

# Fast dynamics perturbation analysis for prediction of protein functional sites

Dengming Ming<sup>1,2</sup>, Judith D. Cohn<sup>1,3</sup>, and Michael E. Wall<sup>\*1,3,4</sup>

<sup>1</sup>Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

<sup>2</sup>School of Life Sciences, Nanjing University, Nanjing, Jiangsu Province, China

<sup>3</sup>Bioscience Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

<sup>4</sup>Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

\*Corresponding author

Email: Dengming Ming – [dming@nju.edu.cn](mailto:dming@nju.edu.cn); Judith D. Cohn – [jcohn@lanl.gov](mailto:jcohn@lanl.gov); Michael E. Wall – [mewall@lanl.gov](mailto:mewall@lanl.gov)

## Abstract

**Background.** We present a fast version of the dynamics perturbation analysis (DPA) algorithm to predict functional sites in protein structures. The original DPA algorithm finds regions in proteins where interactions cause a large change in the protein conformational distribution, as measured using the relative entropy  $D_x$ . Such regions are associated with functional sites.

**Results.** The Fast DPA algorithm, which accelerates DPA calculations, is motivated by an empirical observation that  $D_x$  in a normal-modes model is highly correlated with an entropic term that only depends on the eigenvalues of the normal modes. The eigenvalues are accurately estimated using first-order perturbation theory, resulting in a  $N$ -fold reduction in the overall computational requirements of the algorithm, where  $N$  is the number of residues in the protein. The performance of the original and Fast DPA algorithms was compared using protein structures from a standard small-molecule docking test set. For nominal implementations of each algorithm, top-ranked Fast DPA predictions overlapped the true binding site 94% of the time, compared to 87% of the time for original DPA. In addition, per-protein recall statistics (fraction of binding-site residues that are among predicted residues) were slightly better for Fast DPA. On the other hand, per-protein precision statistics (fraction of predicted residues that are among binding-site residues) were slightly better using original DPA. Overall, the performance of Fast DPA in predicting ligand-binding-site residues was comparable to that of the original DPA algorithm.

**Conclusions.** Compared to the original DPA algorithm, the decreased run time with comparable performance makes Fast DPA well-suited for implementation on a web server and for high-throughput analysis.

## Background

Prediction of protein functional sites is a key aspect of protein function prediction [1], and can be an important step in identifying small-molecule interactions for drug discovery [2]. It can also potentially be used as a pre-processing step to reduce the search space in computational docking algorithms. There are many methods to predict functional sites—here we emphasize those that make use of analysis of protein structure and dynamics. Existing protein structure analysis methods are based on diverse principles, including: association of functional sites with surface clefts that have extreme values of volume [3-6] or other shape descriptors [7-11]; identifying spatial clusters of methyl probes that exhibit energetically favorable interactions with the protein [12]; association of functional sites with charged surface residues either in unfavorable electrostatic environments [13] or with anomalous predicted pH titration curves [14]; identifying spatial clusters of residues whose diversity appears to be correlated with changes in protein function [15, 16]; defining structural features (e.g. motifs) associated with functional sites [17-22]; identifying residues that are on average close to other residues in the protein (closeness centrality) [23-25]; and machine-learning prediction of functional sites/residues using sequence, structure, and chemical features from training sets [26-28]. Principles of methods that consider protein dynamics include association of functional sites with: hinge regions [29, 30]; regions where the harmonic vibrations are largely determined by high-frequency modes [31]; intrinsically disordered regions that are highly mobile in the absence of a molecular interaction partner [32]; and residues where mutations cause a large change in the couplings of local perturbations to remote, local changes in the distribution of folded vs. unfolded states of the protein [33]. Information from complementary methods may be integrated for functional site prediction [34, 35].

We recently developed an additional approach to prediction of protein functional sites that is based on analysis of protein dynamics [36-39]. To help motivate the approach, we note that cellular functions are regulated by molecular interactions that alter protein activity. To enable such control, protein activity, and therefore protein conformational distributions, must be susceptible to alteration by molecular interactions at functional sites. In other words, protein activity should be controllable by allosteric effects (allostery).

Weber [40] recognized the importance of considering changes in the *full conformational distribution* to understand allostery, as opposed to considering mechanistic changes among *discrete, well-defined structural states* in earlier models due to Monod, Wyman, and Changeux [41]; and Koshland, Nemethy, and Filmer [42]. Weber's perspective is well-aligned with more recent emphases on the need to consider allostery from a global thermodynamic/statistical perspective [33, 38, 39, 43-45]. It is also well-aligned with modern rate theories based on the control of protein activity by dynamical transitions among conformational substates [46], as originally suggested by spectroscopic assays of ligand-binding at low-temperature [47, 48].

Given the above considerations, we hypothesized that protein functional sites might tend to evolve at control points where interactions cause a large change in the protein conformational distribution [39]. To test this hypothesis, we developed a method called

dynamics perturbation analysis (DPA) to quantify changes in protein conformational distributions due to molecular interactions [38, 39], examined 305 protein structures from the GOLD docking test set [49], and found that interactions at small-molecule binding sites cause a relatively large change in protein vibrations [37].

Motivated by these results, we developed a DPA-based algorithm that successfully predicts small-molecule binding sites at locations where interactions cause a large change in protein vibrations [37]. This method was evaluated in Ref. [37] using 305 proteins in the GOLD docking test set of protein-ligand structures [49]. For the test, only the top-ranked functional site was selected and was used to predict the location of the ligand-binding site. This is a relatively strict requirement; in other published methods for predicting functional sites (see, e.g., Ref. [11]), performance often is evaluated by allowing for any of several predicted functional sites to overlap a known ligand-binding site. The method produced at least one predicted functional site for 287 of the 305 proteins in the test set. In 87% of cases (250 proteins), at least one predicted residue was in the ligand-binding site. The recall of binding-site residues (percentage of binding-site residues found among the predicted residues) was at least 30% for 80% of cases, and was at least 50% for 76% of the cases. The precision of the predicted residues (percentage of predicted residues found among the binding-site residues) was at least 30% for 68% of the cases, and was at least 50% for 44% of the cases. The statistical significance of the overlaps was assessed using a null model in which surface residues were randomly selected. Using the null model, a P-value was calculated to evaluate predictions for the 250 proteins in which at least one predicted residue was in the ligand-binding site. The P-value estimated the probability of obtaining a precision at least as high as the observed precision by randomly selecting surface residues (see Ref. [37] for details). For 87% of the cases, the P-value was  $10^{-3}$  or smaller, indicating a statistically significant overlap. The performance of the DPA method compared favorably to that of a cleft analysis method for predicting ligand-binding residues.

