results confirm that interactions at small-molecule binding sites cause a large change in the protein conformational distribution, and motivate using DPA for large-scale prediction of functional sites in proteins. They also suggest that natural selection favors proteins whose activities are capable of being regulated by molecular interactions.

Introduction

Bochemical regulation is fundamental to the cell's ability to maintain homeostasis, orchestrate developmental processes, and adapt to environmental changes. Regulation of protein activity is especially important for regulation of cellular functions. Because regulation is such an important feature in biological systems, it is interesting to contemplate its role in protein evolution.

One important mechanism of protein regulation is allosteric regulation, in which molecular interactions influence protein activity through changes in protein structure. In traditional models of allosteric regulation, proteins adopt a limited number of conformations, each of which may have a different activity.[1; 2] However, it has been long recognized that protein structures fluctuate in the cell, and that protein regulation involves changes in the full protein conformational distribution.[3] Indeed, the conformational distribution is known to be a key determinant of protein activity,[4] and is a key element in rate theories.[5] The recent development of a theoretical framework to quantify changes in protein conformational distributions was motivated by these considerations.[6; 7]

For allosteric regulation to work, the protein conformational distribution must be susceptible to alteration by interactions at an allosteric site. In addition, to prevent spurious activity in the absence of specific target molecules, newly synthesized proteins should be biased towards inactive conformations, and interactions in the active site should bias the protein towards conformations that are more active. Does Nature favor regulatable proteins? If so, then, as was suggested by a study of lysozyme,[6] we expect

interactions in protein functional sites to cause a large change in the conformational distribution, facilitating the ability of molecular interactions to change protein activity.

Here we examine 305 protein structures from the GOLD docking test set[8] and investigate whether interactions at small-molecule binding sites cause a large change in the protein conformational distribution. We present a computational method, called dynamics perturbation analysis (DPA), to identify sites at which interactions have a large allosteric potential $D_{\mathbf{x}}$, which is the Kullback-Leibler divergence between protein conformational distributions with and without an interaction.[6; 7] We use DPA to analyze proteins in the test set, and determine whether $D_{\mathbf{x}}$ values for points in the neighborhood of ligand-binding sites are high compared to random points. We then develop a method to predict functional sites in proteins, and evaluate the method using proteins in the test set. The performance of the method is compared to that of a cleft analysis method. The results have important implications for prediction of functional sites in proteins, and in considering whether Nature favors regulatable proteins.

Dynamics perturbation analysis

Dynamics perturbation analysis (DPA) is based on a method previously used to analyze changes in fluctuations of a protein complex for random protein-ligand interactions.[6] In DPA, a protein is decorated with M surface points that interact with neighboring protein atoms. The protein conformational distribution $P^{(0)}(\mathbf{x})$ is calculated in the absence of any surface points, and M protein conformational distributions $P^{(m)}(\mathbf{x})$ are calculated for the protein interacting with each point m. As in a recent study of allosteric effects in trypsinogen,[7] the conformational distributions are calculated using a coarse-grained model of molecular vibrations, and the distributions $P^{(m)}(\mathbf{x})$ are isolated from

models of the protein in complex with the surface points. The Kullback-Leibler divergence $D_{\mathbf{x}}^{(m)}$ between $P^{(0)}(\mathbf{x})$ and $P^{(m)}(\mathbf{x})$ is calculated for each point m[6; 7] (Eq. (1)), and is used as a measure of the change in the protein conformational distribution upon interacting with point *m*. The measure $D_{\mathbf{x}}^{(m)}$ is called the allosteric potential of the interaction of point *m* with the protein.

For a given protein structure, evenly distributed surface points were generated by using the program MSMS.[9] A probe radius of 1.5 Å and a triangulation density of 1.0 vertex/Å^2 were used in running MSMS. The vertex entries were used as surface points.

Protein fluctuations were modeled using the elastic network model (ENM).[10; 11; 12; 13] In the ENM, alpha-carbon atoms are extracted from an atomic model of a protein, and an interaction network is generated by connecting springs between all atom pairs separated by a distance less than or equal to a cutoff distance r_c. Each spring has the same force constant γ, is aligned with the separation between the connected atoms, and has an equilibrium length equal to the equilibrium distance between the atoms. Where possible, we used a cutoff value r_c = 8.5 Å. In several cases, however, a value r_c = 8.5 Å resulted in more-than the expected number of six zero-frequency vibrations. In these cases, the value of r_c was repeatedly increased by 1 Å until there were only six zero-frequency modes. Calculations of $D_{\mathbf{x}}^{(m)}$ are independent of the choice of γ.

The interaction between the protein and a surface point *m* was modeled by connecting springs of force constant γ_s between the surface point and all protein atoms within a cutoff distance r_s of the surface point. The protein coordinates were not modified in modeling the interaction. To make the magnitude of the effect of the surface point on the protein larger (*i.e.*, more comparable to what might be expected from an interaction with

an extended ligand), we increased both the force constant and cutoff distance with respect to the values used for protein atoms. The increases elevated the magnitudes of values of $D_{\mathbf{x}}^{(m)}$ and were empirically found to enhance the statistical significance of the results below. In practice we found that statistically significant results were obtained using $\gamma_s = 12\gamma$ and $r_s = r_c + 5.5$ Å, which are the values that were used in the calculations below.

Calculation of the allosteric potential

The allosteric potential $D_{\mathbf{x}}^{(m)}$ is defined as the Kullback-Leibler divergence between the unperturbed and perturbed protein conformational distributions,

$$D_{\mathbf{x}}^{(m)} = \int d^{3N}\mathbf{x}\left(\log\frac{P^{(m)}(\mathbf{x})}{P^{(0)}(\mathbf{x})}\right)P^{(m)}(\mathbf{x}), \qquad (1)$$

where **x** describes the configuration of the N protein atoms in the protein, and $P^{(0)}(\mathbf{x})$ and $P^{(m)}(\mathbf{x})$ are as defined above. The significance of the Kullback-Leibler divergence has been previously discussed in the context of quantifying equilibrium density fluctuations and the nonequilibrium relaxation of polymer conformational distributions;[14] and in the context of comparing protein conformational distributions with and without a ligand bound[6; 7] Similar to the latter case, in the present context, $D_{\mathbf{x}}^{(m)}$ is essentially the free energy change of the protein upon attaching the point m minus the mean relative energy of protein conformations in the presence of point m, where each energy is measured relative to the energy of the same conformation in the absence of the point.

