BIOMOLECULAR DYNAMICS REVEALED BY ELASTIC NETWORK MODELS AND THE STUDY OF MECHANICAL KEY SITES FOR LIGAND BINDING

by

Lee-Wei Yang

BS, National Taiwan University, 1997

MS, National Tsing Hua University , 1999

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Lee-Wei Yang

It was defended on

June 13, 2005

and approved by

Signature	**Printed Name**	**Date Signed**
_____ (Committee member)	Dr. Michael Cascio, Assistant Professor	_____
_____ (Committee member)	Dr. Jeffry Madura, Associate Professor	_____
_____ (Committee member)	Dr. Hagai Meirovitch, Professor	_____
_____ (Committee member)	Dr. Alan Russell, Professor	_____
_____ (Dissertation Director)	Dr. Ivet Bahar, Professor	_____

BIOMOLECULAR DYNAMICS REVEALED BY ELASTIC NETWORK MODELS AND THE STUDY OF MECHANICAL KEY SITES FOR LIGAND BINDING

Lee-Wei Yang, PhD

University of Pittsburgh, 2005

The Gaussian network model (GNM) can be used as a first approximation for describing the fluctuation dynamics of proteins, the limits of applicability and the range of validity of the model parameters need to be established. A systematic analysis of the GNM predictions is done within the scope of this thesis, and the potential utility of GNM for elucidating structure-dynamics-function relations in enzymes is explored. The application of the GNM to a set of 183 non-homologous proteins shows that it can predict the X-ray crystallographic temperature factors more precisely than full-atomic normal mode analysis (NMA) does. Furthermore, the application to 1250 non-redundant proteins indicates that the GNM predictions agree better with NMR solution data, than X-ray crystallographic, and measurements taken at high diffraction temperatures. A systematic study of 98 enzymes that belong to different enzyme classes (EC) shows that catalytic residues are distinguished by their restricted mobilities in the global modes. The amplitudes of their fluctuations rank in the lowest 7% range amongst the rank-ordered mobilities of all residues. Catalytic residues also bear more restricted mobilities than their 4 flanked neighbors in sequence and this feature holds for more than 70% of the examined catalytic residues, suggesting a communication between chemical activity and molecular mechanics. The observed restricted mobility of catalytic residues is used as a criterion for identifying active sites of enzymes in a newly developed algorithm (COMPACT). The method shows a high sensitivity and a moderate-to-low specificity for a set of representative monomeric

enzymes. All the false-positives predicted by COMPACT are found to be highly conserved, suggesting that their finely tuned dynamics results from evolutionary pressure. These particular sites are proposed to serve as alternative drug binding targets. We have implemented this tool in *i*GNM, a database of protein dynamics. Protein dynamics stored in *i*GNM or computed from the online calculation server (*o*GNM) have assisted in identifying possible silver ion binding residue in creatinase and describing the loop mobilities of low-fidelity DNA polymerase. Over all, this dissertation supports the view that protein structures have been designed to undergo conformational changes that are required for their biological functions.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. DYNAMICS AND FUNCTION

1.1.1. Conformational dynamics: a bridge between structure and function

With recent advances in sequencing genomes, it has become clear that the canonical sequence-to-function paradigm is far from being sufficient. Structure has emerged as an important source of additional information required for understanding the molecular basis of observed biological activities. Nowadays, the accumulated structural information presents a unique opportunity for high-throughput assessment of structure → function relation. Several groups are now engaged in protein structure characterization (Service, 2000), and the size of the Protein Data Bank (PDB)(Berman et al., 2000) is growing exponentially. The elucidation of structures permits us to develop structure-based tools for characterizing dynamics. A wealth of theoretical (Berendsen et al., 2000; Abseher et al. 2000) and experimental (Wand et al., 2001; Goodman et al., 2000) studies provide evidence for the close link between dynamics and function (Frauenfelder et al., 1998; Stock, 1999).

Yet, advances in structural genomics have now demonstrated that structural knowledge is not sufficient for understanding the molecular mechanisms of biological function either. The connection between structure and function presumably lies in *dynamics*, suggesting an encoding paradigm of sequence to structure to dynamics to function.

1.1.2. Functional motions are cooperative fluctuations near the protein native state.

While accurate sampling of conformational space is a challenge for macromolecular systems, the study of protein dynamics benefits from a great simplification: proteins have uniquely defined native structures under physiological conditions, and they are functional only when folded into their native conformation. Therefore, while the motions of macromolecules in solution are quite complex and involve transitions between an astronomical number of conformations, those of proteins near native state conditions are much simpler, as they are confined to a subset of conformations, or microstates, near the folded state. These microstates usually share the same overall fold, secondary structural elements and even tertiary contacts within individual domains. Typical examples are the open and closed forms of enzymes, usually adopted in the unliganded and liganded states, respectively. Exploring the *fluctuation dynamics* of proteins near native state conditions is a first step towards gaining insights about the molecular basis and mechanisms of their function; and fluctuation dynamics can be treated to a good approximation by linear models – such as Normal Mode Analysis (NMA).

Another distinguishable property of protein dynamics – in addition to confinement to a small subspace of conformations – is the *collective* nature of residue fluctuations. The fluctuations are indeed far from random, involving the correlated motions of large groups of atoms, residues, or even entire domains or molecules whose concerted movements underlie biological function. An analytical approach that takes account of the collective coupling between all residues is needed, and again NMA emerges as a reasonable first approximation.

1.2. TWO APPROACHES FOR DESCRIBING PROTEIN DYNAMICS: MOLECULAR DYNAMICS SIMULATIONS AND NORMAL MODE ANALYSIS

1.2.1. Molecular dynamics simulations

Not surprisingly, a major endeavor in recent years has been to develop models and methods for simulating the dynamics of proteins, and relating the observed behavior to experimental data. In the last decades, Molecular Dynamics (MD) simulations have proven to be a promising strategy for generating conformational trajectories for macromolecules in order to visualize the correlation of their dynamics to the biological functions (Brooks et al., 1983; Bahar et al., 1998a; Temiz et al., 2002; Ming et al., 2003a; Cui et al., 2004). In principle, the MD trajectories can be subjected to a Principal Component Analysis (PCA) to decompose motions into different modes (Berendsen et al., 2000). The motions induced by the modes with low frequencies, delineate the global dynamics of the molecules and bear functional significance (Kitao et al., 1999; Berendsen et al., 2000).

The process of extracting the dominant collective modes, or the essential dynamics (Amadei et al., 1993; de Groot et al., 1998) from fluctuations observed in MD trajectories—also called principal component analysis (PCA; Kitao et al., 1991), or the molecular optimal dynamics coordinates analysis (Garcia et al., 1996)—is now an established computational means of studying proteins' dynamics (Kitao et al., 1999). The major shortcoming of this approach is the sampling inefficiency of MD simulations. The sampling problem becomes increasingly important as the size of the investigated molecular system increases, as shown by projecting the MD trajectory onto the first few computers (Clarage et al., 1995; Caves et al., 1998). In general,

multiple independent runs are needed for assessing equilibrium and convergence (Caves et al., 1998; Smith et al., 2002).

MD has also been impeded by the memory and time cost of the computation. The simulations at full atomic scale usually give atomic trajectories in the sub-nanosecond to nanoseconds time range (Leach, 2001), which usually fail to give a suitable time frame to observe biologically meaningful movements such as domain motions, protein folding or allosteric conformation changes that take from nanoseconds to microseconds (Clarage et al., 1995). Hence, a more simplified model is needed to obtain dynamics in large time scale by trading off detailed atomic trajectories and noises.

1.2.2. Normal Mode Analysis (NMA)

A view that emerges from many studies is that proteins possess a tendency, encoded in their three-dimensional (3D) structures, to reconfigure into functional forms, i.e. each native structure tends to undergo conformational changes that facilitate its biological function. An efficient method for identifying such functional motions is NMA, a method that has found widespread use in physical sciences for characterizing molecular fluctuations near a given equilibrium state. The utility of NMA as a physically plausible and mathematically tractable tool for exploring protein dynamics has been recognized for the last 20 years (Brooks et al., 1983; Go et al., 1983). With recent increases in computational power and speed the application of NMA to proteins has gained renewed interest and popularity.

The method has later been extended into a quasiharmonic oscillator approximation which utilizes, as input, the fluctuations (auto- and cross-correlations) observed in MD simulations, thus including the effects of anharmonicity (Karplus et al., 1981; Levy et al., 1984; de Groot BL et al., 1998). The major weakness of the analytical approaches appears, on the other hand, to be their inadequacy to account for the anharmonic motions or multimeric transitions driven by the slowest collective mode observed in MD.

1.2.3. Limitations of existing computational approaches

As mentioned above, MD simulations and PCA-based analyses at full atomic scale are usually held back by computing time and sampling inefficiencies (Clarage et al., 1995). Localized or fast processes accessed by MD convey little information on the collective dynamics of large structures. NMA, on the other hand, demands pre-equilibrium energy minimization of the structure of interest and sophisticated full-atomic force field, which complicates its application to large bio-complexes and supramolecular assemblies. The passages to longer times and larger scales necessitate the adoption of less detailed models, hence the use of continuum models for solvent (Kollman et al., 2000; Wang et al. 2000), Poisson-Boltzmann for electrostatics (Archontis et al. 2001; Baker et al. 2001), and more recently elastic network (EN)-based models (Tirion, 1996; Bahar et al., 1997a; Hinsen et al., 1999; Doruker et al., 2000; Delarue et al., 2002; Ming et al., 2002a, b; Tama et al. 2002b) for global machinery.

1.3. SIMPLIFIED NMA MODELS: ELASTIC NETWORK (EN) MODELS

To simulate the relatively longer time scale dynamics of large biomolecules of biomolecular assemblies, and yet maintain computational efficiency, lower resolution models have been adopted where groups of residues are clustered into unified sites (Kurkcuoglu et al., 2004; Doruker et al.,2002), or rigid blocks (RTB and BNM; Tama et al., 2000 and Li et al., 2002 respectively). Related methods, such as quantized elastic deformational model (QEDM), effectively quantize the shape of the structure without directly identifying specific residues or groups of residues (Ming et al., 2002a; Tama et al. 2002b). A reduction in the number of nodes by one order of magnitude increases the computation speed by three orders of magnitude since NMA computing time scales with N^3. Notably, the global motions computed by such coarse-grained NMA maintain their fundamental characteristics that can be related to functional mechanisms (Doruker et al., 2002).

The Gaussian Network model (GNM) is probably the simplest among these EN-based models. This EN model has been originally introduced at the residue level (Bahar et al., 1997a; Haliloglu et al., 1997), inspired by the full atomic NMA of Tirion. In this seminal work, Tirion demonstrated that the mode dispersion and slowest modes obtained for G-actin with a uniform harmonic potential do not practically differ from those obtained by NMA using a detailed (Charmm) force field (Tirion, 1996). Despite its simplicity, the GNM and its extension, the Anisotropic Network Model (ANM; Doruker et al., 2000; Atilgan et al., 2001), or similar coarse-grained EN models combined with NMA (Hinsen et al., 1999; Tama et al., 2001; Li et al., 2002),

have proven to provide insightful information on biomolecular dynamics and found widespread use since then for elucidating the dynamics of proteins and their complexes.

Significantly, the simplified NMAs with EN models have been applied to deduce both the machinery and conformational dynamics of large structures and assemblies including HIV reverse transcriptase (Bahar et al, 1999b; Temiz et al., 2002), hemagglutinin A (Isin et al., 2002), aspartate transcarbamylase (Thomas et al., 1999), F1 ATPase (Cui et al., 2004), RNA polymerase (Van Wynsberghe et al. 2004), an actin segment (Ming et al., 2003a), GroEL-GroES (Keskin et al., 2002a), the ribosome (Tama et al., 2003b; Wang et al., 2004), and viral capsids (Tama et al., 2002a; 2005; Rader et al., 2005).

The theoretical foundations of the GNM will be presented in the next section, along with a few applications that illustrate its utility. The following questions will be addressed. What is the Gaussian Network Model? What are the underlying assumptions? How is it implemented? Why and how does it work? How does the GNM analysis differ from NMA applied to EN models? Or what are the advantages and limitations compared to coarse-grained NMA? What are the most significant applications and prospective utilities of the GNM, or the EN models in general?

1.4. THE GAUSSIAN NETWORK MODEL

1.4.1. Theory

1.4.1.1. A minimalist model for fluctuation dynamics

Most analytical treatments of complex systems dynamics entail a compromise between physical realism and mathematical tractability. A challenge is to identify the simplest, yet physically plausible, model that retains the physical and chemical characteristics that are needed for the time and length scales of interest. Clearly, as the size and length scales of the processes of interest increase, it becomes unnecessary to account for many of the microscopic details in the model. The inclusion of these microscopic details could, on the contrary, tend to obscure the dominant patterns characterizing the biological function of interest.

GNM was proposed by Bahar, Atilgan and Erman (Bahar et al., 1997a) within such a minimalist mindset to explore the role and contribution of *purely topological constraints*, defined by the 3D structure, on the collective dynamics of proteins. Inspired by the seminal work of Flory and collaborators applied to polymer gels (Flory, 1976), each protein is modeled by an EN (Figure *1-1*), the dynamics of which is entirely defined by network topology. The position of the nodes of the EN are defined by the C^{α}-atom coordinates, and the springs connecting the nodes are representative of the bonded and non-bonded interactions between the pairs of residues located within an interaction range, or cutoff distance, of r_c. The cutoff distance is usually taken as 7.0 Å, based on the radius of the first coordination shell around residues observed in PDB structures (Miyazawa et al., 1985; Bahar et al., 1997b).

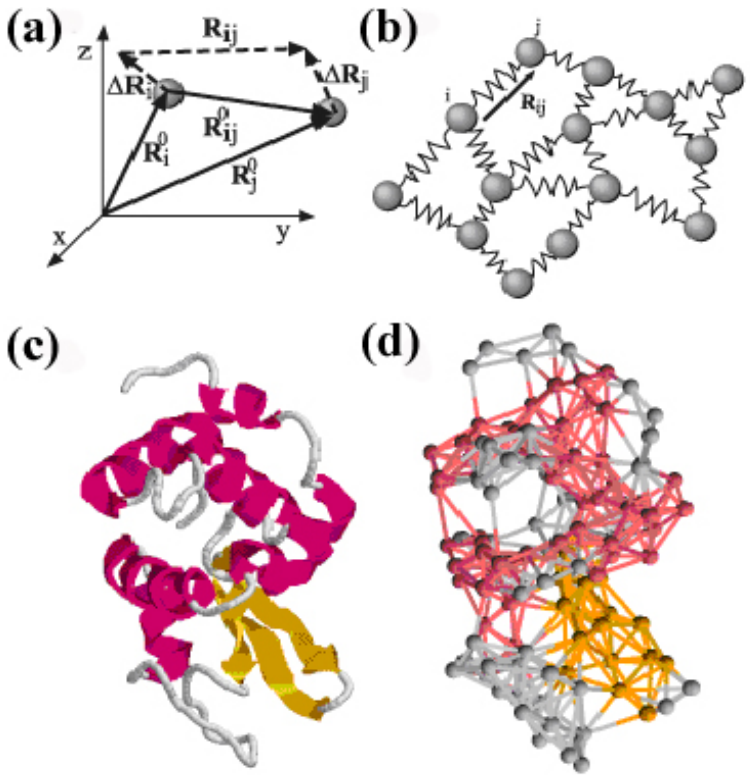

Figure 1-1 Description of the Gaussian network model (GNM).
(a) Schematic representation of the equilibrium positions of the i^{th} and j^{th} nodes, $\boldsymbol{R}_i^0$ and $\boldsymbol{R}_j^0$, with respect to a laboratory-fixed coordinate system (xyz). The instantaneous fluctuation vectors, $\Delta\boldsymbol{R}_i$ and $\Delta\boldsymbol{R}_j$, are shown by the dashed arrows, as well as the instantaneous separation vector $\boldsymbol{R}_{ij}$ between the positions of the two residues. $\boldsymbol{R}_{ij}^0$ is the equilibrium distance between nodes i and j. **(b)** In the elastic network of GNM every residue is represented by a node and connected to spatial neighbors by uniform springs. These springs determine the N-1 degrees of freedom in the network and the structure's modes of vibration. **(c)** Three dimensional image of hen egg white lysozyme (PDB ID:1hel; Wilson et al., 1992) showing the C^α trace. Secondary structure features are indicated by pink for helices and yellow for β-strands. **(d)** Using a cutoff value of 10Å, all connections between C^α nodes are drawn for the same lysozyme structure to indicate the nature of the elastic network analyzed by GNM.

1.4.1.2. GNM assumes the fluctuations to be isotropic and Gaussian

If we denote the equilibrium position vectors of a node, i, by $\boldsymbol{R}^0{}_i$, and the instantaneous position by $\boldsymbol{R}_i$, the fluctuations, or deformations, from this mean position can then be defined in terms of the fluctuation vector $\Delta\boldsymbol{R}_i = \boldsymbol{R}_i - \boldsymbol{R}_i^0$. The fluctuations in the distance vector $\boldsymbol{R}_{ij}$ between residues

i and *j*, can in turn be expressed as $\Delta \boldsymbol{R}_{ij} = \boldsymbol{R}_{ij} - \boldsymbol{R}_{ij}^0 = \Delta \boldsymbol{R}_j - \Delta \boldsymbol{R}_i$ (Figure *1-1a*). By assuming that these fluctuations are *isotropic* and *Gaussian* we can write the potential of the network of *N* nodes (residues), V_{GNM}, in terms of the components ΔX_i, ΔY_i and ΔZ_i of $\Delta \boldsymbol{R}_i$, as

$$V_{GNM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N} \Gamma_{ij}\left[\left(\Delta X_i - \Delta X_j\right)^2 + \left(\Delta Y_i - \Delta Y_j\right)^2 + \left(\Delta Z_i - \Delta Z_j\right)^2\right]\right] \qquad (1\text{-}1)$$

Here Γ_{ij} is the ijth element of the Kirchhoff (or connectivity) matrix of inter-residue contacts defined by

$$\Gamma_{ij} = \begin{cases} -1 & \textbf{if } i \neq j \quad \textbf{and } \boldsymbol{R}_{ij} \leq r_c \\ 0 & \textbf{if } i \neq j \quad \textbf{and } \boldsymbol{R}_{ij} > r_c \\ -\sum_{j, j \neq i} \Gamma_{ij} & \textbf{if } i = j \end{cases} \qquad (1\text{-}2)$$

and γ is the force constant taken to be uniform for all network springs. Expressing the *X-, Y-* and *Z-* components of the fluctuation vectors $\Delta \boldsymbol{R}_i$ as three *N*-dimensional vectors $\Delta \boldsymbol{X}$, $\Delta \boldsymbol{Y}$ and $\Delta \boldsymbol{Z}$, equation 1-1 simplifies to

$$V_{GNM} = \frac{\gamma}{2}\left[\Delta \boldsymbol{X}^T \boldsymbol{\Gamma} \Delta \boldsymbol{X} + \Delta \boldsymbol{Y}^T \boldsymbol{\Gamma} \Delta \boldsymbol{Y} + \Delta \boldsymbol{Z}^T \boldsymbol{\Gamma} \Delta \boldsymbol{Z}\right] \qquad (1\text{-}3)$$

Here $\Delta \boldsymbol{X}^T$, $\Delta \boldsymbol{Y}^T$ and $\Delta \boldsymbol{Z}^T$ are the row vectors $[\Delta X_1 \ \Delta X_2 \ ..... \ \Delta X_N]$, $[\Delta Y_1 \ \Delta Y_2 \ ..... \ \Delta Y_N]$ and $[\Delta Z_1 \ \Delta Z_2 \ ..... \ \Delta Z_N]$, respectively. The total potential can alternatively be expressed as

$$V_{GNM} = \frac{\gamma}{2}\left[\Delta \boldsymbol{R}^T (\boldsymbol{\Gamma} \otimes \boldsymbol{E}) \Delta \boldsymbol{R}\right] \qquad (1\text{-}4)$$

where $\Delta \boldsymbol{R}$ is the 3N-dimensional vector of fluctuations, $\Delta \boldsymbol{R}^T$ is its transpose, $\Delta \boldsymbol{R}^T = [\Delta X_1 \ \Delta Y_1 \ ..... \ \Delta Z_N]$, $\boldsymbol{E}$ is the identity matrix of order 3, and $(\boldsymbol{\Gamma} \otimes \boldsymbol{E})$ is the direct product of $\boldsymbol{\Gamma}$ and $\boldsymbol{E}$,

found by replacing each element Γ_{ij} of $\boldsymbol{\Gamma}$ by the 3×3 diagonal matrix $\Gamma_{ij}\boldsymbol{E}$. One should note that by construction the eigenvalues for this $3N\times3N$ matrix, $\boldsymbol{\Gamma}\otimes\boldsymbol{E}$, are 3-fold degenerate. This degeneracy arises from the isotropic assumption, further explored in the next section.

1.4.1.3. Statistical mechanical foundations of the GNM

What we are primarily interested in is determining the mean-square fluctuations of a particular residue, *i*, or the correlations between the fluctuations of two different residues, *i* and *j*. These respective properties are given by

$$\left\langle \Delta\boldsymbol{R}_i \bullet \Delta\boldsymbol{R}_i \right\rangle = \left\langle \Delta X_i^{\,2} \right\rangle + \left\langle \Delta Y_i^{\,2} \right\rangle + \left\langle \Delta Z_i^{\,2} \right\rangle \tag{1-5}$$

and

$$\left\langle \Delta\boldsymbol{R}_i \bullet \Delta\boldsymbol{R}_j \right\rangle = \left\langle \Delta X_i \Delta X_j \right\rangle + \left\langle \Delta Y_i \Delta Y_j \right\rangle + \left\langle \Delta Z_i \Delta Z_j \right\rangle \tag{1-6}$$

Thus, if we know how to compute the component fluctuations $\left\langle \Delta X_i^{\,2} \right\rangle$ and $\left\langle \Delta X_i \Delta X_j \right\rangle$ then we know how to compute the residue fluctuations and their cross-correlations.

In the GNM, the probability distribution of all fluctuations, $P(\boldsymbol{\Delta R})$ is *isotropic* (equation 1-7) and *Gaussian* (equation 1-8), i.e.

$$P(\boldsymbol{\Delta R}) = P(\boldsymbol{\Delta X}, \boldsymbol{\Delta Y}, \boldsymbol{\Delta Z}) = p(\boldsymbol{\Delta X})\, p(\boldsymbol{\Delta Y})\, p(\boldsymbol{\Delta Z}) \tag{1-7}$$

and

$$\begin{aligned} p(\boldsymbol{\Delta X}) &\propto \exp\left\{ -\frac{\gamma}{2k_B T} \boldsymbol{\Delta X}^T \boldsymbol{\Gamma} \boldsymbol{\Delta X} \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\Delta X}^T \left(\frac{k_B T}{\gamma} \boldsymbol{\Gamma}^{-1} \right)^{-1} \boldsymbol{\Delta X} \right) \right\} \end{aligned} \tag{1-8}$$

where k_B is the Boltzmann constant and T is the absolute temperature. Similar forms apply to $p(\Delta \boldsymbol{Y})$ and $p(\Delta \boldsymbol{Z})$. $\Delta \boldsymbol{X} = [\Delta X_1\ \Delta X_2\ \ldots\ \Delta X_i\ \ldots\ \Delta X_N]$ is therefore a multidimensional Gaussian random variable with zero mean and covariance $\frac{k_B T}{\gamma}\boldsymbol{\Gamma}^{-1}$ in accord with the general definition (Papoulis et al., 1965)

$$W(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Xi}) = \frac{1}{(2\pi)^{N/2}\left|\boldsymbol{\Xi}\right|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Xi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1-9}$$

for multidimensional Gaussian (normal) probability density function associated with a given N-dimensional vector $\boldsymbol{x}$ having mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Xi}$. Here, the term in the denominator, $(2\pi)^{N/2}|\boldsymbol{\Xi}|^{1/2}$, is the partition function that ensures the normalization of $W(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Xi})$ upon integration over the complete space of accessible $\boldsymbol{x}$, and $|\boldsymbol{\Xi}|$ is the determinant of $\boldsymbol{\Xi}$. Similarly, the *normalized* probability distribution $p(\Delta \boldsymbol{X})$ is

$$p(\boldsymbol{\Delta X}) = \frac{1}{Z_X}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\Delta X}^T\left(\frac{k_B T}{\gamma}\boldsymbol{\Gamma}^{-1}\right)^{-1}\boldsymbol{\Delta X}\right)\right\} \tag{1-10}$$

where $Z_{\boldsymbol{X}}$ is the partition function given by

$$Z_X = \int \exp\left\{-\frac{1}{2}\left(\boldsymbol{\Delta X}^T\left(\frac{k_B T}{\gamma}\boldsymbol{\Gamma}^{-1}\right)^{-1}\boldsymbol{\Delta X}\right)\right\} d\boldsymbol{\Delta X} \quad = (2\pi)^{N/2}\left|\frac{k_B T}{\gamma}\boldsymbol{\Gamma}^{-1}\right|^{1/2} \tag{1-11}$$

The same expression is valid for $Z_{\boldsymbol{Y}}$ and $Z_{\boldsymbol{Z}}$ such that the overall GNM partition function (or configurational integral) becomes

$$Z_{GNM} = Z_X Z_Y Z_Z \ = (2\pi)^{3N/2}\left|\frac{k_B T}{\gamma}\boldsymbol{\Gamma}^{-1}\right|^{3/2} \tag{1-12}$$

Now we have the statistical mechanical foundations to write the expectation values of the residue fluctuations, $\left\langle \Delta X_i^2 \right\rangle$, and correlations, $\left\langle \Delta X_i \Delta X_j \right\rangle$. It can be verified that the $N \times N$ covariance matrix $<\Delta \boldsymbol{X}\, \Delta \boldsymbol{X}^T>$ is equal to $\frac{k_B T}{\gamma} \boldsymbol{\Gamma}^{-1}$, using the statistical mechanical average*

$$<\Delta \boldsymbol{X}\, \Delta \boldsymbol{X}^T> = \int \Delta \boldsymbol{X}\, \Delta \boldsymbol{X}^{\mathrm{T}}\; p(\Delta \boldsymbol{X})\, d\, \Delta \boldsymbol{X} \;=\; \frac{k_B T}{\gamma} \boldsymbol{\Gamma}^{-1} \qquad \text{(1-13)}$$

Because

$$<\Delta \boldsymbol{X}\, \Delta \boldsymbol{X}^T> = <\Delta \boldsymbol{Y}\, \Delta \boldsymbol{Y}^T> = <\Delta \boldsymbol{Z}\, \Delta \boldsymbol{Z}^T> = (1/3) <\Delta \boldsymbol{R}\, \Delta \boldsymbol{R}^T>, \qquad \text{(1-14)}$$

we obtain

$$\begin{aligned} \left\langle \Delta \boldsymbol{R}_i^2 \right\rangle &= \frac{3 \boldsymbol{k_B T}}{\gamma} (\boldsymbol{\Gamma}^{-1})_{ii} \\ \left\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \right\rangle &= \frac{3 \boldsymbol{k_B T}}{\gamma} (\boldsymbol{\Gamma}^{-1})_{ij} \end{aligned} \qquad \text{(1-15)}$$

as the mean-square (ms) fluctuations of residues and correlations between residue fluctuations. It should be noted that the assumption of *isotropic fluctuations* (equation 1-8) is intrinsic to the GNM. Thus the 3-*N* dimensional problem (equation 1-4) can be reduced to an *N*-dimensional one described by equation 1-15.