The original DPA algorithm is a highly innovative approach that performs well. However, the computational requirements limit the utility of the original method. For example, it takes about an hour to analyze a 150-residue protein domain using DPA, and the method doesn't scale well to larger systems. Here, we report an improved algorithm based on use of first-order perturbation theory that will facilitate the use of DPA in high-throughput scenarios and increase its utility, e.g., for web server applications. The algorithm, called Fast DPA, enables a dramatic decrease in the time required to predict protein functional sites, with performance that is comparable to the original DPA algorithm.

## Methods

### *Dynamics perturbation analysis*

Our overall approach for predicting functional sites is based on a method called dynamics perturbation analysis (DPA) [37, 39]. In DPA, a protein is decorated with  $M$  surface points that interact with neighboring protein atoms, as illustrated for Protein Data Bank entry 1JEF [50] in Fig. 1. The protein conformational distribution  $P(\mathbf{x})$  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  $m$ . The conformational distributions are calculated using a coarse-grained model of molecular vibrations, and the distributions  $P^{(m)}(\mathbf{x})$  are calculated from models of the protein in complex with each surface point. The relative entropy, or Kullback-Leibler divergence [51],  $D_{\mathbf{x}}^{(m)}$  between  $P(\mathbf{x})$  and  $P^{(m)}(\mathbf{x})$  is calculated for each point  $m$ , and is used as a measure of the change in the protein conformational distribution upon interacting with point  $m$ :

$$D_{\mathbf{x}}^{(m)} = \int d^{3N} \mathbf{x} P^{(m)}(\mathbf{x}) \ln \frac{P^{(m)}(\mathbf{x})}{P(\mathbf{x})} \quad (1)$$

In the present case (unlike in other useful biological applications [52-56]), the relative entropy is not just an *ad hoc* measure; rather, it has real biophysical significance [36, 57]:  $k_B T D_{\mathbf{x}}^{(m)}$ , where  $T$  is the temperature and  $k_B$  is Boltzmann's constant, is the free energy required to change the protein conformational distribution from an equilibrium distribution  $P(\mathbf{x})$  to a non-equilibrium distribution  $P^{(m)}(\mathbf{x})$ .

Thus far, DPA calculations have most often been performed using a simple model of protein vibrations—the elastic network model (ENM) [58-61]. In the ENM,  $C_{\alpha}$  atoms are extracted from an atomic model of a protein, and an interaction network is generated by connecting springs between all atom pairs  $(i, j)$  separated by a distance less than or equal to a cutoff distance  $r_c$ . Each spring has the same force constant  $\gamma$ , is aligned with the separation between the connected atoms, and has an equilibrium length equal to the distance  $d_{ij}$  between the atoms in the initial model. Thus, the potential energy is given by  $U(\mathbf{x}) = \gamma/2 \sum_{i>j} \varepsilon_{ij} (\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij})^2$ , where  $\varepsilon_{ij}=1$  if atoms  $i$  and  $j$  are connected, and  $\varepsilon_{ij}=0$  otherwise. The interaction between the protein and a surface point  $m$  is modeled by connecting springs of force constant  $\gamma_s$  between the surface point and all protein atoms within a cutoff distance  $r_s$  of the surface point. The protein coordinates are not modified in modeling the interaction. The dynamics are defined using normal mode analysis of the model. In this model, the reference distribution  $P(\mathbf{x})$  is given by

$$P(\mathbf{x}) = \prod_{i=1}^{3N, \lambda_i \neq 0} \left( \frac{\lambda_i}{2\pi k_B T} \right)^{1/2} e^{-\frac{1}{2k_B T} \lambda_i |(\mathbf{x} - \mathbf{x}_0) \cdot \mathbf{v}_i|^2} \quad (2)$$

In Eq. (2),  $N$  is the number of atoms in the protein;  $\mathbf{x}_0$  is the equilibrium structure; and  $\lambda_i$  and  $\mathbf{v}_i$  are the  $i^{\text{th}}$  eigenvalue and eigenvector of the Hessian  $\mathbf{H}$ :  $h_{ij} = \partial^2 U / \partial x_i \partial x_j \big|_{\mathbf{x}_0}$ . The perturbed distribution  $P^{(m)}(\mathbf{x})$  is similar to Eq. (2), but substituting the eigenvalues and eigenvectors  $\bar{\lambda}_i^{(m)}$  and  $\bar{\mathbf{v}}_i^{(m)}$  of the pseudo-Hessian  $\bar{\mathbf{H}}^{(m)}$  for  $\lambda_i$  and  $\mathbf{v}_i$ .  $\bar{\mathbf{H}}^{(m)}$  is derived from the full Hessian  $\mathbf{H}^{(m)}$  for the protein model in the presence of the surface point  $m$ :

$$\mathbf{H}^{(m)} = \begin{pmatrix} \mathbf{H}_P^{(m)} & \mathbf{G}^{(m)} \\ \mathbf{G}^{(m)T} & \mathbf{H}_S^{(m)} \end{pmatrix}. \quad (3)$$

The sub-matrix  $\mathbf{H}_p^{(m)}$  couples the protein coordinates, the sub-matrix  $\mathbf{H}_s^{(m)}$  couples the test-point coordinates, and the submatrix  $\mathbf{G}^{(m)}$  couples the protein to the test point. In terms of these matrices,  $\bar{\mathbf{H}}^{(m)}$  is given by [38]

$$\bar{\mathbf{H}}^{(m)} = \mathbf{H}_p^{(m)} - \mathbf{G}^{(m)} \mathbf{H}_s^{(m)-1} \mathbf{G}^{(m)T}. \quad (4)$$

Using expressions for  $P(\mathbf{x})$  and  $P^{(m)}(\mathbf{x})$ , Eq. (1) becomes [38, 39]

$$D_{\mathbf{x}}^{(m)} = \frac{1}{2} \sum_{i=7}^{3N} \left( \log \frac{\bar{\lambda}_i^{(m)}}{\lambda_i} + \sum_{j=7}^{3N} \frac{\lambda_j}{\bar{\lambda}_i^{(m)}} |\bar{\mathbf{v}}_i^{(m)} \cdot \mathbf{v}_j|^2 - 1 \right). \quad (5)$$

The first six modes involve zero eigenvalues and are ignored in the sums. Equation (5) is the central equation that enables DPA.

To use DPA to predict functional sites, we make use of the fact that, empirically, the distribution of  $y = D_{\mathbf{x}}^{(m)}$  values on the surface of a protein calculated using Eq. (5) is observed to obey an extreme value distribution (Fig. 2),

$$\rho(y) = \frac{1}{\beta} e^{\frac{y-\mu}{\beta} - e^{\frac{y-\mu}{\beta}}}. \quad (6)$$

First, DPA is performed on a protein and the distribution of  $D_{\mathbf{x}}^{(m)}$  values is modeled using Eq. (6). Points with  $D_{\mathbf{x}}^{(m)}$  values in the upper 96% of the modeled distribution are selected and are spatially clustered. The clusters are ranked according to the mean value of  $D_{\mathbf{x}}^{(m)}$  within the cluster, and all clusters are considered to be potentially associated with a functional site. Finally, residues in the neighborhood of the clusters are selected and form the basis for functional site predictions.