The distribution $P^{(0)}(\mathbf{x})$ is determined by the eigenvalues $\omega_i^{(0)^2}$ and eigenvectors $\mathbf{v}_i^{(0)}$ of the Hessian matrix $\mathbf{H}^{(0)}$, whose elements are defined as $H_{ij}^{(0)}\big|_{\mathbf{x}_0} = \partial^2 U^{(0)}/\partial x_i \partial x_j\big|_{\mathbf{x}_0}$. $U^{(0)}(\mathbf{x})$ is the potential energy of a configuration **x** evaluated using the ENM in the absence of

surface points, and $\mathbf{x}_0$ is the equilibrium configuration of the protein. As we have previously shown,[7] the distribution $P^{(m)}(\mathbf{x})$ is determined by the eigenvalues $\overline{\omega}_i^{(m)2}$ and eigenvectors $\overline{\mathbf{v}}_i^{(m)}$ of a pseudo-Hessian matrix $\overline{\mathbf{H}}^{(m)}$, which has the same dimensionality as $\mathbf{H}^{(0)}$. $\overline{\mathbf{H}}^{(m)}$ is obtained from the full Hessian $\mathbf{H}^{(m)}$ of the ENM in the presence of the surface point $\mathbf{x}_s^{(m)}$, which is decomposed as follows

$$\mathbf{H}^{(m)} = \begin{pmatrix} \mathbf{H}_p & \mathbf{G} \\ \mathbf{G}^T & \mathbf{H}_s^{(m)} \end{pmatrix}. \tag{2}$$

The sub-matrix $\mathbf{H}_p$ couples protein coordinates, $\mathbf{H}_s^{(m)}$ couples surface-point coordinates, and $\mathbf{G}$ couples coordinates between the protein and the surface point. In terms of these sub-matrices, the pseudo-Hessian $\overline{\mathbf{H}}^{(m)}$ is given by

$$\overline{\mathbf{H}}^{(m)} = \mathbf{H}_p - \mathbf{G}\mathbf{H}_s^{(m)-1}\mathbf{G}^T. \tag{3}$$

Eq. (3) was independently derived both in Ref. [7] and by Zheng & Brooks in Ref. [15]. The value of $D_{\mathbf{x}}^{(m)}$ may then be calculated as[6; 7]

$$D_{\mathbf{x}}^{(m)} = \sum_{i=1}^{3N}{}' \left(\log \frac{\overline{\omega}_i^{(m)}}{\omega_i^{(0)}} + \frac{1}{2k_B T}\omega_i^{(0)2} \left| \Delta\mathbf{x}_0^{(m)} \cdot \mathbf{v}_i^{(0)} \right|^2 + \frac{1}{2} \sum_{j=1}^{3N}{}' \frac{\omega_j^{(0)2}}{\overline{\omega}_i^{(m)2}} \left| \overline{\mathbf{v}}_i^{(m)} \cdot \mathbf{v}_j^{(0)} \right|^2 - \frac{1}{2} \right). \tag{4}$$

Where the primed sums are carried out over all nonzero modes (all but 6 zero-frequency rigid-body modes). In Eq. (4), $\Delta\mathbf{x}_0^{(m)} = \mathbf{x}_0^{(m)} - \mathbf{x}_0$ is the difference between the equilibrium conformation of the protein in the presence and absence of the point *m*; in the present study, we do not consider the surface point to have any influence on the equilibrium conformation, making the second term of Eq. (4) equal to zero.

Results

Analysis of lysozyme

We initially applied DPA to turkey egg-white lysozyme (Protein Data Bank entry 1JEF[16]). The all-atom protein model was used to generate surface points. For normal modes calculations, alpha-carbon coordinates for the protein were extracted and used without modification for calculations both with and without surface points. Equation (4) was applied explicitly, requiring that the pseudo-Hessian in Eq. (3) be diagonalized for each point *m*.

The calculation for lysozyme, which has 129 amino acids, took about 24 minutes on a 3 GHz Pentium 4. Assuming diagonalization scales like the square of the number of residues, and the number of surface points scales like the 2/3 power of the number of residues, the calculation time should scale with the 8/3 power of the number of residues. Based on this back-of-the-envelope calculation, we expect that a protein with twice as many amino acids as lysozyme would take about $2^{8/3} \times 24 = 152$ minutes. Indeed, application of DPA to a 260-residue portion of a NADH-dependent nitrate reductase (Protein Data Bank entry 2CND[17]) took 162 minutes, which is close to the expected length of time. Using first-order perturbation theory, we have calculated estimates of $\sum_{i=1}^{3N}{}' \log \frac{\overline{\omega}_i^{(m)}}{\omega_i^{(0)}}$, which is just the entropic contribution to $D_{\mathbf{x}}^{(m)}$, in as little as 1/50 of the time required to calculate $D_{\mathbf{x}}^{(m)}$. Such an approach shows promise in accelerating the below algorithm for prediction of functional sites (unpublished results).

Consistent with results obtained using an all-atom model of lysozyme,[6] we found that values of $D_{\mathbf{x}}^{(m)}$ are elevated in the neighborhood of the tri-N-acetyl-D-glucosamine (tri-NAG) binding site (Fig. 1). Interestingly, the distribution of $y = D_{\mathbf{x}}^{(m)}$ values is empirically well-fit by a probability density $\rho(y)$ given by

$$\rho(y) = \frac{1}{\beta} e^{\frac{y-\mu}{\beta} - e^{\frac{y-\mu}{\beta}}}, \tag{5}$$

which is an extreme value distribution of width β centered on μ (Fig. 2). The fit was obtained using standard nonlinear least-square methods. Later we describe how the extreme value distribution model was used to predict functional sites.

Analysis of the test set

We then applied DPA to 305 protein structures in the GOLD docking test set.[8] Calculations were performed in the same manner as for lysozyme (for the larger proteins in the test set, surface points were evaluated in parallel using several processors to decrease computation time). We wished to quantitatively assess the tendency for $D_{\mathbf{x}}^{(m)}$ values to be elevated in the neighborhood of ligand-binding sites, and developed the following statistical analysis for this purpose. For each protein *k*, we selected the set of *L* surface points $\mathcal{L}$ within 6 Å of any non-hydrogen atom in the ligand. For each selected surface point $l \in \mathcal{L}$, we then calculated the fraction P_{kl}^{+} of all surface points *m* that have a value of $D_{\mathbf{x}}^{(m)}$ at least as high as $D_{\mathbf{x}}^{(l)}$ in protein *k*. A total score for the set was calculated as $z_k = \prod_{l \in \mathcal{L}} P_{kl}^{+}$. The smaller the value of z_k, the more elevated the values of $D_{\mathbf{x}}^{(m)}$ are in the neighborhood of the binding site. To determine the statistical significance of a value z_k, we wished to calculate the probability that a random selection of the same number of

surface points yields a score z_k or smaller. This probability is very nearly the same as the probability $P^-(z_k)$ of obtaining a value z_k or smaller for a product of L independent random variables uniformly distributed over the range [0,1], which is given by the distribution $P^-(z) = z\sum_{l=1}^{L}\frac{(-\ln z)^{l-1}}{(l-1)!}$ (Appendix). We therefore used the following P-value to quantify the statistical significance of the collected values of $D_{\mathbf{x}}^{(m)}$ in the neighborhood of a functional site:

$$P_k^- = z_k\sum_{l=1}^{L}\frac{(-\ln z_k)^{l-1}}{(l-1)!}, \qquad (6)$$

In 14 of the 305 proteins, the ligand was buried and was not close to any of the surface points. We used the remaining 291 proteins to evaluate the tendency for $D_{\mathbf{x}}^{(m)}$ to be elevated in the neighborhood of the ligand-binding site. Results are illustrated in Fig. 3. For 95% of proteins, the P-value P_k^- is 10^{-3} or lower, indicating that the elevation of $D_{\mathbf{x}}^{(m)}$ in the neighborhood of functional sites is statistically significant.