---

* Note that solving equation 1-13 involves the ratio of the multidimensional Gaussian counterparts for the two integrals $\int \exp\{-ax^2\} dx = \frac{1}{2}\sqrt{\left(\frac{\pi}{a}\right)}$ and $\int x^2 \exp\{-ax^2\} dx = \frac{\sqrt{\pi}}{4} a^{-3/2}$ in the range (0, ∞) such that $\left\langle x^2 \right\rangle = \frac{\sqrt{\pi}}{4} a^{-3/2} \Big/ \frac{1}{2}\sqrt{\left(\frac{\pi}{a}\right)} = \frac{1}{2a}$. For the simplest case of a single spring, subject to harmonic potential $\frac{1}{2}\gamma x^2$, $a = \frac{\gamma}{2k_B T}$ and $<x^2> = \frac{k_B T}{\gamma}$.

1.4.1.4. Influence of local packing density

The diagonal elements of the Kirchhoff matrix, Γ_{ii}, are equal to the residue coordination numbers, z_i ($1 \leq i \leq N$), which represent the degree of the EN nodes in graph theory. Thus z_i is a direct measure of *local packing density* around the i^{th} residue. To better understand this, it is possible to express $\boldsymbol{\Gamma}$ as a sum of two matrices $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$, consisting exclusively of the diagonal and off-diagonal elements of $\boldsymbol{\Gamma}$, respectively. Using these two matrices, $\boldsymbol{\Gamma}^{-1}$ may be written as

$$\boldsymbol{\Gamma}^{-1} = [\boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_2]^{-1} = [\boldsymbol{\Gamma}_1 (\mathbf{E} + \boldsymbol{\Gamma}_1^{-1} \boldsymbol{\Gamma}_2)]^{-1} = (\mathbf{E} + \boldsymbol{\Gamma}_1^{-1} \boldsymbol{\Gamma}_2)^{-1} \boldsymbol{\Gamma}_1^{-1}$$

$$= (\mathbf{E} - \boldsymbol{\Gamma}_1^{-1} \boldsymbol{\Gamma}_2 + ...) \boldsymbol{\Gamma}_1^{-1} = \boldsymbol{\Gamma}_1^{-1} - \boldsymbol{\Gamma}_1^{-1} \boldsymbol{\Gamma}_2 \boldsymbol{\Gamma}_1^{-1} + ... \qquad (1\text{-}16)$$

if one assumes that the invariants of the product $(\boldsymbol{\Gamma}_1^{-1} \boldsymbol{\Gamma}_2)$ are small compared to those of the identity matrix, $\mathbf{E}$, which is a valid approximation for protein Kirchhoff matrices. Thus, the information concerning local packing density and distribution of contacts is dominated by the diagonal matrix, $\boldsymbol{\Gamma}_1^{-1}$, which is the leading term in a series expansion for $\boldsymbol{\Gamma}^{-1}$ in equation 1-16. Consequently, application of equation 1-15 indicates that $< (\Delta \mathbf{R}_i)^2 >$ scales with $[\boldsymbol{\Gamma}_1^{-1}]_{ii} = 1/z_i$, to a first approximation. Thus the *local packing density* as described by the coordination numbers is an important structural property contributing to the mean square (ms) fluctuations of residues (Halle, 2002). However these coordination numbers represent only the leading contribution and not the entire set of dynamics described by equation 1-15.

1.4.2. Applications of GNM

1.4.2.1. Equilibrium fluctuations

The ms fluctuations of residues are experimentally measurable (e.g. X-ray crystallographic B-factors, or root mean-square (rms) differences between different models from NMR), and as such,

have often been used as an initial test for verifying and improving computational models and methods. Beginning with the original GNM paper (Bahar et al., 1997a), several applications have demonstrated that the fluctuations predicted by the GNM are in good agreement with experimental B-factors (Bahar et al., 1998b; Bahar, 1999a, b; Keskin et al., 2000a; Atilgan et al., 2001; Kundu et al., 2002). The B-factors are related to the expected residue fluctuations and calculated according to

$$B_i = (8\pi^2/3) < (\Delta \boldsymbol{R}_i)^2 > = (8\pi^2 k_B T/\gamma) [\boldsymbol{\Gamma}^{-1}]_{ii} \qquad (1\text{-}17)$$

Figure *1-2a* illustrates the agreement between the B-factors predicted by the GNM (solid curve) and those calculated from experimental data (open circles) for an example protein, ribonuclease T1 (RNase T1), where $\boldsymbol{\Gamma}$ has been constructed from the C^{α} coordinates for RNase T1 deposited in the PDB. Panel B compares the rms fluctuations predicted by the GNM and those observed across the 20 NMR models deposited in the PDB for reduced disulfide-bond formation facilitator (DsbA; Wilson et al., 1992). The correlation coefficient between the GNM results and experimental data for these two example proteins are 0.769 and 0.823 in the respective panels a and b. An extensive comparison of experimental and theoretical (GNM) B-factors for a series of PDB structures by Phillips and coworkers has shown that GNM calculations yield an average correlation coefficient ($R_{corr,avg}$) of 0.59 with experimental B-factors over 113 non-redundant high-resolution structures (Kundu et al., 2002). The agreement is even more prominent provided that the contacts between neighboring molecules in the crystal form are taken into account, which gives a $R_{corr,avg}$ of 0.66. The agreement with NMR data is also remarkable, pointing to the consistency between the fluctuations undergone in solution and those inferred from X-ray structures.

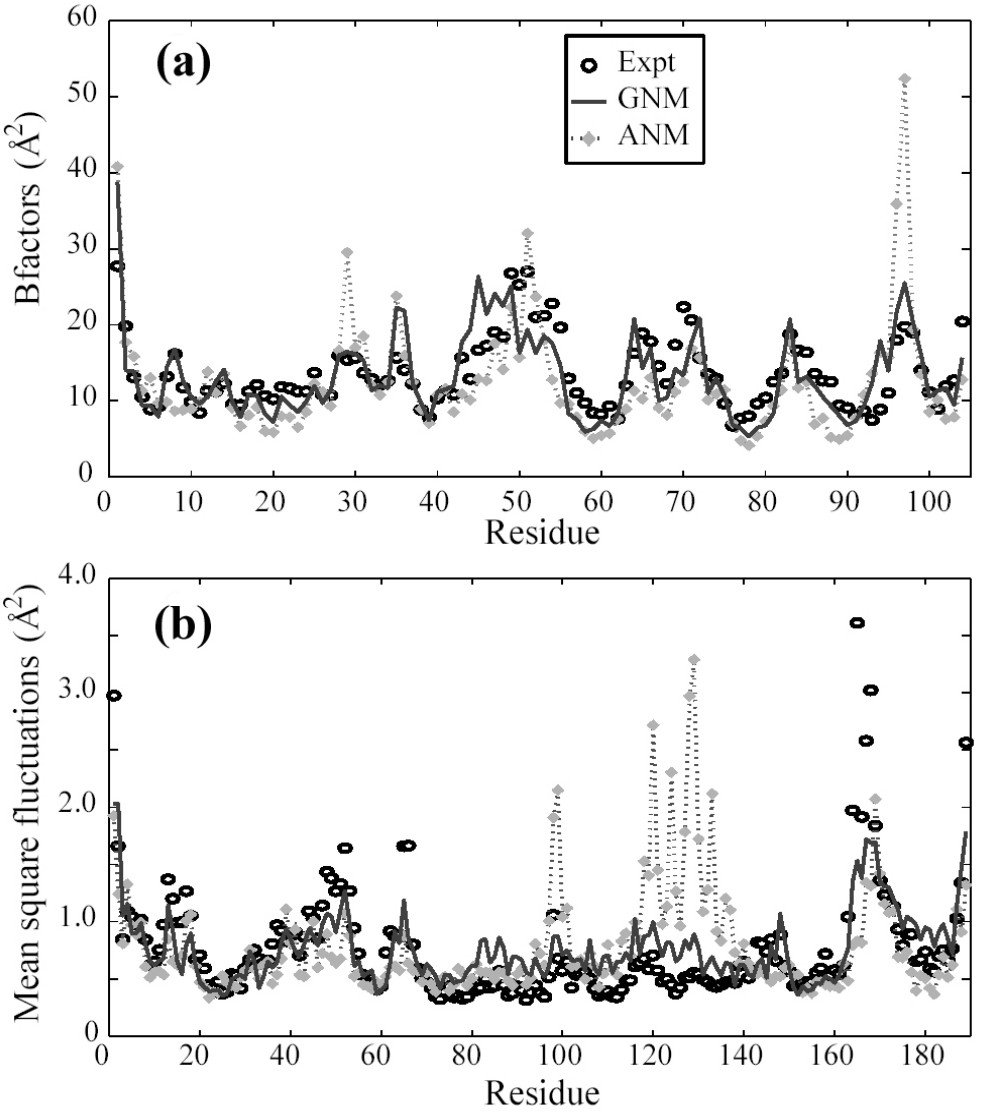

Figure 1-2 Compare fluctuations predicted by GNM and ANM with experimental observations. **(a)** Experimental X-ray crystallographic B-factors (open circles) reported for ribonuclease T1 (PDB ID:1bu4; Loris et al., 1999) plotted with calculated values from GNM (solid line) and ANM (dotted line) against residue number. **(b)** Root mean square (rms) deviation between C^{α} coordinates of NMR model structures (open circles) deposited for the reduced disulphide-bond formation facilitator (DsbA) in the PDB file 1a24 (Wilson et al., 1992).

1.4.2.2. Mode decomposition: Physical meaning of slow and fast modes

A major utility of the GNM is the ease of obtaining the collective modes of motion accessible to structures in native state conditions. The GNM normal modes are found by transforming the Kirchhoff matrix into a product of three matrices, the matrix $\boldsymbol{U}$ of the eigenvectors $\boldsymbol{u}_i$ of $\boldsymbol{\Gamma}$, the diagonal matrix $\boldsymbol{\Lambda}$ of eigenvalues λ_i, and the transpose $\boldsymbol{U}^T = \boldsymbol{U}^{-1}$ of the unitary matrix $\boldsymbol{U}$ as in equation 1-18.

$$\boldsymbol{\Gamma} = \boldsymbol{U}\,\boldsymbol{\Lambda}\,\boldsymbol{U}^T \qquad (1\text{-}18)$$

The eigenvalues are representative of the *frequencies* of the individual modes, while the eigenvectors define the *shapes* of the modes. The first eigenvalue, λ_1, is identically zero with the corresponding eigenvector comprised of elements with a magnitude equal to a constant, $1/\sqrt{N}$, indicative of an absence of internal motions in this *zero* mode. The vanishing frequency reflects the fact that the molecule can be translated rigidly without any potential energy change.

Combining equations 1-15 and 1-18, the cross-correlations between residue fluctuations can be written as a sum over the *N-1* nonzero modes ($2 \le k \le N$) using

$$\left\langle \Delta\boldsymbol{R}_i \bullet \Delta\boldsymbol{R}_j \right\rangle = \left(3\boldsymbol{k}_B\boldsymbol{T}/\gamma\right)\left[\boldsymbol{\Gamma}^{-1}\right]_{ij} = \left(3\boldsymbol{k}_B\boldsymbol{T}/\gamma\right)\left[\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\right]_{ij} = \left(3\boldsymbol{k}_B\boldsymbol{T}/\gamma\right)\sum_k \left[\lambda_k^{-1}\boldsymbol{u}_k\boldsymbol{u}_k^T\right]_{ij} \qquad (1\text{-}19)$$

This permits us to identify the correlation, $\left[\Delta\mathbf{R} \bullet \Delta\mathbf{R}\right]_k$ contributed by the k^{th} mode as

$$\left[\Delta\boldsymbol{R}_i \bullet \Delta\boldsymbol{R}_j\right]_k = \left(3\boldsymbol{k}_B\boldsymbol{T}/\gamma\right)\lambda_k^{-1}[\,\boldsymbol{u}_k\,]_i[\,\boldsymbol{u}_k\,]_j \qquad (1\text{-}20)$$

where $[\boldsymbol{u}_k]_i$ is the i^{th} element of $\boldsymbol{u}_k$. Because $\boldsymbol{u}_k$ is normalized, the plot of $[\boldsymbol{u}_k]_i^2$ against residue index, *i*, yields the normalized distribution of mean-square fluctuations of residues in the k^{th} mode, shortly referred to as the *k^{th} mode shape* (Figure *1-3a*). Because the residue fluctuations are

related to the experimental temperature (B-factors) by equation 1-17, these elements of $\mathbf{u}_k$ reflect the residue mobilities in the k^{th} mode.

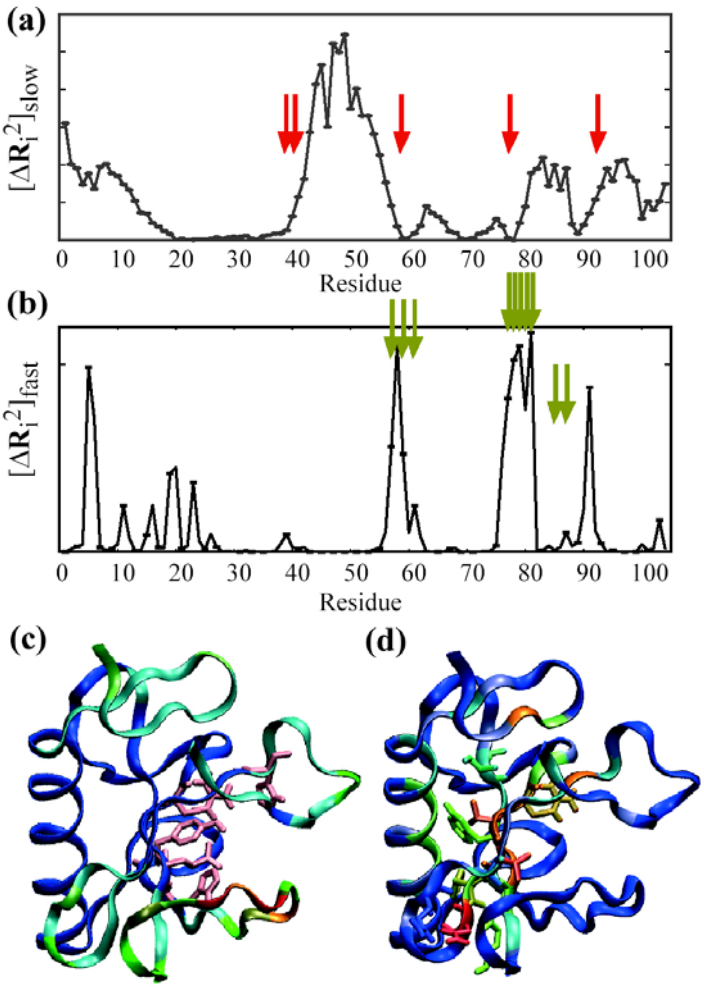

Figure 1-3 Physical meaning of slow and fast modes in GNM.
(a) Distribution of squared displacements of residues in the slowest mode as a function of residue index for ribonuclease T1 (RNase T1). The red arrows identify local minima that correspond to five experimentally identified catalytic residues: Tyr38, His40, Glu58, Arg77, and His92. (**b)** Distribution of squared displacements averaged over the ten fastest modes for the same protein. Here the arrows indicate the residues shown by hydrogen/deuterium exchange to be the most protected and thus important for reliable folding. A majority of these critical folding residues appear as peaks in the fast modes. (**c)** Color-coded mapping of the slowest mode (**a**) onto the 3D C^{α} trace of RNase T1 (PDB ID: 1bu4; Loris et al., 1999) where red is most mobile and blue least mobile. The side chains of the five catalytic residues are shown in pink surrounding the nucleotide binding cavity. (**d)** A similar color-coded mapping of the fluctuations of the ten fastest modes (**b**) onto the C^{α} trace. Here the side chains of the ten most protected residues from hydrogen deuterium exchange experiments are drawn explicitly showing that most of them are calculated to be mobile (red). The images in C and D were generated using VMD (Humphrey et al., 1996).

Note that the factor λ_k^{-1} plays the role of a statistical weight, which suitably rescales the contribution of mode *k*. This ensures that the slowest mode has the largest contribution. In addition to their significant contribution, the slowest motions are in general also those having the highest *degree of collectivity*. Many studies have shown that the shapes of the slowest modes indeed reveal the mechanisms of *cooperative* or *global* motions, and the most constrained residues (minima) in these modes play a critical role, such as a hinge-bending center, that govern the correlated movements of entire domains (Bahar et al., 1998a; Bahar et al., 1999b; Jernigan et al.,1999; Thomas et al., 1999; Keskin et al., 2000; Jernigan et al., 2000; Keskin et al., 2002a, b; Temiz et al., 2002; Xu et al., 2003; Sluis-Cremer et al., 2004; Wang et al., 2004). It is important to note that although these motions are slow, they involve substantial conformational changes distributed over several residues. The fastest modes, on the other hand, involve the most tightly packed and hence most severely constrained residues in the molecule. Their high frequency does not imply a definitive conformational change, because they cannot effectively relax within their severely constrained environment. On the contrary, they enjoy extremely small conformational freedom, on a *local* scale, by undergoing fast, but small amplitude fluctuations.

Figure *1-3* illustrates the contrast between the degree of collectivity for the slowest and fastest modes for an example protein, RNase T1. As in this case, the slow modes involve almost the entire molecule as indicated by the broad, delocalized peaks in panel a. The relative potential motion predicted by this mode is plotted onto the 3-d structure in Figure *1-3c*, color-coded such that minima are blue and maxima are red. For RNaseT1, five red arrows are drawn in Figure *1-3a* to indicate the residues identified as part of the catalytic site (Y38, H40, E58, R77 and H92; Loris et al., 1999). With the exception of H92, these five residues are located near minima in the slow

(global) mode shape (Figure *1-3a*) and their side chains are shown to be spatial neighbors (pink tubes) in the 3-d plot of this protein *(*Figure *1-3c*).

In contrast, the fastest modes are highly localized, with mode shapes that usually involve only a few peaks, as in Figure *1-3b*. These peaks refer to the residues that have a high concentration of local energy and are tightly constrained in motion. It has been noticed that these residues are often conserved across species and may form the folding nuclei (Demirel et al., 1998; Rader 2004a, b). In the application to RNase T1, the ten most protected residues (57, 59, 61, 77-81, 85, and 87), as identified by hydrogen-deuterium exchange experiments (Mullins et al., 1997), are indicated by gold arrows in Figure *1-3b* and shown with their side chains in the 3-d structure, color-coded such that minima are blue and maxima are red (Figure *1-3d*). As illustrated, many of these residues involve interactions between different strands of the central β-sheet, suggesting their potential involvement in the folding of RNase T1.

1.4.2.3. The utility of GNM beyond protein dynamics

Studying proteins with the GNM provides more than the dynamics of individual biomolecules. GNM has indeed been used for identifying the common traits among the equilibrium dynamics of proteins (Kundu et al., 2002), the influence of native state topology on stability (Burioni et al., 2004), the localization properties of protein fluctuations (Wu et al., 2003) or the definition of protein domains (Kundu et al., 2004a, b). Additionally, GNM has been used to identify residues most protected during hydrogen-deuterium exchange (Bahar et al., 1998c; Jaravine et al., 2000), critical for folding (Ortiz et al., 2000; Micheletti et al., 2002b; Micheletti et al., 2004), conserved

among members of a given family (Chen et al., 2004; Yang et al., 2005a), or involved in ligand binding (Micheletti et al., 2002a; Yang et al., 2005a).

1.5. ANISOTROPIC NETWORK MODEL (ANM)

1.5.1. What is ANM?

The anisotropic network model (ANM) has been originally introduced as an alternative EN model to address the deficiencies of the GNM (Doruker et al., 2000; Atilgan et al., 2001), mainly to compute the directions of the modes of motions, in addition to their sizes. In the ANM, the total potential of the structure is defined as

$$V_{ANM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\left(R_{ij} - R_{ij}^{0}\right)^2 H\left(r_c - R_{ij}\right)\right] \qquad (1\text{-}21)$$

where $H(r_c - R_{ij})$ is the Heavyside step function equal to 1 if the argument is positive, and zero otherwise. $H(r_c - R_{ij})$ selects all residue pairs within the cutoff separation of r_c. In the GNM, on the other hand, the potential is given by

$$V_{GNM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\left(\boldsymbol{R}_{ij} - \boldsymbol{R}_{ij}^{0}\right)^2 H\left(r_c - R_{ij}\right)\right] \qquad (1\text{-}22)$$

1.5.2. How does GNM differ from ANM?

Equation 1-22 looks very similar to equation 1-21, with the major difference that the vectors $\boldsymbol{R}_{ij}$ and $\boldsymbol{R}_{ij}^{0}$ in equation 1-22 are replaced by distances (scalars), R_{ij} and R_{ij}^{0}. This means

that the potential which depended upon the dot product between the fluctuation vectors in the GNM,

$$\begin{aligned}(\boldsymbol{R}_{ij} - \boldsymbol{R}_{ij}^0) \bullet (\boldsymbol{R}_{ij} - \boldsymbol{R}_{ij}^0) &= R_{ij}^{\ 2} + (R_{ij}^0)^2 - 2R_{ij}R_{ij}^0 \cos(\boldsymbol{R}_{ij}, \boldsymbol{R}_{ij}^0) \\ &= R_{ij}^{\ 2} + (R_{ij}^0)^2 - 2\left(X_{ij}X_{ij}^0 + Y_{ij}Y_{ij}^0 + Z_{ij}Z_{ij}^0\right)\end{aligned} \qquad (1\text{-}23)$$

now (in the ANM) depends upon their scalar product,

$$\begin{aligned}(R_{ij} - R_{ij}^0)(R_{ij} - R_{ij}^0) &= R_{ij}^{\ 2} + (R_{ij}^0)^2 - 2R_{ij}R_{ij}^0 \\ &= R_{ij}^{\ 2} + (R_{ij}^0)^2 - 2\left[X_{ij}^2 + Y_{ij}^2 + Z_{ij}^2\right]^{1/2}\left[(X_{ij}^0)^2 + (Y_{ij}^0)^2 + (Z_{ij}^0)^2\right]^{1/2}\end{aligned} \qquad (1\text{-}24)$$

Because the scalars R_{ij} and $R_{ij}^{\ 0}$ depend upon their components in a non-quadratic form, it is natural to end up with anisotropic fluctuations upon taking the second derivatives of the potential with respect to the displacements along the *X-*, *Y-* and *Z-* axes as is done in NMA. Using equations 1-23 and 1-24, the difference between these two potentials is

$$V_{GNM} - V_{ANM} = \gamma\left[\sum_{i,j}^{N} R_{ij}R_{ij}^0\left(1 - \cos(\boldsymbol{R}_{ij}, \boldsymbol{R}_{ij}^0)\right)H\left(r_c - R_{ij}\right)\right] \qquad (1\text{-}25)$$

i.e. the two potentials are equal only if $\cos(\boldsymbol{R}_{ij}, \boldsymbol{R}_{ij}^{\ 0}) = 1$, i.e. $\boldsymbol{R}_{ij} = c\boldsymbol{R}_{ij}^{\ 0}$ or $\Delta\boldsymbol{R}_i = c\Delta\boldsymbol{R}_j$ where c is a scalar.

Physically, this means that in addition to changes in inter-residue distances (springs), any change in the *direction* of the inter-residue vector $\boldsymbol{R}_{ij}^{\ 0}$ is also being resisted or penalized in the GNM potential. On the contrary, the ANM potential depends exclusively on the *magnitudes* of the inter-residue distances and does not penalize any such changes in orientation. It is conceivable that within the densely packed environment of proteins, orientational deformations may be as important as translational (distance) ones, and a potential that takes account of the energy dependence associated with the *internal orientational* changes (i.e. V_{GNM}) is physically more meaningful than one exclusively based on distances (V_{ANM}). Not surprisingly, ANM used in a

NMA has been observed to give rise to excessively high fluctuations compared to the GNM results or experimental data (Figure *1-2*), and hence necessitated the adoption of a higher cutoff distance for interactions (Atilgan et al., 2001). With a higher cutoff distance, each residue is connected to more neighbors in a more constrained and consolidated network.

Inasmuch as V_{GNM} is physically more realistic, one might prefer to adopt the GNM, rather than the ANM for a coarse-grained NMA. However, this greater realism comes at a price. Because the GNM describes the dynamics within an N-dimensional configurational space as opposed to a $3N$-dimesional one of ANM, the residue fluctuations predicted by the GNM are intrinsically isotropic. Thus GNM cannot provide information regarding the individual components: $\Delta X^{(k)}$, $\Delta Y^{(k)}$ and $\Delta Z^{(k)}$, of the deformation vectors $\Delta \boldsymbol{R}^{(k)}$ associated with each mode, k, but rather predicts the magnitudes, $|\Delta \boldsymbol{R}^{(k)}|$, induced by such deformations. The conclusion is that GNM is more accurate, and should be chosen when evaluating the deformation magnitudes, or the distribution of motions of individual residues. However, ANM is the only possible (less realistic) model when it comes to assessing the directions or mechanisms of motions. That the fluctuations predicted by the GNM correlate better with experimental B-factors than those predicted by the ANM has been observed and confirmed in a recent systematic study of Phillips and coworkers (Kundu et al., 2002). The dotted curves in Figure *1-2* illustrate the ANM results, and provide a comparison of the level of agreement (with experimental data) usually achieved by the two respective models. The correlation coefficients between the GNM results and experimental data are 0.769 and 0.823 in the panels A and B, respectively, whereas their ANM counterparts are 0.639 and 0.261. We note that the two sets of computed results are themselves correlated (0.756 and 0.454, respectively), which can be expected from the similarity of the underlying models.