### *Fast dynamics perturbation analysis*

Fast DPA is based on a simple empirical observation: for dynamics defined by normal modes, the total value of  $D_{\mathbf{x}}$  in Eq. (5) is highly correlated with just the first (entropic) term,

$$D_{\mathbf{x}}^{\lambda, (m)} = \frac{1}{2} \sum_{i=7}^{3N} \log \frac{\bar{\lambda}_i^{(m)}}{\lambda_i}. \quad (7)$$

Hereafter we refer to  $D_{\mathbf{x}}^{\lambda, (m)}$  simply as  $D_{\mathbf{x}}^{\lambda}$ . Observation of this correlation motivates the use of  $D_{\mathbf{x}}^{\lambda}$  as a surrogate for  $D_{\mathbf{x}}$  in DPA, and, because  $D_{\mathbf{x}}^{\lambda}$  only involves eigenvalues, creates an avenue for accelerating DPA. The acceleration arises because the eigenvalues of the normal modes of the protein in the presence of test points are well-approximated using first order perturbation theory. In this approximation, the pseudo-Hessian  $\bar{\mathbf{H}}^{(m)}$  of the protein in the presence of point  $m$  is written as the Hessian  $\mathbf{H}$  of the protein in the absence of the ligand plus a perturbation term  $\delta\bar{\mathbf{H}}^{(m)}$ :

$$\bar{\mathbf{H}}^{(m)} = \mathbf{H} + \delta\bar{\mathbf{H}}^{(m)}, \quad (8)$$

where the expression for  $\bar{\mathbf{H}}^{(m)}$  is as in previous studies [37, 38]. To estimate the eigenvalues of  $\bar{\mathbf{H}}^{(m)}$ , we use the canonical first-order perturbation theory expression,

$$\lambda_i^{(m)} \approx \lambda_i + \mathbf{v}_i^T \delta \bar{\mathbf{H}}^{(m)} \mathbf{v}_i, \quad (9)$$

where  $\lambda_i$  is the  $i$ th eigenvalue of  $\mathbf{H}$ .

The Fast DPA algorithm is the same as the original DPA algorithm, except instead of using values of  $D_{\mathbf{x}}$ , the analysis is based on values of  $D_{\mathbf{x}}^\lambda$  estimated using perturbation theory. (It is possible to evaluate all terms in Eq. (5) using first-order perturbation theory, but doing so would not accelerate the method because the computational cost is comparable to that of solving the full eigenvalue problem in original DPA.)

### *Implementation of Fast DPA*

Our implementation of DPA and Fast DPA here follows our previous implementation of DPA for functional site prediction [37]. Given an input PDB structure, MSMS [62] was run with a 1.5 Å probe radius and a triangulation density of 1 vertex per Å<sup>2</sup> to generate test points on the surface of the protein. As when using original DPA to predict functional sites, perturbations were calculated using every other point in the MSMS output (we also tried using every point, but this led to decreased performance in the precision measures). The cutoff  $r_c$  for interactions between protein C $_{\alpha}$  atoms was 8.5 Å. For some proteins, this cutoff yielded more than six zero-frequency modes, indicating that the network of springs was too sparse (for example, if only one spring connects two domains, then free rotations about the spring yield two additional zero-frequency modes). In these cases, the connectivity of the elastic network model was increased by incrementing  $r_c$  in 1 Å steps until the additional zero-frequency modes were eliminated. The cutoff  $r_s$  for interactions between a test point and the protein was 14 Å, and the interaction strength between a test point and protein atoms was  $\gamma_s = 12\gamma$ , or 12 times the strength of the interaction between two protein atoms. Results are independent of the value of  $\gamma$ .

### *Implementation of functional site prediction using DPA*

To predict functional sites, the distribution of  $y = D_{\mathbf{x}}^{(m)}$  values was fit using Eq. (6). Points with  $D_{\mathbf{x}}^{(m)}$  values in the upper 96% of the distribution were selected and spatially clustered using the OPTICS algorithm [63] with a distance threshold of 6 Å and a minimum of 3 points per cluster. C $_{\alpha}$  atoms within 6 Å of any point in a cluster were selected and were used to define predicted functional sites. The sites were ranked according to the mean value of  $D_{\mathbf{x}}^{(m)}$  within the corresponding cluster of points. Only the top-ranked predicted site was used for the evaluation of performance described below.

## **Results and discussion**

### *Results that motivate Fast DPA*

To motivate the use of  $D_{\mathbf{x}}^\lambda$  instead of  $D_{\mathbf{x}}$  for DPA, we analyzed proteins from the GOLD test set. We found that  $D_{\mathbf{x}}$  is highly correlated with  $D_{\mathbf{x}}^\lambda$  for these cases; Fig. 3 illustrates the agreement for four proteins. This is not a trivial result mathematically (see Eqs. (5) and (7))—it means that  $\sum \log(\bar{\lambda}_i^{(m)}/\lambda_i)$  is highly correlated with  $\sum_i \sum_j |\bar{\mathbf{v}}_i^{(m)} \cdot \mathbf{v}_j|^2 \lambda_j / \bar{\lambda}_i^{(m)}$ .

To motivate the use of perturbation theory to estimate  $D_x^\lambda$ , we compared the true eigenvalues to those estimated using perturbation theory for proteins in the GOLD test set. Because in our model the strength of the spring that connects the test points to the protein is 12 times the strength of the spring that connects protein atoms to each other (Methods), it was not obvious that first-order perturbation theory would yield reasonable estimates of eigenvalues. However, we had hoped for success based on the fact that we were only adding a single test point to the model, compared to, typically,  $o(100)$  protein  $C_\alpha$  atoms. As illustrated for lysozyme in Fig. 4, we did find that Eq. (9) approximates well the true eigenvalues obtained by diagonalization of  $\mathbf{H}^{(m)}$ . Finally, we found that  $D_x$  calculated using original DPA was highly correlated with  $D_x^\lambda$  calculated using Fast DPA, as illustrated for four proteins in Fig. 5.

### *Evaluation of Fast DPA for prediction of functional sites*

The above results motivated us to develop the Fast DPA algorithm for prediction of protein functional sites (Methods). Through use of first-order perturbation theory, Fast DPA replaces matrix diagonalization by matrix-vector multiplication for each test point (Eq. (9)). Because matrix diagonalization requires  $o(N^3)$  operations, and matrix-vector multiplication requires  $o(N^2)$  operations, we expected Fast DPA to run  $N$ -fold faster than the original DPA. We found this to be the case (Fig. 6): the original DPA scales roughly as  $N^{3.45}$ , while fast DPA scales roughly as  $N^{2.29}$ , yielding a factor of  $N^{1.16}$  decrease in the time required to perform Fast DPA vs. DPA (here,  $N$  is the number of residues in the protein).