Prediction of binding sites using DPA

We suspected that points with high values of $D_{\mathbf{x}}^{(m)}$ could be used to predict the locations of functional sites, and developed an algorithm for this purpose. The algorithm works as follows. First, DPA is performed on a protein. Then, the statistics of $D_{\mathbf{x}}^{(m)}$ values is modeled using an extreme value distribution. Points with significantly high values of $D_{\mathbf{x}}^{(m)}$ are selected and are spatially clustered. The clusters are ranked according to the mean value of $D_{\mathbf{x}}^{(m)}$ within the cluster, and points in the highest-ranked cluster are predicted to be associated with a functional site. Finally, residues in the neighborhood of the highest-ranked cluster are selected and are predicted to reside within the functional

site. We used this algorithm to predict functional sites for proteins in the test set and examined the overlap of the predictions with the ligand-binding sites.

Consistent with the analysis of lysozyme, the $D_{\mathbf{x}}^{(m)}$ values for the test-set proteins indicate that the statistics are well-described by an extreme-value distribution (Fig. 4). To select points with significantly high values of $y = D_{\mathbf{x}}^{(m)}$, we selected an operating point *C* of the cumulative distribution $c(y) = 1 - e^{-e^{\frac{y-\mu}{\beta}}}$, fitted μ and β using the actual distribution of $D_{\mathbf{x}}^{(m)}$ for the protein, and calculated a lower threshold *Y* on $D_{\mathbf{x}}^{(m)}$ as follows:

$$Y = y(C) = \beta \ln\left[-\ln(1-C)\right] + \mu \,. \qquad (7)$$

We found that a value $C = 0.96$ yielded a high overlap of our predictions with the ligand-binding sites in the test set (see below).

Points with $D_{\mathbf{x}}^{(m)} > Y$ were clustered spatially using the Ordering Points to Identify the Clustering Structure (OPTICS) algorithm.[18] Parameters were selected such that at least 3 other surface points are within 6 Å of each point in the cluster. Using this clustering criterion combined with $C = 0.96$ resulted in at least one cluster for 287 of the 305 proteins, yielding a prediction rate of 94% for the test set. The mean value of $D_{\mathbf{x}}^{(m)}$ for each cluster was calculated and was used to rank the clusters; for each protein, the rank-1 cluster was identified as the cluster with the highest mean value.

Protein alpha-carbons within 6 Å of any of the points in the rank-1 cluster were selected and were used to identify the set of R_P residues $\mathcal{R}_P$ that are predicted to reside in a functional site. These were compared with the set of R_L residues $\mathcal{R}_L$ that are in the neighborhood of the ligand found in complex with the protein in the test set: the intersection is the set of $R_{P\cap L}$ residues $\mathcal{R}_P \cap \mathcal{R}_L$ found in both the predicted set and the

ligand set. The overlap of the predictions with the ligand-binding site was assessed using the precision $R_{P\cap L}/R_P$ and the recall $R_{P\cap L}/R_L$. (Detailed information about residues found near DPA clusters and ligand-binding site residues is available online at http://public.lanl.gov/mewall/dpa).

Figure 5 depicts a typical rank-1 cluster in the neighborhood of a ligand-binding site. Statistics of the overlaps from the entire test set are illustrated in Figs. 6 and 7. As mentioned above, predictions were made for 287 of the 305 proteins. In 87% of cases (250 proteins), at least one predicted residue was in the ligand-binding site. The recall was at least 0.3 for 80% of cases, and was at least 0.5 for 76% of the cases (Fig. 6). The precision was at least 0.3 for 68% of the cases, and was at least 0.5 for 44% of the cases (Fig. 7). These performance measures depend on the value of the threshold C; for example, the dependence of the 0.5-level precision and recall statistics on C is illustrated in Fig. 8. The value $C = 0.96$ was chosen to yield a relatively high precision with little cost in the recall.

The statistical significance of the overlaps was assessed using a null model in which surface residues were randomly selected. A list of all surface residues for a protein was found by selecting all residues whose alpha-carbons are within 6 Å of at least one surface point. The number R_S of such residues was then used to calculate the probability of finding at least $R_{P\cap L}$ residues in the ligand-binding site by randomly selecting R_P residues from R_S surface residues:

$$P_{null} = \sum_{n=R_{P\cap L}}^{\min(R_P,R_L)} \binom{R_L}{n} \Big/ \binom{R_S}{R_P} = \sum_{n=R_{P\cap L}}^{\min(R_P,R_L)} \frac{R_L!R_P!(R_S-R_P)!}{n!(R_L-n)!R_S!}. \quad (8)$$

We calculated the P-value P_{null} for all cases in which there was an overlap of at least one residue between the predicted residues and the ligand-binding-site residues, 250 cases in all. Results are shown in Fig. 9. For 87% of the cases, P_{null} is 10^{-3} or smaller, indicating that there is a statistically significant overlap.

Prediction of binding sites using cleft analysis

To provide some context for the performance of the DPA prediction algorithm, we compared the DPA algorithm to an algorithm based on cleft analysis. Cleft analysis was chosen because it is commonly used to identify ligand-binding sites, and, like DPA, it only requires structure information as an input, and does not require sequence information. The algorithm used was based on standard software for cleft analysis, SURFNET[19] (including the programs SURFNET, SURFACE, SURFPLOT and MASK), which was originally developed to study the association of large clefts with positions of ligands on the protein surface. For rigorous comparison to our method, we used the output of SURFNET to locate residues near the largest cleft in a protein. SURFNET version 1.4 was used with default parameter values to analyze atomic coordinate files, resulting in a volume-ranked list of cleft locations. Then, SURFACE was used to analyze the largest cleft, and SURFPLOT was used to generate a set of surface points that surrounded the cleft. Finally, MASK was used to convert the coordinates of the surface points to Protein Data Bank format. The output of MASK was used to select residues in precisely the same manner as for clusters of points with high values of $D_{\mathbf{x}}^{(m)}$ (**Prediction of binding sites using DPA**).

Before using the above method to predict ligand-binding site residues in the GOLD test set, we confirmed that SURFNET yielded predictions of ligand-binding positions that

were similar to those found in an early application of SURFNET by Laskowski et al.[20] Indeed, examination of the position of the ligand for the set of 67 proteins from the Laskowski et al. study yielded the same results as those reported in Ref. [20], with only three exceptions: in structures of proteinase K (Protein Data Bank entry 1PEK[21]) and actylcholinesterase (entry 2ACK[22]), the ligand was found in the second-largest cleft instead of the largest cleft; and in a structure of aconitase (entry 8ACN[23]), the ligand was found in the largest cleft instead of the third-largest cleft. These exceptions are most likely due to changes in the SURFNET software after the original study was conducted (Roman Laskowski, personal communication).

We then applied the cleft analysis method to analyze the 305 proteins in the GOLD test set, yielding predictions for 303 of the proteins. Through qualitative visualization of the results, we found that the largest cleft often not only overlapped but also extended beyond the ligand-binding region (Fig. 5). This observation is supported by a statistical analysis of the predictions: the recall of ligand-binding residues for the cleft algorithm is high compared to that of the DPA algorithm (Fig. 6), and the precision is low by comparison (Fig. 7). Analysis using the null model supports these results (Fig. 9): for 62% of the 278 proteins with at least one overlapping residue, P_{null} is 10^{-3} or smaller, compared to 87% of 250 proteins using the DPA algorithm. Application of the DPA algorithm to this test set therefore provided more statistically significant overlaps than did the cleft analysis algorithm.