1.6. APPLICATIONS OF GNM/ANM IN STRUCTURAL BIOLOGY

1.6.1. Applicability to supramolecular structures

A major advantage of the GNM is its applicability to large complexes and assemblies. The size of the Kirchhoff matrix is $N \times N$ for a structure of N residues, as opposed to the size $3N \times 3N$ of the equivalent Hessian matrix for a residue-level EN NMA (or ANM). The resulting computational time requirement for GNM analysis is then about 3^3 times shorter than for ANM, which in turn is about 8^3 times shorter than for NMA at atomic scale (assuming 8 atoms on the average per residue). This enormous decrease in computational time permits us to use the ANM, and certainly the GNM, for efficiently exploring the dynamics of supramolecular structures (Keskin et al., 2002a; Rader et al., 2004a).

Due to limitations in computational memory and speed, efforts to analyze large structures of $\sim 10^5$ residues rely upon further coarse-graining of the structure of interest. This type of coarse-graining is now the standard approach, having been implemented in several forms by various research groups including hierarchical coarse-graining (HCG; Doruker et al., 2002), discussed below; rotations-translations of blocks (RTB; Tama et al., 2000) or block normal mode (BNM; Li et al., 2002) and substructure-synthesis method (SSM; Ming et al., 2003b).

For both GNM and ANM, it has been demonstrated that an HCG scheme where clusters of residues and their interactions, as opposed to individual pairs of residues, are considered as the EN nodes (as opposed to individual residues), can successfully reproduce the essential features of the full-residue GNM/ANM calculations (Doruker et al., 2002). The global dynamics of hemagglutinin A were obtained at least two orders of magnitude faster than standard GNM/ANM by coarse-graining to the level of every 40^{th} residue (N/40). (Doruker et al., 2002) Notably, the minima in the global mode shapes, which refer to key regions that coordinate the collective dynamics, could be *exactly* reproduced by the N/40 coarse-graining.

Figure *1-4* illustrates the application of GNM to the wild type 70S ribosome from *Escherichia coli* (Vila-Sanjurjo et al., 2003). The calculations were performed by considering a single node for each amino acid (on the C^{α} atom) and each nucleotide (on the P atom), yielding a total of 10,453 nodes (residues and nucleotides). Because the diameter of the A-form RNA double helix is 20 Å, a larger cutoff distance is required to correctly identify base-paired nucleotides solely by their P-atom positions (Bahar et al., 1998b). To ensure adequate connectivity, two cutoff distances were adopted, 9.0 Å if both atoms were C^{α} and 21.0 Å if one or both were phosphorous, analogous to the recent ANM analysis of ribosome (Wang et al., 2004). Panels A and B illustrate the slowest (non-zero) mode shape as a color-coded 3-d structure and against the residue index. The coloring emphasizes the distinct difference between the motions of the 50S (red) and 30S (blue) subunits in this mode and indicates an anti-correlated motion of one subunit with respect to the other. This type of anti-correlated motions (i.e. ratcheting of one subunit with respect to the other) has been observed by cryo-EM (Frank et al., 2000).

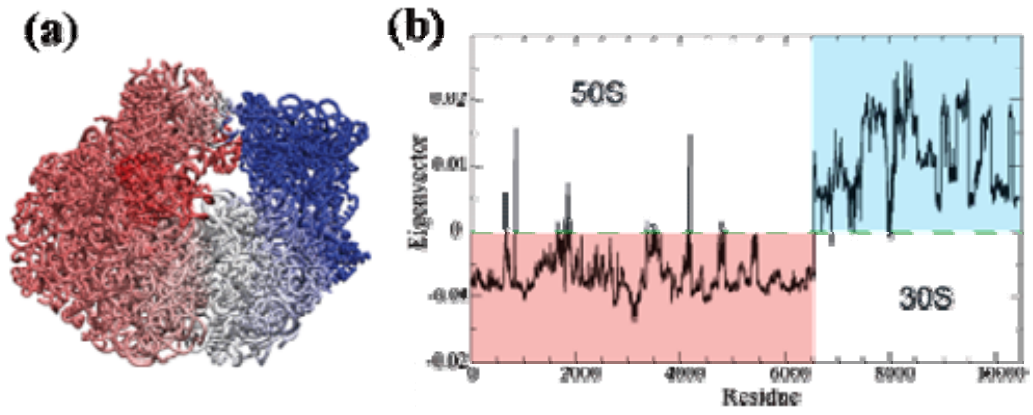

Figure 1-4 Application of GNM to the 70S ribosome structure.
The calculations were performed on the wild type 70S ribosome from E. coli (PDB IDs: 1pnx and 1pny; Vila-Sanjurjo et al., 2003). **(a)** The slowest non-zero mode for the 70S ribosome colored from -1 (red) to +1 (blue) is mapped onto the 3-d structure indicating a dramatic break at the interface between the two subunits (50S and 30S). This image was generated using VMD. (**b)** The slowest non-zero mode plotted versus the residue number. Residues in the 50S subunit (blue) exhibit one direction of motion that is opposed to the motion in the 30S subunit (red).

1.6.2. Other applications

1.6.2.1. Flexible docking

A major application of normal modes is the identification of potential conformational changes, e.g. of enzymes upon ligand binding (Tama et al., 2001; Delarue et al., 2002). In particular, it has been shown that over half of 3800 known protein motions (inferred from different forms of the same protein deposited in PDB) can be approximated by perturbing the original structures along the direction of their two low-frequency normal modes (Krebs et al., 2002). Such results suggest that the protein structures may have evolved to accommodate or facilitate biologically functional conformational changes. The functional mechanisms are indeed more readily accessible provided that they coincide with the smoothest ascent directions in the neighborhood of the global energy minimum, i.e. those along the lowest frequency modes. The fact that the observed changes

coincide with those predicted by the slowest NMA modes should not be a coincidence but a design principle favored by nature. Building on the notion that NMA can be used to identify potential motions induced by binding, a computationally tractable way to generate a set of docking targets has been proposed (Delarue et al., 2002).

1.6.2.2. Cryo-EM structure modeling

Recently there have been several applications of NMA to low-resolution cryo-electron microscopy (cryo-EM) structure modeling. Cryo-EM data are naturally low-resolution, being reconstructed by averaging over multiple images of many molecules from several different angles. Additionally, the imaged molecules often undergo structural changes along with vibrations making it very difficult to extract high-resolution structural information. Several groups (Ming et al, 2002a; Tama et al., 2002b; Delarue et al., 2004) have constructed EN models of pseudo-atomic representations for particular cryo-EM maps and calculated the resulting distortions due to normal modes as an aid in the refinement of the raw cryo-EM data to produce higher-resolution structural information. Alternatively, a procedure for the flexible docking of atomic or residue level structures into cryo-EM has been suggested by using the NMA mode shapes calculated for either these pseudo-atomic EN or homology-based structures (Delarue et al., 2004; Tama et al., 2004a, b; Hinsen et al., 2005).

1.6.2.3. Steering MD simulations and exploring non-equilibrium dynamics.

As discussed above, the low-frequency modes from NMA are able to capture the *collective* dynamics of proteins. A recent application of this fact is to steer MD simulations along these dominant modes of motion using hybrid methods that combine MD and harmonic modes (Zhang

et al., 2003; He et al., 2003; Tatsumi et al., 2004). Specifically a hybrid MD-NMA simulation protocol has been implemented where motions along the direction of the slowest few modes are coupled to a temperature bath and thus amplified to study the unfolding and large-scale domain motions of peptides and proteins (Zhang et al., 2003; He et al., 2003). The inverse of this approach, namely, that the normal modes of a protein can be extracted from an applied driving force in a MD simulation (Kaledin et al., 2004) has also recently been shown.

Drawing on similar insights, it has been suggested that one can minimize steric clashes and interpolate between two conformations of a protein using the modes from an EN model (Kim et al., 2002) to characterize this transition. Because the harmonic approximation of NMA remains valid only near the equilibrium structure, an alternative method to escape the local minima surrounding the native state involves the iterative calculation of successive EN models deformed along one or several low-frequency modes (Miyashita et al., 2003). This method allows "cracking" or partial unfolding of the underlying EN structure suggesting that such unfolding or "proteinquakes" may be coupled to collective motions (Itoh et al., 2004; Miyashita et al., 2005).

1.6.2.4. High throughput examination of families of proteins

Fold families such as globins (Maguid et al., 2005), and protein superfamiles (Leo-Macias et al., 2005) in general have been compared using NMA-based methods to identify common and distinctive structural-dynamic features. For the test case of proteases, these salient dynamic features from GNM calculations combined with data mining techniques in an unsupervised learning technique have been shown to identify the highly conserved catalytic triad (Chen et al., 2004). More recently the minima in the slowest modes (global hinge centers) have been shown to

be co-localized near catalytic residues in a representative set of enzymes (Yang et al., 2005). These results indicate that there is a great deal of information about functional residues to be extracted from the comparative EN-based NMA of protein family members.

1.6.2.5. Databases/servers of molecular motion

The logical extension of the protein family analysis is the compilation and update of web accessible databases housing NMA-based calculations for all available protein structures. Several such databases have been constructed including *i*GNM (Yang et al., 2005), ProMode (Wako et al., 2004), ElNémo (Suhre et al., 2004a), WEBnm@ (Hollup et al., 2005), and MolMovDB (Alexandrov et al., 2005) that allow users to browse pre-calculated data and/or submit structures for NMA. These current developments and applications will be discussed in more details at Chapter 4.

2. JUSTIFICATION OF THE APPLICABILITY OF GAUSSIAN NETWORK MODEL AND PARAMETER REFINEMENTS

2.1. ABSTRACT

In this chapter, we contest the idea of introducing residue specificities in the GNM, after a thorough examination of its consequences. A modified GNM model (*SPGNM*), which allows the attractive potentials only between the residue pairs that belong to the same polarity group, is tested. Our results have shown that SPGNM, a model that throws away the springs connecting the residues that have different polarities, gives a decreased agreement with experimental data. We further challenge the idea - '*the more specificities are considered, the better prediction of dynamics can be achieved*.' by comparing GNM predictions for X-ray crystallographic temperature factors (B_{exp}) with conventional NMA results. To our surprise, GNM predicts B_{exp} slightly better than NMA does over a set of 183 monomeric, non-homologous proteins. A correlation coefficient (R_{corr}) value of 0.58 ± 0.16 is obtained between theoretical and experimental data using the non-specific GNM, as opposed to 0.52 ± 0.18 by the examination of the same set by NMA. We further compared GNM predicted Mean Square Deviations (MSD) between NMR structural models and B_{exp} values of X-ray structures deposited in the Protein Data Bank,. We found that GNM predictions exhibit a stronger correlation with NMR MSD, compared to their correlation with the B_{exp} of X-ray structures over 181 non-homologous families, each of which contain both an NMR structure and an X-ray structure. The R_{corr} between

GNM and NMR MSD is 0.65 ± 0.22 while it is only 0.53 ± 0.19 between GNM and X-ray crystallographic B_{exp} values. The data suggests that the GNM yields better agreement with experiments if the measurements are done under less constrained environment. For 1250 non-homologous proteins independently examined by GNM, it was found that the correlation coefficient (R_{corr}) between B_{theo} and B_{exp} is a function of cutoff distance of inter-residue interaction adopted in the model, as well as the X-ray diffraction temperature (XDT). The cutoff distance of 15Å is found to give the best average correlation for all the 1250 proteins that were determined by X-ray at an XDT ≥ 70K, and gives the same R_{corr} as the optimal result when the cutoff is chosen over a range from 7.3 to 15 Å at the XDT ≥ 297K for 59 proteins. These data suggest that a larger cutoff distance covering the second coordination shell in the neighborhood of amino acids in folded structures gives a better R_{corr} for proteins in the relatively lower XDT regime. On the other hand, R_{corr} is the same for a wide range of cutoffs as the measurement is obtained at higher XDT, presumably larger than protein glass-transition temperature. In the study of protein Penicillopepsin (E.C. 3.4.23.20; E.C. stands for Enzyme Class; PDB ID: 1BXO), we see how the choice of the cutoff distances affects the mobility of a functionally critical residue, Tyr75. A careful selection of GNM cutoffs may therefore be needed for improving the accuracy of the predictions. A new criterion to assess the improvement of the model is discussed later in this chapter.

2.2. INTRODUCTION

The Gaussian Network Model (GNM) is a highly simplified model for examining protein dynamics, as introduced and discussed in the INTRODUCTION. Despite its simplicity, GNM has proven to yield results in good quantitative and qualitative agreement with experimental data and MD simulations (Bahar et al., 1998a; 1998c; 1999b; Demirel et al., 1998; Bahar and Jernigan, 1998; 1999; Haliloglu and Bahar, 1999; Jaravine et al., 2000; Kundu et al., 2002; Rader A.J., 2004b; Kurt et al., 2003; Wu et al., 2003; Erkip and Erman, 2004; Burioni et al., 2004; Kundu et al., 2004b; 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 (B_{exp}), 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.

This and other studies based on elastic network (EN) models lend support to the view that proteins possess intrinsic mechanical characteristics uniquely defined by their particular architecture, regardless of their chemical properties. The current model, GNM, gives similar large-scale dynamic behavior as similar architecture is assumed by the sequence (Keskin et al., 2000). Detailed biological functions are dictated by the local chemistry within a relatively small number of mechanical frameworks recruited by sequences. We would like to examine if there is any change in dynamics as we single-mutate a residue in the sequence. Also, a scaffold maybe used for multiple purposes and subtle functional differences are mediated by local motions, which involve more specific interactions.

To achieve this goal, our first attempt is to include specificities in the GNM by grouping amino acids into two categories – polar (P) or hydrophobic (H). We then compare the ability of the GNM and the modified models, termed specific GNM (*SPGNM)* in this chapter, to predict B_{exp}.

In the past, the theoretical computation of temperature factors (B_{theo}) has been a computationally expensive task. B_{theo}, which is the sum of all the eigenvalue-weighted normal modes, demands computing time that scales with N^3 to decompose the Kirchhoff matrix of inter-residue contacts and extract the N-1 GNM modes for a structure of N nodes (or N residues is a single-residue-per-node representation is adopted). To minimize the computing time requirement for moderate-to-large size proteins, researchers chose to approximate this sum by that of a subset of modes (Suhre and Sanejouand, 2004a). However, the error incurred in this approximation grows with the size of the protein, and becomes significant if a fixed number of modes are taken into account. The quality of the agreement between B_{theo} and B_{exp} may be viewed as an important measure for assessing the validity of the molecular motions predicted by NMA-based models. Hence, we have developed a new approach, PowerB, which takes advantage of power method (Mendelsohn, 1957) to minimize this computational cost of evaluating all modes for large structures. The B_{theo} values can be elegantly computed by subtracting the zero mode from the pseudo-inverse Kirchhoff matrix, as will be explained in the METHOD section.

Also, we have conducted a series of studies to compare GNM predictions with NMA results, the Mean Square Deviations (MSD) of residues between NMR models, the X-ray crystallographic B

factors using models where a range of cutoff distances are tested. The need for new criteria to assess the performance of the EN models is discussed in the end of the chapter.

2.3. METHOD

2.3.1. Residue-specific GNM (SPGNM)

According to the Venn diagram proposed for classifying amino acids (Taylor, 1986), we group amino acids according to their polarity into two groups: Polar (P) group, which comprises amino acids N, Q, S, T, H, D, E, C, R and K, and Hydrophobic (H) group, which comprises amino acids I, L, V, F, M, Y, W, A, P and G. A(i) denotes amino acid type of residue i. Its value is either H or P (e.g. if residue i is Val, A(i) = H). The cutoff distances are selected according to the residue type. The Kirchhoff matrix is thus modified as follows:

$$\Gamma_{ij} = \begin{cases} -1 & if \quad \{(A(i) = A(j)) and (i \neq j) and (r_{ij} \leq r_{cij})\} \quad or \quad \{abs(i-j) = 1\} \\ 0 & if \quad \{A(i) \neq A(j)\} or \{(i \neq j) and (r_{ij} > r_{cij})\} \\ -\sum_{i \neq j} \Gamma_{ij} & if \quad i = j \end{cases}$$

where r_{cij} is the cutoff distance of interaction between the C^{α}s, The cutoff distances are taken as

r_{cij} =7.3 Å for $A(i) = A(j) =$ H and r_{cij} =5.5 Å for $A(i) = A(j) =$ P;

We set the potential of H-P pairs to zero and assumed that the same force coupling H-H pairs and P-P pairs in different contact ranges. The condition abs(i-j) = 1 ensures that a spring connects sequential amino acids regardless of their type. Hence, in the modified Kirchhoff matrix, only

the elements representing the HH or PP interactions within given contact distances and the sequential neighbors are assigned -1 whereas the remaining off-diagonal elements are zeros.

2.3.2. B factor calculation. PowerB method

2.3.2.1. Power method

We start from a short review of conventional power method (Mendelsohn, 1957). Let us consider a real matrix $\boldsymbol{A}$ of size NxN, which has n linear eigenvectors $\boldsymbol{V}_{i=1,..,\ N}$ associated with the eigenvalues $|\lambda_N| > |\lambda_{N-1}| \geq |\lambda_{N-2}| \geq ... \geq |\lambda_1|$. λ_N is the dominant eigenvalue. The eigenvalues and eigenvectors are related as

$$\boldsymbol{A}\boldsymbol{V}_i = \lambda_i \boldsymbol{V}_i \qquad 1 \leq i \leq N \qquad (2\text{-}1)$$

The decomposition of a symmetric matrix $\boldsymbol{A}$ (such as the Kirchhoff matrix $\boldsymbol{\Gamma}$) ensures real and orthogonal eigenvectors such that

$$\boldsymbol{V}_i \cdot \boldsymbol{V}_N = 0 \text{ for } i \neq N$$

$$\boldsymbol{V}_N \cdot \boldsymbol{V}_N = 1$$

Any N-dimensional vector $\boldsymbol{X}_o$ (given by rand(N,1) in the code) can be written as

$$\boldsymbol{X}_o = \sum_{i=1}^{N} C_i \boldsymbol{V}_i \ , C_i \neq 0 \qquad (2\text{-}2)$$

By premultiplying both sides by $\boldsymbol{A}$, we obtain

$$\boldsymbol{A}\boldsymbol{X}_{\mathrm{o}} = \sum_{i=1}^{N} C_i \boldsymbol{A}\boldsymbol{V}_i = \sum_{i=1}^{N} C_i \lambda_i \boldsymbol{V}_i \qquad (2\text{-}3)$$

And repeating (2-3) for k-1 times, we obtain

$$\boldsymbol{A}^{\mathrm{k}}\boldsymbol{X}_{\mathrm{o}} = \sum_{i=1}^{N} C_i \lambda_i^k \boldsymbol{V}_i = \lambda_N^k \sum_{i=1}^{N} C_i (\frac{\lambda_i}{\lambda_N})^k \boldsymbol{V}_i \qquad (2\text{-}4)$$

For large k, the ratio $(\frac{\lambda_i}{\lambda_N})^k \to 0$ for i = 1, .., *N*-1, and equation (1-4) is reduced to one surviving term in the summation, i.e.

$$\boldsymbol{A}^{\mathrm{k}}\boldsymbol{X}_{\mathrm{o}} \approx \lambda_N^k C_N \boldsymbol{V}_N \qquad (2\text{-}5)$$

This leads us to the normalized $\boldsymbol{V}_{\mathrm{n}}$

$$\boldsymbol{V}_{\mathrm{n}} = \frac{\boldsymbol{A}^{\mathrm{k}}\boldsymbol{X}_{\mathrm{o}}}{\left|\boldsymbol{A}^{\mathrm{k}}\boldsymbol{X}_{\mathrm{o}}\right|} \qquad (2\text{-}6)$$

Using equation (2-5), it is also correct to state that

$$\boldsymbol{A}^{\mathrm{k+1}}\boldsymbol{X}_{\mathrm{o}} \approx \lambda_N^{k+1} C_N \boldsymbol{V}_N$$

which gives the relationship

$$\frac{\boldsymbol{A}^{\mathrm{k}+1}\boldsymbol{X}_0}{\lambda_N^{\mathrm{k}+1}} = C_N\boldsymbol{V}_N = \frac{\boldsymbol{A}^{\mathrm{k}}\boldsymbol{X}_0}{\lambda_N^{\mathrm{k}}}$$

The inner product of both sides of the above equation with a random vector $\boldsymbol{Y}_{N\times 1}$ and rearrangement of the result, yields

$$\frac{\boldsymbol{A}^{\mathrm{k}+1}\boldsymbol{X}_0 \cdot \boldsymbol{Y}}{\boldsymbol{A}^{\mathrm{k}}\boldsymbol{X}_0 \cdot \boldsymbol{Y}} = \frac{\lambda_N^{\mathrm{k}+1}}{\lambda_N^{\mathrm{k}}} = \lambda_N \qquad (2\text{-}7)$$

as long as $\boldsymbol{Y}_{N\times 1}$ is not perpendicular to $\boldsymbol{A}^{\mathbf{k+1}}\boldsymbol{X}_{\mathrm{o}}$. This way, the dominant eigenvalue and –vector are obtained.

One can derive the subdominant eigenvalues and –vectors through a deflation process.
The inner product of both sides of (2-2) by $\boldsymbol{V}_N$ gives

$$\boldsymbol{X}_0 \cdot \boldsymbol{V}_N = \sum_{i=1}^{N} C_i(\boldsymbol{V}_i \cdot \boldsymbol{V}_N) = C_N$$

and we define

$$\boldsymbol{X}_o' = \boldsymbol{X}_o - C_N\boldsymbol{V}_N = \sum_{i=1}^{N-1} C_i\boldsymbol{V}_i$$

Repeating the above described process, (2-3) through (2-7) for $\boldsymbol{X}_0$', we evaluate the second largest eigenvalue λ_{N-1} and eigenvector $\boldsymbol{V}_{N-1}$. Any pair of (λ_k, $\boldsymbol{V}_k$) can thus be obtained recursively from previously known pairs (λ_i = 1, .., k-1, $\boldsymbol{V}_i$ = 1..k-1). Note that the larger $\left|\frac{\lambda_k}{\lambda_{k-1}}\right|$ is, the faster we reach convergence.

2.3.2.2. Evaluation of B-Factors with the Power Method

The fact that the diagonal elements of the Kirchhoff matrix are equal to the summation of off-diagonal elements in the same row/column reduces the rank of the matrix by one, which therein contributes to a zero eigenvalue.

$$\boldsymbol{\Gamma} = 0 \times \boldsymbol{u}_o \boldsymbol{u}_o^T + \sum_{i=1}^{N-1} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T \qquad (2\text{-}8)$$

The pseudo-inverse of the Kirchhoff matrix is obtained after eliminating the zero eigen mode:

$$\boldsymbol{\Gamma}^{-1} = \sum_{i=1}^{N-1} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T \qquad (2\text{-}9)$$

The diagonal elements of the pseudo-inverse Kirchhoff scales with B_{theo} (Bahar et al., 1997a) as

$$B_{theo} = \frac{8\pi^2 k_B T}{\gamma} \Gamma_{ii}^{-1} = \frac{8\pi^2 k_B T}{\gamma} \left(\sum_{k=1}^{N-1} \frac{1}{\lambda_k} \boldsymbol{u}_k \boldsymbol{u}_k^T\right)_{ii} \qquad (2\text{-}10)$$

In principle, one has to compute and add up the contribution of all modes in order to evaluate B_{theo}. However, if we perturb the Kirchhoff matrix by adding a small number ε, on one of its elements, say Γ_{11}, $\boldsymbol{\Gamma}$ becomes invertible.

$$\boldsymbol{\Gamma}\big|_{\Gamma_{11}+\varepsilon} = \delta \times \boldsymbol{u}_o \boldsymbol{u}_o^T + \sum_{i=1}^{N-1} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

$$\boldsymbol{\Gamma}\big|^{-1}_{\Gamma_{11}+\varepsilon} = \frac{1}{\delta} \times \boldsymbol{u}_o \boldsymbol{u}_o^T + \sum_{i=1}^{N-1} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T$$

Equation (2-10) can be therefore re-written as

$$B_{theo} = \frac{8\pi^2 k_B T}{\gamma} \Gamma_{ii}^{-1} = \frac{8\pi^2 k_B T}{\gamma} (\boldsymbol{\Gamma}\big|^{-1}_{\Gamma_{11}+\varepsilon} - \frac{1}{\delta} \times \boldsymbol{u}_o \boldsymbol{u}_o^T)_{ii} \quad (2\text{-}11)$$

where δ is a small number (of the order of 10^{-7} if ε = 0.0001). In general, δ-1 >> λ-1, such that we can extract the first dominant eigenvalue δ^{-1} at a fast convergence rate. We term this fast algorithm to obtain B_{theo} as *PowerB*.

2.3.3. Examination of a non-homologous protein set

We examined the fluctuations of residues in a non-homologous protein set retrieved from PDB-REPRDB (Noguchi et al., 1997; http://mbs.cbrc.jp/pdbreprdb-cgi/reprdb_menu.pl). We selected only the structures resolved by X-ray crystallography with a resolution ≤ 2.4 Å, and R-factors ≤ 0.3, and those having no chain discontinuities in the reported PDB files, and number of residues ≥ 40, excluding membrane proteins. Chains are classified into families. Members of each family

have sequence identity ≤ 30% and structural RMSD ≥ 10 Å with members of the other families. 1930 families were obtained. We further deleted those cases where the single protein chains (such as 1JJ2) comprise multiple families to leave 1804 families to which PowerB was applied with a series of cutoffs. We further eliminated the structures that contained more than five nucleotides, those that did not report B_{exp} values, and those create the 'eigenvalue error' in GNM computations caused by missing atom coordinates in the structure (see Chapter 4, Figure 4-2) or contained C^{α} atoms that are assigned multiple positions as solved by X-ray, which resulted in 1250 families. The entire list of the resulting set proteins can be found in Table S1-1 of Supplementary material*.

2.3.4. Comparison of NMA results with GNM predictions

Time average fluctuations, computed by NMA, for 183 monomeric proteins that are included in the *Pro*Mode DB (http://promode.socs.waseda.ac.jp/) (Wako and Endo, 2002; Wako et al. 2003; Wako et al. 2004) were kindly provided by Dr. Wako (personal communication). Each protein belongs to a different SCOP family. The structures are pre-equilibrated first and then a NMA is performed in the coordinate system of dihedral angles after the work of Go and collaborators (Wako et al., 1995). Due to the sophisticated energy minimization process as well as approximately six degrees of freedom (rotatable bonds on the backbone and sidechain) of each residue considered, *Pro*Mode DB currently reports NMA results for relatively small proteins having < 300 residues in view of the time and memory cost of the computation. The smallest and the largest proteins in this study contain 77 and 245 residues, respectively. The B_{theo}, predicted by GNM, were also retrieved from *i*GNM (Yang et al., 2005 and also Chapter 3) for the same set of 183 proteins to compare with the NMA results and experimental values.