Because  $D_x$  calculated using original DPA and  $D_x^\lambda$  calculated using Fast DPA are highly correlated (Fig. 5), we expected the performance of Fast DPA in predicting functional site residues to be comparable to that of the original DPA. We analyzed the performance of the algorithm on the 305-protein GOLD test set [49], which was used to evaluate the original DPA algorithm [37]. Each prediction has an associated recall (fraction of residues in the binding site that are among those in the rank-1 prediction) and precision (fraction of rank-1 predicted residues that are among those in the binding site). To evaluate performance statistically, we use (1) the fraction of binding sites for which the recall is greater than or equal to a minimum value, and (2) the fraction of fraction of rank-1 predictions for which the precision is greater than or equal to a minimum value.

Figure 7 compares the performance of Fast DPA using different thresholds of the extreme value distribution, and is equivalent to Fig. 8 in Ref. [37]. The nominal threshold of 0.96 indicated in this figure is equivalent to that chosen for original DPA. Fig. 8 compares the performance of Fast DPA with original DPA for different thresholds. When the threshold is 0.96 or smaller, the recall statistics of Fast DPA tend to be better, and the precision statistics of original DPA tend to be better. When the threshold is 0.97 or higher, original DPA outperforms Fast DPA in both precision and recall statistics.

At the nominal threshold value of 0.96, the performance of Fast DPA is comparable to that of original DPA. At this threshold, original DPA yielded 287 rank-1 predictions for the test set (rate of 94%), whereas Fast DPA yielded 267 rank-1 predictions (rate of 86%) (Table 1). However, Fast DPA makes 251 predictions that have at least one residue that

overlaps the binding site, while original DPA makes 250 such predictions, yielding a higher rate of locating binding sites for rank-1 Fast DPA predictions (94%) than for original DPA (87%) (Table 1). The recall statistics tend to be a bit better for Fast DPA (Table 1, Fig. 9), and the precision statistics tend to be better for original DPA (Table 1, Fig. 10).

## **Conclusions**

Use of Fast DPA enables functional site predictions to be performed  $N$ -fold faster than original DPA, with comparable performance in predicting residues in functional sites. The acceleration will facilitate optimization of Fast DPA for functional site predictions. Calculations that once took hours using DPA now may be performed in a matter of minutes, making practical the use of DPA via a web server. Indeed, high-throughput analysis using Fast DPA has already produced over 60,000 predicted functional sites for about 50,000 protein domains in the SCOP database [64] (J.D. Cohn, D. Ming, and M.E. Wall, in preparation). These predictions will provide a rich source of information for developing hypotheses concerning mechanisms of protein function.

## **Authors' contributions**

DM implemented the Fast DPA algorithm, tested its performance, and helped to draft the manuscript. JC provided assistance with databases and automation. MW conceived of the study, coordinated the work, and drafted the manuscript. All authors read and approved the final manuscript.

## **Acknowledgments**

Supported by the US Department of Energy through contract DE-AC52-06NA25396. We thank James Faeder for reading the manuscript.

## Figure legends

**Figure 1.** Application of Dynamics Perturbation Analysis (DPA) to predict protein functional sites. *Left.* In this example, the surface of lysozyme (PDB entry 1JEF [50], yellow cartoon) is decorated with test points (533 spheres at a density of 1 point per  $\text{\AA}^2$ ), and the degree to which the test points individually perturb the protein conformational distribution is calculated (temperature-coded coloring of the spheres). A tri-NAG molecule (purple wireframe) binds in the active site. Warm-colored spheres indicate where the perturbation is large. *Center.* Points where the perturbation is largest are selected and clustered (green spheres). *Right.*  $C_\alpha$  atoms within 6  $\text{\AA}$  of the DPA cluster are selected, and the associated residues define the predicted functional site (16 residues). For comparison,  $C_\alpha$  atoms within 6  $\text{\AA}$  of the tri-NAG are selected; we use the associated residues to define the actual functional site (7 residues). The overlapping residues (6 residues) are shown in orange; there are 10 predicted residues that do not exactly match the functional site (green), and there is 1 functional site residue that is not among the predicted residues (purple, in the helix on the right hand side).

**Figure 2.** Distribution of  $D_x^{(m)}$  values for 4859 points on the surface of lysozyme 1JEF (the number of points was increased in this case to evaluate the fit). The distribution is well-fit by an extreme value distribution (Eq. (6)) with parameters  $\mu = 23.07$  and  $\beta = 8.45$  (solid line). By examining the cumulative distribution (dashed line), the fit is used to find surface points that lie within the upper 96% of the distribution; these points are used to predict functional sites.

**Figure 3.** Values of  $D_x$  (y-axis) and  $D_x^\lambda$  (x-axis) calculated using original DPA are plotted for four PDB entries (values of the Pearson correlation,  $C$ , between the two, are listed here parenthetically): a) 1AEC, from an actinidin-E-64 complex [65] ( $C = 0.988$ ); b) 1FKI, from a FKBP complex [66] (0.989); c) 1JEF, from a lysozyme complex [50] (0.992); and d) 1STP, from a biotin complex [67] (0.989).

**Figure 4.** Eigenvalues (used for calculation of  $D_x^\lambda$ ) that are estimated using perturbation theory (filled triangles) are a good approximation to the true eigenvalues of a lysozyme elastic network model (open circles).

**Figure 5.** Values of  $D_x$  calculated using original DPA (y-axis) and  $D_x^\lambda$  calculated using Fast DPA (x-axis) are plotted for four PDB entries (values of the Pearson correlation between the two are listed here parenthetically): a) 1AEC (0.981); b) 1FKI (0.982); c) 1JEF (0.981); d) 1STP (0.980).

**Figure 6.** Comparison of run times for DPA (upwards-pointing triangles) vs. Fast DPA (downwards-pointing triangles) for various protein sizes. The inset shows the ratio of run times for various protein sizes.

**Figure 7.** Comparison of Fast DPA performance using different thresholds of the extreme value distribution (Eq. (6)). The y-axis is either the fraction of proteins for which

a prediction is made (squares), the fraction of binding sites with a recall of at least 0.5 (circles), or the fraction of predictions with a precision of at least 0.5 (triangles). The threshold is indicated on the x-axis; the 0.96 threshold used for Figs. 9 and 10 is indicated using a vertical dashed line.

