Bioinformatics Advance Access published April 28, 2005

*i*GNM: A Database of Protein Functional Motions Based on Gaussian Network Model

Lee-Wei Yang,[1,*] Xiong Liu,[2] Christopher J. Jursa,[2] Mark Holliman,[1] A.J. Rader,[1] Hassan A. Karimi,[2] and Ivet Bahar[1]

[1]Department of Computational Biology, School of Medicine, and [2]Department of Information Science and Telecommunications, University of Pittsburgh, Pittsburgh, PA 15213

*To whom correspondence should be addressed

Running Title. Database housing protein dynamics based on GNM

Abstract

Motivation. Knowledge of structure is not sufficient for understanding and controlling protein function. Function is a dynamic property. While protein structural information has been rapidly accumulating in databases, little effort has been invested to date towards systematically characterizing protein dynamics. Recent success of analytical methods based on elastic network models, and in particular the Gaussian Network Model (GNM), permits us to perform a high throughput analysis of proteins' collective dynamics.

Results. We computed the GNM dynamics for 20,058 structures from the Protein Data Bank, and generated information on the equilibrium dynamics at the level of individual residues. The results are stored on a web-based system called *i*GNM, and configured so as to permit users to visualize or download the results through a standard web browser using a simple search engine. Static and animated images for describing the conformational mobility of proteins over a broad range of normal modes are accessible, along with an online calculation engine available for newly deposited structures. A case study of the dynamics of twenty non-homologous hydrolases is presented to illustrate the utility of the *i*GNM database for identifying key residues that control the cooperative motions and revealing the connection between collective dynamics and catalytic activity.

Availability. http://ignm.ccbb.pitt.edu/

Contact. lwy1@pitt.edu

Introduction

With the rapid accumulation of protein structures in the Protein Data Bank (PDB) (Berman et al., 2000) it has become evident that structural information *per se* is not sufficient for gaining insights into the mechanisms of function. Protein function is a dynamic property. It is closely related to conformational mechanics, which, in turn, is largely dictated by the equilibrium (native) structure. It is now widely recognized that efficient computational methods and tools are needed for understanding the dynamics, and thereby controlling the function of proteins and their complexes.

Time cost of molecular dynamics simulations has been a major drawback for a systematic computational characterization of protein dynamics. This motivated efforts for developing efficient, but physically realistic, methods for deriving dynamic properties based on structure. Recent success of analytical methods based on Normal Mode Analysis (NMA) combined with Elastic Network (EN) models after the original studies of Tirion (1996); Bahar and coworkers (Bahar et al., 1997; Atilgan et al., 2001), Hinsen (Hinsen, 1998; Hinsen and Kneller, 1999) and Tama (Tama and Sanejouand, 2001) is paving the way for overcoming the computational limitations and making a rapid assessment of proteins collective motions (Tama, 2003; Ma, 2004).

Among the EN models of different complexities, the simplest is the Gaussian Network Model (GNM) (Bahar et al., 1997; Haliloglu et al., 1997). The GNM is entirely based on inter-residue contact topology in the folded state; it requires no *a priori* knowledge of empirical energy

parameters, in accord with the original proposition of Tirion (1996). Most importantly, it lends itself to *a unique, closed mathematical solution for each structure*.

An important feature of the GNM is the possibility of dissecting the observed motion into a collection of normal modes. The GNM mode analysis is similar, but simpler and more efficient than conventional NMA (see Methods). The slowest modes usually provide information on the collective motions relevant to biological function (Hinsen and Kneller, 1999; Kitao and Go, 1999; Tama and Sanejouand, 2001), as demonstrated in many applications. Despite its simplicity, the GNM has proven to yield results in good quantitative and qualitative agreement with experimental data and MD simulations (Bahar et al., 1998a; 1998b; 1999; Demirel et al., 1998; Bahar and Jernigan, 1998; 1999; Bahar, 1999; Haliloglu and Bahar, 1999; Jaravine et al., 2000; Kundu et al., 2002; Rader A.J. and Bahar, 2004; Kurt et al., 2003; Wu et al., 2003; Erkip and Erman, 2004; Burioni et al., 2004; Kundu et al., 2004; Lattanzi, 2004; Liao and Beratan, 2004; Micheletti et al., 2004; Temiz et al. 2004). Experimental data that have been compared and successfully reproduced with the GNM include X-ray crystallographic B-factors, H/D exchange protection factors or free energies of exchange, order parameters from ^{15}N-NMR relaxation, hinge regions and correlations between domain motions inferred from the comparison of the different forms of a given protein, key residues whose mutations have been observed to impede function or folding. The accumulating evidence that supports the utility of the GNM as an efficient tool for a first estimation of the machinery of proteins and their complexes led us to the construction of iGNM, a DB of GNM results compiled for >20,000 PDB structures.

The earliest attempt to establish a collection of biomolecular motions was the Database (DB) of

Macromolecular Movements (MolMovDB; http://molmovdb.mbb.yale.edu/molmovdb/), originally known as the DB of Protein Motions, constructed by Gerstein and collaborators (Echols et al., 2003). Two main features of MolMovDB are the *visualization* and *classification* of molecular motions according to their size and their mechanism. The displayed animations require the knowledge of starting and ending conformational states between which the molecule moves. About 17,000 movies are available in the DB, generated by morphs interpolating between pairs of known structures of proteins and RNA molecules, and refined by X-PLOR (Brünger, 1993) and CNS (Brünger et al., 1998). Another resource offered by Gerstein's lab is the use of a simplified NMA to display the biomolecular motions in the low frequency modes (Krebs et al., 2002).

A similar online calculation tool based on a simplified NMA combined with the RTB (Rotations-Translations of Blocks) algorithm (Tama et al., 2000) has been developed by Sanejouand and coworkers (elNémo; http://igs-server.cnrs-mrs.fr/elnemo/) presenting up to 100 slowest modes of studied structures (Suhre and Sanejouand, 2004a). This website provides information on the degree of collectivity of each predicted mode, as well as the overlap with experimentally observed change in conformation. Additionally, the implementation of normal mode perturbed models as templates for diffraction data phasing through molecular replacement is discussed (Suhre and Sanejouand, 2004b).

A more extensive study has been conducted by Wako and coworkers where the normal modes have been generated using the ECEPP/2 force field (Nemethy et al., 1883), and collected in the *Pro*Mode DB (http://promode.socs.waseda.ac.jp/) (Wako and Endo, 2002; Wako et al. 2003;

Wako et al. 2004) for nearly 1400 single chain proteins from the PDB. The structures are subjected therein to a detailed energy minimization prior to NMA computation. The NMA is performed in the coordinate system of dihedral angles after the work of Go and collaborators (Wako et al., 1995), such that each residue is subject to approximately six degrees of freedom (rotatable bonds on the backbone and sidechain), assuming the bond rotations to be independent. *Pro*Mode DB has been restricted to relatively proteins having less than 200 residues in view of the time cost of energy minimization. Finally, a recent effort in this direction is the Molecular Vibrations Evaluation Server (MoViES; http://ang.cz3.nus.edu.sg/cgi-bin/prog/norm.pl), constructed by Chen and coworkers (Cao et al., 2004), for NMA of proteins and DNA/RNA containing up to 4000 heavy atoms, in a full atomic framework. The results can be obtained in seven days via email.

Despite all these attempts, a DB of *predicted* mobilities for *all* PDB structures, ranging from small enzymes to *large complexes and assemblies* in a unified framework is lacking. In this paper we discuss a new *i*nternet-based system, *i*GNM, recently developed to address this need and to release the results from GNM computations applied to PDB structures.