2.3.5. Comparison of GNM predictions with X-ray crystallographic B_{exp} factors, and with NMR data

Structures determined by both X-ray and NMR are listed in the PDB (Berman et al., 2000; http://www.rcsb.org/pdb/XrayAndNmr.html). There were 263 families listed as of Feb 2nd 2005. Each family comprises proteins that have sequence similarity larger than 95% between one another for a stretch containing at least 100 amino acids. One X-ray structure and one NMR structure from each family were selected. By selecting the NMR structures that contained at least 3 models, 197 families were retrieved. We further removed 12 'eigen error' structures and 4 structures that contained unrealistic B_{exp} data, which resulted in 181 families (see Table S1-2 in supplementary materials*) for the analysis. The Mean Square Deviation (MSD) for each residue in a given set of NMR models for a given structure was calculated and compared with the B_{theo} predicted by the GNM. The MSD for residue *i* is defined as

$$\mathrm{MSD}_i = \frac{\sum_{k=1}^{m} \left| \boldsymbol{r}_{i,k} - \bar{\boldsymbol{r}}_i \right|^2}{m}$$

where $\bar{\boldsymbol{r}}_i = \frac{\sum_{k=1}^{m} \boldsymbol{r}_{i,k}}{m}$, $\boldsymbol{r}_{i,k}$ is the coordinate vector of residue *i* in the *k*th model and *m* is the total number of models in the structure. Note that GNM was performed on the first model in the NMR structure file. The average correlation coefficients between the MSD_i and $B_{theo,i}$ values, as well as those between $B_{exp,i}$ and $B_{theo,i}$ for the 181 examined proteins and their subsets were computed and compared.

2.4. RESULTS

2.4.1. PowerB – speed And accuracy

Application of PowerB to 10 proteins with sizes ranging from 150 to 7350 residues showed the accuracy and high efficiency of the method (Figure 2-1). B_{theo} obtained by PowerB has exactly same values as those computed using the conventional singular value decomposition (SVD) subroutine from Numerical Recipes (Press et al., 1992) for all the proteins with a correlation coefficient of unity for each of them. The SVD subroutine computing (real) time is observed to scale with the 3.8th power of the number of residues (N) (t_{SVD} (seconds) = 4.17E-10×$N^{3.77}$) while PowerB shows a time dependence proportional to approximately N^2 (t_{PB} = 2.29E-5×$N^{1.94}$), which can be explained by the fact that the computing time for matrix inversion, also scaling as N^2, dominates the overall computation. This reduction in computation time is especially prominent for large structures such as GroEL Protein (PDB ID: 1KP8; 7350 nodes). The conventional approach (SVD) took longer than 63 hrs for getting B_{theo} which can be obtained within 16 mins by PowerB.

2.4.2. SPGNM

Inspired by the classical HP model (Chan et al. 1998), which is widely used in protein folding simulations, we modified GNM by (1) assigning attractive interactions to H-H and P-P pairs only, and neglecting the interactions between H-P pairs and (2) considering that interactions of P-P and H-H pairs occur in particular distance ranges, characteristic of the type of amino acids.

We repeated the above described computations with the objective of seeing if this will give any better B_{exp} predictions.

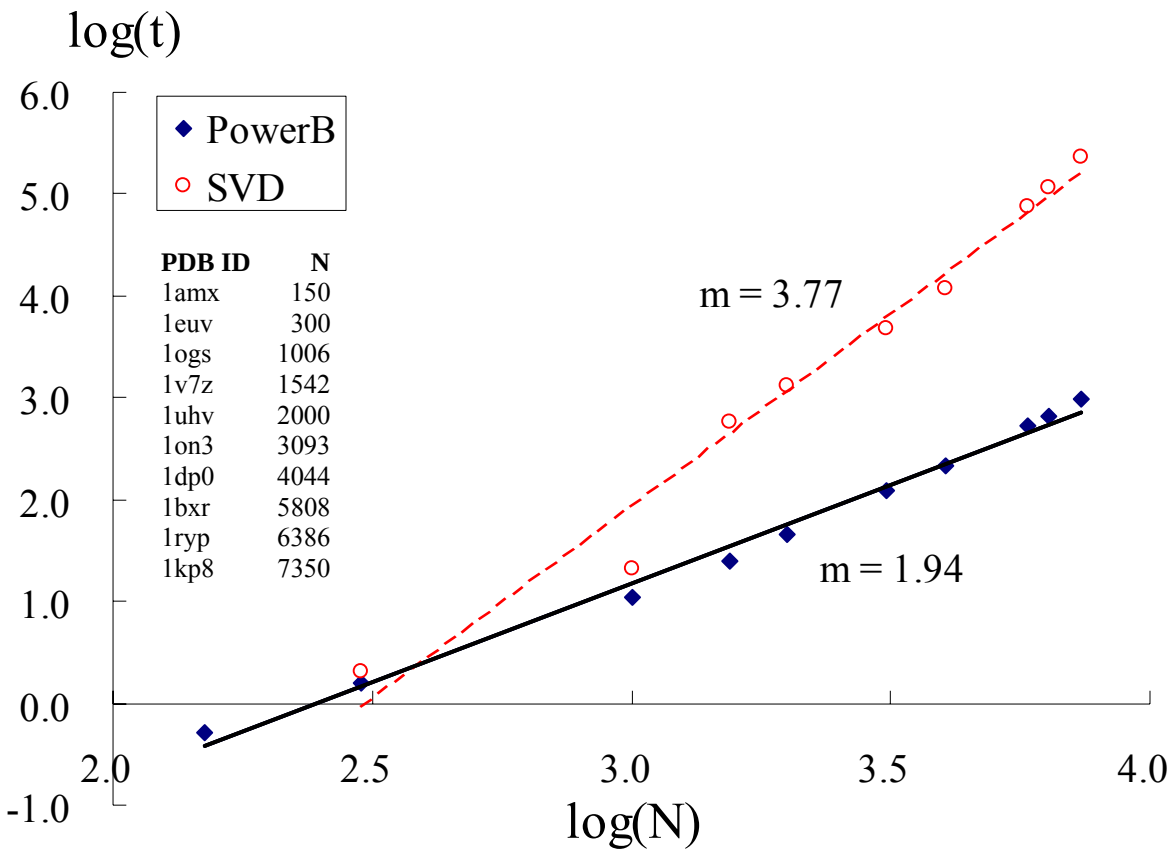

Figure 2-1 Improvement in computation time using PowerB method.
The logarithmic computing time is plotted as a function of logarithmic residue number *N*. Theoretical B-factors (B_{theo}) were computed by conventional approach (SVD) and by PowerB method. *m* denotes the slope of the regression lines in log-log plot. SVD shows that time (in units of seconds) scales with the 3.8^{th} power of *N* while PowerB gives a time increase scaling approximately with *N* square, which reduces the computing time dramatically for large structures such as GroEL (PDB ID: 1KP8; 7350 nodes). The conventional approach (SVD) took longer than 63 hrs to compute B_{theo} which could be computed within 16 minutes by PowerB. The data points used in this analysis are listed on the left. The correlation coefficient between the B-factors computed from SVD and those from PowerB is equal to 1.

We repeated the above described computations with the objective of seeing if this will give any better B_{exp} predictions.

Bahar and Jernigan reported structure-derived potentials in 1996 and 1997 (Jernigan et al., 1996; Bahar et al., 1997b) and found that the most favorable attractive potentials between hydrophobic groups occur in range between 4 and 6 Å and those between polar and charged groups occur in a closer interval between 2 and 4 Å, suggesting a stronger interaction exist between P-P pairs. The potential between hydrophobic and polar groups monotonically decreases with increasing inter-residue separations. We termed this modified model SPGNM. This model will first be compared with GNM in their abilities to predict B_{exp}. If the agreement with experiments is higher than that achieved by the GNM, we could then expect to further improve the model and parameters to represent the perturbations in fluctuation dynamics induced by alterations in local chemistry.

The theoretical B-factors (B_{theo}) of 1250 proteins (see Supplementary, Table S2-1*), retrieved from the PDB-REPRDB (Noguchi et al., 1997), were calculated by PowerB approach in different models and compared with the B_{exp} values. The results are shown in Figure 2-2. As we can see, the original GNM, using two different cutoffs – 7.3 and 15 Å, outperforms SPGNM models, and exhibits a higher agreement (R_{corr}) with experimental B-factors. These results do not support the view of further pursuing this type of inclusion of specificities in the GNM.

We then step back and ask ourselves again if it is true that more specific models do not improve the B_{exp} prediction? The failure of the HP model may lie in its simplicity. Can we generalize this finding to models containing more specific features of the 20 different types of amino acids? To answer this question, we would like to see how much, if any, a full atomic NMA with a detailed force field would outperform GNM in predicting B_{exp}?

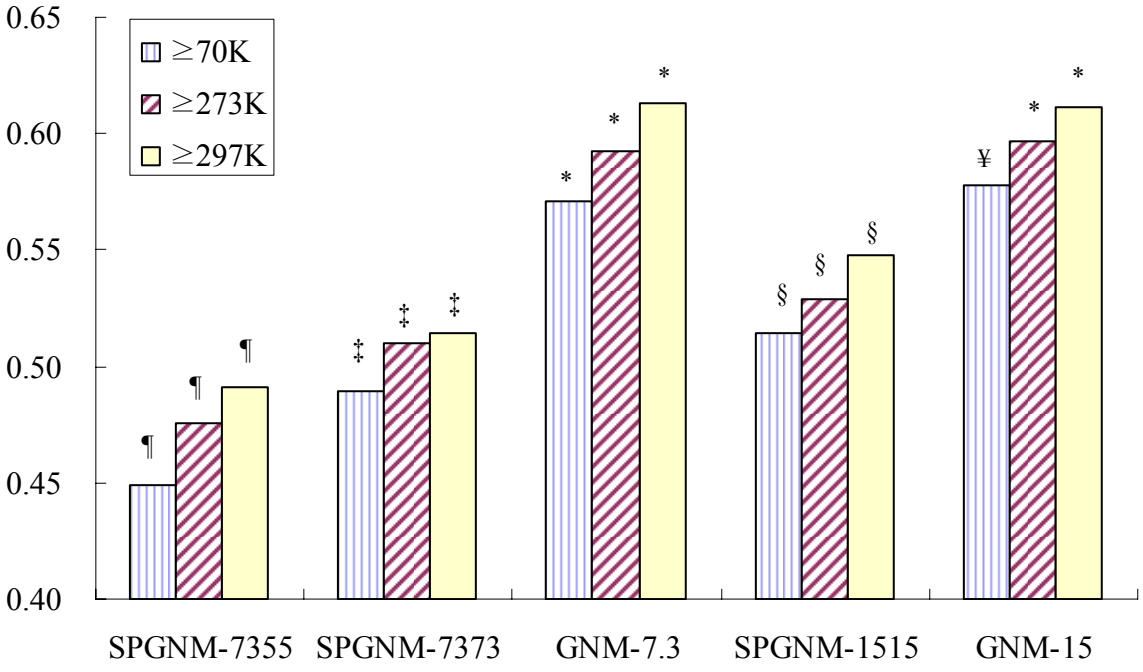

Figure 2-2 The average correlation coefficients (R_{corr}) between B_{theo} and B_{exp} over 1250 non-homologous proteins.

The model SPGNM-7355 uses cutoff 7.3 Å for HH pairs and 5.5 Å for PP pairs. SPGNM-7373 and SPGNM-1515 use 7.3 Å and 15 Å respectively for both HH pairs and PP pairs while they set a zero potential for HP pairs. GNM-7.3 and GNM-15 are the original GNM models that use 7.3 and 15 Å cutoffs respectively. All the 1250 proteins have X-ray diffraction temperatures (XDTs) above 70K. The average R_{corr} values of the 1250 proteins are shown by the vertical stripe columns. The average R_{corr} for 235 proteins that have XDT ≥ 273K, and for 59 proteins that have XDT ≥ 297K are indicated in the diagonal stripe and solid columns respectively. The symbols above the group columns refer to the statistical significance of the differences between these values. The groups referring to the same XDT (same color) are considered statistically identical by paired student T-test if the symbols above the group columns are the same. Groups with different symbols have different mean values.

2.4.3. Comparison of NMA and GNM results

NMA results computed for 183 monomeric, non-homologous proteins were kindly provided by Dr. Hiroshi Wako, the creator of *Pro*Mode, a database where equilibrium dynamics computed by full atomic NMA has been collected. The performance of NMA in predicting B_{exp} could provide

guidance as to how far GNM predictions can be improved by introducing specificities into its coarse-grained framework. The assumption here is that *the more specificities are considered, the better prediction of B_{exp} can be achieved*.

The results are presented in Figure 2-3. To our surprise, we find that the GNM predictions agree with B_{exp} better than NMA do. The R_{corr} of NMA and B_{exp} is 0.549 while GNM and B_{exp} is 0.582 and 0.575 for cutoff 7.3 and 15 Å respectively. The difference between NMA and GNM is statistically significant. However, the changes in the cutoff distances used in GNM do not affect the quality of predictions in this case, as indicated by the same symbol in Figures 2-3.

This finding led us to reconsider the rationale that GNM can be improved by introduction of residue specificities. At least, we have seen the upper bound of the predicting power for B_{exp} in current elastic network models. On the other hand, we have to re-define what we mean 'improvement' in the model. Some alternative experimental benchmarks are needed as we assess the performance of GNM. Or, new criteria, directly related to protein functions, have to be established so as to verify if the time average fluctuations or mode mobilities, predicted by the GNM are functionally meaningful. The present results also warn us about future attempts in introducing amino acid specificities into the simplified elastic network models such as GNM, ANM, βNM (Micheletti et al., 2004) or Hinsen's NMA model in MMTK (Hinsen, 2000) when the performance of the model is assessed by their ability to reproduce B_{exp}. Not surprisingly, one can find the correlation coefficient between NMA and GNM to be quite high (0.735 for cutoff7.3 Å and 0.783 for cutoff 15 Å). It is very interesting to notice that GNM using cutoff 15 Å bears closer agreement with NMA than GNM with a 7.3 Å cutoff. This could be explained by the fact

that a longer range of interaction, compare to atomic NMA, is incorporated into the elastic network. At this point, we deem appropriate not to pursue any further our attempts for including residues specificities in the GNM to give better B-factor predictions.

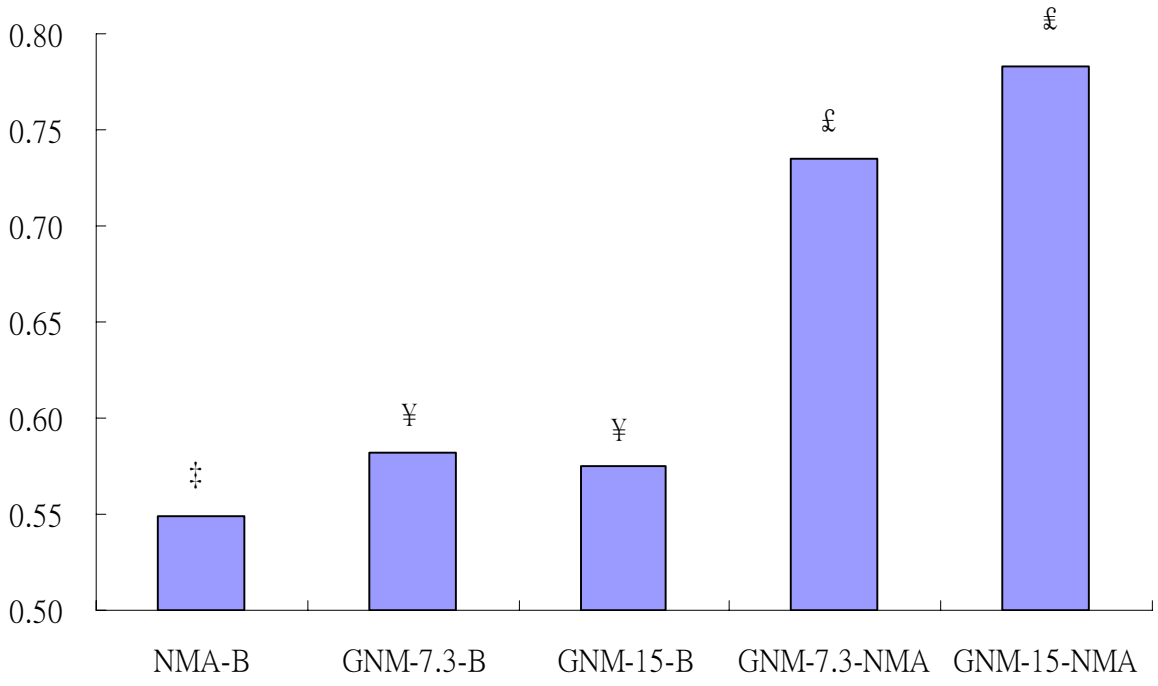

Figure 2-3 Comparison of NMA and GNM predictions.
The predictions of B_{theo} from NMA and GNM using two cutoff distances are compared with their experimental counterparts, B_{exp}, for 183 test proteins NMA-B, GNM-7.3-B and GNM-15-B are the average R_{corr} between B_{exp} and B_{theo} from NMA, GNM using 7.3 Å cutoff and GNM using 15 Å cutoff, respectively, for 183 monomeric proteins. Correlations of NMA and GNM predictions using two different cutoff distances are also shown on the rightmost two columns. The groups whose mean R_{corr} values are statistically identical, as examined by paired student T-test, are indicated by the symbols above the group columns.

2.4.4. GNM predictions for X-ray crystallographic B-factors and mean-square deviations from NMR models

Apart from B_{exp}, we are also interested in assessing how GNM predictions agree with the Mean Square Deviations (MSDs) of NMR models deposited in the PDB for a given protein and in finding alternative benchmarks to validate the performance of GNM in general. As a result, we initiate a systematic analysis to compare GNM predictions with the MSDs of NMR models. We also would like to know how good the agreement with NMR results is in comparison to the agreement between GNM predictions and B_{exp} from X-ray Structures.

181 families (Supplementary, Table S2-3*) of proteins were selected as described in the METHOD section. One X-ray structure and one NMR structure in each family were downloaded. So, there are 181 X-ray structures and 181 NMR structures. The GNM analyses are applied for the entire 362 structures and MSD of the 181 NMR structures were computed. The first model in each NMR structure is taken for the GNM B_{theo} calculation, which is again obtained from PowerB algorithm. The data have shown that the average R_{corr} of B_{theo} from GNM and MSD from NMR structures over 181 proteins is 0.65 while it is only 0.53 for that of B_{theo} and B_{exp} (X-ray structures) over the other 181 proteins. The difference is statistically significant. Due to some X-ray members that are much larger in size than their NMR partners within the same families; we selected 82 families the members of which were both monomeric structures with a size difference less than 40 residues. This should exclude the possibilities that the differences in correlation coefficients that might be imparted by the size discrepancy between the members. We note that the GNM is theoretically expected from central limit theorem to give poorer predictions for smaller proteins. The data confirm again that the fluctuation behavior predicted by the GNM

does give a better agreement with NMR MSD (than X-ray B-factors), scoring a R_{corr} value of 0.66 while the X-ray $B_{exp,}$ give an R_{corr} value of 0.57. Again, the difference of the two is statistically significant according to the result of paired student T-test.

2.4.5. Parameter refinement

As we know from the INTRODUCTION, the GNM has two adjustable parameters. One is the spring constant γ, which is usually adjusted to match the absolute size of experimental B_{exp} and the other is the cutoff distance, defining the connectivity of the network. The latter is less influential in terms of qualitative effect on the profile of fluctuations, specially in the low frequency regime. However, there have been only a few rigorous studies, carried out systematically, to examine how the selection of cutoff distance affects GNM predictions except for Kundu's work in 2002 on exploring this effect over 113 non-homologous proteins (Kundu *et al*., 2002). Their results suggested an optimal cutoff of 7.3 Å, which is consistent with the first coordination shell range of 6.8-7 Å (Miyazawa and Jernigan, 1985; Bahar and Jernigan 1997) in the neighborhood of amino acids in folded structures. With a fast B-factor calculation algorithm (PowerB) at hand, we would like to revisit this issue by varying the cutoff distances over a wider range and by performing the computations for a larger set of proteins.

The representative set of proteins of 1250 non-homologous proteins (Supplementary material, Table S2-1*) was selected to study the effect of the cutoff distance on R_{corr} of B_{theo} and B_{exp}. The computation was carried out using PowerB. An average constant $k_BT/\gamma = 1.10 \pm 0.50$ Å^2 was obtained in reasonable agreement with Kundu's result, $k_BT/\gamma = 0.87 \pm 0.46$ Å^2, over 113 monomeric proteins (Kundu *et al*., 2002). As we can see in Figure 2-4, an optimal cutoff of 15 Å

was observed as opposed to the local maximum at 7.3 Å for all the proteins that have X-ray diffraction temperature (XDT) above 70K. When we select the 235 proteins that have XDT ≥ 273K, we can see that the range of cutoffs from 7.3 Å to 15 Å give statistically identical R_{coff} except for the case as R_c = 10 Å. However, if one selects only those proteins with an XDT ≥ 297K, which results in a subset of 59 proteins, the wide range of the cutoff 7.3 Å to 15 Å are observed to give the same R_{coff}. The inverse coordination number $1/\Gamma_{ii}$ (denoted as 1/contact in Figure 2-4), the leading term of serial expansion of B_{theo}, is also compared with GNM results using different cutoff distances. We can see how GNM improves R_{corr} compared to this first approximation. Figure 2-4 shows that the predictions from GNM with all of the different cutoffs give a better R_{corr} with experiments than $1/\Gamma_{ii}$ does.

2.4.6. Other criteria to assess the performance of GNM: Protein Penicillopepsin as an illustrative example

Penicillopepsin is a hydrolase that has a broad specificity similar to that of pepsin A (Khan et al., 1998). If one examines its dynamics using GNM with two different cutoffs – 7.3 and 15Å, it can be seen that the mobility at the catalytic residue Tyr75 has a dramatic change in its lower modes (Figure 2-5). At cutoff 7.3 Å, Tyr75 appears highly mobile as evidenced by the red color in the ribbon diagram or mounting on a high peak in the mobility plot (the upper and lower panel in Figure 2-5a, respectively). Interestingly, these two cutoffs give approximately the same R_{corr} between B_{exp} and B_{theo} (0.596 for cutoff 7.3Å and 0.603 for cutoff 15Å) despite the disappearance of the peak near Tyr75. This suggests that we could miss the mobility changes in some functional residues if we were only focusing on the overall R_{corr} values.

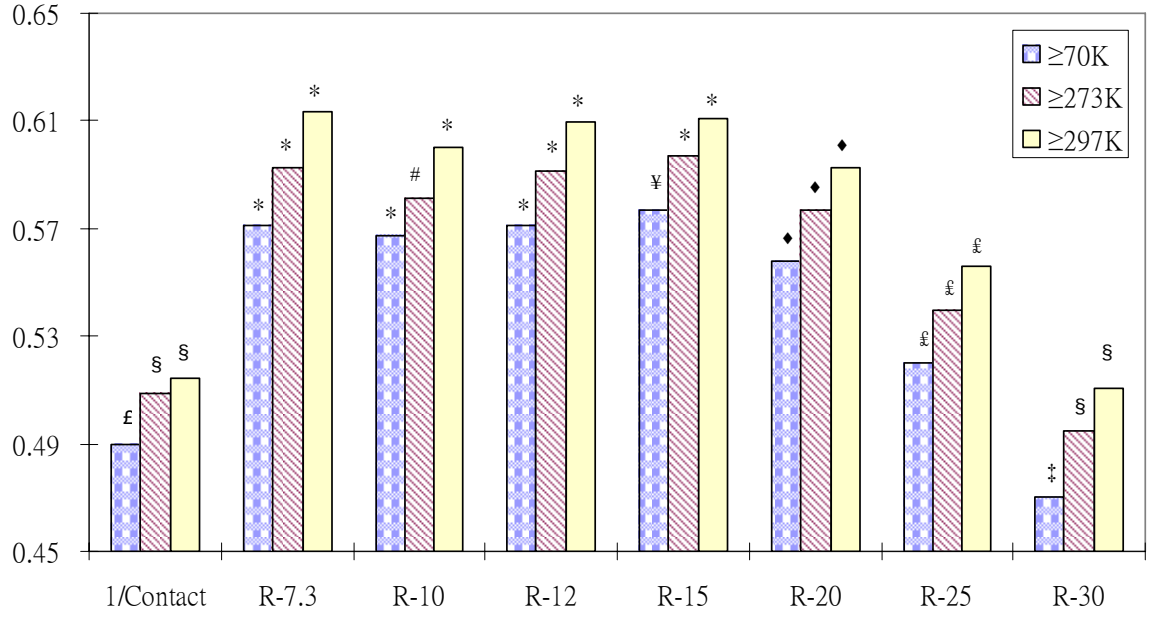

Figure 2-4 Average R_{corr} as a function of cutoff distance and XDT, over 1250 proteins.
The correlation coefficient (R_{corr}) between B_{theo} and B_{exp} is shown as a function of cutoff distance and XDT for 1250 non-homologous proteins. The first set of results (labeled 1/Contact) refers to the correlation R_{corr} between B_{exp} and the inverse residue coordination number ($1/\Gamma_{ii}$) averaged over all proteins. Γ_{ii} is the coordination number of residue *i*, an indication of local packing density (Halle, 2002). The value is determined with a GNM cutoff 7.3 Å in this case. The average R_{corr}, computed by PowerB approach using different cutoff distances, were obtained for a subset of proteins of which the X-ray diffraction data were collected at different temperature ranges of XDT - ≥ 70K, ≥ 273K and ≥ 297K. The mean R_{corr} values of different groups are considered statistically identical by paired student T-test if the symbols above the group columns are the same. Groups with different symbols have different mean values.

In Chapter 2, we will show that a strong coupling exists between enzyme catalytic residues and key mechanical sites that are distinguished by their low mobilities in the slower modes. Since this restricted mobility of catalytic sites in the GNM slow modes is a general effect observed for enzymes across all the classes, this observation could serve as a new criterion, directly related to protein function, to assess the optimal model parameters.