**Figure 8.** Comparison of Fast DPA vs. original DPA precision and recall statistics at different thresholds of the extreme value distribution (Eq. (6)). The curves are similar to precision-recall curves: the y-axis is the fraction of binding sites with a recall of at least 0.5, and the x-axis is the fraction of binding sites with a precision of at least 0.5. Fast DPA values are indicated using open squares, and original DPA is indicated using filled squares. Points corresponding to a threshold of 0.96 are indicated using arrows.

**Figure 9.** Comparison of recall of binding-site residues using DPA vs. Fast DPA for 287 (number of predictions using DPA) or 267 (number of predictions using Fast DPA) proteins in the 305-protein GOLD test set. The y-axis indicates the fraction of proteins with a recall at least as high as the value on the x-axis (y-values should be read from the top of each step).

**Figure 10.** Comparison of precision of predicted residues using DPA vs. Fast DPA (see also Fig. 9). The y-axis indicates the fraction of proteins with a precision at least as high as the value on the x-axis (y-values should be read from the top of each step).

## Tables

**Table 1.** Performance statistics for Fast DPA and original DPA using a threshold of 0.96.

	Rank-1 Predictions <sup>a</sup>	Any match <sup>b</sup>	Recall $\geq 0.3^c$	Precision $\geq 0.3^d$	Recall $\geq 0.5^c$	Precision $\geq 0.5^d$
Original	287	0.87	0.80	0.68	0.76	0.44
Fast	267	0.94	0.86	0.65	0.75	0.38

<sup>a</sup> Number of proteins for which at least one DPA cluster was produced, out of 305 total.

<sup>b</sup> Fraction of rank-1 predictions that have at least one overlapping residue with the binding site.

<sup>c</sup> Fraction of binding sites for which the recall was at least 0.3 or 0.5.

<sup>d</sup> Fraction of predictions for which the precision was at least 0.3 or 0.5.

## References

1. Ofra Y, Punta M, Schneider R, Rost B: **Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery.** *Drug Discov Today* 2005, **10**(21):1475-1482.
2. Campbell SJ, Gold ND, Jackson RM, Westhead DR: **Ligand binding: functional site location, similarity and docking.** *Curr Opin Struct Biol* 2003, **13**(3):389-395.
3. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J: **CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues.** *Nucleic Acids Res* 2006, **34**(Web Server issue):W116-118.
4. Hendlich M, Rippmann F, Barnickel G: **LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.** *J Mol Graph Model* 1997, **15**(6):359-363, 389.
5. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM: **A method for localizing ligand binding pockets in protein structures.** *Proteins* 2006, **62**(2):479-488.
6. Laskowski RA: **SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.** *J Mol Graph* 1995, **13**(5):323-330, 307-328.
7. Coleman RG, Burr MA, Souvaine DL, Cheng AC: **An intuitive approach to measuring protein surface curvature.** *Proteins* 2005, **61**(4):1068-1074.
8. Coleman RG, Sharp KA: **Travel depth, a new shape descriptor for macromolecules: application to ligand binding.** *J Mol Biol* 2006, **362**(3):441-458.
9. Hendrix DK, Kuntz ID: **Surface solid angle-based site points for molecular docking.** *Pac Symp Biocomput* 1998:317-326.
10. Nayal M, Honig B: **On the nature of cavities on protein surfaces: application to the identification of drug-binding sites.** *Proteins* 2006, **63**(4):892-906.
11. Xie L, Bourne PE: **A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites.** *BMC Bioinformatics* 2007, **8 Suppl 4**:S9.
12. Laurie AT, Jackson RM: **Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.** *Bioinformatics* 2005, **21**(9):1908-1916.

13. Elcock AH: **Prediction of functionally important residues based solely on the computed energetics of protein structure.** *J Mol Biol* 2001, **312**(4):885-896.
14. Ondrechen MJ, Clifton JG, Ringe D: **THEMATICS: a simple computational predictor of enzyme function from structure.** *Proc Natl Acad Sci U S A* 2001, **98**(22):12473-12478.
15. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257**(2):342-358.
16. Yao H, Mihalek I, Lichtarge O: **Rank information: a structure-independent measure of evolutionary trace quality that improves identification of protein functional sites.** *Proteins* 2006, **65**(1):111-123.
17. Wallace AC, Borkakoti N, Thornton JM: **TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites.** *Protein Sci* 1997, **6**(11):2308-2323.
18. Shulman-Peleg A, Nussinov R, Wolfson HJ: **Recognition of functional sites in protein structures.** *J Mol Biol* 2004, **339**(3):607-633.
19. Stark A, Russell RB: **Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures.** *Nucleic Acids Res* 2003, **31**(13):3341-3344.
20. Stark A, Shkumatov A, Russell RB: **Finding functional sites in structural genomics proteins.** *Structure* 2004, **12**(8):1405-1412.
21. Liang MP, Brutlag DL, Altman RB: **Automated construction of structural motifs for predicting functional sites on protein structures.** *Pac Symp Biocomput* 2003:204-215.
22. Barker JA, Thornton JM: **An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis.** *Bioinformatics* 2003, **19**(13):1644-1649.
23. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanelly D, Venger I, Pietrokovski S: **Network analysis of protein structures identifies functional residues.** *J Mol Biol* 2004, **344**(4):1135-1146.
24. Thibert B, Bredesen DE, del Rio G: **Improved prediction of critical residues for protein function based on network and phylogenetic analyses.** *BMC Bioinformatics* 2005, **6**:213.
25. Chea E, Livesay DR: **How accurate and statistically robust are catalytic site predictions based on closeness centrality?** *BMC Bioinformatics* 2007, **8**:153.

26. Ofra Y, Rost B: **ISIS: interaction sites identified from sequence.** *Bioinformatics* 2007, **23**(2):e13-16.
27. Gutteridge A, Bartlett GJ, Thornton JM: **Using a neural network and spatial clustering to predict the location of active sites in enzymes.** *J Mol Biol* 2003, **330**(4):719-734.
28. Wei L, Altman RB: **Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function.** *J Bioinform Comput Biol* 2003, **1**(1):119-138.
29. Ma B, Wolfson HJ, Nussinov R: **Protein functional epitopes: hot spots, dynamics and combinatorial libraries.** *Curr Opin Struct Biol* 2001, **11**(3):364-369.
30. Yang LW, Bahar I: **Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes.** *Structure* 2005, **13**(6):893-904.
31. Haliloglu T, Keskin O, Ma B, Nussinov R: **How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues.** *Biophys J* 2005, **88**(3):1552-1559.
32. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK: **Intrinsic disorder and functional proteomics.** *Biophys J* 2007, **92**(5):1439-1456.
33. Liu T, Whitten ST, Hilser VJ: **Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble.** *Proc Natl Acad Sci U S A* 2007, **104**(11):4347-4352.
34. Rossi A, Marti-Renom MA, Sali A: **Localization of binding sites in protein structures by optimization of a composite scoring function.** *Protein Sci* 2006, **15**(10):2366-2380.
35. Petrova NV, Wu CH: **Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties.** *BMC Bioinformatics* 2006, **7**:312.
36. Wall ME: **Ligand binding, protein fluctuations, and allosteric free energy.** *AIP Conf Proc* 2006, **851**:16-33.
37. Ming D, Wall ME: **Interactions in native binding sites cause a large change in protein dynamics.** *J Mol Biol* 2006, **358**:213-223.
38. Ming D, Wall ME: **Allostery in a coarse-grained model of protein dynamics.** *Phys Rev Lett* 2005, **95**:198103.