Discussion

We used dynamics perturbation analysis to examine a test set of hundreds of proteins, and performed a rigorous statistical analysis of the results. The major conclusion is that ligand-binding sites in the test set are located at control points that enable large changes in protein dynamics: for 95% of proteins, the degree of elevation of $D_{\mathbf{x}}^{(m)}$ values in the neighborhood of functional sites would be expected only once in a thousand random samples. Most of the ligands in the test set are small molecules, and we expect that small-molecule binding sites in other proteins would also tend to be located at dynamical control points. The implications of our results for sites of large-molecule interactions, such as protein-protein interactions, are currently unknown; the present study motivates a larger survey to examine the association of different types of functional sites with dynamical control points in a wide variety of proteins.

Another major finding is that DPA can predict functional sites in proteins. We found that the DPA algorithm yielded predicted residues that had a significant overlap with the ligand-binding-site residues in the test set. There were some exceptions, however: in 37 cases, residues in the neigborhood of the rank-1 cluster had no overlap with the ligand-binding site. What can be said of these exceptions? Because proteins in the test set might have functional sites in addition to the ligand-binding site, we expect some of the predictions to have a high degree of overlap with alternative functional sites. Analysis of specific cases supports this idea: (1) streptavidin is in a dimeric form in which only one of the two monomers has a ligand bound; the rank-2 cluster is at the ligand-binding site on one monomer, and the rank-1 cluster is at an equivalent site on the other monomer (Fig. 10). (2) In a trypsinogen complex, the rank-2 cluster is at the ligand-binding site, and the rank-1 cluster is at an alternative site (Fig. 11). In other cases, like in porcine

synovial collagenase, the rank-2 cluster is at the ligand-binding site, and the rank-1 cluster lies at the interface between two domains; by splitting the protein into two domains, the position of the rank-1 cluster moves to the ligand-binding-site (Fig. 12). In addition, as all of these examples suggest, rank-2 clusters are often associated with the ligand-binding site: 17 of the 37 cases are of this type, with an additional two cases in which it is the rank-3 cluster that is in the binding site, and one case in which it is the rank-4 cluster.

Recently, in a study of a set of 98 enzymes, Yang & Bahar[24] found that catalytic residues tend to be associated with structural hinge regions. For each enzyme, an elastic network model was used to simulate harmonic vibrations, and the two lowest nonzero-frequency modes were analyzed. Catalytic residues were found to be associated with the sequence neighborhood of the residue whose mean amplitude of vibration over these two modes is the smallest in the protein, *i.e.*, they tended to be located in a hinge region with respect to the low-frequency modes. Yang & Bahar[24] reported similar temperature factors for models of liganded and unliganded enzymes, which at first glance appears to hint at an inconsistency with the present study. However, the quantity $D_{\mathbf{x}}^{(m)}$ used here measures differences in the entire conformational distribution, whereas their study only considered changes in temperature factors. Because we have previously shown that $D_{\mathbf{x}}^{(m)}$ can be large even when differences in temperature factors are small,[7] the studies are not inconsistent – rather, they represent complementary approaches to quantifying the relation between protein dynamics and functional sites. It would be interesting to conduct a more detailed study of the relations among functional sites, structural hinges, and the dynamical control points that were the subject of this study.

The performance of the DPA algorithm in predicting ligand-binding sites for the GOLD test set compared favorably to an algorithm based on cleft analysis, yielding fewer true positives on the one hand, but fewer false positives and more statistically significant overlaps with ligand-binding sites on the other hand. However, it is important to note that, although SURFNET was developed to locate clefts where binding interactions might occur, it was not explicitly developed for the present application of predicting specific residues that contact the ligand.[19] In addition, for comparison to the DPA algorithm, the cleft analysis algorithm used here only made use of structure information; recently, evolutionary conservation patterns have been used in combination with SURFNET to trim clefts and obtain a better overlap with the volume of a bound ligand.[25] Finally, although we were able to reproduce published results using SURFNET, we have not rigorously tuned the cleft analysis algorithm for optimal performance in predicting ligand-binding residues. Therefore, it might be possible to achieve better performance than we have presented here.

It is important to note that all of the analyses in this study were performed on protein structures obtained from a protein-ligand complex and were used without modification (e.g., energy minimization). Therefore, although the present results demonstrate the utility of DPA in predicting ligand-binding sites for protein conformations that are consistent with ligand interactions, the ability of DPA to predict ligand-binding sites for ligand-free protein structures that exhibit a significant mean conformational change upon binding a ligand remains to be tested. Future studies are needed to determine whether binding sites are detectable using the mean conformation of the ligand-free protein structure, or whether it will be necessary to consider alternative structures, e.g., by

performing DPA on an ensemble of structures sampled from simulations of thermal fluctuations.

Looking beyond the present work, DPA will be able to contribute to the goals of predicting which ligands bind to a protein, predicting which residues in a binding site are functionally most important, and predicting what functions those important residues carry out. For each of these tasks, initial application of DPA may be used to focus efforts on a small number of dynamical control points instead of the entire protein surface, saving computing time. In addition, DPA uses protein dynamics information that is complementary to information used by other protein structure and sequence analysis methods, and might therefore be integrated with other methods to increase the accuracy of protein-function prediction methods. It will be interesting to integrate DPA, cleft analysis, amino-acid conservation, and other types of information to make more accurate predictions about functional sites in proteins.

Ultimately, detailed examinations of changes in conformational distributions will be required for a complete mechanistic understanding of allosteric regulation. More generally, however, proteins whose activities are allosterically regulated must have conformational distributions that are susceptible to alteration by molecular interactions. Our results support this general observation. They also motivate a perspective in which naturally occurring protein folds are controllable designs with intrinsically preferred locations for functional sites. Specific residues in these sites provide different protein activities and target specificities, but the overall architecture of the protein dictates their preferred locations to optimize their coupling to protein dynamics. In this perspective, the

greater the potential for interactions at a site to change the conformational distribution, the more likely it is that the site will evolve as a locus for controlling protein activity.

In summary, the present evidence for the tendency of functional sites to be located at dynamical control points supports a scenario in which Nature favors regulatable proteins. It will be fascinating to see how this perspective evolves within the context of our deepening understanding of protein function and evolution.

Acknowledgement

We are grateful to William Bruno for suggesting the extreme value distribution as a model, and to members of the Protein Inference Group at Los Alamos for discussions. This work was supported by the Department of Energy.

Appendix

The probability density $\rho_z(z)$ of the product $z = \prod_{i=1}^{N} P_i$ of *N* uniformly distributed random variables P_i over the range [0,1] is given by

$$\rho_z(z) = \int_0^\infty dP_1 \theta(1-P_1) \int_0^\infty dP_2 \theta(1-P_2) \ldots \int_0^\infty dP_N \delta(z - P_1 \ldots P_N) \theta(1-P_N), \tag{A1}$$

where $\delta(x)$ is the Dirac delta function, and $\theta(x)$ is the unit step function: $\theta(x) = 0$, $x < 0$; $\theta(x) = 1$, $x \geq 0$. The delta function may be rewritten as

$$\delta(z - P_1 \ldots P_N) = \frac{\delta\left(P_N - \dfrac{z}{P_1 \ldots P_{N-1}}\right)}{P_1 \ldots P_{N-1}}, \tag{A2}$$

which, substituted into Eq. (A1), yields

$$\rho_z(z) = \int_0^\infty dP_1 \theta(1-P_1) \int_0^\infty dP_2 \theta(1-P_2) \ldots \int_0^\infty \frac{dP_{N-1}}{P_1 \ldots P_{N-1}} \theta\left(1 - \frac{z}{P_1 \ldots P_{N-1}}\right). \tag{A3}$$