The current version of *i*GNM consists of three modules: DB Engine, GNM Computations Engine, and Visualization Engine. The DB Engine is presented here, which contains visual and quantitative information on the collective modes predicted by the GNM for 20,058 structures deposited in the PDB prior to September 15, 2003. The goal of constructing the DB Engine has been to provide information on the dynamics of *all* proteins beyond those experimentally provided by B-factors (for X-ray structures) or root-mean-square fluctuations (NMR structures),

or by interpolation between existing PDB structures. We have developed an internet-based query system to allow users to retrieve information through a simple search engine by entering the PDB identifier of the protein structure of interest. The retrieved data are viewed by a Chime plug-in (for 3-D visualization) or a Java applet (for graphics). The output includes: the equilibrium fluctuations of residues and comparison with X-ray crystallographic B-factors, the sizes for residue motions in different collective modes, the cross-correlations between residue fluctuations, or domain motions in the collective modes, the identity of residues that assume a key mechanical role (e.g. hinge) in the global dynamics, and thereby function, of the molecule, as well as those potentially participating in folding nuclei/cores (Bahar et al., 1998; Rader and Bahar, 2004). In addition to retrieving the data stored in the DB, the user has the ability to compute the GNM dynamics for newly deposited structures through an automated online calculation server.

A case study, the dynamics of hydrolases, is presented here, to illustrate the use of *i*GNM data for inferring functional information. The catalytic sites in a set of hydrolases are located as residues participating in low mobility regions in the global modes, which could serve as a new prediction criterion to locate catalytic pockets from a given enzyme structure. *i*GNM is accessible at http://ignm.ccbb.pitt.edu/ .

Systems and Method

Model: GNM

The GNM is built on the statistical mechanical theory developed by Flory and coworkers for describing the fluctuation dynamics of polymer networks (Flory, 1976; Mattice and Suter, 1994).

Accordingly, the structure is modeled as an elastic network, the nodes of which of are the amino acids usually represented by their α–carbons, and uniform springs of force constant γ connect the pairs of α–carbons located within an interaction cutoff distance r_c. The dynamics of this network is fully defined by the N x N connectivity (or Kirchhoff) matrix of inter-residue contacts, $\mathbf{\Gamma}$. The off diagonal elements of $\mathbf{\Gamma}$are defined as $\Gamma_{ij} = -1$ if the distance between residues i and j, R_{ij}, is shorter than r_c, and zero otherwise; and the i^{th} diagonal terms is the degree of node i, or the coordination number of residue i. $\mathbf{\Gamma}$ contains the same information as contact maps. The statistical thermodynamics of the network are controlled by the Hamiltonian (Bahar et al., 1998)

$$H = (\gamma/2)\,[\mathbf{\Delta X}\,\mathbf{\Gamma}\,\mathbf{\Delta X}^T + \mathbf{\Delta Y}\,\mathbf{\Gamma}\,\mathbf{\Delta Z}^T + \mathbf{\Delta Z}\,\mathbf{\Gamma}\,\mathbf{\Delta Z}^T] \qquad (1)$$

where $\mathbf{\Delta X}$, $\mathbf{\Delta Y}$ and $\mathbf{\Delta Z}$ are the N-dimensional vectors of the X-, Y- and Z- components of the fluctuation vectors $\{\mathbf{\Delta R}_1, \mathbf{\Delta R}_2, \ldots, \mathbf{\Delta R}_N\}$ of the N residues in the examined protein. The mean-square fluctuations of residue i scale with the i^{th} diagonal element of the inverse of $\mathbf{\Gamma}$ (Bahar et al., 1997; Haliloglu et al., 1997), as

$$\langle(\Delta R_i)^2\rangle = (3kT/\gamma)\,[\mathbf{\Gamma}^{-1}]_{ii}\,, \qquad (2)$$

and the cross-correlations $\langle\Delta\mathbf{R}_i \bullet \Delta\mathbf{R}_j\rangle$ scale with the ij^{th} off-diagonal elements of $\mathbf{\Gamma}^{-1}$.

The fluctuation dynamics of the structure results from N-1 superposed GNM modes. The modes can be extracted by the eigenvalue decomposition $\mathbf{\Gamma} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}^T$ where $\mathbf{U}$ is an orthogonal matrix whose columns $\mathbf{u}_k$ ($1 \le k \le N$) are the eigenvectors of $\mathbf{\Gamma}$, and $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues λ_k. The kth eigenvector reflects the *shape* of the kth mode as a function of residue

index *i*; the *k*th eigenvalue represents its frequency (Haliloglu et al., 1997; Bahar, 1999).

Structures

All the structures deposited in PDB as of Sept. 15, 2003 have been downloaded (22,549 of them) and subjected to GNM analysis. A file parser was implemented to eliminate structures composed of (1) predominantly DNA or RNA molecules, (2) carbohydrates, small organic compounds or short peptides containing less than 15 residues, which eliminated 6.2% of the structures, and (3) 4.8% of the originally downloaded structures that yielded unrealistic mode shapes due to their incomplete and/or inaccurate coordinates deposited in the PDB. Figure 2 gives a schematic description of such an occurrence where a portion of the network is 'disconnected'. For a given fully connected structure Γ has rank *N-1* and its eigenvalue decomposition yields *N-1* non-trivial eigenvalues and one zero eigenvalue. However, more than one zero eigenvalue was obtained for the disconnected networks.

We generated the GNM results for 20,058 structures, after filtering out the above listed cases. The examined structures cover a broad range of size, including for example, large proteins such as contractile protein of insect flight muscle (PDB: 1o1c), with 11,730 amino acids. The size distribution of the examined structures is shown in *Figure 1*.

Computations

The eigenvalue decomposition of Γ is the most time-consuming part of the computations. We have recently implemented the BLZPACK package (Marques, O., 1995) based on the Lanczos algorithm, which permits us to efficiently extract subsets of interesting modes at either end of the

vibrational spectrum. This package reduces the computing time by at least three orders of magnitude in the case of large proteins.

Results

Output Files

Eleven output files can be accessed for each query structure (Fig 3a). Users can retrieve the generated output files for structures of interest by simply entering the 4-digit PDB ID in the search engine, http://ignm.ccbb.pitt.edu/FileDownload.htm. A brief description of the output files and/or the type of information that can be extracted is presented below.

(i) Contact topology (".ca", ".cont" and ".eigen")

The residue types, sequence numbers, α-carbon coordinates and temperature factors reported in the PDB and used in the GNM are listed in the files with suffix ".ca". The size of the protein, defined by the number of α carbons (N) included in the computations, is listed in last line of the file. The ".cont" file lists the contact number (the number of adjacent neighbors within a cutoff r_c = 7.3 Å) for each residue. A large contact number refers to a constrained environment that limits or inhibits the residue mobility. The ".eigen" file lists the *N-1* non-zero eigenvalues λ_k in descending order, starting from the fastest mode ($k = N-1$), and the zero eigenvalue λ_0 is listed as the last element. Any value of the order of 10^{-6} or lower is deemed as zero. The structures with the above described 'discontinuity' yielded more than one zero eigenvalue, which were captured in the corresponding '.eigen' files, and were removed from the DB.

(ii) Time-average fluctuations and their correlations (".bfactor" and ".cc")

The theoretical temperature factor (B_i) predicted by the GNM is proportional to the inverse Kirchhoff matrix and also to the summation of all modes as

$$\boldsymbol{B_i} = (8\pi^2 k_B T/\gamma)\left[\Gamma^{-1}\right]_{ii} = (8\pi^2/3)\sum_{k=1}^{N-1}\lambda_k^{-1}[\boldsymbol{u}_k]_i[\boldsymbol{u}_k]_i \qquad (3)$$

Eq 3 follows from eq 2 and the definition $B_i = (8\pi^2/3) <(\Delta R_i)^2>.[\boldsymbol{u}_k]_i$ designates the i^{th} element (corresponding to i^{th} residue) of the k^{th} eigenvector. The ".bfactor" file contains the experimental B_i values of α–carbon atoms (if available in the PDB) and the corresponding theoretical B_i values for each residue. *Figure 4* panel ***c*** illustrates the comparison of the two sets of B_i values, as a function of residue index, for a query protein, phospholipase 2 (1BK9; Zhao, et al., 1998) shown in panel ***a***. A correlation coefficient of 0.72 between the experimental (yellow curve) and theoretical results (red curve) is obtained.