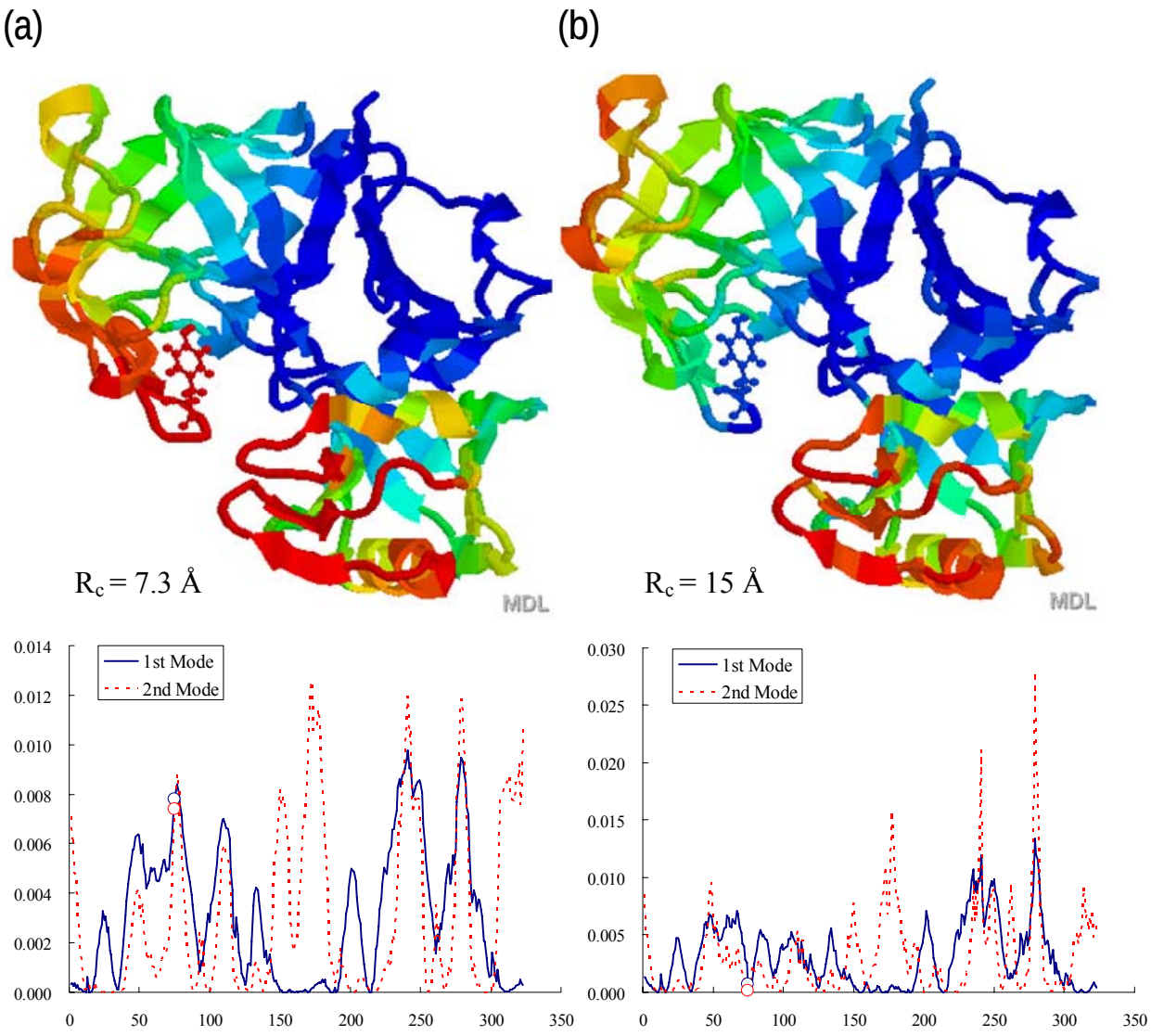

Figure 2-5 Dependence of the slowest two modes accessed by 1BXO on cutoff distances.
The slowest two modes accessed by 1BXO are changed with the selection of cutoff distances. The ribbon diagrams are color-coded according to the residue mobilities in the upper panels. The residues are colored blue→ green→ yellow→ orange→ red in the order of increasing mobilities. The catalytic residue, Tyr75, is shown in ball-and-stick. Residue mobilities against residue index are plotted in the bottom panels. The 1st and 2nd slowest normal modes are shown in solid and dash lines, respectively. The mobility of Tyr75 in the two modes are marked by open circles. This GNM analysis was performed with two cutoff distances, 7.3 and 15 Å.

In Chapter 2, we will show that a strong coupling exists between enzyme catalytic residues and key mechanical sites that are distinguished by their low mobilities in the slower modes. Since this restricted mobility of catalytic sites in the GNM slow modes is a general effect observed for enzymes across all the classes, this observation could serve as a new criterion, directly related to protein function, to assess the optimal model parameters.

2.5. DISCUSSION

Why does GNM agree better than NMA with B_{exp}? The difference between GNM and ANM could still be attributed to their different potentials. As discussed in the INTRODUCTION, GNM potential that takes account of the energy dependence associated with the internal orientational changes (i.e. V_{GNM}), and it is physically more meaningful than one exclusively based on the magnitude of distances (V_{ANM} or V_{NMA}). Hence, GNM could give better B_{exp} predictions despite of its isotropic assumption. Full atomic NMA potentials are more complicated, and uncertainties in energy functions and parameters may be partly responsible for their limited performance.

The better agreement of GNM with NMR MSDs, compared to that with X-ray B_{exp} is another interesting result. Since this study is performed over proteins within the same family, the difference in R_{corr} between the two proteins in the same family mainly results from different experimental approaches. It seems that the measurements done in solution (NMR) show better agreement with GNM, compared to those measured in densely packed crystals (X-ray). In Kundu's study (Kundu *et al*., 2002), crystal contacts were considered when comparing the B_{exp}

with GNM result. A better agreement is obtained when crystal contacts from the neighbor molecules are considered than those are omitted. The correlation coefficient with experiments increases 7% (from 0.58 to 0.65; Kundu *et al*., 2002) in this case. We did not consider the crystal contacts in this study. Hence, the fact that R_{corr} value between NMR and GNM surpasses that of X-ray and GNM by 9% over 82 protein pairs could be partly due to the crystal contacts in the X-ray structures.

Additionally, we adopted a wide range of distances from 7.3 to 30 Å, in the examination of the effect of cutoff distance on GNM results, and found that, in general, R_c values from 7.3 to 15 Å give approximately the same R_{corr} although R_c = 15 Å gives the best correlation at low XDT. Higher XDT values results in a better correlation. The enhancement is quite noticeable. It seems that, at an XDT higher than the glass transition temperature of proteins, roughly 200K (Ringe and Petsko, 2004), the topology of inter-residue contacts rigorously included in the GNM plays a dominant role. To summarize the findings, we believe that the dynamics exhibited by a protein molecule can be satisfactorily described by the GNM as a first approximation. And the predictions provided by GNM would be expected to be more accurate if the measurements refer to a higher temperature and less constrained environment (e.g. solution).

In the illustrative study of the penicillopepsin (1BXO), the dramatic change in the mobility of Tyr75, as an increased cutoff distance is imposed, can be explained by its relatively larger coordination number (CN) compared to its sequential neighbors. For instance, the CNs for SER72, ILE73, SER74, *TYR75*, GLY76, ASP77, GLY78 and SER79 are 9, 8, 8, *7*, 5, 4, 6 and 10, respectively, for R_c = 7.3 Å. However, the CN for the same residues changes to 53, 60, 45, *54*, 53,

43, 28 and 41 as the cutoff increases to 15 Å. One can easily see that the CN of Tyr75 changes from being relatively low to relatively high compared to its neighbors as R_c increases. This is due to the fact that Tyr75 sits on a loop that points into the center of an open pocket surrounded by the rest of the structure. As the cutoff increases, the surrounded neighbors start to be 'visible' or 'sensible'. The special topological arrangement in which Tyr75 is embedded, and therefore the resulting constrained dynamics renders this residue functionally critical. As a result, 15 Å could be a 'safer' choice when we apply GNM to a middle size protein based on the new criterion that *catalytic residues in enzymes have a constrained dynamics*, which will be discussed extensively in the next chapter. However, one should note that, in general, cutoffs from 7.3 to 15 Å give qualitatively the same GNM mode profile especially in the low frequency mode. Only in a few cases did we observe a mobility change in a group of residues as that we see in 1BXO. More discussion regarding this will be presented in the discussion of *COMPACT* algorithm in the next chapter.

[*] All the Supplemental materials can be found at http://ignm.ccbb.pitt.edu/Lee-thesis.zip

3. COUPLING BETWEEN CATALYTIC SITES AND COLLECTIVE DYNAMICS: A REQUIREMENT FOR MECHANOCHEMICAL ACTIVITY OF ENZYMES[‡]

3.1. ABSTRACT

A subtle interplay between chemical kinetics and molecular mechanics results in enzyme activities. This interplay requires a communication between catalytic residues and key mechanical regions of the enzyme. Here, we conducted a systematic study on 98 enzymes representative of different enzyme classes using GNM to reveal the existence of coordination between catalytic function and conformational dynamics. The result showed that more than 70% of the catalytic residues in the examined monomeric enzymes are co-localized with the global hinge centers predicted by the GNM. Moreover, 94% (87/93) of the examined enzymes have at least one global hinge center in their active site. If one normalizes the fluctuations in each protein and rate the most mobile residue as 100% and the least one as 0%, a low *translational mobility* (< 7%) is observed for the catalytic residues consistent with the fine-tuned design of enzymes to achieve precise mechano-chemical activities. The odds ratio showed an average 3.9-fold enhancement in the probabilities of finding a catalytic residue in the key mechanical regions (or hinge sites) compared to that of randomly finding one such residue in a given enzyme. On the other hand, the ligand-binding residues enjoy a moderate flexibility to accommodate the incoming substrates, although they exhibit a tendency to closely neighbor the catalytic sites.

[‡] Reprinted from *Structure* **13**(6), Yang, L-.W. and Bahar, I. "Coupling between Catalytic Site and Collective Dynamics: A requirement for Mechanochemical Activity of Enzymes" p893-904 Copyright (2005) with permission from Elsevier.

Nevertheless, highly mobile ligand-binding residues are occasionally observed in the case of sites that bind a wide range of ligands or sites that serve as part of the proton-shuttling machinery. We utilized these findings in an algorithm for enzyme active site prediction, which is based on low resolution structural constraints and dynamic fingerprints. The method shows a high sensitivity and a moderate specificity for a set of representative monomeric enzymes across all the six enzyme classes. All the false positives predicted by this algorithm turn out to be highly conserved residues, suggesting their dynamics to be associated with evolutionarily optimized functional requirements. These findings could serve as new criteria for assessing drug binding residues, and reduce the computational time and memory cost of substrate docking searches.

3.2. INTRODUCTION

Understanding the relationship between protein structure and biochemical function is of utmost importance for effective design or inhibition of proteins. Despite the rapidly increasing number of known structures and the advances in techniques for probing activity, relatively few studies have systematically investigated the connection between catalytic function and conformational dynamics. While several groups have examined the molecular dynamics of individual enzymes, only recently is conformational dynamics being recognized as a mechanism that supports catalytic activity (Benkovic et al., 2003; Daniel et al., 2002; Diaz et al., 2003; Luo et al., 2004; Ringe et al., 2004; Tousignant et al., 2004; Agarwal et al., 2004; Eisenmesser et al., 2002; Clark, 2004; Kohen et al., 1999; Wolf-watz et al., 2004).

Kern has given compelling evidence on the coupling between enzymatic kinetics and molecular dynamics in a *quantitative* detail (Eisenmesser et al., 2002) In this study, the transverse auto-relaxation rates (R2), sensible to the *slow* conformational changes (micro- to milliseconds), have been measured for an enzyme, human cyclophilin A(CypA). Measurements of R2 revealed that during the catalysis of CypA, some amino acids (the catalytic residues) undergo microsecond conformational exchanges. Kern divided the catalytic scheme into 3 states – free enzyme (E), enzyme bound with *trans*-formed substrate (ES_{trans}) and enzyme bound with *cis*-formed substrate (ES_{cis}). There are mass flows between two given states with certain forward and backward rates. The flow from E to ES_{trans} or ES_{cis} reflects the substrate binding affinity and the flow between ES_{trans} and ES_{cis} gives a catalytic rate. Kern was able to derive the rate constants from measuring the increased value of R2 upon substrate binding (R_{ex}) as a function of the increase in substrate concentration. Those rates were determined to be in hundred microseconds range, which agreed well with the measurements on the enzyme conformational changes. If we assume this catalysis to be described by a simpler Michaelis-Menten scheme, the parameter K_{cat}, revealed by the rate constants for the flows between ES_{trans} and ES_{cis}, and K_M, revealed by the binding kinetic constants associated with the flow from E to ES_{trans} and ES_{cis}, have been indirectly determined in this study. In other words, the conformational dynamics for correct enzymatic catalysis should occur in a frequency range proportional to the enzyme turnover rate or to K_{cat}/K_M, which can be experimentally measured, as proven by Kern's study.

Thornton and collaborators recently created a dataset (CATRES) in which structural and physico-chemical data on 615 catalytic residues have been compiled (Bartlett et al., 2002). The catalytic residues in the dataset are defined according to well-defined criteria and experimental

data reported for 176 non-homologous enzymes. Properties compiled in CATRES include amino acid type, secondary structure, solvent accessibility, flexibility, conservation and quaternary structure and function. In particular, attention is invited to the low temperature factors of catalytic residues as well as their preferred coiled conformations. Also, the data have shown that the relative solvent accessibility (RSA) of catalytic residues is low; catalytic residues indeed tend to be localized within 'clefts' that have moderate access to solvent, while they are not deeply buried inside the proteins. Using a neural network algorithm and spatial clustering, Thornton and coworkers were able to predict the catalytic sites of a number of test enzymes with an accuracy rate of 69% (Gutteridge et al., 2003). From the weighting scheme used in their neural network algorithm, it can be inferred that the conservation of amino acids, the residue types, especially the charged residues, RSA and cleft types play an important role (in the order of importance) but the secondary structure features and the depth (extent of the burial) have trivial impact on the accuracy of active site predictions. More recently, a new resource, the Catalytic Site Atlas (CSA) database (http://www.ebi.ac.uk/thornton-srv/databases/CSA/), has been made available by the same group (Porter et al., 2004). CSA contains both hand-curated CATRES entries *and* homologous entries generated by multiple sequence alignments, covering about 27% of all enzymes structures presently deposited in the Protein Data Bank (PDB) (Berman et al., 2000).

Recently, Ma and coworkers invited attention to the possibility of accurately describing proteins dynamics in the absence of amino acid sequence and atomic coordinates (Ming et al., 2002a; 2002b) . The major point is to take rigorous account of the protein architecture, described by the inter-residue contact topology, using an elastic network (EN) formalism (Bahar et al., 1997a; Atilgan et al., 2001). This and other studies based on EN models lend support to the view that

proteins possess mechanical characteristics uniquely defined by their particular architecture, regardless of their chemical properties. It also raises other questions. To what extent are these structure-induced mechanical properties functional? Is there any coupling between conformational mechanics and chemical activity? Can we identify potentially functional residues by merely examining the enzyme dynamics?

We present here the results from a set of 98 non-redundant, non-homologous enzymes, 24 of which are inhibitor-bound enzymes extracted from the PDB (Set 1; Table 3-1), and 74 are monomeric enzymes taken from CATRES (Set 2). Set 1 provides information on 104 catalytic residues, and 159 ligand-binding residues, and Set 2, on 253 catalytic residues.

Figure *3-1* illustrates the distribution of mean-square (ms) fluctuations, as exhibited by the experimental temperature factors, for all residues (panel *a*), catalytic residues (panel *b*), ligand-binding residues (panel *c*) in this set, as well as the distribution of these enzymes among the six EC classes (panel *d*). The B-factors scale with ms fluctuations as $B_i = 8\pi^2 <(\Delta R_i^2)>/3$, where the subscript $1 \leq i \leq N$ refers to the residue position along the sequence. We already note upon comparison of the distributions in panels *a* and *b* that catalytic residues tend to have smaller fluctuations compared to the average behavior. The origin of this behavior will be clarified below by examining the involvement of active sites in the collective motions predicted by an EN model, the Gaussian network model (GNM) (Bahar et al., 1997a).

Table 3-1 Correlation between functional sites from experiments and computations.

PDB [a]	Protein name	Size [b]	Experimental data [c] Catalytic res	Ligand-binding res	Theoretical results [d] The global hinge centers
10GS	Human glutathione S-transferase P1-1	2 x 209	7, 8, 13, 38, 44, 51, 52, 64, 65, 98	A, B: 10, 108	A, B: 47-50
1A16	Aminopeptidase P	440	260, 271 ,354, 361, 383, 406	350, 404	168-181
1A30	HIV-1 protease	2 x 99	A25, A30, B25	A27, A29, A48-A50, B23,B81,B84	A, B: 25-28, 47-54
1A3B	Human α-thrombin heavy chain	245 +14	57, 195	60A, 60D, 189, 194, 215, 219	95-102, 121-123, 132-138, 158-176, 198-208, 212-220, 221-228
1A42	Human carbonic anhydrase II	260	64, 92, 94, 96, 119	106, 131, 198, 199, 200, 202	44-53, 142-148, 186-191, 210-215, 243-245
1A47	CGTase	683	101, 141, 228, 230, 258, 328, 329	197, 371	131-148, 247-262, 496-510
1A5I	Plasminogen activator	244	57, 102, 156, 195	194	90-105, 120-123, 135-141, 155-161, 183-192, 194-209
1A5V	Asv integrase	54-199	64, 121, 157	62, 119, 154, 155, 158	62-67, 76-82, 153-158
1AEC	Actinidin	218	25	19 ,24, 26,66,68,69, 162	7-19, 113-115
1AL8	Glycolate oxidase	359	24, 108, 129, 257	161, 254	80-106, 150-161, 225-258
1ARZ	*E Coli* dihydrodi-picolinate reductase	4 x 273	B-D:159, 160, 163	B-D:12,13, 16, 17, 34, 39, 81 ,84 ,88, 102,104, 127, 129, 169, 170	A, B: 134-195, 197-239 C, D: 147-164, 189-216
1B3N	β-ketoacyl carrier protein synthase	412	163,398-401	107, 108, 111, 193, 198,202, 303, 340, 342	41-56, 145-219
1B6A	Methionine aminopeptidase 2	110-478	231	219, 328, 331, 339, 340, 376, 444, 447	163-271, 363-381, 445-462
1BGQ	N-Terminal domain of yeast Hsp90	214	40,44,79,80,84,92,93 98, 123, 124,171,173	34, 83, 124, 171	27-42, 82-93, 127-141, 149-165
1BH6	Subtilisin DY	275	32, 64, 221	99-101 ,125- 127, 155	20-26, 122-126, 204-207, 214-217
1BVV	Endo-1,4-β–xylanase	185	69, 78, 172	9, 80, 112, 116, 166	59-109, 128-140, 162-177
1BLC	β-lactamase	31-290	70	69, 234	65-72, 206-215
1BR6	Ricin	268	80, 81, 121,123, 177, 180	78	14-33, 45-52, 168-180
1BIO	Complement factor D	16-243	57, 102, 195	189, 214, 218	122-124, 136-153, 155-160
1BK9	Phospholipase A2	134	48, 52, 99	5, 9, 30, 45, 49	3-22, 43-54, 100-111
1BXO	Penicillopepsin	323	33, 213	75, 216	146-180
1CP3	Apopain	35 +227	121, 122, 161-165	64, 205, 207, 209, 214	169-195, 261-274
1CQQ	Human rhinovirus 3C protease	180	40, 71, 145, 147	142, 143, 144, 161, 165, 170	61-63, 70-72, 86-89
1CR6	Murine soluble epoxide hydrolase	2 x 544	A, B: 333, 334, 465, 495, 523	A, B: 381	A,B : 225-241

[a] References: 10GS:(Oakley et al., 1997); 1A16:(Wilce et al., 1998); 1A30:(Louis et al., 1998); 1A3B:(Zdanov et al., 1993); 1A42:(Stams et al., 1998); 1A47:(Wind et al., 1998); 1A5I:(Renatus et al., 1997); 1A5V:(Lubkowski et al., 1998); 1AEC:(Varughese et al., 1992); 1AL8:(Stenberg et al., 1997); 1ARZ:(Scapin et al., 1997); 1B3N:(Moche et al., 1999); 1B6A:(Liu et al., 1998); 1BGQ:(Roe et al., 1999); 1BH6:(Eschenburg et al., 1998); 1BVV:(Sidhu et al., 1999); 1BLC:(Chen et

al., 1992); 1BR6:(Yan et al., 1997);1BIO:(Jing et al., 1998); 1BK9:(Zhao et al., 1998); 1BXO:(Khan et al., 1998); 1CP3:(Mittl et al., 1997); 1CQQ:(Matthews et al., 1999); 1CR6:(Argiriadi et al., 1999).

(b) 1A3B has two subunits of 14 and 245 residues. 1A5V, 1B6A and 1BLC PDB coordinates refer to the indicated ranges.

(c) The underlined residues are computed to have mobility scores < 0.10.

(d) Hinge residues with mobility scores < 0.05, at the crossover between positive and negative displacements in mode 1.

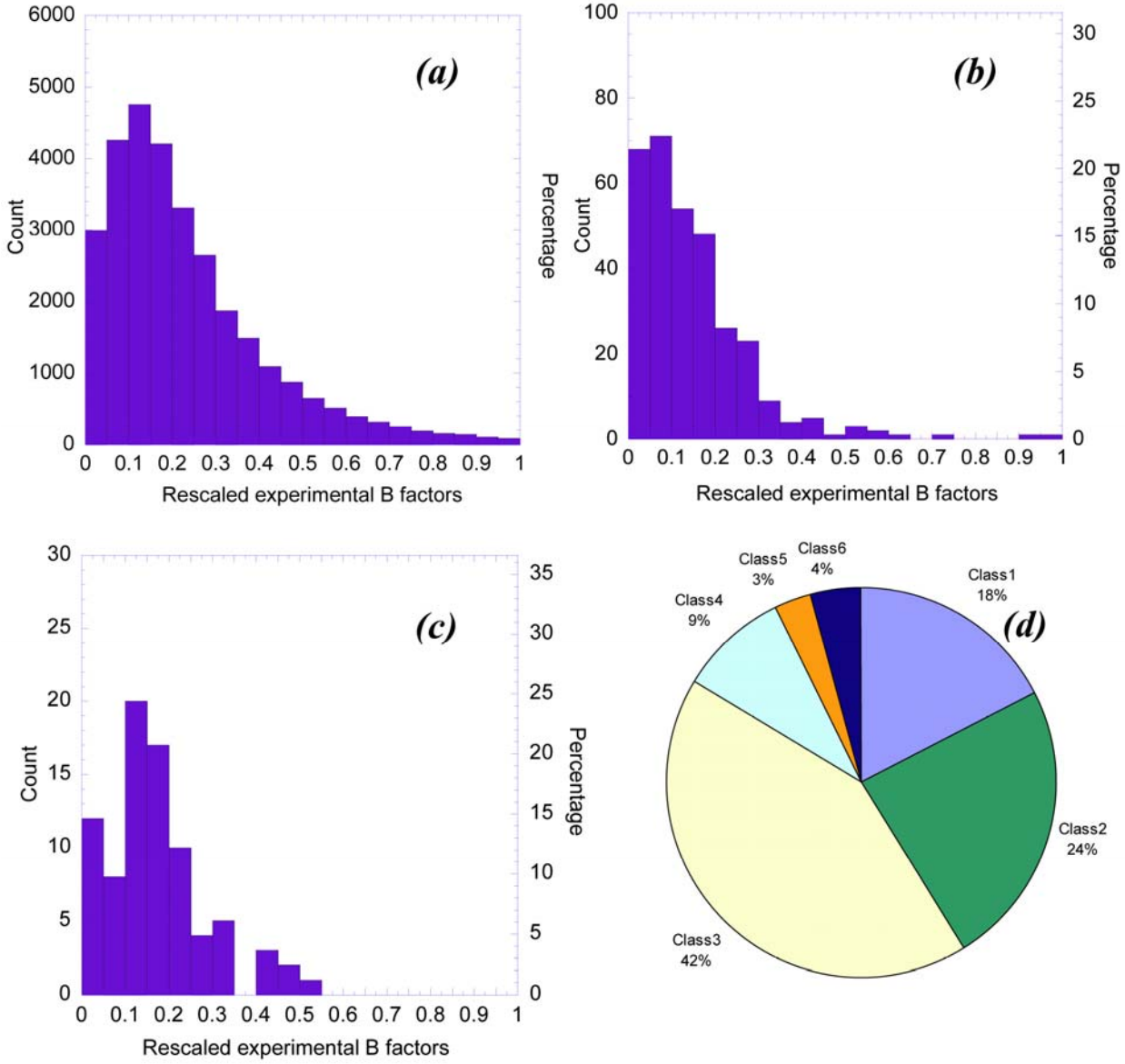

Figure 3-1 Distribution of temperature (B-) factors for average residues and catalytic residues.
Distribution of temperature (B-) factors for ***(a)*** all (30419) residues in the examined dataset of 93 monomeric enzymes (Table 1 in Supplementary materials), ***(b)*** the 324 catalytic residues of these enzymes, and ***(c)*** and the 82 ligand- binding residues of the subset of 19 monomeric enzymes in Set 1. The abscissa refers to the B-factors, divided into 20 intervals of equal size, and the ordinate indicates the number of counts in each interval. The original

Gaussian-like distribution of all residues is skewed towards low B-factors in the case of catalytic residues, and shows the same tendency but to a weaker extent in the case of ligand-binding residues. The mean values are $\langle B \rangle$ = 0.24 for 'all' residues, 0.14 for catalytic residues and 0.17 for inhibitor-binding residues. Panel ***(d)*** shows the distribution of enzyme classes (E.C.) in the observed sets of proteins.

Hinge-bending flexibility has been pointed out in several studies to be an important mechanism that underlies functional changes in protein conformations (Bahar et al., 1998a; Banks et al., 1979; Frauenfelder et al., 1998; Falcon et al., 1999; Hirano et al., 2002; Levitt et al., 1985; Pang et al., 2003; Xiang et al., 2001; Zhang et al., 2003; Ma et al., 1998; McCammon et al., 1976; Cregut et al., 1998; Sinha et al., 2001). Hinge motions may be instrumental in facilitating ligand binding (Bahar et al., 1998a; Towler et al., 2004), in mediating allosteric effects (Xu et al., 2003) or fine-tuning function (Gutteridge et al., 2003). We have shown for HIV-1 reverse transcriptase (RT), for example, that the hinge residues in the p66 palm subdomain form a stable anchoring region about which the thumb and fingers enjoy rotational mobility (Bahar et al., 1999b). The *mechanical* role of the p66 palm goes hand-in-hand with its *biochemical* function, inasmuch as RT catalyzes nucleotide addition in the p66 palm, and not surprisingly RT inhibitors bind the palm to interfere with the global motions (Temiz and Bahar, 2002).