The last step function in Eq. (A3) may be rewritten as

$$\theta\left(1 - \frac{z}{P_1 \ldots P_{N-1}}\right) = \theta(P_1 \ldots P_{N-1} - z) = \int_z^1 du_{N-1} \delta(u_{N-1} - P_1 \ldots P_{N-1}), \tag{A4}$$

which, through use of Eq. (A2) and substitution into Eq. (A3), yields

$$\rho_z(z) = \int_z^1 \frac{du_{N-1}}{u_{N-1}} \int_0^\infty dP_1 \theta(1-P_1) \int_0^\infty dP_2 \theta(1-P_2) \ldots \int_0^\infty \frac{dP_{N-2}}{P_1 \ldots P_{N-2}} \theta\left(1 - \frac{u_{N-1}}{P_1 \ldots P_{N-2}}\right). \tag{A5}$$

Repeated application of Eqs. (A4) and (A2) eventually yields

$$\rho_z(z) = \int_z^1 \frac{du_{N-1}}{u_{N-1}} \int_{u_{N-1}}^1 \frac{du_{N-2}}{u_{N-2}} \ldots \int_{u_2}^1 \frac{du_1}{u_1}, \tag{A6}$$

which, when integrated, yields

$$\rho_z(z) = \frac{(-\ln z)^{N-1}}{(N-1)!}. \qquad \text{(A7)}$$

The probability $P^-(z')$ that the product *z* is less than or equal to z' is then given by

$$P^-(z') = \int_0^{z'} dz \rho_z(z) = z' \sum_{n=1}^{N} \frac{(-\ln z')^{n-1}}{(n-1)!} \quad , \qquad \text{(A7)}$$

which is the P-value expression used in the text.

References

1. Koshland, D. E., Jr., Nemethy, G. & Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **5**, 365-85.

2. Monod, J., Wyman, J. & Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J Mol Biol* **12**, 88-118.

3. Weber, G. (1972). Ligand binding and internal equilibria in proteins. *Biochemistry* **11**, 864-78.

4. Austin, R. H., Beeson, K. W., Eisenstein, L., Frauenfelder, H. & Gunsalus, I. C. (1975). Dynamics of ligand binding to myoglobin. *Biochemistry* **14**, 5355-73.

5. Frauenfelder, H. & Wolynes, P. G. (1985). Rate theories and puzzles of hemeprotein kinetics. *Science* **229**, 337-45.

6. Ming, D. & Wall, M. E. (2005). Quantifying allosteric effects in proteins. *Proteins* **59**, 697-707.

7. Ming, D. & Wall, M. E. (2005). Allostery in a coarse-grained model of protein dynamics. *Phys Rev Lett* **95**, 198103.

8. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727-48.

9. Sanner, M. F., Olson, A. J. & Spehner, J. C. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305-20.

10. Tirion, M. M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters* **77**, 1905-1908.

11. Bahar, I., Atilgan, A. R. & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des* **2**, 173-81.

12. Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins* **33**, 417-29.

13. Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O. & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* **80**, 505-15.

14. Qian, H. (2001). Relative entropy: free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Phys Rev E Stat Nonlin Soft Matter Phys* **63**, 042103.

15. Zheng, W. & Brooks, B. R. (2005). Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: myosin versus kinesin. *Biophys J* **89**, 167-78.

16. Harata, K. & Muraki, M. (1997). X-ray structure of turkey-egg lysozyme complex with tri-N-acetylchitotriose. Lack of binding ability at subsite A. *Acta Crystallogr D Biol Crystallogr* **53**, 650-657.

17. Lu, G., Lindqvist, Y., Schneider, G., Dwivedi, U. & Campbell, W. (1995). Structural studies on corn nitrate reductase: refined structure of the cytochrome b reductase fragment at 2.5 A, its ADP complex and an active-site mutant and modeling of the cytochrome b domain. *J Mol Biol* **248**, 931-48.

18. Ankerst, M., Breunig, M. M., Kriegel, H. P. & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *Proceedings of the ACM SIGMON International Conference on Management of Data* **28**, 49-60.

19. Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **13**, 323-30, 307-8.

20. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci* **5**, 2438-52.

21. Betzel, C., Singh, T. P., Visanji, M., Peters, K., Fittkau, S., Saenger, W. & Wilson, K. S. (1993). Structure of the complex of proteinase K with a substrate analogue hexapeptide inhibitor at 2.2-Å resolution. *J Biol Chem* **268**, 15854-8.

22. Ravelli, R. B., Raves, M. L., Ren, Z., Bourgeois, D., Roth, M., Kroon, J., Silman, I. & Sussman, J. L. (1998). Static Laue diffraction studies on acetylcholinesterase. *Acta Crystallogr D Biol Crystallogr* **54**, 1359-66.

23. Lauble, H., Kennedy, M. C., Beinert, H. & Stout, C. D. (1992). Crystal structures of aconitase with isocitrate and nitroisocitrate bound. *Biochemistry* **31**, 2735-48.

24. Yang, L. W. & Bahar, I. (2005). Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure* **13**, 893-904.

25. Glaser, F., Morris, R. J., Najmanovich, R. J., Laskowski, R. A. & Thornton, J. M. (2006). A method for localizing ligand binding pockets in protein structures. *Proteins* **62**, 479-88.

26. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci* **20**, 374.

27. Lamers, M. B., Antson, A. A., Hubbard, R. E., Scott, R. K. & Williams, D. H. (1999). Structure of the protein tyrosine kinase domain of C-terminal Src kinase (CSK) in complex with staurosporine. *J Mol Biol* **285**, 713-25.

28. Weber, P. C., Pantoliano, M. W., Simons, D. M. & Salemme, F. R. (1994). Structure-based design of synthetic azobenzene ligands for streptavidin. *J Am Chem Soc* **116**, 2717-2724.

29. Bode, W., Walter, J., Huber, R., Wenzel, H. R. & Tschesche, H. (1984). The refined 2.2-Å (0.22-nm) X-ray crystal structure of the ternary complex formed by bovine trypsinogen, valine-valine and the Arg15 analogue of bovine pancreatic trypsin inhibitor. *Eur J Biochem* **144**, 185-90.

30. Li, J., Brick, P., O'Hare, M. C., Skarzynski, T., Lloyd, L. F., Curry, V. A., Clark, I. M., Bigg, H. F., Hazleman, B. L., Cawston, T. E. & et al. (1995). Structure of full-length porcine synovial collagenase reveals a C-terminal domain containing a calcium-linked, four-bladed beta-propeller. *Structure* **3**, 541-9.

Figure captions

Figure 1. Visualization of lysozyme surrounded by 553 surface points. Lysozyme is rendered as a yellow ribbon, and surface points are rendered as spheres temperature-coded according to the value $D_{\mathbf{x}}^{(m)}$. Values of $D_{\mathbf{x}}^{(m)}$ are elevated in the neighborhood of the binding site of the tri-NAG ligand (magenta wireframe). In this and other figures, RASMOL[26] was used to visualize molecular structures.