39. Ming D, Wall ME: **Quantifying allosteric effects in proteins.** *Proteins* 2005, **59**(4):697-707.
40. Weber G: **Ligand binding and internal equilibria in proteins.** *Biochemistry* 1972, **11**(5):864-878.
41. Monod J, Wyman J, Changeux JP: **On the nature of allosteric transitions: a plausible model.** *J Mol Biol* 1965, **12**:88-118.
42. Koshland DE, Jr., Nemethy G, Filmer D: **Comparison of experimental binding data and theoretical models in proteins containing subunits.** *Biochemistry* 1966, **5**(1):365-385.
43. Hilser VJ, Thompson EB: **Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins.** *Proc Natl Acad Sci U S A* 2007, **104**(20):8311-8315.
44. Pan H, Lee JC, Hilser VJ: **Binding sites in Escherichia coli dihydrofolate reductase communicate by modulating the conformational ensemble.** *Proc Natl Acad Sci U S A* 2000, **97**(22):12020-12025.
45. Gunasekaran K, Ma B, Nussinov R: **Is allostery an intrinsic property of all dynamic proteins?** *Proteins* 2004, **57**:433-443.
46. Frauenfelder H, Wolynes PG: **Rate theories and puzzles of heme protein kinetics.** *Science* 1985, **229**(4711):337-345.
47. Austin RH, Beeson K, Eisenstein L, Frauenfelder H, Gunsalus IC, Marshall VP: **Dynamics of carbon monoxide binding by heme proteins.** *Science* 1973, **181**(99):541-543.
48. Austin RH, Beeson KW, Eisenstein L, Frauenfelder H, Gunsalus IC: **Dynamics of ligand binding to myoglobin.** *Biochemistry* 1975, **14**(24):5355-5373.
49. Jones G, Willett P, Glen RC, Leach AR, Taylor R: **Development and validation of a genetic algorithm for flexible docking.** *J Mol Biol* 1997, **267**(3):727-748.
50. Harata K, Muraki M: **X-ray structure of turkey-egg lysozyme complex with tri-N-acetylchitotriose. Lack of binding ability at subsite A.** *Acta Crystallogr D Biol Crystallogr* 1997, **53**(Pt 6):650-657.
51. Kullback S, Leibler RA: **On information and sufficiency.** *Annals of Math Stats* 1951, **22**:79-86.
52. Bhasi K, Zhang L, Brazeau D, Zhang A, Ramanathan M: **Information-theoretic identification of predictive SNPs and supervised visualization of genome-wide association studies.** *Nucleic Acids Res* 2006, **34**(14):e101.

53. Sterner B, Singh R, Berger B: **Predicting and Annotating Catalytic Residues: An Information Theoretic Approach.** *J Comput Biol* 2007, **14**:1058-1073.
54. Igarashi Y, Aoki KF, Mamitsuka H, Kuma K, Kanehisa M: **The evolutionary repertoires of the eukaryotic-type ABC transporters in terms of the phylogeny of ATP-binding domains in eukaryotes and prokaryotes.** *Mol Biol Evol* 2004, **21**(11):2149-2160.
55. Liu X, Zhang LM, Guan S, Zheng WM: **Distances and classification of amino acids for different protein secondary structures.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **67**(5 Pt 1):051927.
56. del Sol Mesa A, Pazos F, Valencia A: **Automatic methods for predicting functionally important residues.** *J Mol Biol* 2003, **326**(4):1289-1302.
57. Qian H: **Relative entropy: free energy associated with equilibrium fluctuations and nonequilibrium deviations.** *Phys Rev E* 2001, **63**(4 Pt 1):042103.
58. 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**(1):505-515.
59. Bahar I, Atilgan AR, Erman B: **Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.** *Fold Des* 1997, **2**(3):173-181.
60. Hinsen K: **Analysis of domain motions by approximate normal mode calculations.** *Proteins* 1998, **33**(3):417-429.
61. Tirion MM: **Large amplitude elastic motions in proteins from a single-parameter, atomic analysis.** *Physical Review Letters* 1996, **77**(9):1905-1908.
62. Sanner MF, Olson AJ, Spehner JC: **Reduced surface: an efficient way to compute molecular surfaces.** *Biopolymers* 1996, **38**(3):305-320.
63. Ankerst M, Breunig MM, Kriegel HP, Sander J: **OPTICS: ordering points to identify the clustering structure.** *Proceedings of the ACM SIGMOD International Conference on Management of Data* 1999, **28**:49-60.
64. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**(4):536-540.
65. Varughese KI, Su Y, Cromwell D, Hasnain S, Xuong NH: **Crystal structure of an actinidin-E-64 complex.** *Biochemistry* 1992, **31**(22):5172-5176.

66. Holt DA, Luengo JI, Yamashita DS, Oh HJ, Konialian AL, Yen HK, Rozamus LW, Brandt M, Bossard MJ, Levy MA *et al*: **Design, synthesis, and kinetic evaluation of high-affinity FKBP ligands and the X-ray crystal-structures of their complexes with FKBP12.** *J Am Chem Soc* 1993, **115**:9925-9938.
67. Weber PC, Ohlendorf DH, Wendoloski JJ, Salemme FR: **Structural origins of high-affinity biotin binding to streptavidin.** *Science* 1989, **243**(4887):85-88.