We focus here on the low frequency motions, also called *global motions* (as opposed to local motions subject to high frequency modes), and ask if or how the global dynamics and enzymatic function are correlated. The dominant role of the slow modes in effectuating the *functional* motions has been suggested in early normal mode analyses (NMAs) (see for example Karplus et al., 1983; Go et al., 1983) and confirmed in many studies (e.g. Tama et al., 2000; Kitao et al., 1999). Our approach is to determine the slowest modes for each enzyme, examine if the catalytic residues and ligand-binding sites are distinguished by any patterns in these modes.

The hypothesis on which our study is based is that a general dynamic pattern exists for the active sites of enzymes. We further ask how we can predict enzyme active sites using this information along with our knowledge of structural and chemical properties of proteins. In the later part of this chapter, we demonstrate an early attempt to predict enzyme active sites using information of global mode mobility and coarse-grained structural features. A new algorithm – COnformational-Mobility-based Prediction for enzyme ACTive sites (*COMPACT*) was developed for this purpose. Recognition of such dynamic patterns could serve as an additional criterion for identifying potentially functional sites, and pinpointing the relationship between dynamics and function in a more concrete manner, an issue that becomes increasingly important with progresses in structural genomics and proteomics.

3.3. METHOD

3.3.1. Sample proteins

Our dataset consists of two sets of enzymes. First, all ligand-protein complexes available in the PDB were downloaded. Structures having higher than 90% sequence identity were removed; the remaining >100 structures were reduced to 24 (Set 1, Table 3-1) after requiring (i) the availability of explicit experimental data on inhibitor-binding and catalytic residues, (ii) the size of the inhibitor to be small-to-moderate (up to 35 heavy atoms), and (iii) all atomic coordinates to be deposited except those at the truncated domains that do not interfere with the catalytic site. Set 2 consists of 74 non-homologous, monomeric proteins extracted from CATRES (Bartlett et al., 2002). Three of these have a substrate composed of less than 10 residues (PDB identifiers:

2PHK, 8PCH and 8TLN). The complete dataset of enzymes is given in the Supplementary Material, Table S3-1 (http://ignm.ccbb.pitt.edu/Lee-thesis.zip).

3.3.2. Definition of catalytic residues

According to the definition introduced by Bartlett et al. (Bartlett et al., 2002), a given residue is catalytic if *(i)* it is directly involved in a catalytic function, *(ii)* it affects the residues or water molecules directly involved in catalysis, *(iii)* it can stabilize a transient intermediate, or *(iv)* it interacts with a substrate or cofactor that facilitates the local chemical reaction. These criteria were adopted for defining the catalytic residues in set 2. Those in set 1 were identified either from *(i)* experimental data that explicitly indicate the involvement in catalytic function, or from *(ii)* the label 'SITE' in the PDB entry. We note that not all PDB files of enzymes include these labels, hence the need to examine the literature. This definition was confirmed to point to the same 'active' amino acids when applied to set 1, except for the inclusion of 1-2 additional residues in a few cases.

The inhibitor-binding sites listed in Table 3-1 are those reported in previous experimental studies to bind the inhibitor (ligand). They may, or may not, overlap with an active site.

3.3.3. Defining catalytic/inhibitory residues in two illustrative examples

Penicillopepsin (PDB code:1BXO)

The Protein Data Bank file 1BXO is a structure comprising a penicillopepsin, an aspartic protease, and a macrocycle pentapeptide inhibitor, PPi4. The aspartic protease family, which carries the essential proteolytic functions in many human pathogens, utilizes two aspartic acids and a deprotonated water to mediate the electron transfer between the catalytic residues and the

substrate, which in turn triggers the breakage of the scissile peptide bond of the substrate (Khan et al., 1998).

The general catalytic mechanism is demonstrated through an example of rhizopuspepsin in Figure 3-2 (reproduced from Suguna's work; Suguna et al., 1987). First, a water molecule coordinated by the catalytic aspartic acids D35 (corresponding to D33 in penicillopepsin) and D218 (corresponding to D213 in penicillopepsin) is deprotonated by the carboxyl oxygen of D218, which generates a nucleophilic hydroxyl group (Figure 3-2a). The carbonyl oxygen of the scissile bond forms a hydrogen bond with the $O^{\delta 1}$ atom of D35, polarizing the carbonyl bond and making the carbon atom more susceptible to the nucleophilic attack by the hydroxyl ion. A tetrahedral carbonate intermediate is thus formed at the onset of the peptide bond cleavage (Figure 3-2b). A weak hydrogen bond between the carbonyl oxygen of the P2 residue of the substrate and one of the hydrogens of the peptide nitrogen positions the hydrogen, such that the opposite direction becomes accessible to accept a proton from the $O^{\delta 2}$ of D218, which breaks the peptide bond and convert the catalytic aspartic acids back to their original hydrated state (Figure 3-2c).

In the case of penicillopepsin, the substrate is replaced with the noncovalently bound inhibitors, PPi3 and PPi4, mimicking the tetrahedral intermediate in the transition state along the reaction pathway (Figure 3-3; reproduced figure; Khan et al., 1998). The P2 asparagine on PPi3 and PPi4 are coordinated in position through a hydrogen bond between the $N^{\delta 2}$ atom of the asparagines and the $O^{\gamma 1}$ atom of T216 in both PPi3 and PPi4. Y75, a highly conserved residue in the flap region of all aspartic proteinases, forms part of the S1 pocket that interacts with the substrate at

the P1 position that contains a leucine residue. Here, T216 and Y75 are considered as inhibitory residues due to their contribution in interacting with specific inhibitors.

In the CSA, the catalytic information on 1BXO was not obtained manually from literature survey. Instead, it is derived from sequence alignment against its sequential homologs whose catalytic residues are also annotated through sequence alignment. In addition to D33 and D213, threonine 34, serine 36 and threonine 216 were also found to be potentially active site residues. However, warnings have been prompted in the CSA stating that the catalytic site annotation transferred by sequence alignment could be problematic. This problem has been revealed at the slight difference in protein functions that are conventionally indicated by 4-digit E.C. codes, although the differences usually lie in the last one or two digits. In this case, we conform to the original literature and annotate D33 and D213 as the catalytic residues for penicilopepsin.
Based on the GNM analysis, 46 residues fulfill the requirement of belonging to minima of type I or type II with a mobility score $M_{i1,2} < 0.05$; D33 and D213 are found to take part in this set of selected residues. Thus, the *odds ratio* for 1BXO is equal to (the number of catalytic residues present in the set "Selected" by GNM/ total number of "Selected" residues) / (total no. of catalytic residues/size of the protein) = (2/46)/(2/323) = 7.02 (Table 3-3); "Selected" is defined as the residues that belong to minima of type I and II with $M_{i1,2} < 0.05$.

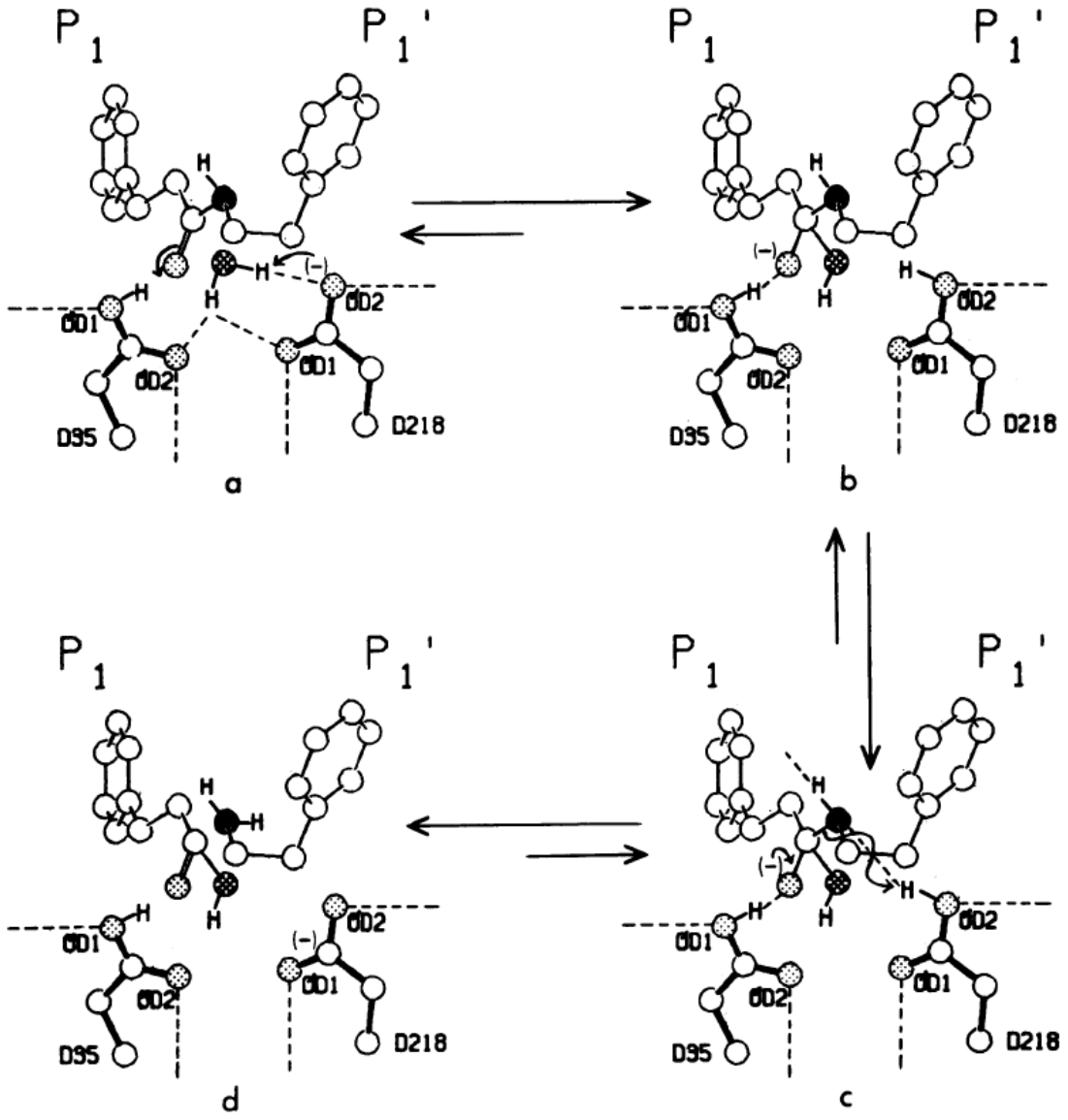

Figure 3-2 A proposed catalytic mechanism for rhizopuspepsin.
D35 is Asp-35; D218 is Asp-218; OD1 is $O^{\delta 1}$; OD2 is $O^{\delta 2}$.

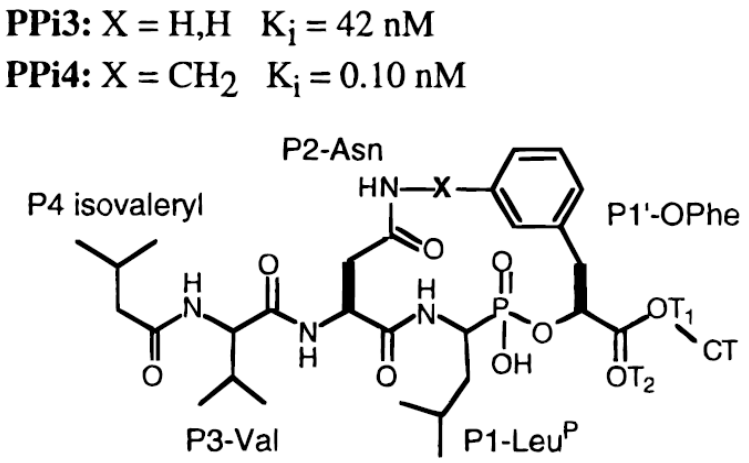

Figure 3-3 The chemical structures of PPi3 and PPi4.
PPi3 and PPi4 are cyclic pentapeptide inhibitors with the side chains of P1' and P2 being intramolecularly bonded.

Endo-1, 4-xylanase (PDB code:1BVV)

The endo-1, 4-xylanase of family G/11 is one of the enzymes responsible for xylan hydrolysis in many organisms. The catalytic mechanism is known as a *retaining mechanism* involving two carboxylic acids, one of which serves as a general base and the other serves as a general acid. The proposed mechanism is illustrated in Figure 3-4 (reproduced from Figure 1 of McCarter's work in 1994; McCarter and Withers, 1994). First, a glycosyl-enzyme intermediate is formed through an anomeric linkage between the C^1 atom of the glucoside and the general base as the general acid protonates the –OR leaving group. Subsequently, an adjacent water molecule comes in between the two carboxylic acids and protonates the general acid. The resulting nucleophilic

hydroxyl group forms a covalent bond with the C^1 atom of the glucoside and replaces the anomeric bond. In the case of endo-1, 4-xylanase, E78 is the general base and E172 is the general acid as shown in Figure 3-5 (reproduced from Figure 1 of Sidhu et al., 1999). CSA annotates these two as the catalytic residues as well according to Davies' work in 1995 (Davies G. and Henrissat B., 1995). However, in Sidhu's experiments (Sidhu et al., 1999), point mutation of Tyr69 into phenylalanine abolishes the enzyme activity entirely. The catalytic importance of Y69 is also confirmed by the mutagenesis study of Tyr298 in β-glucosidase, corresponding to the Y69 in the xylanase. The catalytic activity of the mutant is reduced by 2000 fold compared to that of the native enzyme. Based on structural analysis, it is observed that a bifurcated hydrogen bond is formed between H^{η} of Y69 and the endocyclic oxygen (O^5) of the xylose residue as well as between H^{η} of Y69 and $O^{\varepsilon 2}$ of the catalytic nucleophile, Glu78 (Sidhu et al., 1999). Hence, it is very likely that an asymmetric protonation occurs during the catalytic transition that favors the hydrogen bond on the side of H^{η};Y69 - $O^{\varepsilon 2}$;E78. This in turn facilitates the attack of the nucleophilic water on the anomeric carbon (the C^1 atom of the xylose residue), and results in the displacement of the α-anomeric linkage with the hydroxyl group. Based on this information, we consider Y69 also to be a catalytic residue for 1BVV.

A few additional interactions are involved in the binding or better positioning of the inhibitor but not directly in mediating the glucosyl bond breakage in the enzyme-glucoside intermediate complex, BCX-2FXb (2-deoxy-2-fluoro xylobiosyl enzyme intermediate of *Bacillus circulans* xylanase). These are either contributed from the hydrogen bonds (Arg112, Pro116 and Tyr166) or by the hydrophobic interaction (Trp9) (Figure 3-6). These residues are defined as inhibitory residues in the present study.

There are 57 residues that are identified by GNM to belong to minima of type I and II with $M_{i1,2}$ < 0.05; Y69 is the only catalytic residue among them. Residues E78 and E172, on the other hand, have $M_{i1,2}$ < 0.05 but do not belong to the minima I/II. Thus, the *odds ratio* for 1BVV is equal to (1/57)/(3/185) = 1.08 (Table 3-3);.

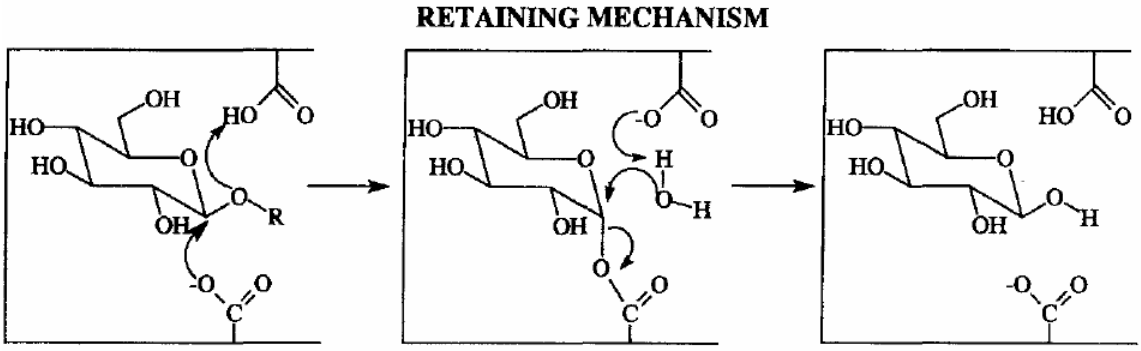

Figure 3-4 The proposed mechanism for retaining glucosidases.

3.3.4. Collectivity

The degree of collectivity of mode *k* (Tama et al., 2001) is computed from

$$\Omega_k = (1/N) \exp \{- \sum_k [\boldsymbol{u}_k]_i^2 \, ln \, [\boldsymbol{u}_k]_i^2 \} \qquad (3\text{-}1)$$

The k^{th} eigenvector ($\boldsymbol{u}_k$) of $\boldsymbol{\Gamma}$ gives the profile of residue displacements along the mode *k*. N is the number of residues. The lower frequency modes usually have higher collectivity than the higher frequency modes. We examine the spatial position of the catalytic and inhibitor-binding residues relative to that of the *global hinge regions* associated with the slowest modes. The global hinges are identified from the crossover between the positive and negative elements of the eigenvectors.

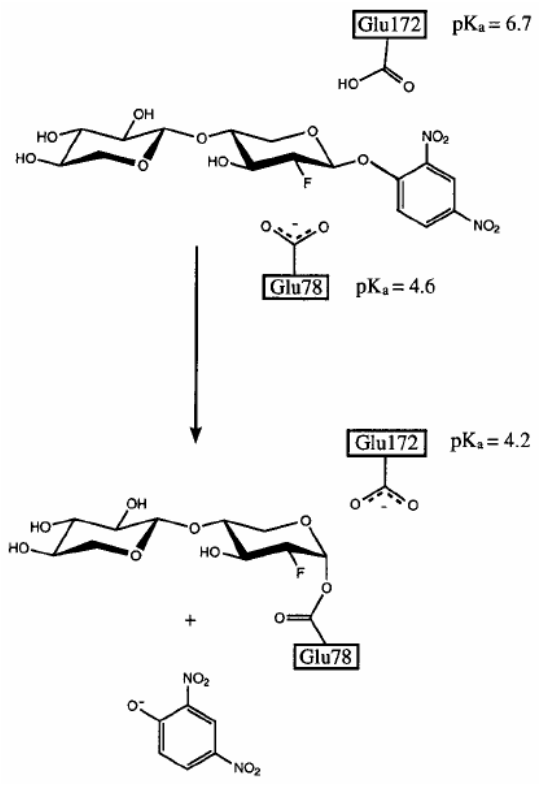

Figure 3-5 Description of the formation of the glycosyl-enzyme intermediate.
The fluorine atom at position 2 of saccharide slows down both the formation and the hydrolysis of the intermediate. The 2, 4-dinitrophenyl leaving group polarizes the anomeric carbon upon leaving and facilitates the first nucleophilic attack by Glu78, allowing the intermediate to form.

3.3.5. Illustration of GNM analysis and comparison with experiments

Figure *3-7* illustrates the computations for an example protein, endo-1,4-β-xylanase (1BVV) (Oakley et al., 1997) from Set 1. The eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are plotted in panel *a*. The global hinge centers in the modes 1 and 2 are located at the crossover between the positive and negative

displacements of the respective eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. Four groups of global hinge residues are shared by the two modes: T67-G70, V81-S84, R89-P90, and Y166-M169, indicated by the

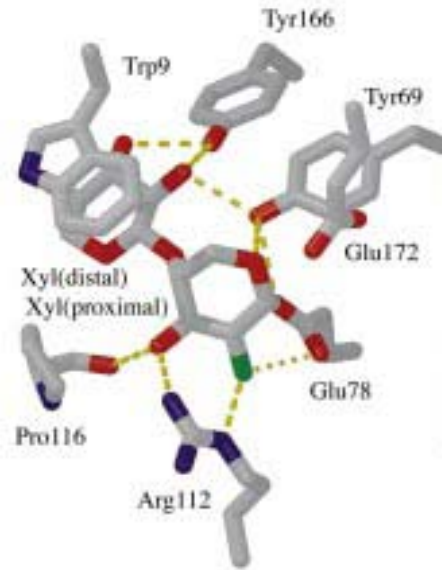

Figure 3-6 2FXb group covalently bound with E78.
2FXb (2-deoxy-2-fluoro xylobiosyl) group is surrounded by the catalytic residues of E78, E172 and Y69 and the inhibitory residues Trp9, R112, P116 and Y166. The fluoride atom is shown in green, oxygen in red, and nitrogen in blue along with the gray carbon backbone.

arrows. Panel *b* shows the ribbon diagrams colored from red (most mobile) to blue (most rigid) according to the square fluctuations of residues in the slowest two modes, and panel *c* displays the structural regions exhibiting opposite direction fluctuations in these modes, colored red (positive) and blue (negative) in the two modes. The hinge residues lie at the interface between these anticorrelated regions. The catalytic and/or inhibitor-binding residues reported in the literature are labeled in panel *a*. Note that mode 1 essentially sets in motion an extended loop at the entrance of the catalytic pocket permitting the opening/closing of the catalytic site, while

mode 2 engages a larger portion of the structure. The respective collectivities of modes 1 and 2 calculated from Eq. 3-1 are $\Omega_1 = 0.246$ and $\Omega_2 = 0.576$.

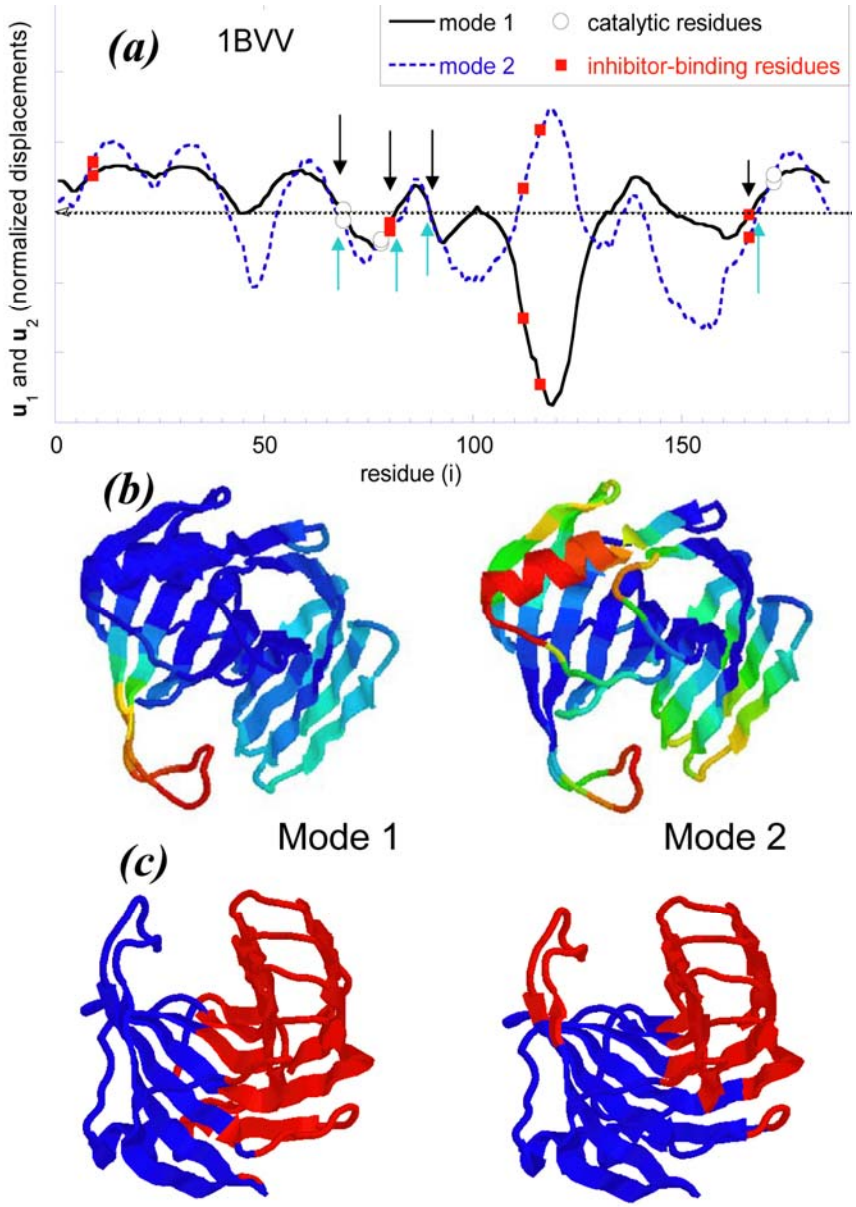

Figure 3-7 Distribution of residue displacements along global modes for endo-1,4-β-xylanase
(a) Distribution of displacements along slowest modes 1 (solid) and 2 (dotted), as a function of residue index, computed for endo-1,4-β-xylanase. Catalytic and inhibitor binding residues are indicated by the respective labels 'O' and '■'. The crossover regions between negative and positive motions are the predicted global hinge sites, G41-W42, Y53-N54, T67-G70, V81-S84, R89-P90, T110-T111, F125-T126, Q133-P137, S140-N41, T147-N148, Y166-M169

(by combining both modes), four of which indicated by arrows are common to modes 1 and 2. ***(b)*** ribbon diagrams colored blue-green-yellow-orange-red in the order of increasing mobility of residues along modes 1 and 2. The (normalized) mobilities, $[u_k]_i^2$, are directly found from the values in panel ***a***, squared, ***(c)*** the regions subject to opposite direction movements in modes 1 and 2, deduced from panel ***(a)***. Regions colored red and blue correspond to '+' and '-' displacements, respectively, along the first (left) or second (right) mode axes.

Of interest is the comparison of experimentally known residues with the hinge sites predicted by the GNM. Three catalytic residues have been reported for 1BVV, one of which (Y69) coincide with a global hinge residue and the other two (E78 and E172) are positioned close to the hinge centers V81-S84 and Y166-M169, supporting the view of a coupling (communication) between catalytic and mechanically important sites.

The inhibitor binding residues, on the other hand, show a more varied behavior: one (Y166) coincides with a global hinge residue, two (Y80 and R112) are first sequential neighbors to hinge residues V81 and T111 (mode 2 only), while P116 is rather exposed, distinguished by a high flexibility. The latter may indeed be instrumental for efficient recognition of substrates. These results suggest that ligand-binding may involve a network of residues that includes both highly flexible 'recognition' sites as well as constrained residues establishing the communication with the mechanically and chemically active sites.

3.3.6. COnformational-Mobility-based Prediction of enzyme ACTive sites (COMPACT)

Residues that precisely lie at minima in the mobility plots (see the plots in Figures 3-9a) are referred to as key mechanical residues of type I, and their first and second neighbors along the sequence are described as types II and III, respectively. In *COMPACT*, each minimum comprises 5 residues : i-2, i-1, i, i+1 and i+2 (i falls in the range $3 \le i \le N$-2) where $M_{i\text{-}2,\text{av}} \ge M_{i\text{-}1,\text{av}} \ge M_{i,\text{av}} \le M_{i+1,\text{av}} \le M_{i+2,\text{av}}$ and $M_{i,\text{av}}$ has to be less than 0.1. Residue i is the type I key mechanical residue.