Figure 2. Distribution of $D_{\mathbf{x}}^{(m)}$ values (labeled as AP values) for 4859 points on the surface of lysozyme (the number of points was increased in this case to evaluate the fit). The distribution is well-fit by an extreme value distribution with parameters $\mu = 23.07$ and $\beta = 8.45$ (Pearson correlation coefficient of 0.992). The fit is used to find the 96% upper bound of $D_{\mathbf{x}}^{(m)}$ for the surface points; this bound is used as the threshold to select high-$D_{\mathbf{x}}^{(m)}$ points for use in predicting functional sites.

Figure 3. Statistical significance of elevated values of $D_{\mathbf{x}}^{(m)}$ in functional sites. The distribution of P-values P_k^- (calculated in bins of width 2 in log units) is shown for 291 proteins in the GOLD docking test set.

Figure 4. The distribution of $D_{\mathbf{x}}^{(m)}$ values for proteins in the test set is well-modeled using an extreme-value-distribution. Shown is a histogram of Pearson correlation coefficients calculated between the model and the data for all proteins in the test set. Also

shown is the histogram for just the subset of proteins for which ligand-binding-site predictions were made.

Figure 5. Illustration of DPA and cleft analysis applied to the tyrosine kinase domain of human C-terminal Src kinase (Protein Data Bank entry 1BYG[27]). In each panel, the protein is represented using yellow ribbons, and the residues in the neighborhood of the ligand are represented using magenta ribbons and wireframes. Individual panels show, in green coloring, (*Left*) a thick wireframe representation of the ligand; (*Center*) spheres centered on points in the rank-1 DPA cluster; and (*Right*) a thin wireframe representation of the surface surrounding the largest cleft. The results illustrated here are typical of other proteins in the test set (Figs. 6, 7): using DPA algorithm, the recall of the ligand-binding-site residues is 0.70, and the precision of the predicted residues is 0.42, while using the cleft analysis algorithm, the recall is 1.0 and the precision is 0.22.

Figure 6. Recall of the ligand-binding-site residues from the predicted residues. Results for 287 proteins in the test set for which the DPA algorithm produced predictions (solid line) are compared to results for 303 proteins in the test set for which the cleft analysis algorithm produced predictions (dashed line).

Figure 7. Precision of predicted residues with respect to the ligand-binding-site residues. Results for 287 proteins in the test set for which the DPA algorithm produced predictions (solid line) are compared to results for 303 proteins in the test set for which the cleft analysis algorithm produced predictions (dashed line).

Figure 8. Dependence of the DPA algorithm prediction performance on the threshold *C*. The fraction of proteins for which the predictions have at least 50% recall, and that at 50% precision, are plotted for values of *C* between 0.8 and 0.99. Also plotted is the fraction of proteins for which a prediction is made for given threshold; there are fewer predictions for higher values of *C*. The value $C = 0.96$ yields a relatively high precision with little cost in either the recall or the total prediction rate.

Figure 9. Statistical significance of the overlaps of predicted residues with ligand-binding-site residues. For each protein, a P-value (corresponding to the probability in a null model of finding at least as many ligand-binding-site resides as does the prediction algorithm) is calculated; the resulting distribution of P-values is shown here. For the DPA algorithm (solid line), a total of 250 proteins in the test set were considered; and for the cleft analysis algorithm (dashed line), a total of 278 proteins were considered. (In each case, only proteins for which the algorithm yielded at least one residue in the ligand-binding site were considered.)

Figure 10. Biotin-binding protein streptavidin (yellow ribbons, Protein Data Bank entry 1SRH[28]). The rank-2 cluster (blue points) is closely associated with the the 2-[(4'-hydroxyphenyl)-azo]benzoate ligand on one monomer (magenta wireframe), and the rank-1 cluster (green points) is near an equivalent binding site on the other monomer (the ligand at this site is absent in the GOLD test set but is present in the crystal structure). A similar result was found for Protein Data Bank entry 1SRF.[28]

Figure 11. Complex formed by bovine trypsinogen (yellow ribbon), bovine pancreatic trypsin inhibitor, (grey ribbon) and an Ile-Val ligand (magenta wireframe, Protein Data Bank entry 4TPI[29]). The rank-1 cluster (green points) is located near the interface between the protein and the inhibitor. The rank-2 cluster is located at the ligand-binding site.

Figure 12. Porcine synovial collagenase (Protein Data Bank entry 1FBL[30]). (*Left*) The N-terminal catalytic domain (yellow ribbon) is linked to a C-terminal domain (grey ribbon). The rank-1 cluster (green points) is located at the interface between the two domains. The rank-2 cluster (blue points) is associated with the ligand (magenta wireframe) in the catalytic domain. (*Right*) After isolating the catalytic domain, the rank-1 cluster (green points) is located at the ligand-binding site.

Figure 1

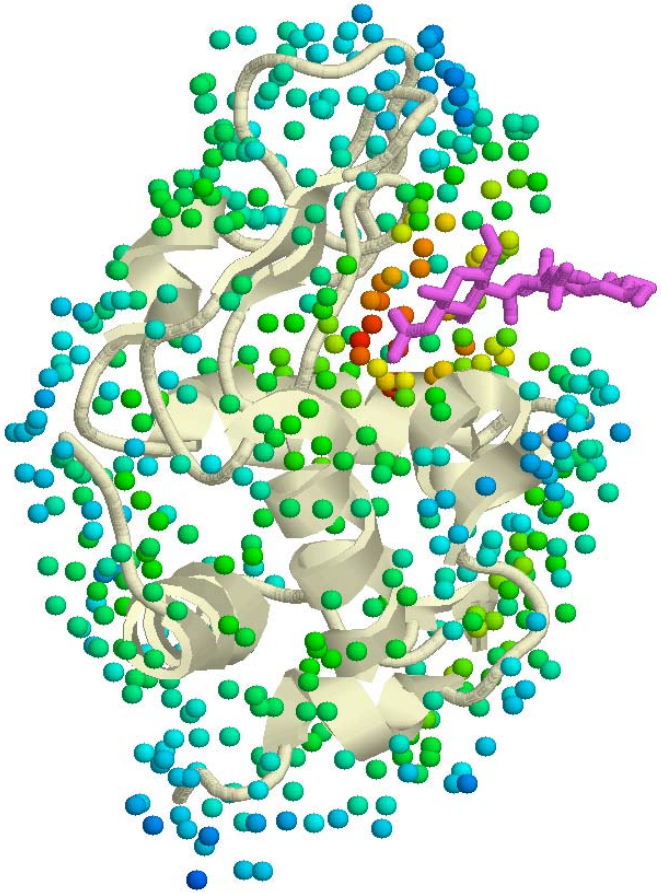

Figure 2

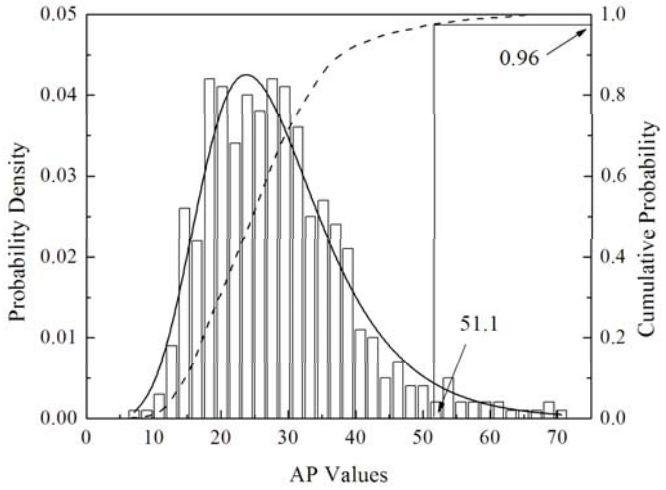
Probability Density
Cumulative Probability
AP Values
0.96
51.1
0.05
0.04
0.03
0.02
0.01
0.00
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70