The predicted cross-correlations $<\Delta\mathbf{R}_i \bullet \Delta\mathbf{R}_j>$ between the fluctuations of residues *i* and *j* are listed in the '.cc' files. These are reported for small-to moderate size proteins (N <290) due to memory constraints. The data in these files are used to construct the color-coded correlation maps (called CCplot) (*Figure 4**d***). $<\Delta\mathbf{R}_i \bullet \Delta\mathbf{R}_j>$ values are normalized between -1 and 1, by dividing them by $[<(\Delta R_i)^2><(\Delta R_j)^2>]^{½}$. A value of -1 refers to perfectly anticorrelated (i.e. concerted but in opposite direction) fluctuations undergone by residues *i* and *j* (colored blue in the map), and +1 refers to fully correlated motions (colored red).

(iii) Mobilities in normal modes (".sloweigenvector", ".slowmodes" and ".slowav")

The shapes of slowest 20 modes ($[\boldsymbol{u}_k]_i^2$, $1 \le k \le 20$, as a function of residue index *i*) are given in

the ".slowmodes" file, and the corresponding eigenvectors, $\boldsymbol{u}_k$, in the ".sloweigenvector" file. Each row in these files corresponds to a given residue, and each column to a different mode, starting from the slowest (global) mode. We note that the eigenvectors are orthonormal, and consequently the k^{th} mode shape represents the normalized *distribution* of residue mobilities (square displacements) induced in mode k. The joint effect of modes 1 and 2 to mobilities can be found in the ".slowav" file. The entries therein refer to the weighted average

$$[(\Delta R_i)^2]_{1\text{-}2} = (\lambda_1^{-1} + \lambda_2^{-1})^{-1} (\lambda_1^{-1} [\boldsymbol{u}_1]_i [\boldsymbol{u}_1]_i + \lambda_2^{-1} [\boldsymbol{u}_2]_i [\boldsymbol{u}_2]_i) \qquad (4)$$

(iv) Global hinge residues at crossovers between positive and negative elements of $\boldsymbol{u}_1$

The positive and negative elements of $\boldsymbol{u}_k$ refer to residues moving in opposite direction along mode k. Of interest are the residues at the passage between positive and negative elements of slowest modes, which presumably act as *hinges* between the oppositely moving clusters of residues. The ".sloweigenvector" files thus provide information on the identity of the residues that play a mechanically critical role in the global modes.

(v) Peaks in high frequency modes (".fasteigenvector", ".fastmodes", ".fast10av")

The shapes of fastest 20 modes ($[\boldsymbol{u}_k]_i^2$, $N\text{-}20 \le k \le N\text{-}1$, as a function of residue index i) are given in the ".fastmodes" file, and the corresponding eigenvectors, $\boldsymbol{u}_k$, in the ".fasteigenvector" file, similarly to their slow mode counterparts. We note that, contrary to the slow mode shapes, the fast modes are highly localized and exhibit sharp peaks at certain residues. The cumulative mode shape for the fastest 10 modes is presented in the file ".fast10av". The peaks in the latter file are indicative of potential folding nuclei or conserved residues important for stability (Demirel et al., 1998; Rader and Bahar, 2004).

Query and Visualization

*i*GNM allows users to conveniently query and visualize GNM output files. By typing the PDB ID in the 3-D Visualization Module (http://ignm.ccbb.pitt.edu/3D_GNM.htm) users can view and compare the ribbon diagrams of the query structures color-coded according to the mobilities of residues in the slowest or fastest 20 modes. Likewise, the B-factors Visualization Module (http://ignm.ccbb.pitt.edu/BFactors.htm) provides access to ribbon diagrams colored by the mean-square fluctuations predicted and observed for *all* modes (B-factors).

In addition to queries using PDB IDs, *i*GNM is integrated with PDB SearchLite query interface for keyword-based queries (http://ignm.ccbb.pitt.edu/PDB_Integration.htm). By typing keywords related to the biological macromolecules of interest, users can browse PDB records and *i*GNM output files for a given protein family in an integrated environment (Fig 3**b**).

The data visualization uses a Java applet in the Java 2 Runtime Environment (Sun Microsystems, Inc. http://java.sun.com/) and Chime (MDL Information Systems, Inc. http://www.mdli.com/) to produce interactive mobility plots and structure animations. These cross-platform software tools can be freely downloaded and easily installed. Chime, as a browser plug-in, allows users to manipulate color-coded structure in atomic details. The Java applet displays the residue mobility in a pop-up window (graph). The user can point the cursor to the positions of interest (minima or maxima) on the graphs and see the corresponding residue number and relative fluctuations. Links to the raw *i*GNM data, PDB, PDBsum, SCOP and CATH are also included for user references.

Online Calculations

Currently (Jan 25, 2005) the PDB contains 29,326 structures. The *i*GNM DB has processed 22,549 of them, and generated results for 20,058. When the user performs a search for a PDB ID the Database Engine is checked first for that structure's GNM files. If the structure's results are found, the results are displayed to the user through the Visualization Engine. For those PDB structures that are not included, an interface to perform online calculations is provided at http://ignm.ccbb.pitt.edu/gnmwebserver/index2.html.

The online calculation module is a three-tier architecture, where the user's browser communicates with *i*GNM, and the server communicates with the PDB server (Fig 5). This server takes as input the 4-digit PDB ID, searches the PDB, and if the structure is found it then retrieves the file and runs the GNM calculations on it. Once the calculation is complete the results are passed to the Visualization Engine for graphical presentation to the user.

Future additions to *i*GNM will include an automatic update module for checking the PDB for new structures, downloading the structure files, running the GNM calculations on the structure files, and updating the Database with the newly computed GNM results. When this automatic update module is completed, the online calculation server will be reconfigured to utilize structure files submitted by users over the web. This will allow users to submit their own structure files for online GNM calculations, and allow then to view the results through the Visualization Engine. These additions are currently in the design and testing phases.

A Case Study: Interplay between Dynamics and Chemistry

An application to a family of enzymes illustrating the utility of *i*GNM DB is presented here. To this aim, twenty single-chain hydrolases that exhibit a range of functions (EC number) and structural sub-classes (CATH) were selected from the DB (Table 1). Of these, thirteen are taken from the catalytic residues dataset compiled by Thornton and collaborators (Bartlett et al., 2002), and seven (indicated by the asterisks) are additional hydrolases retrieved from the PDB. The catalytic sites are required to meet one of the following criteria set forth by Thornton and coworkers: (i) they are directly involved in catalytic function, (ii) they affect residues or water molecules that are directly involved in catalysis, (iii) they can stabilize a transient intermediate, or (iv) they interact with a substrate or cofactor that facilitates the local chemical reaction. The amino acids which simply bind substrates or ligands are not necessarily catalytic residues.

Participation of catalytic residues in the collective modes

Experimental B factors (in ".bfactor" files) and the weighted average mobilities $[(\Delta R_i)^2]_{1\text{-}2}$ (in ".slowav") extracted from *i*GNM for the catalytic residues of the examined hydrolases are listed in the last two columns of Table 1. To make a quantitative assessment across the complete set, both the B-factors and mobilities of residues in a given mode were normalized in the range [0, 1]. The distributions of $[(\Delta R_i)^2]_{1\text{-}2}$ are displayed in *Figure 6* for two proteins from the examined set, phospholipase A2 (1BK9) and protein tyrosine phosphatase (1YTW; Fauman, et al., 1996). The catalytic residues are indicated by the arrows in the Figure 6.