Here, $M_{i,\mathrm{av}} = M_{i,1}/\lambda_1 + M_{i,2}/\lambda_2$ (3-2)

Let R_i represent the distance between the mass center of the protein and a given residue i. A small R_i value will usually correspond to a residue occupying a more central (interior) position, and the coordination number z_i corresponding to that residue would be expected to be relatively high. Thus, an inverse proportionality between R_i and z_i would be expected. The normalized r_i, equal to $R_i N^{-1/3}$, for each minimum (potential catalytic residues) is plotted against Z_i, which is the cumulative z_i that equals $\sum_{k=i-2}^{i+2} z_k$, in Figure 3-8 on the right panel. A residue may be viewed as an 'outlier' if its coordination number is much smaller than that expected from its R_i by the regression line. This residue could enjoy a relatively higher solvent exposure despite being tightly packed. So, the relative position of minima and the regression line provide information on the solvent accessibilities of potential catalytic sites.

The minima in the global mode shapes that have a high-to-moderate Z_i or the 'outliers' that are distinguished by relatively low Z_i (for a given r_i) are selected for further scoring. We divide the region between the highest and the lowest Z_i (see the legend of Figure 3-8) into four zones – Zone I, II , III and IV. In each zone, those minima are ranked according to their Δr_i value where $\Delta r_i = f(Z_i) - r_i$ and $f(Z_i)$ is the function of the regression line. The residue with the largest Δr_i is top-ranked in each zone. A subset of top-ranking residues are selected for further clustering and scoring, which consists of the top-ranking 4/5 residues in zone IV, the top-ranking 4/5 residue below the regression line in zone III, the top-ranking 2/5 residues below the regression line in zone II, and one residue below the regression line in zone I. The number of minima to be selected in a given zone is rounded to its next integer value. For example, if there are 11 minima

in zone IV, the first 9 (instead of 8.8) with the highest Δr_i values will be selected. The selected n_1 minima are rank-ordered according to their Δr_i values and a deviation score, DS(i), is given by

$$DS(i) = (n_1 - i + 1)/n_1 \text{ , } i = 1, .n_1 \qquad (3\text{-}3)$$

where $\Delta r_i > \Delta r_{i+1}$, i = 1,. n_1-1

The clustering property of a given minimum is indicated by the summation of its pairwise distances from the other minima. d(i, j) denotes the linear distance between minimum i and j. We rank the summation of pairwise distances for minimum i, D(i) and give clustering score, CS(i), is given by

$$CS(i) = (n_1 - i + 1)/n_1, \; i = 1,.n_1 \qquad (3\text{-}4)$$

where $D(i) < D(i+1)$, i = 1,. n_1-1

and $D(i) = \Sigma_j \, d(i, j)$, j=1..n_1

As we can see, both DS(i) and CS(i) are normalized between 0 and 1. We further rank the residues by their FS(i) score, which is the sum of CS(i) and DS(i). The smallest of 12 or 3/4 of the minima with the highest FS score will be selected. If the number is less than 9, a re-clustering procedure is applied and minima are ranked according to their new FS scores.

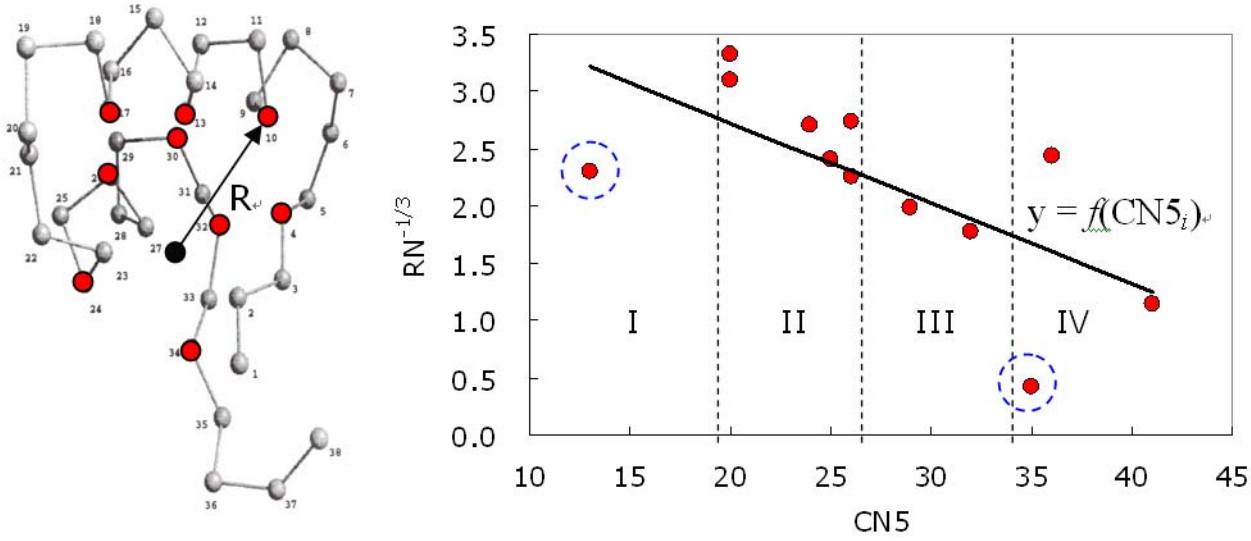

Figure 3-8 The *COMPACT* algorithm

The left panel provides a coarse-grained presentation of this illustrative protein. The gray nodes indicate the C^{α} atoms of the residues. The black dot denotes the mass center and each red dot indicates the type I minimum with $M_{i,\mathrm{av}} \leq 0.1$ centered at residue *i*. $CN5_i$ is the summation over the coordination numbers of residues *i*-2, *i*-1, *i*, *i*+1 and *i*+2. R stands for the distance between the mass center of the protein and a given residue. On the panel right, R at each minimum is plotted against its CN5. R is normalized by dividing it by $N^{1/3}$. Blue dash circles indicate the minima that are far away from the regression line, implying locally high solvent accessibilities. Three vertical dash lines divide the minima into four regions: Zones I, II, II and IV. Each region spans one quarter of the difference between the highest and the lowest CN5 values. 4/5 of the total minima in Zone IV, 4/5 of the total minima underneath the regression line in Zone III, 2/5 of the total minima underneath the regression line in Zone II and 1 minimum underneath the regression line in Zone I are selected for further spatial clustering and scoring, after rank-ordering the residues based on their departure from the regression line. See the text for more details.

3.4. RESULT AND DISCUSSION

3.4.1. Catalytic residues coincide or communicate with global hinge regions

Figure *3-9a* displays the *fluctuation profiles* in the global modes of motion for a few enzymes selected from set 1. Fluctuation profiles for the complete set of 98 enzymes can be accessed at http://ignm.ccbb.pitt.edu/. The abscissa represents the residue index, and the curve displays the

distribution of residue fluctuations (squared) in the slowest modes predicted by the GNM. Peaks indicate the most mobile regions, and minima are those regions anchored in space, some of which act as global hinge residues (shown by the arrows). Global hinge residues are at the interfaces between domains, or clusters of residues, which move in opposite directions in the global modes. We will refer to minima in the slow modes as *key mechanical sites*.

Figure *3-9a* suggests that most of the active residues tend to occupy minima in the fluctuation profiles. Notably, the catalytic residues preferably coincide with, or sequentially neighbor, key mechanical sites regardless of the enzyme function or size. Panel *b* displays the color-coded ribbon structures corresponding to the proteins in panel *a*. The dark blue regions (minima in the slow modes) point to the residues subject to the strongest constraints in the global modes. Although these residues are not contiguous along the sequence, they usually cluster in space so as to consolidate the anchor/hinge region that coordinates the global movements. Examination of their structural properties and context shows that they are not necessarily coiled regions or domain linkages, but may occasionally occur in secondary structural motifs, such as kinks in helices.

3.4.2. Quantitative assessment of mobilities in the global modes

In order to make a quantitative assessment of the dynamics of active residues or ligand-binding in relation to key mechanical residues, we assigned a mobility score M_{ik} to each residue i in the k^{th} mode. M_{ik} is the square fluctuation normalized with respect to the most mobile residue in the k^{th} mode of the particular enzyme. The highest peak in the slow mode profile of each enzyme (Figure *3-9a*) is thus assigned a mobility score of unity and the lowest, zero. Additionally, we examined the relative mobilities along the sequence, because a given residue may appear rigid

due to the constraints imposed on its bonded neighbors. Residues that precisely lie at local minima are referred to as key mechanical residues of type I, and their first and second neighbors are described as types II and III, respectively.

The *global* mobility scores corresponding to the more collective mode among the two slowest modes were computed for all catalytic and inhibitor-binding residues of our dataset, which led to the distributions shown in Figure *3-10.* The scores for the individual residues can be accessed in the *i*GNM database at http://ignm.ccbb.pitt.edu/ (Yang et al., 2005). The distributions in Figure *3-10* panels *a* and *b* may be compared to the respective distributions in panels *b* and *c* of Figure *3-1*, which reveals that the skewed distribution of ms fluctuations indicated by the B-factors is further pronounced when attention is confined to the mobilities in the most collective modes. 78% of the catalytic residues show mobility scores below 0.10 in the slow modes (66% below 0.05) as can be seen from the inset in panel *a*. In contrast, the same interval (< 0.10) in the B-factors distribution (i.e. all modes) contains 43% of catalytic residues and 25% of all residues. The insets in Figure 3-10 display the cumulative percentage of catalytic residues (panel ***a***), and inhibitor-binding residues (panel *b*) in different mobility ranges, compared to those observed for all residues in the same mobility ranges. We also compare the predictions from global modes, vs all modes, which clearly demonstrates the high propensity of catalytic residues to have low mobilities in the global mode profiles. The low frequency mode shapes thus provide a means of discriminating potentially active sites.

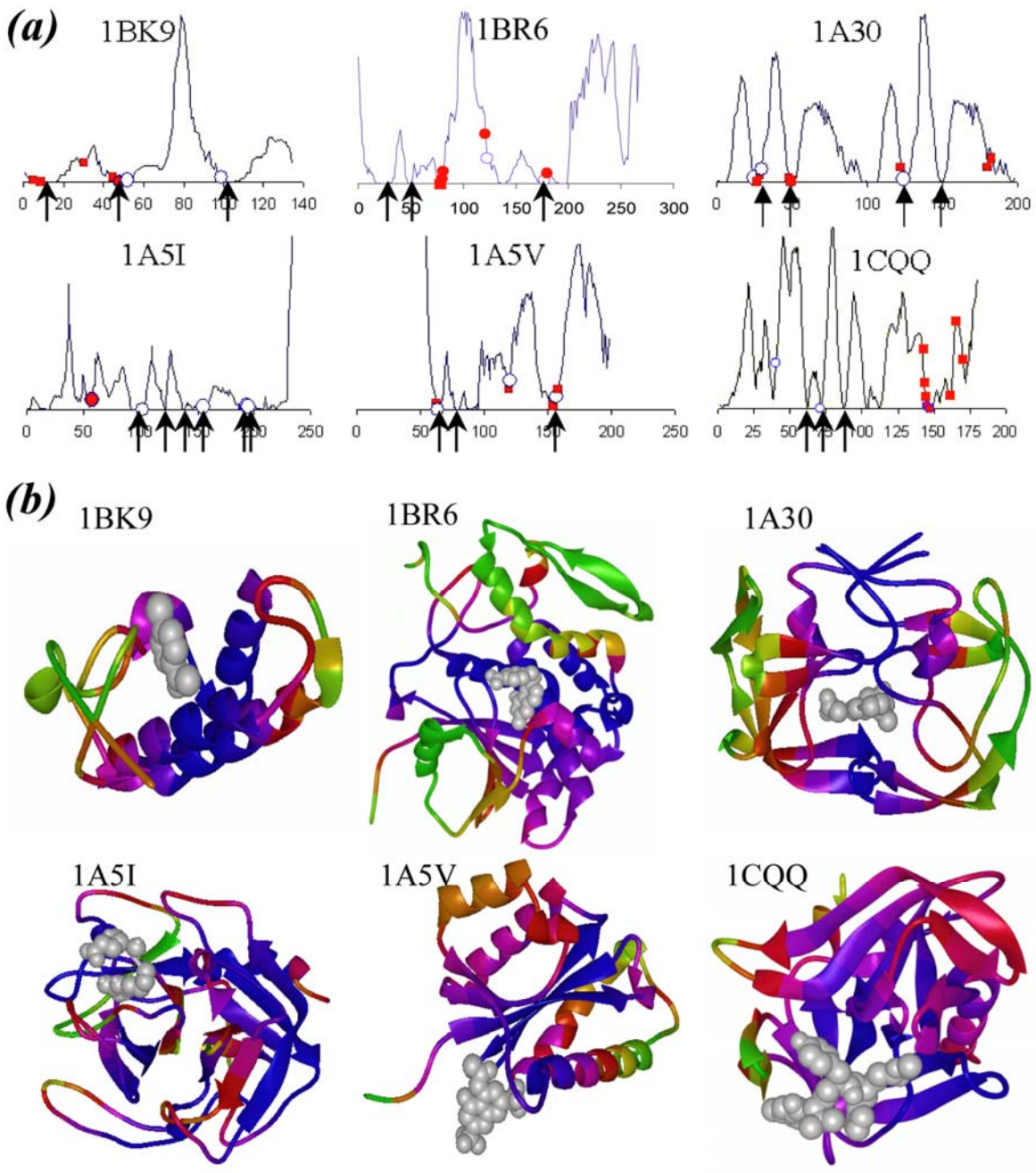

Figure 3-9 Fluctuation profiles and Color-coded ribbon diagrams for 6 representative enzymes
(a) Fluctuation profiles in the global mode ($k = 1$) and position of catalytic and inhibitor binding residues illustrated for six enzymes from Set 1. Residues involved in catalytic function are marked in (O), inhibitors binding sites are in (■), and residues serving both catalytic and inhibitor binding functions are marked in (●). Arrows indicate the hinge sites. ***(b)*** Color-coded ribbon diagrams showing the localization of inhibitors (gray ball-and-stick) near the most constrained (blue) regions. See Table 1 for the list of chemically (from experiments) and mechanically (from computations) key residues for all ligand/inhibitor- bound enzymes in our dataset.

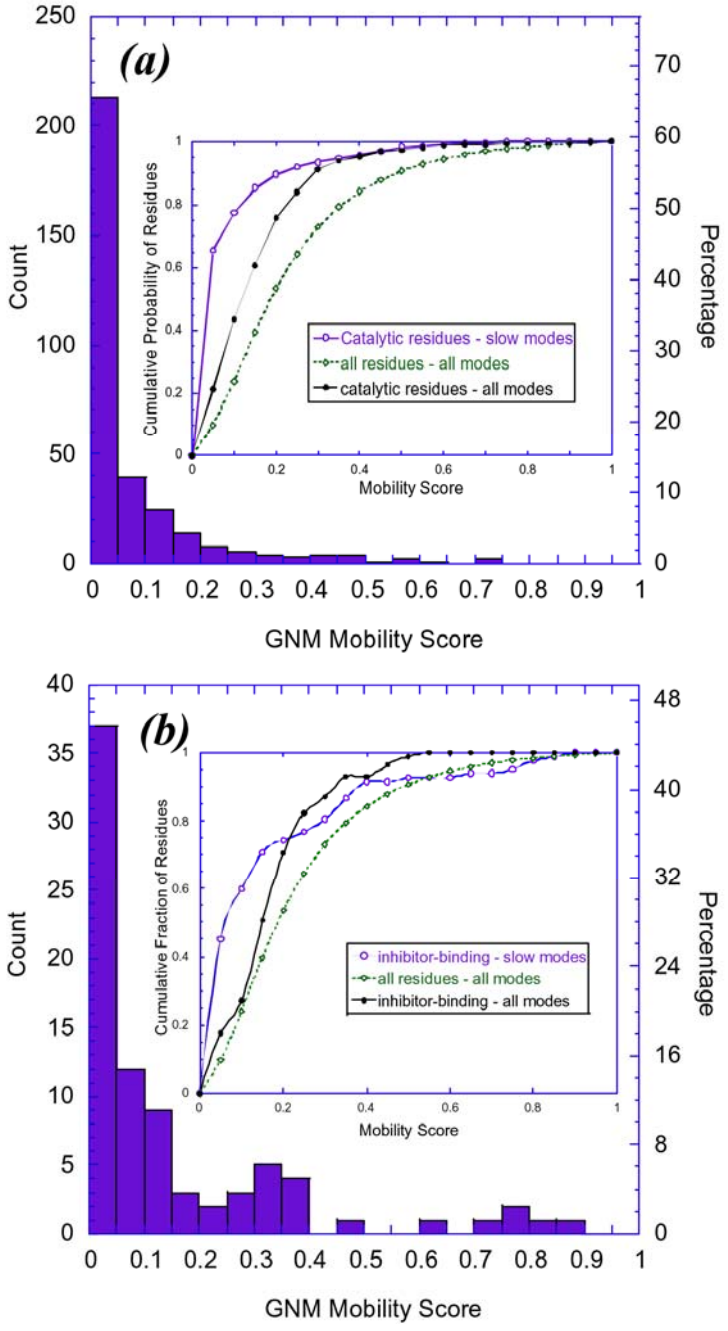

Figure 3-10 Distribution of mobilities predicted by the GNM slow modes analysis for ***(a)*** 325 catalytic residues, and ***(b)*** 81 ligand-binding residues in the examined set of 98 enzymes. The abscissa refers to the mobility scores of the residues found in the most collective GNM mode. The inset displays the cumulative fraction of catalytic residues (panel ***a***), and inhibitor-binding residues (panel ***b***) in different ranges of mobility based on the slow modes, compared to the cumulative fraction of catalytic residues and all residues in all modes and (see the labels).

The results found by averaging the scores over all proteins are presented in *Table 3-2*. $\langle M_1 \rangle_{cat}$ and $\langle M_2 \rangle_{cat}$ refer to the average behavior of catalytic residues in the slowest modes 1 and 2, and $\langle M_1 \rangle_{lig}$ and $\langle M_2 \rangle_{lig}$ are their counterparts for ligand-binding residues. The averages were calculated by evaluating first an average over catalytic residues for each protein, and then averaging over all proteins, which removes the biases arising from the different numbers of functional sites per protein. These results show that the catalytic residues possess highly suppressed mobilities in the first two slowest (dominant) modes. The low mobilities indicate their participation in (or close proximity to) the key mechanical sites of the molecules. These residues are for the most part fixed/anchored in space while the other regions undergo motions about them. The low mobility of the catalytic residues was already apparent in their B-factors, although this effect was less pronounced due to the superimposition of all modes in the B-factors. Extraction of the global modes shows that the reduced mobilities are essentially caused by their constrained global dynamics, rather than local packing effects.

Table 3-1 compares the *chemically active* (catalytic and ligand-binding) residues identified in previous experimental studies (columns 4 and 5) and the global hinge centers (column 6) predicted by present computations. The underlined residues in columns 4 and 5 are those, among the experimentally reported chemically active residues, that are indicated by the GNM to be key mechanical sites of types I – III and to have mobility scores < 0.10. The close correspondence between chemical activity and mechanical role is evident by the large fraction of underlined residues. Column 6 lists the global hinge centers distinguished by mobility scores below 0.05 in the slowest mode. These residues are proposed here to be *critical residues* from mechanical point

of view, which could serve as targets for interfering with the dynamics of the enzyme, yet may or may not being involved in the active site.

Table 3-2 Mobility scores (x 100) for catalytic and ligand-binding residues

	Enzymes	ACTIVE SITES			LIGAND-BINDING SITES		
		$\langle M_1 \rangle_{cat}$	$\langle M_2 \rangle_{cat}$	$\langle M_B \rangle_{cat}$	$\langle M_1 \rangle_{lig}$	$\langle M_2 \rangle_{lig}$	$\langle M_B \rangle_{lig}$
All (set1)	Average over proteins	6.72	5.94	12.55	12.15	9.72	16.72
	Standard deviation	8.60	9.91	7.44	13.37	10.31	7.24
All (set2)	Average over proteins	9.06	6.75	13.77			
	Standard deviation	9.40	8.46	9.09			
Monomers (set1)	Average over proteins	4.55	3.47	11.49	11.06	8.89	15.25
	Standard deviation	5.22	4.67	7.55	12.82	10.42	6.94
Multimers (set1)	10GS (dimer)	21.30	41.08	24.45	27.31	23.05	25.62
	1A30 (dimer)	10.13	5.13	11.56	5.24	6.94	24.38
	1CP3 (dimer)	35.90	2.73	21.51	39.33	3.71	22.21
	1CR6 (dimer)	12.98	18.11	20.90	12.92	17.60	18.00
	1ARZ (dimer)	0.10	0.13	12.13	2.81	5.46	27.15

3.4.3. Ligand-binding residues enjoy higher mobility despite their close proximity to catalytic sites

The ligand-binding sites exhibit higher flexibility ($\langle M_{1\text{-}2} \rangle_{lig} \approx 0.11$ and $\langle M_B \rangle_{lig} \approx 0.17$) and larger variations compared to the catalytic sites, although they are also relatively constrained in the global motions. We note that some of the ligand-binding residues also act as catalytic residues. Exclusion of these residues (46 out of 159), leads to an increase in mobility scores

($\langle M_1 \rangle_{lig}$ = 0.148, $\langle M_2 \rangle_{lig}$ = 0.104 and $\langle M_B \rangle_{lig}$ = 0.19) accompanied with larger variations. This suggests that ligand-binding residues occupy proximal positions with respect the catalytic residues but enjoy higher fluctuation, and the latter may indeed be required for efficient recognition of substrate, and optimization of intermolecular interactions. The close proximity of inhibitors to catalytic sites and their moderate mobility suggest that they block the catalytic function by interacting with the fluctuating residues in the entrance of a catalytic pocket for example. Our findings support the observation that regions of high and low structural stabilities participate in binding sites (Freire, 1999).

3.4.4. Dimerization induces new cooperative modes that engage the catalytic site

Among the 98 enzymes presently examined, four (10GS, 1A30, 1CP3 and 1CR6) are dimers, and one (1ARZ) is a tetramer. It is of interest to assess how multimerization affects the mobilities of active sites. A pure monomer set was generated by removing the multimers from the set 1. It is shown in Table 3-2 that the $\langle M_1 \rangle_{cat}$ and $\langle M_2 \rangle_{cat}$ values decrease in this case, as well as their standard deviations, consistent with the higher mobilities of chemically active residues in the multimers (Table 3-2). Multimerization usually provides a means of achieving structural and dynamic properties that would otherwise be inaccessible to the monomers. It is of interest to see if the new structures and structure-induced modes of motion (especially low-frequency global motions) impart stability and/or mechanical properties that affect catalytic residues. It can be anticipated that a high mobility/disorder at the catalytic site might be detrimental from the point of view of the precise regulation and communication ability of the active site.

Figure 3-11 displays the color-coded ribbon diagrams obtained for the 1st (left column) and 2nd (right column) slowest modes of these multimers. Catalytic residues are shown in red ball-and-stick representation, and inhibitor-binding residues are shown in color-coded (according to mobilities) ball-and-stick representation.

The dimers 10GS, 1A30 and 1CP3 exhibit symmetric motions with respect to extended hinge regions that span the entire structures. The hinge-region lies at the interface between the monomers in mode 1 in all three cases, whereas it runs through the cores of the monomers in mode 2, as indicated by the dashed lines. Interestingly, the motions of the monomers in the 2nd mode of the dimers almost exactly reproduce their 1st mode computed for the monomers taken alone (not shown), consistent with the decrease in the wavelength of the 2nd slowest vibrational mode by ½ compared to the 1st; and one expects the higher frequency modes to be even more localized. So, mode 1 is the new (most collective) mode that appears to be exclusively induced in the dimeric form. In 10GS, this mode ensures the localization of the catalytic sites in a mechanically important region. The mobility score of the catalytic residues in this mode is 0.21, which is decreased by a factor of 2 compared to the mobility in mode 2 (0.41), and this coupling to a global hinge region may be anticipated to be functional. A similar effect can be conjectured in 1CR6 where dimerization secures the co-localization of the catalytic site with a global hinge region. In 1A30 and 1CP3, on the other hand, mode 2 already confines the catalytic site in a mechanically key region, as evident from the low mobility scores. Dimerization seems to be a structural, rather than dynamic, requirement in these cases. It permits the catalytic site to be sequestered from solvent in HIV-1 protease (1A30). Finally, the catalytic residues are positioned at the interface between the core domain and two different peripheral domains in the two slowest modes of 1ARZ, suggesting the activation of different domains in different modes.

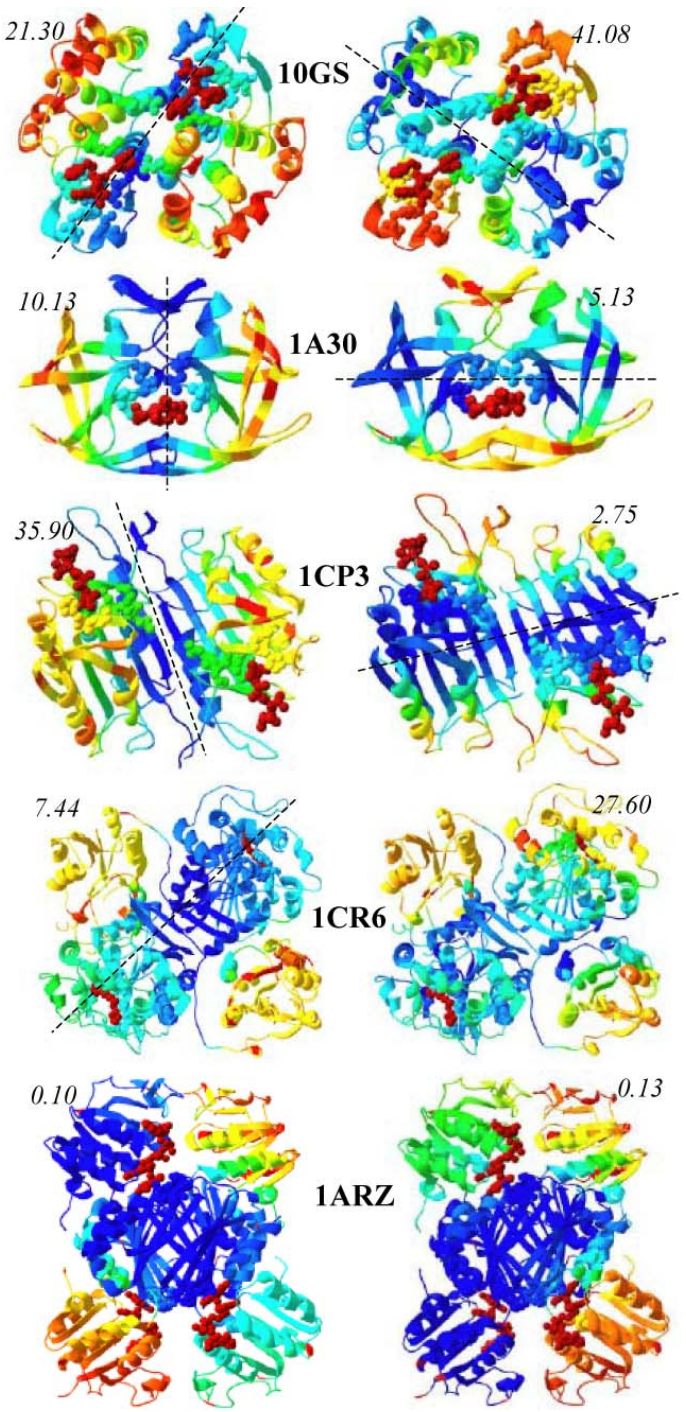

Figure 3-11 Global mode shapes of five multimeric enzymes included in our dataset.
The left and right ribbon diagrams refer to modes 1 and 2, respectively. Catalytic residues are shown by red ball representation. The dashed line indicates the loci of global hinge regions. The mobility scores corresponding to catalytic residues are indicated in each case.