Figure 3

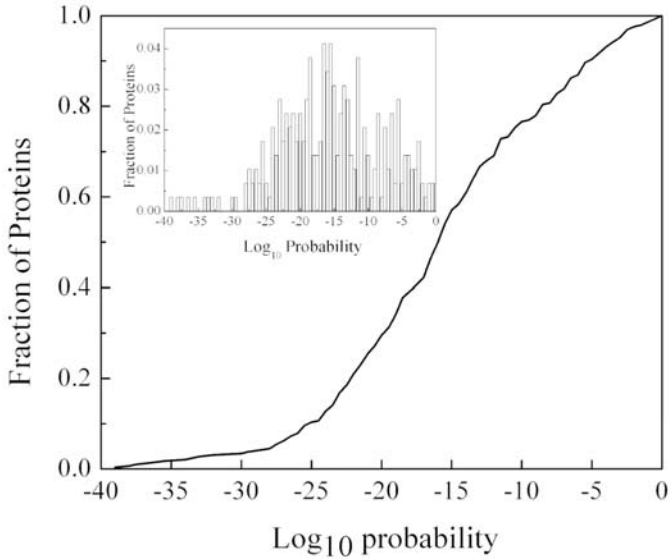
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of Proteins
-40
-35
-30
-25
-20
-15
-10
-5
0
Log10 probability
0.04
0.03
0.02
0.01
0.00
Fraction of Proteins
-40
-35
-30
-25
-20
-15
-10
-5
0
Log10 Probability

Figure 4

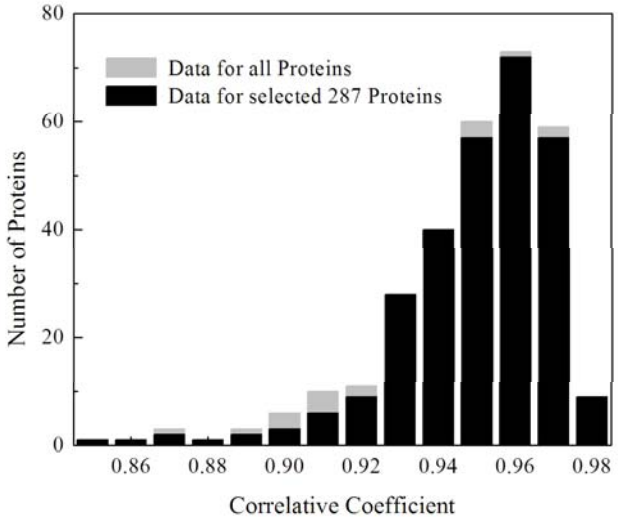
Data for all Proteins
Data for selected 287 Proteins
Number of Proteins
Correlative Coefficient
0
20
40
60
80
0.86
0.88
0.90
0.92
0.94
0.96
0.98

Figure 5

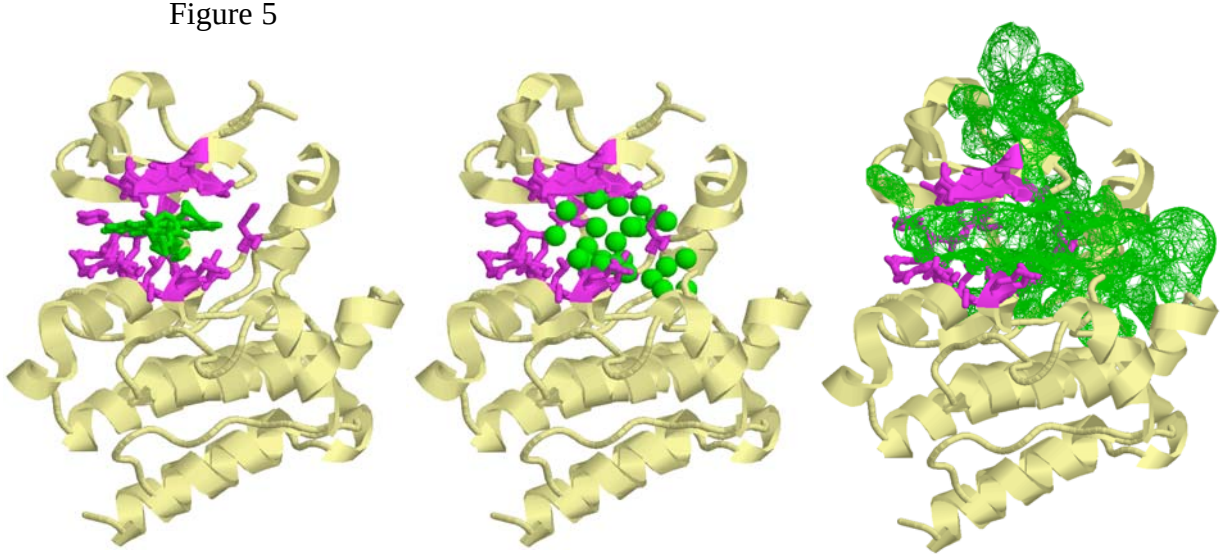

Figure 6

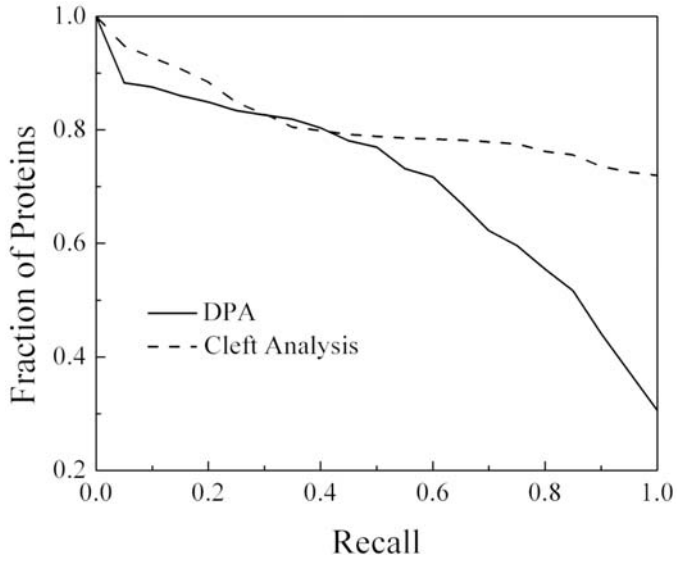
1.0
0.8
0.6
0.4
0.2
Fraction of Proteins
DPA
Cleft Analysis
0.0
0.2
0.4
0.6
0.8
1.0
Recall