The mode shapes in Figure 6 and the values listed in the last two columns of Table 1 show that the catalytic sites occupy regions that are spatially constrained in general (evidenced by low B_i values), and this tendency becomes more pronounced in the examination of the slowest modes

(indicated by even lower $[(\Delta R_i)^2]_{1\text{-}2}$ values). The average $<[(\Delta R_i)^2]_{1\text{-}2}>_{cat}$ over all the catalytic residues of the examined hyrolases is 0.045, and the average $<B_i>_{cat}$ is 0.126, as opposed to the respective averages over *all* residues of 0.180 and 0.244. Thus, the square fluctuations of catalytic residues are reduced by a factor of two on the average compared to other residues, and their mobilities in the slowest modes are further constrained, by a factor four. Such severely constrained regions are usually involved in, or closely communicate with, the mechanically key sites (e.g. hinges, anchoring regions, symmetry centers, etc) that control the collective dynamics of the enzymes. Thus, chemically active residues are found here to also participate in critical sites from conformational mechanics point of view, which invites attention to the functional coupling between catalysis and global dynamics.

Discussion

We generated information on the equilibrium dynamics of 20,058 structures in the reach of covering the entire PDB. The case study of 20 hydrolases sets a simple example of the use of the collective dynamics data to gain insights on the mobilities of catalytic residues and their participation in large scale motions of the overall enzyme. The catalytic residues are shown to preferentially occupy cooperatively constrained regions (minima in slow modes) which might be expected to efficiently transmit the effect of chemical reaction to other regions of the enzymes. This feature, which has also been observed in other families of enzymes (Chen and Bahar, 2004; Yang and Bahar, 2005), may be advantageously used in identifying or designing active sites.

The eigenvalue decomposition of the connectivity matrix Γ is the most expensive task in GNM calculations from computational *time* point of view. We used a singular value decomposition

(SVD) subroutine to this aim (Press et al., 1992), the computing time of which scales with N^3 for a network of N residues. For N < 1,500, the computations are performed within minutes, while the CPU times increased up to 15days in the case of the largest structures, the output of which are compiled and accessible in the DB. While all $N-1$ modes, and the mean-square fluctuations resulting from the superposition of all modes have been compiled to date in the *i*GNM, we have also implemented an alternative algorithm that utilizes the BLZPACK software (Marques, 1995) based on Block Lanczos Method for large structures. The latter evaluates a subset ($1 \le k \le 100$) of dominant (slowest) modes, within a time scale of N^2, i.e. the computing times is more than 3 orders of magnitude shorter than the routine SVD, when structures of $>10^3$ residues are analyzed. The same algorithm will be particularly useful for generating the ANM (anisotropic network model) (Atilgan et al., 2001) data that we plan to include in the near future in the *i*GNM DB.

In a previous study, we have shown that GNM can satisfactorily reproduce the experimentally observed fluctuations and functional motions of proteins complexed with RNA or DNA (Bahar and Jernigan, 1998; Bahar et al., 1999; Temiz and Bahar, 2002), including supramolecular structures like ribosomal complexes(Wang et al., 2004) or viral capsids (Rader et al. 2005). P and O4' atoms of nucleotides are usually adopted as nodes to model the RNA/DNA structures. The choice of these two atoms per nucleotide provides a spatial resolution comparable to that of α-carbons in proteins, and the cutoff distances are reasonably adjusted to account for the longer range interactions of nucleotides. Presently, the *i*GNM DB does not contain the results for such complexes or assemblies containing RNA/DNA components, although a server is presently available at http://ignm.ccbb.pitt.edu/GNM_Online_Calculation.htm, which can generate GNM results for such cases using the nucleotides coordinates reported in the PDB.

Finally, users have to be cautious about two facts: (i) the *i*GNM results reflect the equilibrium dynamics for proteins in their crystal form reported in PDB, and (ii) the method is applicable to fluctuations near the native structure and conformational changes involving the passage over an energy barrier, or other non linear effects on the conformational dynamics cannot be described by the GNM and necessitate more detailed MD simulations. In some cases, the crystallized form may not be the active state of the protein under physiological condition. For instance, PDB entry 1hho contains one half of a hemoglobin (Hb) molecule (two chains) in the crystal asymmetric unit, while the bio-active Hb is a tetramer that can actually be generated by combining 1hho with its crystallographic two-fold axis partner. We are currently designing a new module that will facilitate the retrieval and generation of such user-customized structures that combine the biological units or any structural parts of interest. Finally, we note that the GNM is particularly useful in the case of large structures and complexes/assemblies, while its application to small structures (a network less than 30 nodes) may not be always justifiable. First, small structures are amenable to analysis using more details with full atomic models that take account of their specific interactions. Secondly, the Gaussian approximation for residue fluctuations becomes more accurate with increasing size of the network, as follows from central limit theorem.

As the number of ‘new’ folds deposited in the PDB decreases on a yearly basis, we are close to collecting data for a large fraction of all possible folds. While the biomolecular function overwhelmingly exceeds the number of known folds, the types of large scale conformational motions undergone by biomolecules seem to be relatively limited, similarly to the finite number of folds. The particular fold and its intrinsic global dynamics can presumably offer a versatile

scaffold and mechanism for achieving a diversity of biochemical functions by amino acid substitutions that can accommodate the same fold and global dynamics. *i*GNM resulted from an attempt to collect those dynamic data in a DB framework to enable further exploration and establishment of biomolecular structure-dynamic-function relations.

Acknowledgment . We would like to thank Dr. Rob Bell for his efforts in facilitating high-speed computing hardware for the calculations in this work, Mr. Shannching Chen for resolving the memory allocation problem in sequential computation of 22,549 structures. Partial support by the NSF-ITR grant # EIA-0225636, and the NIH grant # 1 R01 LM007994-01A1 is gratefully acknowledged.

Table 1. Properties of 20 hydrolases subjected to *i*GNM analysis

PDB	E.C.	CATH	Enzyme Name	Catalytic Residues	Slowav(%)	Exp-B(%)
1BK9*	3.1.1.4	1.20.90.10	phospholipase A2	H48,Y52,D99	2.4	14.1
1BWP	3.1.1.47	3.40.50.1110	2-acetyl-1-alkylglycero-phosphocholine esterase	S47,G74,N104,D192,H195	8.8	12.4
1CHD	3.1.1.61	3.40.50.180	Prot-glu methylesterase	S64,T165,H190,M283,D286	3.5	24.7
1RPT	3.1.3.2	3.40.50.1240	High MW acid phosphatase	R11,H12,R15,R79,H257,D258	2.2	N/A
1YTW	3.1.3.48	3.90.190.10	Yersinia protein tyrosine phosphatase	E290,D356,H402,C403,R409,T410	1.3	14.4
1DNK	3.1.21.1	3.60.10.10	Deoxyribonuclease I	E78,H134,D212,H252	6.3	19.6
1BOL	3.1.27.1	3.90.730.10	Ribonuclease T2	H46,E105,H109	4.0	7.2
1BVV*	3.2.1.8	2.60.120.180	G/11 xylanase	Y69,E78,E172	2.3	9.5
1UOK	3.2.1.10	5.1.2991.1	Oligo-1,6-glucosidase	D199,E255,D329	5.2	9.9
1EUG	3.2.2.3	3.40.470.10	Uridine nucleosidase	D64,H187	1.5	20.6
1BR6*	3.2.2.22	5.1.60.1	Ricin	Y80,V81,G121,Y123,E177, R180	0.8	5.7
1A16*	3.4.11.9	5.1.2912.1	Aminopeptidase P	D260,D271,H354,H361,E383,E406	7.0	11.8
1B6A*	3.4.11.18	5.1.3307.1	Met aminopeptidase 2	H231	3.1	7.0
1BIO*	3.4.21.46	5.1.40.1	Hum complemnt factor D	H57,D102,S195	1.2	7.7
9PAP	3.4.22.2	3.90.70.10	Thiol-endopeptidase	Q19,C25,H159,N175	6.1	7.4
1BXO*	3.4.23.20	5.1.750.1	Penicillopepsin	D33,D213	1.4	0.3
8TLN	3.4.24.27	5.1.780.1	Metalloproteinase M4	E143,H231	8.2	14.3
1LBA	3.5.1.28	3.40.80.10	T7 lysozyme	Y46,K128	11.0	27.3
1BTL	3.5.2.6	3.40.710.10	β-Lactamase class A	S70,K73,S130,E166	5.7	18.5
1CTT	3.5.4.5	5.1.590.1	Cytidine deaminase	E104	8.3	19.9
				Average(*)	4.5	12.6

(*)over the average mobility of catalytic residues in each protein

Reference

Atilgan, A. R., Durrell, SR., Jernigan, R. L., Demirel, M. C., Bahar, I. (2001) *Biophys. J.* **80**, 505-515.