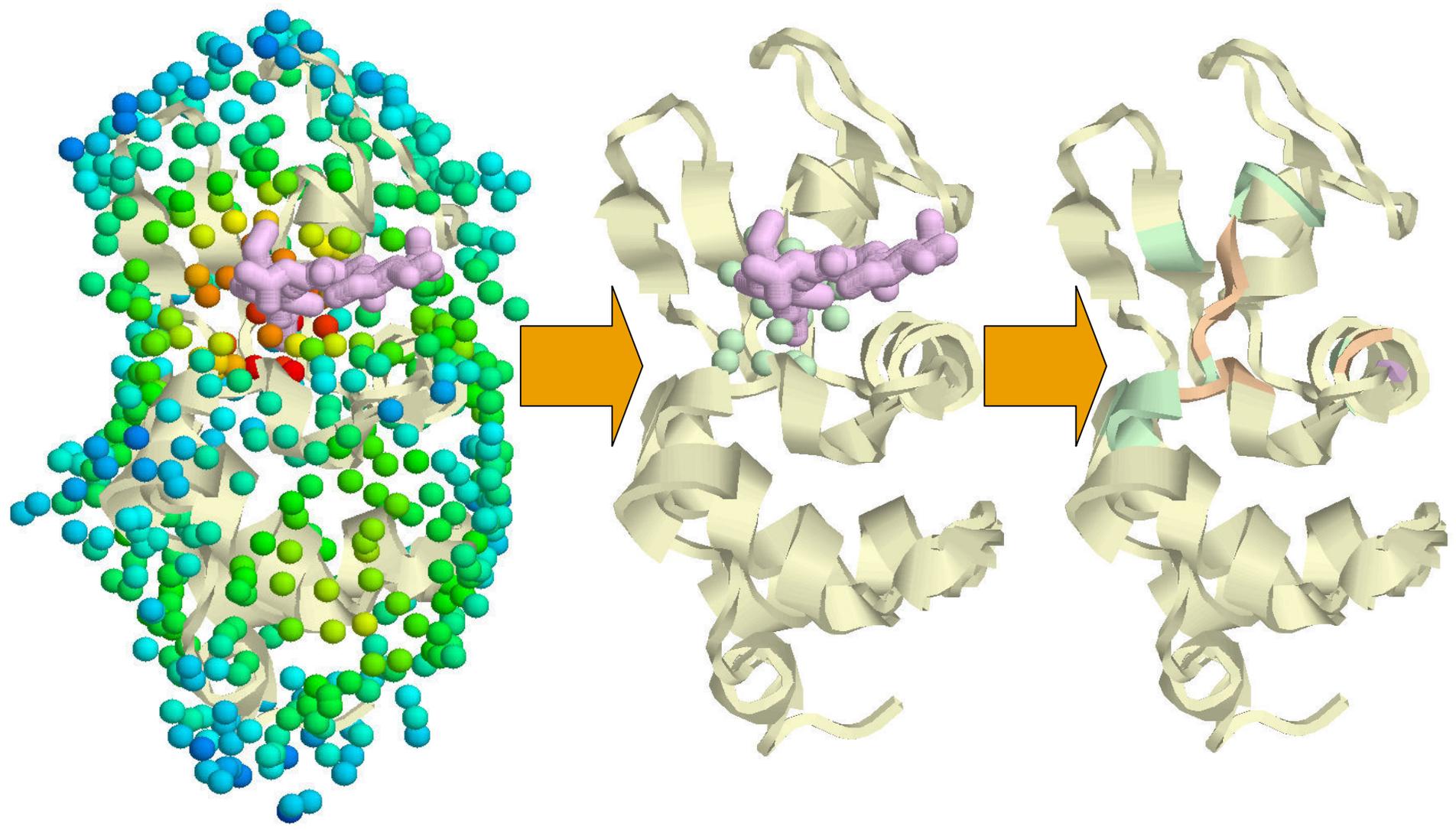


Figure 1

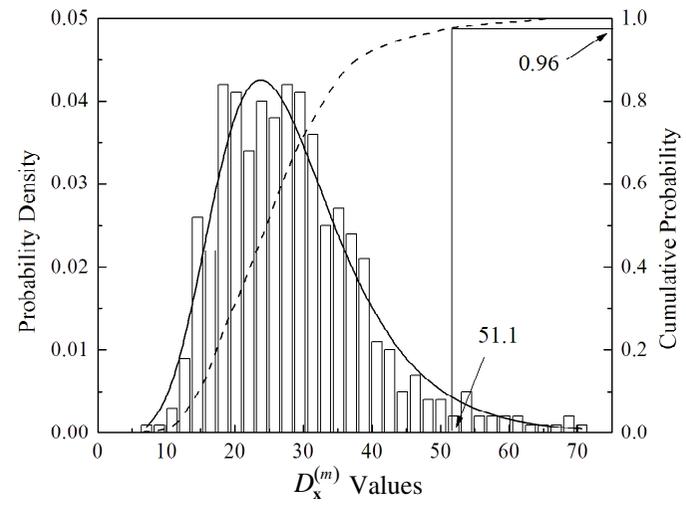


Figure 2

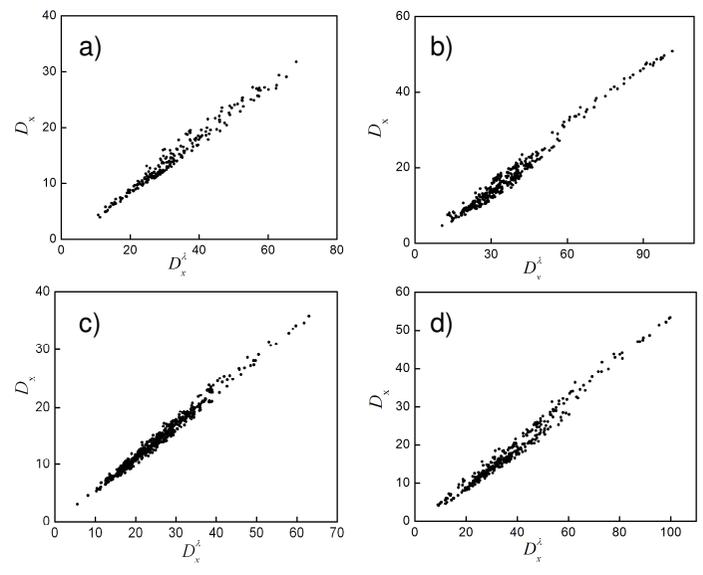


Figure 3

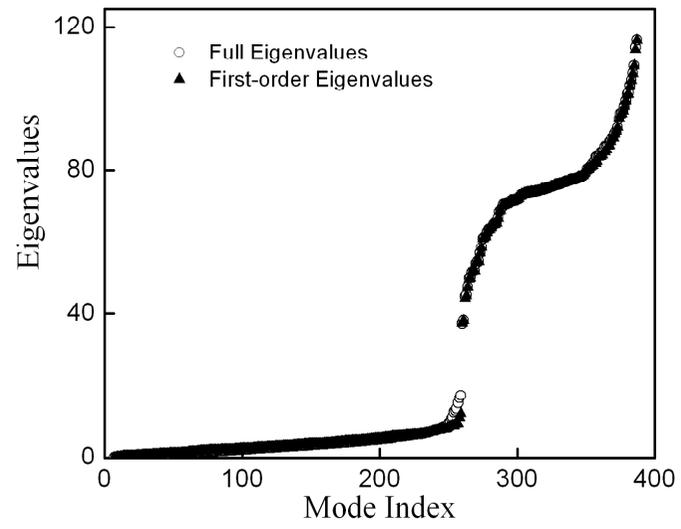


Figure 4

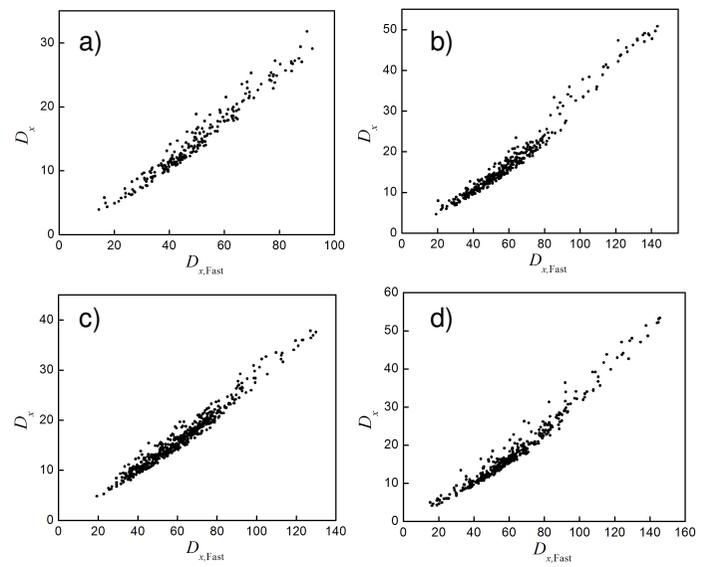


Figure 5