Figure 7

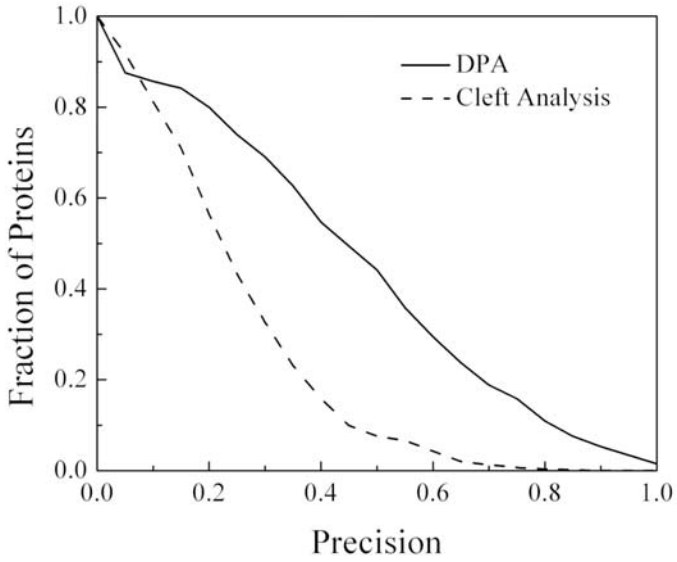
DPA
Cleft Analysis
Fraction of Proteins
Precision
0.0
0.2
0.4
0.6
0.8
1.0

Figure 8

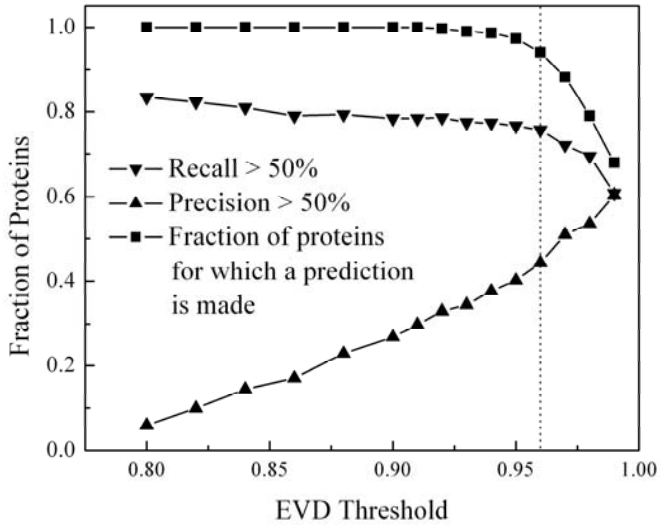
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of Proteins
Recall > 50%
Precision > 50%
Fraction of proteins
for which a prediction
is made
0.80
0.85
0.90
0.95
1.00
EVD Threshold

Figure 9

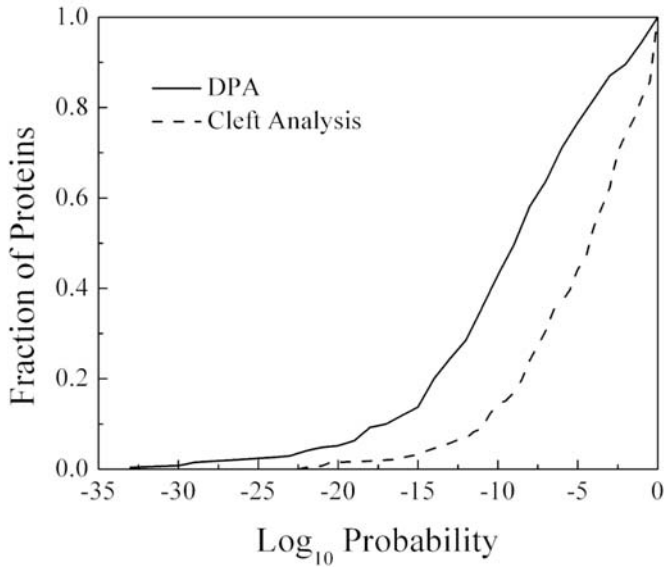
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of Proteins
DPA
Cleft Analysis
-35
-30
-25
-20
-15
-10
-5
0
Log10 Probability

Figure 10

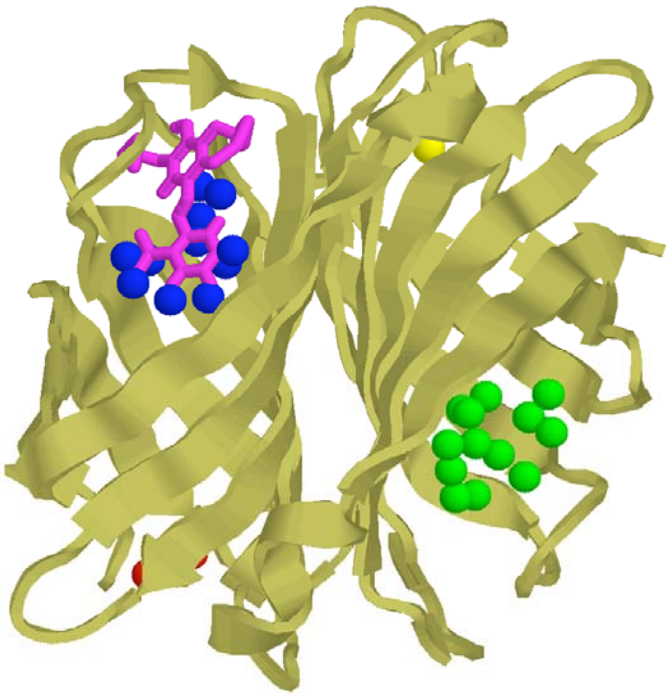

Figure 11

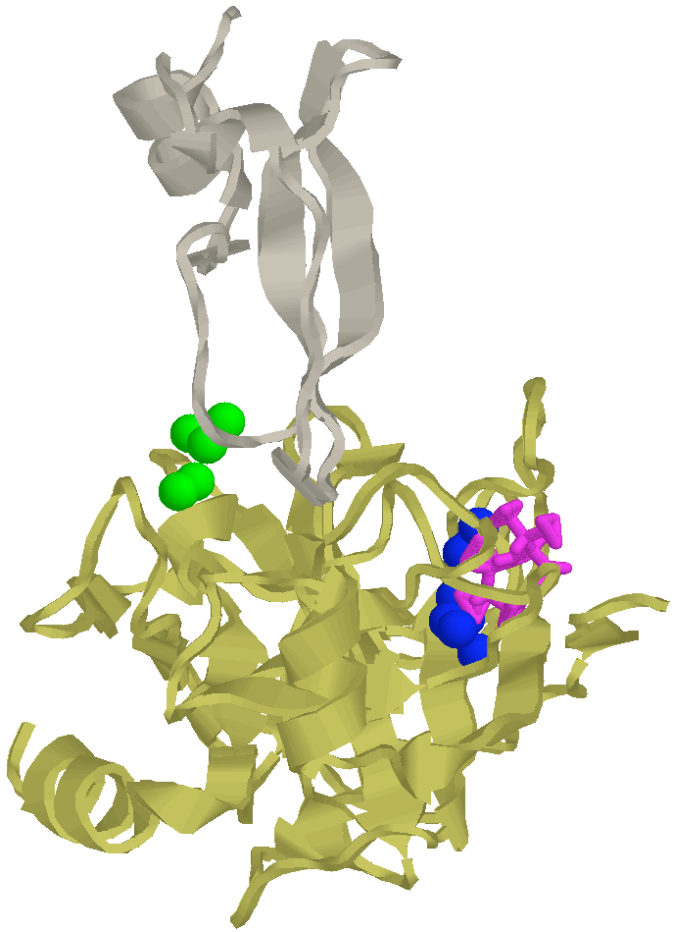

Figure 12