Bahar, I. and Jernigan, R. L. (1998) *J. Mol. Biol.* 281, 871-884.

Bahar, I. and Jernigan, R. L. (1999) *Biochemistry* **38**, 3478-3490.

Bahar, I. (1999) *Rev. Chem. Eng.* **15**, 319-349.

Bahar, I., Atilgan, A. R., Demirel, M. C., Erman, B. (1998a) *Phys. Rev. Lett.* **80**, 2733-2736.

Bahar, I., Atilgan, A. R., Erman, B. (1997) *Fold & Des* 2, 173-181.

Bahar, I., Erman, B., Jernigan, R. L., Atilgan, A. R., Covell, D. G. (1999) *J. Mol. Biol.* 285, 1023-1037.

Bahar, I., Wallqvist, D. G., Covell, D. G., Jernigan, R. L. (1998b) *Biochemistry* 37, 1067-1075.

Bartlett, G., Porter, C.T., Borkakoti, N., Thornton, J. M. (2002) *J. Mol. Biol.* 324, 105-121.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000) *Nucleic Acids Res.* 28, 235-242.

Brooks, B. and Karplus, M. (1983) *Proc. Natl. Acad. Sci. U. S. A.* 80, 6571-6575.

Brünger, A. T. (1993) *Yale University Press, New Haven, USA.*

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S.,

Burioni,R., Cassi,D., Cecconi,F., Vulpiani,A. (2004) *Proteins* **55**, 529-535

Cregut, D., Drin, G., Liautard, J. P., Chiche, L. (1998) *Protein Eng.* 11, 891-900.

Cao, Z.W., Xue, Y., Han, L. Y., Xie, B., Zhou, H., Zheng, C. J., Lin, H. H., Chen, Y.Z. (2004) *Nucleic Acids Res.* 32, 679-685.

Chen, Y.Z. and Prohofsky, E.W. (1992) *Nucleic Acids Res.* 20, 415-419..

Chen, S.C. and Bahar, I. (2004) *Bioinformatics* 20, i77-i85.

Demirel,M.C., Atilgan,A.R., Jernigan,R.L., Erman,B., Bahar,I. (1998) *Protein Sci*. **7**, 2522-2532

Echols, N., Milburn, D., Gerstein, M. (2003) *Nucleic Acids Res.* **31**, 478-482.

Erkip, A. and Erman, B. (2004) *Polymer* **45**, 641-648

Flory, P. J. (1976) *Proc. Roy. Soc. Lond. A.* 351, 351-380.

Fauman, E. B., Yuvaniyama, C., Schubert, H. L., Stuckey, J. A., Saper, M. A. (1996) *J Biol Chem* **271,** 18780-18788

Haliloglu, T. and Bahar, I. (1999) *Proteins* 37, 654-667.

Haliloglu, T., Bahar, I., Erman, B. (1997) *Phys. Rev. Lett.* **79**, 3090-3093.

Hinsen, K. (1998) *Proteins* **33**, 417-429

Hinsen, K. and Kneller, G. R. (1999) *J. Chem. Phys.* **111**, 10766-10769.

Jaravine,V.A., Rathgeb-Szabo,K., Alexandrescu,A.T. (2000) *Protein Sci*. **9,** 290-301

Kitao, A. and Go, N. (1999) *Curr Opin Struct Biol* **9**, 164-169.

Krebs, W.G., Alexandrov, V., Wilson, C.A., Echols, N., Yu, H., Gerstein, M. (2002) *Proteins* **48,** 682-695

Kundu, S., Melton, J. S., Sorensen, D. C., Phillips, G. N. (2002) *Biophys. J.* 83, 723-732.

Kurt,N.; Scott,W.R.; Schiffer,C.A.; Haliloglu,T. (2003) *Proteins* **51**, 409-422

Lattanzi,G. (2004) *Comput Mat Sci* **30**, 163-171

Liao, J.L. and Beratan, D.N. (2004) *Biophys J* **87**, 1369-1377

Ma, J.P. (2004) *Curr Protein Pept Sc* **5**, 119-123

Marques, O. (1995) *BLZPACK: Description and User's Guide*, TR/PA/95/30, CERFACS, Toulouse, France, 1995

Mattice, W. L. and Suter, U. W. (1994) *John Wiley & Sons, Inc., New York.*

McCallum, S. A., Hitchens, T. K., Torborg, C., Rule, G. S. (2000) *Biochemistry* **39**, 7343-7356.

Micheletti,C.; Carloni,P.; Maritan,A. (2004) *Proteins* **55**, 635-645

Nemethy, G., Pottle,M., Scheraga, H. (1983) *J Phys Chem* **87**, 1883-1887

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T. (1992) *Numerical Recipes in Fortran* **Chp 2.6,** 51-62

Rader A.J. and Bahar, I. (2004) *Polymer* **45**, 659-668.

Rader,A.J., Vlad,D.H., Bahar,I. (2005) *Structure,* in press.

Suhre, k. and Sanejouand, Y.-H. (2004) *Nucleic Acids Res.* **32**, 610-614.

Suhre, k. and Sanejouand, Y.-H. (2004) *Acta Crystallogr. , Sect. D* **60**, 796-799.

Tama, F. and Sanejouand, Y.-H. (2001) *Protein Eng.* 14, 1-6.

Tama, F., (2003) *Protein Peptide Lett* **10**, 119-132

Tama, F., Gadea, F. X., Marques, O., Sanejouand, Y. H. (2000) *Proteins* **41**, 1-7.

Temiz,N.A., Meirovitch,E, Bahar,I. (2004) *Proteins* **57**, 468-480

Temiz, N.A. and Bahar,I. (2002) *Proteins* **49**, 61-70.

Tirion, M. M. (1996) *Phys. Rev. Lett.* 77, 1905-1908.

Wako, H., Kato, M., Endo, S. (2004) *Bioinformatics* **20**, 2035-2043.

Wako, H, and Endo, S. (2002) *Genome Informatics* **13**, 519-520.

Wako, H, Kato, M., Endo, S. (2003) *Genome Informatics* **14**, 663-664.

Wako, H., Endo, S., Nagayama, K., Go, N. (1995) *Comp. Phys. Comm.*, **91**, 233-251.

Wang, Y., Rader, AJ, Bahar, I. and Jernigan, RL. (2004) *J. Struct Biol.* **147,** 302-314.

Wu,Y., Yuan,X., Gao,X., Fang,H., Zi,J. (2003) *Phys.Rev.E.Stat.Nonlin.Soft.Matter Phys*. **67**, 041909

Yang, L.W. and Bahar, I. (2005) submitted.

Zhang, Z., Shi, Y., Liu, H. (2003) *Biophys. J.* 84, 3583-3593.

Zhao, H., Tang, L., Wang, X., Zhou, Y., Lin, Z. (1998) *Toxicon* **36**, 875-876

Figure Legends

Fig 1. Distribution of the sizes of PDB structures compiled in the *i*GNM DB. The number N of residues includes the number of amino acids contained in the examined PDB structures. 8.4% (1,701 out of 20,058) of the structures contain more than 10^3 residues. The inset displays the same distribution on a logarithmic plot to show the complete range of protein sizes (up to N = 11,730), each point corresponding to the total number of counts in intervals of size $\Delta N = 50$.