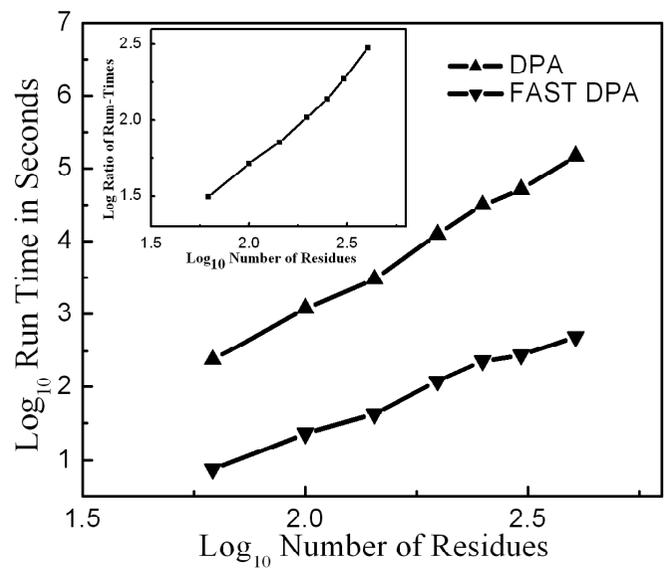


Figure 6

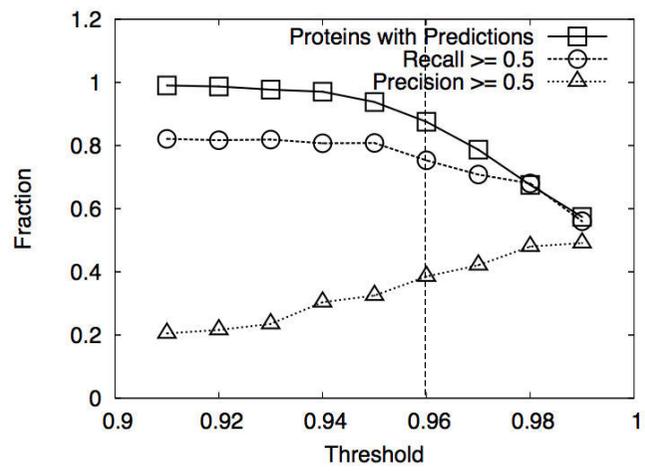


Figure 7

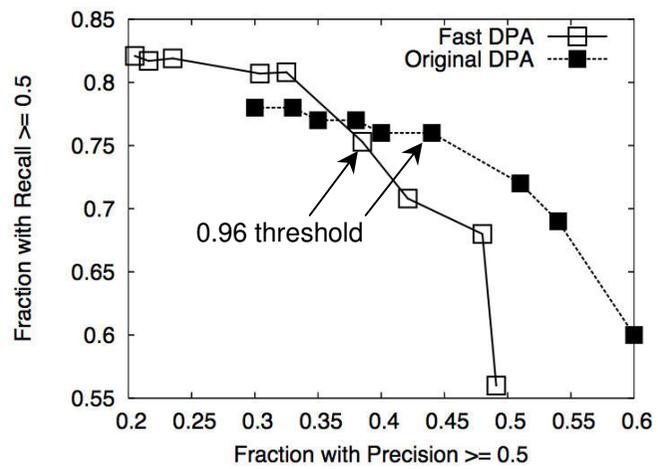


Figure 8

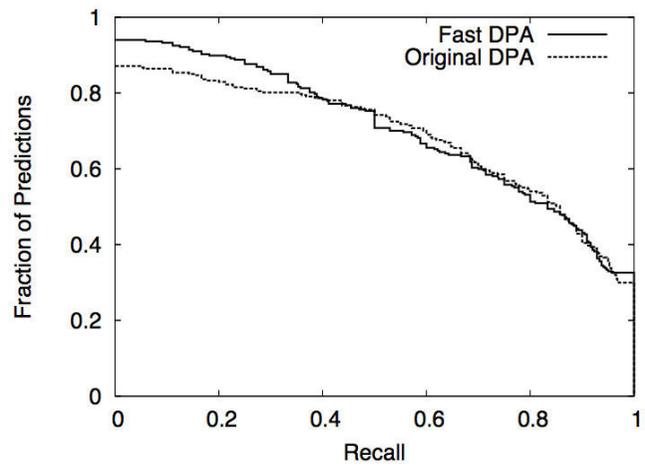


Figure 9

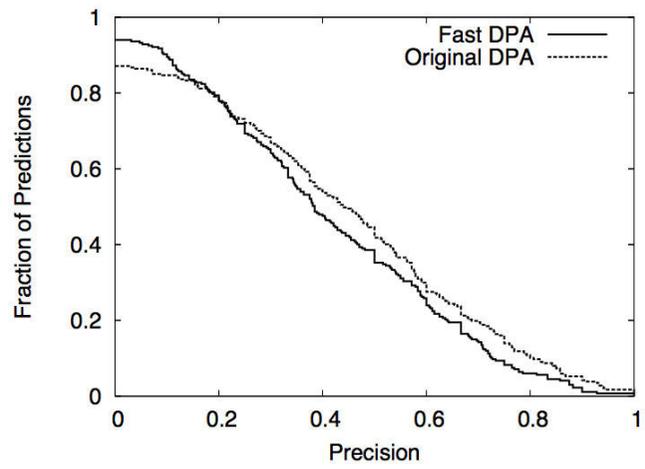


Figure 10