Fig 2. A schematic diagram illustrating how a discontinuity in the PDB sequence/coordinates may lead to more than one zero eigenvalue. In panel ***a***, the coordinates of residue C belonging to the A-B-C-D-E are missing. The distance between residues B and D is larger than the cutoff 7.3 Å, which leads to two independent blocks in the corresponding Kirchhoff matrix $\mathbf{\Gamma}_{DIS}$ and more than one zero eigenvalue in the associated diagonal matrix of eigenvalues $\mathbf{\Lambda}_{DIS}$. In contrast, the continuous tetrapeptide (no gap) in panel ***b*** accurately gives one zero eigenvalue, despite the possibly missing terminal residue E.

Fig 3. ***(a)*** The query engine to retrieve GNM data for 20,058 structures. The PDB identifier (ID) of the protein of interest is entered to retrieve the output files from the *i*GNM. Alternatively, a search with a keyword is made ***(b).*** The results using 'phospholipase' as keyword are shown. The GNM information for all the retrieved structures is tabulated in the right column.

Fig 4. Visualization of GNM dynamics for phospholipase A2 (PDB ID: 1BK9). ***(a)*** Color-coded ribbon diagram (Chime) that illustrates the mobilities in the slowest GNM mode (slow1). The structure is colored from dark blue, green, orange to red in the order of increasing mobility in the

slow mode ***(b)*** The Java applet shows the corresponding mobility plot ($[\boldsymbol{u}_1]_i$ vs. i) with scalable range of view, max/min value information window and pop-up tag to show the residue number and coordinates. ***(c)*** Comparison of experimental and theoretical B_i factors. ***(d)*** Cross-correlation map, i.e. normalized $\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \rangle$ values plotted for residue i(abscissa) and j (ordinate). The fully concerted motion (+1) is colored dark red while the perfect anti-correlated motion (-1) is colored dark blue, and weakly correlated and anticorrelated regions are yellow and cyan, respectively.

Fig 5. *i*GNM currently consists of two standalone servers, one that houses the DB Engine with the Visualization Engine, and the other houses the online calculation module and visualization for structures deposited after Sept 2003. To use the system, the user can choose to view Mobility Ribbon Diagrams, B-Factors, or download GNM results in http://ignm.ccbb.pitt.edu. Upon entering the 4-digit PDB ID the DB Engine is checked for GNM files of the queried structure. If the files are found they are immediately displayed to the user's browser window. If they are not found, the user is offered the option of using the online calculation. If they choose to use this, it will then invoke the search of PDB for the structure file, and if found it retrieves the structure file to the server where the GNM calculations are then performed. Once the calculations are complete the results are displayed to the user's browser window. Future plans for *i*GNM involve implementing an automatic update module for synchronizing the DB Engine with PDB, and an online calculation module where users can submit their own models for GNM calculations.

Fig 6. Mobilities in the slowest two modes $[(\Delta R_i)^2]_{1\text{-}2}$ vs residue index i) and corresponding color-coded structure for phospholipase A2 (1BK9) and *Yersinia* protein tyrosine phosphatase (1YTW). The catalytic residues (listed in Table 1) are shown by the arrows to be lie near the minima in the respective global mobility distribution curves.

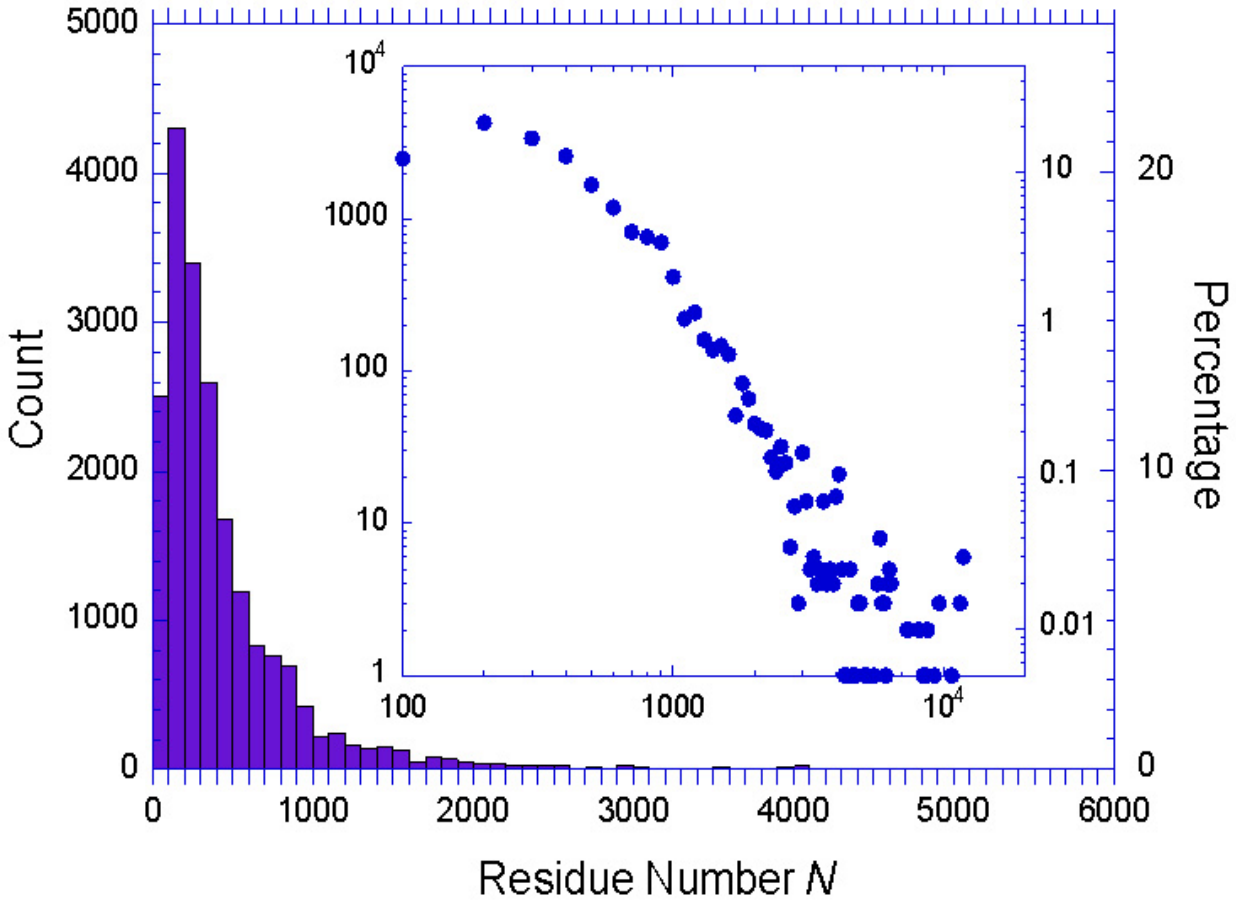
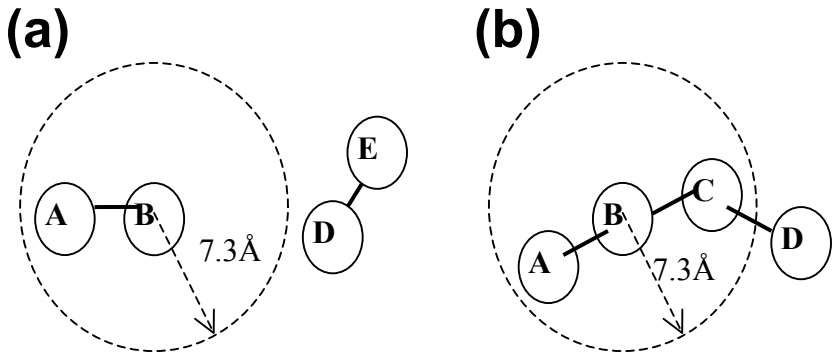

Count
Residue Number N
Percentage
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
10^4
1000
100
10
1
100
1000
10^4
10
1
0.1
0.01
20
10
0

Figure 1



$$\Gamma_{DIS} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \qquad \Gamma_{C} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

$$\Lambda_{DIS} = \begin{bmatrix} 2.0 & 0 & 0 & 0 \\ 0 & 2.0 & 0 & 0 \\ 0 & 0 & 0.0 & 0 \\ 0 & 0 & 0 & 0.0 \end{bmatrix} \qquad \Lambda_{C} = \begin{bmatrix} 3.4 & 0 & 0 & 0 \\ 0 & 2.0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.0 \end{bmatrix}$$

Figure 2

(a)

(b)

234 records have been found. Record 0-20 >>next page

QUERY REPORT & GNM CALCULATION

PDBID	Resolution	Title	GNM
1A2A	2.80 Å	Agkistrotoxin, A Phospholipase A2-Type Presynaptic Neurotoxin From Agkistrodon Halys Pallas	1A2A
1A3D	1.80 Å	Phospholipase A2 (Pla2) From Naja Naja Venom	1A3D
1A3F	2.65 Å	Phospholipase A2 (Pla2) From Naja Naja Venom	1A3F
1AE7	2.00 Å	Notexin, A Presynaptic Neurotoxic Phospholipase A2	1AE7
1AH7	1.50 Å	Phospholipase C From Bacillus Cereus	1AH7
1AII	1.95 Å	Annexin III Co-Crystallized With Inositol-2-Phosphate	1AII
1AIN	2.50 Å	Crystal structure of human annexin I at 2.5 A resolution.	1AIN
1AKN	2.80 Å	Structure Of Bile-Salt Activated Lipase	1AKN
1AOD	2.60 Å	Phosphatidylinositol-Specific Phospholipase C From Listeria Monocytogenes	1AOD
1AOK	2.00 Å	Vipoxin Complex	1AOK
1AQL	2.80 Å	Crystal Structure Of Bovine Bile-Salt Activated Lipase Complexed With Taurocholate	1AQL
1AX9	2.80 Å	Acetylcholinesterase Complexed With Edrophonium, Laue Data	1AX9
1AXN	1.78 Å	The high-resolution crystal structure of human annexin III shows subtle differences with annexin V.	1AXN
1AYP	2.57 Å	A Probe Molecule Composed Of 17-Percent Of Total Diffracting Matter Gives Correct Solutions In Molecular Replacement.	1AYP
1B4W	2.60 Å	Basic Phospholipase A2 From Agkistrodon Halys Pallas- Implications For Its Association and Anticoagulant Activities By X-Ray Crystallography	1B4W

Figure 3

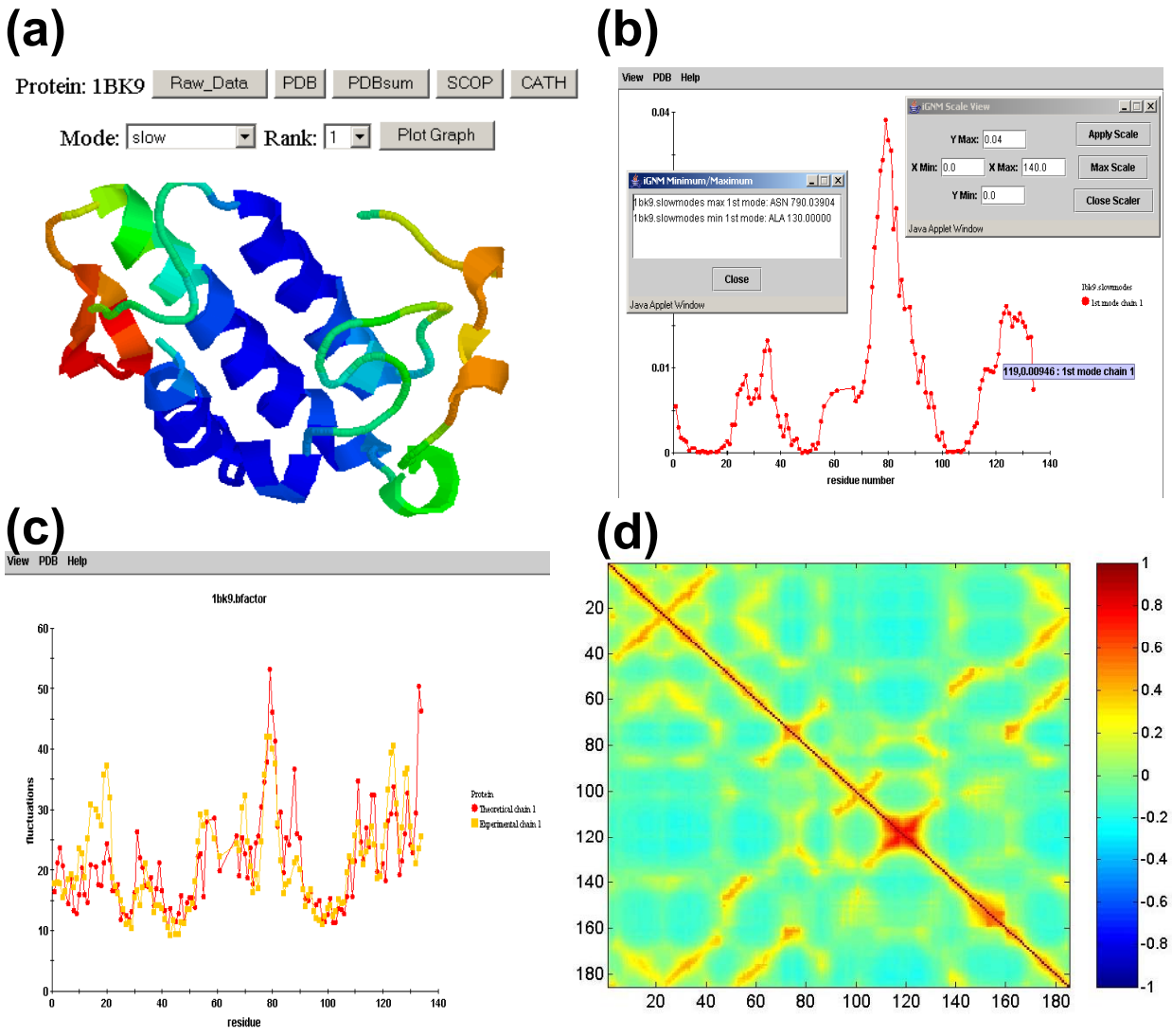

(a)
Protein: 1BK9
Raw_Data
PDB
PDBsum
SCOP
CATH
Mode: slow
Rank: 1
Plot Graph
(b)
View PDB Help
iGNM Scale View
Y Max: 0.04
Apply Scale
X Min: 0.0
X Max: 140.0
Max Scale
Y Min: 0.0
Close Scaler
Java Applet Window
iGNM Minimum/Maximum
1bk9.slowmodes max 1st mode: ASN 790.03904
1bk9.slowmodes min 1st mode: ALA 130.00000
Close
Java Applet Window
1bk9.slowmodes
1st mode chain 1
119,0.00946 : 1st mode chain 1
residue number
(c)
View PDB Help
1bk9.bfactor
fluctuations
residue
Protein
Theoretical chain 1
Experimental chain 1
(d)

Figure 4

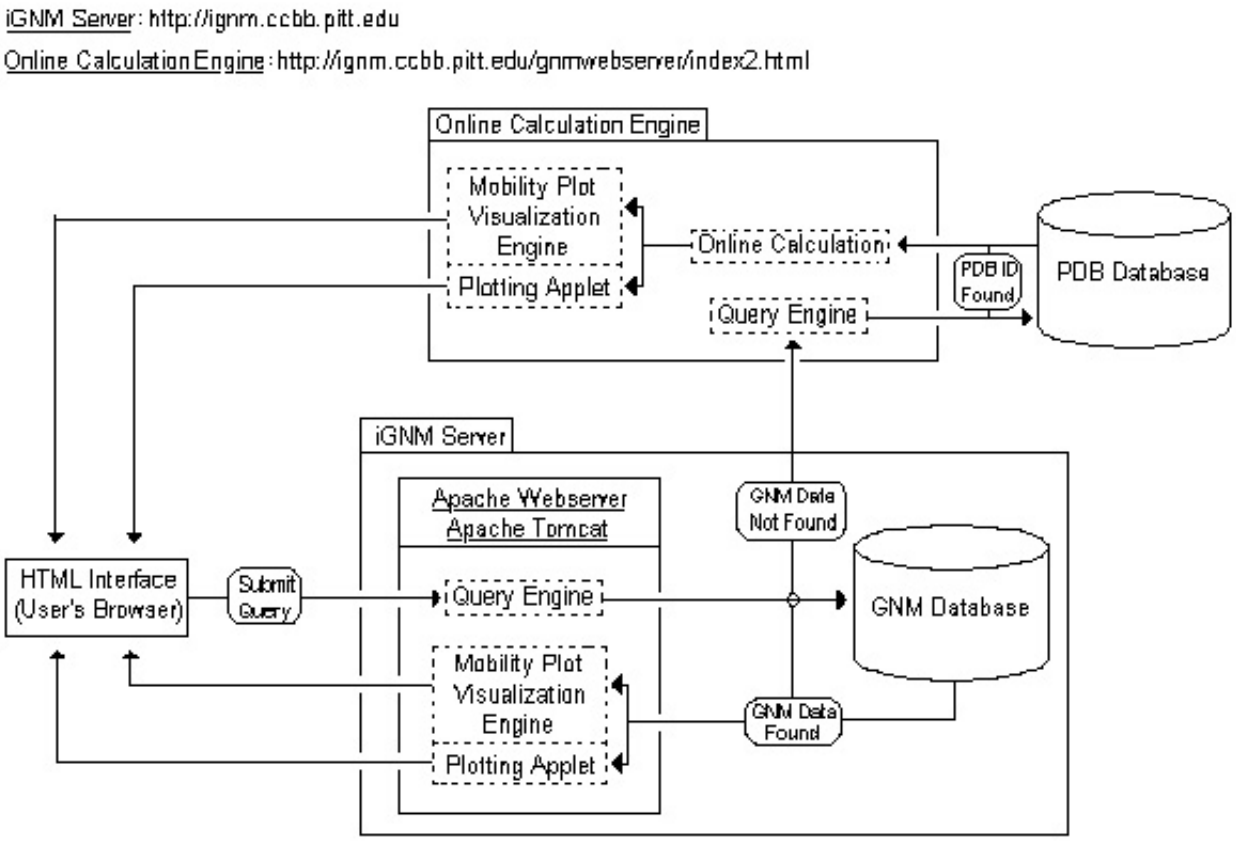

iGNM Server: http://ignm.ccbb.pitt.edu
Online Calculation Engine: http://ignm.ccbb.pitt.edu/gnmwebserver/index2.html
Online Calculation Engine
Mobility Plot Visualization Engine
Plotting Applet
Online Calculation
PDB ID Found
PDB Database
Query Engine
iGNM Server
Apache Webserver
Apache Tomcat
GNM Data Not Found
HTML Interface (User's Browser)
Submit Query
Query Engine
GNM Database
Mobility Plot Visualization Engine
Plotting Applet
GNM Data Found

Figure 5

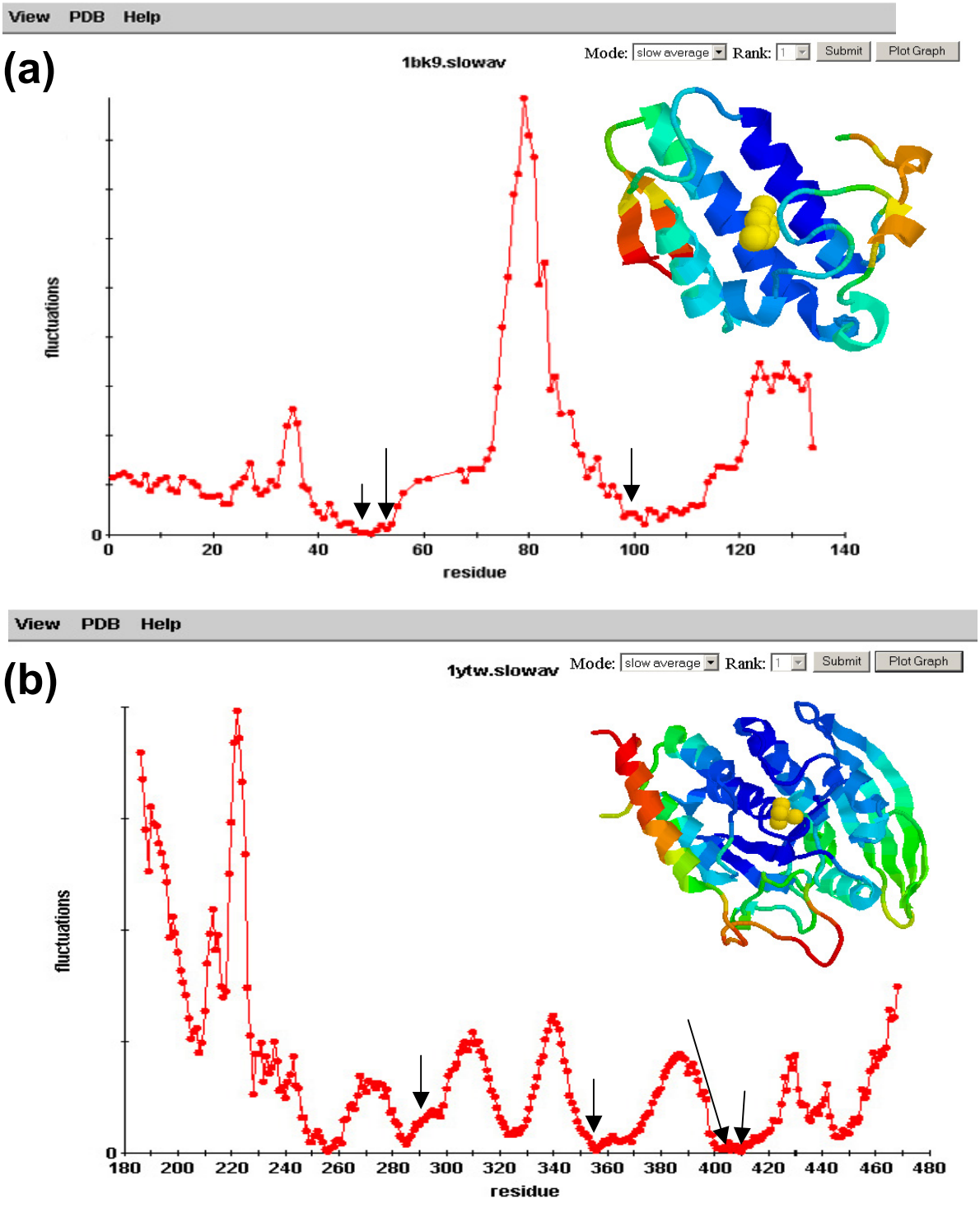

View PDB Help
(a)
1bk9.slowav
Mode: slow average Rank: 1 Submit Plot Graph
fluctuations
0
0 20 40 60 80 100 120 140
residue
View PDB Help
(b)
1ytw.slowav
Mode: slow average Rank: 1 Submit Plot Graph
fluctuations
0
180 200 220 240 260 280 300 320 340 360 380 400 420 440 460 480
residue

Figure 6