3.4.5. Catalytic residues occupy or neighbor key mechanical sites

The correspondence between the loci of the catalytic residues and key mechanical sites, revealed upon comparison of the position of the residues that control the chemical activity (from experiments) and global dynamics (from computations), is a feature of fundamental functional importance that deserves further examination. The ligand-binding residues are found to exhibit a broad range of mobilities. The catalytic residues, on the other hand, are severely constrained. The averages over the 93 monomeric enzymes are $\langle M_1 \rangle_{cat} = 0.085$ and $\langle M_2 \rangle_{cat} = 0.066$ in the 1st and 2nd slowest modes, respectively, as opposed to $\langle M_1 \rangle_{all} = 0.236$ and $\langle M_2 \rangle_{all} = 0.154$ for all residues. From another perspective, 228, out of 325 catalytic residues included in our dataset serve as a key mechanical sites of type I, II, and III (with respective proportions of 107:82:39) upon considering the weighted average of the two slowest modes. Therefore more than 70% of the examined catalytic residues communicate with key mechanical sites, if not directly engaging in a mechanical role, and their ms fluctuations are, on the average, 2-3 times smaller than those of 'average' residues. This reveals a simple but important feature in the design of enzymes: *catalytic activity takes place at cooperatively constrained regions distinguished by suppressed fluctuations in the collective dynamics.* And a corollary is *to select inhibitor target sites from amongst the residues lying at the minima of the global mode shapes.*

One could hypothesize based on these observations that catalytic residues are immobilized in order to protect the delicate arrangement of functional groups. It is important to note, however, that the low mobility of catalytic residues is not a consequence of their being rigidly embedded in a given (functional) domain, but lying near crossover regions between substructures subject to oppositely correlated motions, as illustrated in Experimental Procedures for an example enzyme.

The catalytic residues are therefore localized in/near anchoring (hinge) regions that have limited, if any, translational mobility, while enjoying rotational flexibility to allow for the concerted rearrangements of the surrounding domains. This co-localization may be a requirement for effective coupling between chemical activity and conformational mechanics.

Examination of the catalytic residues (< 30% of the complete set) that do not conform to the general behavior of restricted fluctuations in the global modes shows that the computed mobilities of these particular residues may be attributed to specific requirements. For example, Tyr75 in 1BXO is located at the tip of a "flap" - a β-hairpin loop that forms the catalytic pocket against which inhibitors pack. The large fluctuations of Tyr75 are needed to accommodate the docking of big inhibitors such as PPi3 or PPi4. Another example in His64 in 1A42, which acts as a proton shuttle between zinc-bound solvent and bulk solvent, and switches between 'in' and 'out' conformations depending on the pH. The rotational motions of His64 ensure the occurrence of catalytic reaction at suitable pH.

From another perspective, our analysis shows that 86% (80/93) of the enzymes have at least one key mechanical residues of type II and III; and 94% (87/93) have at least one key mechanical residue of type I, II and III in their active site.

3.4.6. Enzymes are predisposed to couple their chemical and mechanical activities

Several studies have demonstrated that the global mode shapes are insensitive to structural details, but are uniquely defined by the overall 3-dimensional structure (see for example Kitao and Go, 1999). The global mode profiles may indeed be viewed as the signatures of particular

architectures. Consistent with this observation, the inclusion or exclusion of a few contacts are usually inconsequential, because the observed dynamics is a collective property of ~10^3 inter-residue contacts (for a protein of $N = 300$ residues, and an average coordination number of 7 per residue). The restricted motions at the catalytic sites are not due to the presence of substrates at those sites, but are intrinsic mechanical properties of the enzymes themselves irrespective of bound molecules. We also note that active sites are frequently in clefts, which may be functional in excluding water molecules and/or maximizing the contact surface at the ligand binding site.

Figure *3-12* shows the global modes obtained for liganded and unliganded forms of a protein (β-lactamase in panel *a*, and for two different ligand-bound forms of another protein (plasminogen, panel *b*. The close similarity of the two curves in each panel illustrates the robustness of the global modes. We note, however, that inhibitor binding may alter the mobility of a few key sites, while leaving the overall profile almost unchanged, and the maintenance of the fluctuation profile may indeed be critical for inducing or propagating functional motions. The similarity in the global dynamics of liganded and unliganded forms suggests that particular regions of proteins are already predisposed to serve as a catalytic center prior to substrate or inhibitor binding, and substrate binding essentially stabilizes the conformations, or induces the motions, that are intrinsically favored or accessible by the enzyme under native state conditions.

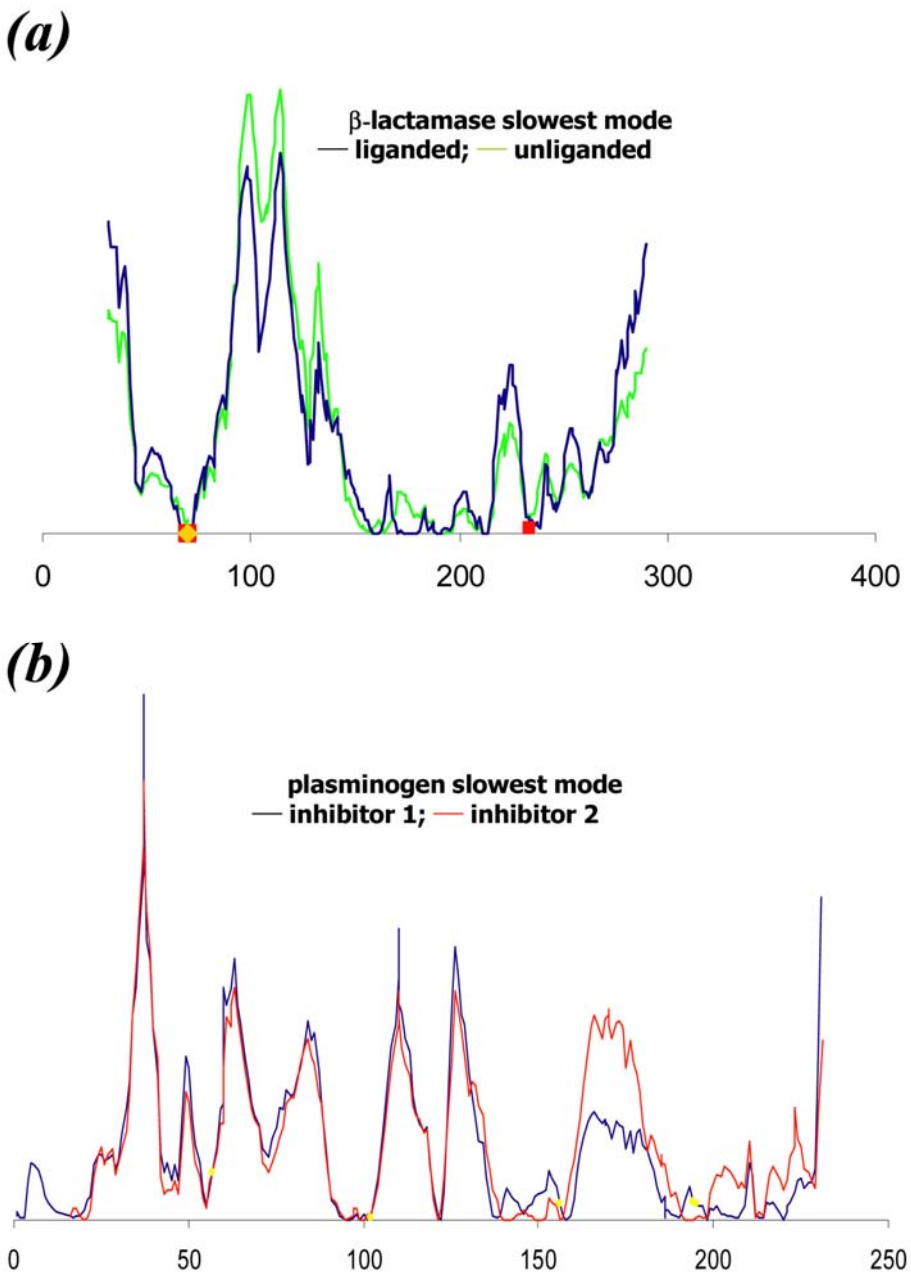

Figure 3-12 Comparison of the dynamics of the liganded and unliganded forms of two enzymes
(a) β-lactamase in liganded (1BLC, black) and unliganded (1BK9, green) forms. ***(b)*** plasminogen activator bound to different ligands. The black curve refers to the complex with inhibitor Glu-Gly-Arg chloromethyl ketone (1A5I), and the red to the complex with the inhibitor 2-(2-hydroxy-5-methoxy-phenyl)-1H-benzoimidazole-5-carboxamidine (1GI9). The critical mechanical sites (minima) are preserved in both forms, illustrating that the global dynamics retains the same qualitative features in the liganded and unliganded forms while the amplitudes of motions may be affected.

Table 3-3 Selected key mechanical residues (minima residues) and the odds ratio.

PDB	Catalytic residues [(a)]	Selected residuals that are of minima type I or II with $M_{i1,2} < 0.05$ [(b)]	OR [(c)]
10GS	7, 8, 13, 38, 44, **51**, 52, **64**, 65, 98	A,B: 48-51, 63-64, 66-67, 70-71, 74, 83-84, 86-88, 90-91, 94	8.3
1A16	260, 271 ,354, **361**, 383, 406	2-3, 170-173, 178-180, 358-359, 361-364	5.6
1A30	**A25**, A30, **B25**	A:4-7, 25-27, 48-49, 51-53, 98-99; B:1, 3, 5, 7-8, 25-27, 49-51, 87, 91, 95-97	4.9
1A3B	**57, 195**	16, 20-22, 27, 42-43, 54-57, 73, 90-91, 93, 95, 97-97, 97, 99, 101-103, 121-123, 139-140, 142-143, 151-152, 155-157, 179, 193-195, 197-198, 200, 208-209, 211-213, 228-229, 231, 233-235	4.8
1A42	64, **92, 94, 96, 119**	48-50, 52-53, 72, 74-75, 77-79, 84-85, 89-98, 100, 102, 104-106, 108, 112-114, 116-117, 119-121, 144-147, 188-190, 212-214, 244-245	4.5
1A47	**101, 141, 228, 230, 258**, 328, 329	1, 3, 5, 7, 9-11, 20-23, 26-27, 31-32, 34, 36-38, 44-45, 47-49, 51, 53-55, 58-59, 62-63, 77-82, 84-85, 87-88, 91, 96-97, 99-101, 103, 105-106, 108-110, 114-116, 119-121, 123-124, 127-128, 136-138, 141-144, 146-151, 153-154, 160-162, 169-170, 198-199, 212, 214-215, 218-219, 222-230, 232-233, 247-248, 253-255, 257-259, 269-270, 272, 274-276, 278-280, 315-317, 320-322, 377, 498-501, 503, 505-506, 517-520, 522-525, 533, 535-536, 541-546, 561-571, 575-576	3.0
1A5I	**57, 102, 156, 195**	1-4, 6, 16-20, 25-27, 29-31, 42-43, 45, 47-48, 54-58, 61-62, 68-74, 77, 79-80, 90-99, 101-103, 110-110, 110-110, 110, 115-117, 121-123, 132-133, 136-139, 142-143, 145-149, 151-152, 156-160, 165, 167-175, 177-178, 184-186, 186-186, 186-186, 186-186, 186-186, 186-186, 186-186, 186, 188-192, 194-195, 197-208, 212-216, 219-221, 221-223, 226-227, 229-230	2.9
1A5V	**64**, 121, **157**	64-66, 73-74, 76, 78-80, 86-87, 89, 91, 93-94, 99-101, 103-105, 108-109, 113, 115, 119-121, 125-126, 130-132, 134-136, 144-145, 147, 150-151, 153-154, 156-157, 192-193, 196-197	5.7
1AEC	**25**	7-8, 10-11, 13, 15, 17-19, 25, 29, 32-34, 36-37, 40, 113-115, 132, 134, 163, 165, 183-184, 213-215	7.5
1AL8	**24**, 108, **129, 257**	3, 12-14, 17-19, 21-24, 70, 74-77, 83-85, 87-88, 91-92, 94-96, 98-99, 101-103, 119, 122-125, 127-129, 155-157, 228-230, 236-237, 240-241, 243, 245-248, 254-255, 257-258, 265, 268-269, 271-272, 275-276, 279-281, 285-286, 307-309, 311-316, 318-321, 323-324	3.2
1ARZ	B-D:**159, 160, 163**	A:143-192, 204-234; B:128, 130-150, 154-221, 235-239; C:194-201, 227-229; D:3-137, 152-180, 183-213, 216-224, 232-273	2.5
1B3N	**163,398-401**	3, 5-8, 10, 13-15, 17-20, 22-23, 29, 31-34, 40, 42, 44-45, 47, 52-54, 76-77, 80-81, 83-85, 87-89, 95-96, 98-99, 102-103, 105, 107, 140, 142-143, 146-148, 154-155, 160-163, 165-166, 168, 170-172, 174-175, 177-179, 183, 185, 187-189, 191, 193-194, 196-199, 201-202, 204-205, 210, 212-216, 218-219, 232-234, 236, 238, 240-244, 246-249, 252-254, 260, 262-263, 268-270, 278, 282-283, 285-290, 300-301, 303, 305, 334-335, 337-338, 340-343, 345, 347-352, 354-355, 357-359, 363-364, 366, 383-384, 386-388, 396-401, 403, 405, 407, 409-410, 412	4.1
1B6A	**231**	110, 112-117, 119-120, 122-123, 125, 127-128, 168-169, 171-172, 174-176, 178-180, 182-184, 187-188, 190-193, 195-199, 201-210, 214-218, 221-224, 228, 230-233, 240-242, 246, 248, 251-255, 257-261, 267-268, 325, 327-328, 334-335, 337-339, 360-362, 364, 366, 368, 370-371, 375-377, 446-448, 455-457, 459-462, 472-473, 475-476, 478	3.3
1BGQ	40,44,**79,80,84,92,93** 98, **123, 124,171**,173	1-5, 12-14, 16-17, 19-20, 27-28, 31-32, 34-37, 52-59, 61-62, 64, 66, 69-70, 72-74, 78-80, 84-90, 92-96, 123-124, 126-127, 130-131, 133-134, 140, 143-151, 153, 156-160, 167, 170-171, 175-182, 184-186, 193-198, 201-202	1.5
1BH6	32, 64, **221**	3-5, 21-23, 26, 68, 71-73, 75-76, 78, 81-82, 84-85, 121, 123-124, 126, 130-132, 136, 139-140, 143-144, 146, 148-150, 171, 173-174, 205-207, 215-217, 221-222, 224-226, 228-229, 231-233, 235-236, 238-240, 243	1.6
1BVV	**69**, 78, 172	3-6, 18, 23-24, 26-27, 43-48, 58, 60, 68-70, 80-82, 89-91, 98-100, 102-105, 131-	1.1

		134, 146-151, 154-158, 165-168, 176-177, 183-185	
1BLC	**70**	68-71, 152-153, 155, 157, 159-162, 169-171, 173, 176-180, 184-186, 188-189, 191-192, 195-196, 208-209, 212-214, 216, 232, 234, 236-239	6.2
1BR6	**80**, 81, 121,123, **177, 180**	13, 17-19, 21-29, 31-32, 34, 47-49, 56, 77-80, 133-134, 136-140, 143-144, 172-180, 191-193, 195-199	2.9
1BIO	**57, 102, 195**	17-20, 26-30, 42-43, 45, 47-48, 54-55, 57-58, 60-62, 64-68, 71-73, 79-85, 90-91, 93-94, 98-103, 110-112, 115, 118, 123-124, 124, 132-133, 137-139, 142-147, 149-151, 157-159, 167-170, 170-172, 176-178, 187-191, 194-195, 197-199, 202, 207-209, 212-214, 217-218, 220-221, 224, 226-230	2.1
1BK9	**48, 52, 99**	5-6, 8-10, 12-14, 16, 21, 41, 44-45, 47-48, 50-54, 99-102, 104-105, 107-108, 110	5.0
1BXO	**33, 213**	1-3, 5, 7, 9-12, 14-15, 33-35, 125-126, 153-155, 158, 160-161, 163-164, 169-171, 179, 181, 185-188, 190-191, 213-216, 304, 311-313, 322-323	7.0
1CP3	**121, 122, 161**-165	A:36-37, 39-40, 48-50, 52, 55-60, 67-68, 71, 74-76, 78-79, 81-82, 84-85, 87-91, 95-97, 116-118, 120-122, 125-128, 130-131, 138-139, 158-161, 195-197, 200-201, 205, 210, 212-213, 215, 217, 219-220, 223-224, 227-228, 230-232, 265-267, 273-274, 276-277; B:35-37, 39-40, 43-44, 48-51, 53-54, 63-65, 67-68, 71-72, 74-76, 80-81, 85, 87-88, 93, 95-96, 111-112, 118-119, 121-122, 125-126, 130, 132-134, 159-160, 162-163, 195-197, 201-202, 205, 209-211, 215-217, 219-221, 223-224, 227-229, 231-232, 265-267, 273-274, 276	0.5
1CQQ	40, **71, 145, 147**	2-3, 6-7, 11, 62, 71-72, 86-88, 103-105, 107-108, 111-113, 144-145, 147-148, 154-155	5.4
1CR6	A, B: 333, 334, 465, 495, 523	A: 227-231, 233, 235, 237-241, 243-245, 247, 249-253, 256-258, 261-262, 269-270, 273-274, 276-287, 292, 294, 296-299, 314-315, 317-319, 321-323, 330, 440, 452-453, 455-456, 525-527, 530-531, 533-535, 537-544; B:226-227, 233-234, 236-237, 254-256, 276-278, 280-281, 283-286, 322	-

(a) The bold-faced residues in the 2nd column are the catalytic residues among those selected of minima type I or II and having mobility scores $M_{i1\text{-}2} \leq 0.05$.
(b) All the residues that are of minima type I or II and have mobility scores ≤ 0.05 in the two slowest mode.
(c) *Odds ratio*(**OR**) is the ratio p/p_0, which is the probability of finding a catalytic residue among minima, relative to that in all residues where p is the fraction of catalytic residues that fulfill the criteria $M_{i1\text{-}2} < 0.05$ and minima type I or II, p_0 is. the fraction of experimentally reported catalytic residues in all the residues of a given protein

3.4.7. Participation in key mechanical sites: a criterion for identifying functional sites

The present analysis suggests that the low mobilities in the global modes can be adopted as a new criterion for discriminating catalytic sites. The utility of this criterion may be assessed by a simple probabilistic analysis. Let us first sort all mechanically key residues of types I and II whose mobility score found from the weighted average of modes 1 and 2, $M_{i,1\text{-}2} < 0.05$. Let us consider the odds ratio p/p_0 of detecting a catalytic site among these key mechanical residues (p), compared to a random search over all residues (p_0). p_0 is simply the fraction of active residues in the examined enzyme. The ratio p/p_0 was computed for all enzymes in Set 1. The results are

listed in the last column of Table 3-3. p/p_0 is found to be 3.9 on the average, with a standard deviation of 2.1, which means the odds of having a catalytic residue among the key mechanical sites is about 4 times higher than the fraction of catalytic residues in the sequence.

3.4.8. Enzyme active site prediction

Our recent study supports the utility of considering global modes profiles in addition to physico-chemical features, for predicting catalytic sites (Chen et al., 2004), an issue that will be further pursued in the future. Here, an interesting question arises following the present study – do all the minima predicted by GNM enjoy the same functional significance in enzymes? The answer is surely negative given that there are 10-50 minima in the global mode of an enzyme (depending on the size of enzymes) while only an average 3.5 (324/93 in this study) catalytic residues in each enzyme. Assuming one minimum contains 1-2 catalytic residues, less than 4 minima out of 10-50 are the 'catalytic minima'. We would like to see how we can combine other structural information to discriminate the chemically functional minima before considering any chemical and phylogenetic properties of residues in the prediction algorithm. As pointed out by Gutteridge and Bartlett (Gutteridge et al., 2003), catalytic residues possess moderate RSA and are usually located in the 'clefts'. We also note that catalytic residues should function in groups, or say, be clustered in space. Hence, the rationale therein is to select minima that cluster in space, and those that exhibit moderate RSA and coordination number (CN).

A *de novo* algorithm, *COMPACT*, was developed to test this idea. Given coarse-grained and chemical-detail-free structures at hand, we exploited the idea that active sites, usually in the clefts, bear moderate solvent accessibility. This feature may be assessed by taking into account

the position of the residue in the R_i vs Z_i plot (Figure 3-8). A naïve scoring scheme is adopted based on (i) the degree of clustering with other minima and (ii) the extent of its Z_i departure from the average behavior (see detailed approaches in METHOD). The results are summarized in Table 3-4. 12 functionally distinct enzymes are selected from the CSA database for the analysis except for 1a5v and 1br6, the catalytic sites of which were acquired from literature survey. A good sensitivity and a low-to-moderate specificity due to many false positives, are observed. In all enzymes, certain catalytic residues are correctly predicted by *COMPACT*, i.e. they are the highest or second highest-ranking minima.

We are also interested in assessing if the minima predicted by *COMPACT* have any evolutional significance. We obtained the normalized conservation score for each residue using the ConSurf server (Glaser et al., 2003). The conservation scores of the correctly predicted residues are high, as expected. Interestingly, the false positives also turn out to be conserved (an average score 7.0 in comparison with the score 5.6 for average residues). This could suggest that the 'wrong predictions' may also bear some mechanical or structural significances even if they do not seem to have any obvious, functional importance. It remains to be further tested to see if the clustered FPs can serve as alternate drug binding sites.

One should note that, in our prediction, we have ignored the chemical properties of the sequence. All amino acids are treated as identical nodes connected by uniform spring regarless of their specificities. No empirical conservation properties or other specific propensities of residues, which appear to play a dominant role in Gutteridge's neural network, were used in the prediction. Of course, inclusion of those properties would narrow down the search space, and help pinpoint

the functional residues. However, besides the focus on predicting the functional residues, we are intrigued to see *why/when* some residues are functional and conserved (instead of knowing *what* residues are catalytic, while *dynamic* features are coming into the scope.

The results summarized in Table 3-4 were obtained with a Cα-Cα cutoff of 7.3Å. However, a recent test demonstrated that the use of a larger cutoff distance can further improve the predictions. Take the protein triosephosphate isomerase (PDB ID:1TPE) as an example, with the new cutoff 15Å. We notice that the results improve from {N11(**2**),K13(**x**),H95(**3**), E167(**6**),G173(**x**)} to {N11(**3**),K13(**3**),H95(**9**), E167(**2**),G173(**/**)}. Two new minima, K13 and G173 cannot be detected with the lower cutoff, appeared while K13 is highly ranked in the end. Another significance of this observation is the establishment of an alternate benchmark to validate/experiment the effects or these cutoffs on GNM results, which may be assessed by relating the performance directly to function in addition to the current comparisons of GNM predictions with experimental B-factors and NMR RMSD.

A larger enzyme set should be tested to further validate the applicability of *COMPACT*. Another interesting verification would be to apply *COMPACT* to clustered electron clouds in the electron density map of an X-ray determined enzyme with known catalytic residues and see if the algorithm is still valid when the type and connectivity of the residues are barely defined in the protein.

Table 3-4 Active site prediction of 12 functionally distinct enzymes using *COMPACT*

PDB	E.C.	Length	Active Site	NMP	Sensitivity	Specificity	Conservation Score (**TP+FP** / **FP** / **All**)
2alr	1.1.1.2	312	Y49(**12**),K79(**2**),D44(**1**),H112(**6**)	12	1	0.33	7 / 6.6 / 5.2
2jcw	1.15.1.1	153	H63(/),R143(**2**)	8	0.5	0.13	5.1 / 4.7 / 5.5
4thi	2.5.1.2	362	C113(**1**), E241(**5**)	8	1	0.25	N / A
1a5v*	2.7.7.49	146	D64(**2**), D121(**9**), E157(**5**)	10	1	0.20	5.5 / 4.6 / 5.5
1br6*	3.2.2.22	268	Y80(**2**), V81(**2**), G121(**8**), Y123(**8**), E177(**1**), R180(**3**)	11	1	0.45	7.7 / 7.2 / 5.1
1bvv	3.2.1.8	185	Y69(**1**), E78(**2**), E172(**5**)	10	1	0.30	6.7 / 6.1 / 5.5
1m4n	4.4.1.14	421	Y145(**12**), D230(**1**), K273(**6**), C400(/)	12	0.75	0.25	7.4 / 7.0 / 5.9
2plc	4.6.1.13	274	H45(**2**), D46(**2**), R84(**1**), H93(x), D278(**4**)	6	0.8	0.67	8 / 8.5 / 6.4
1tpe	5.3.1.1	249	N11(**2**),K13(x),H95(**3**), E167(**6**),G173(x)	9	0.6	0.33	7.2 / 7.3 / 5.2
3pgm	5.4.2.1	230	H8(**2**), R59(**8**), E86(**3**), H179(**1**)	8	1	0.50	8.4 / 8.0 / 5.3
1gim	6.3.4.4	431	D13(**4**), H41(**6**), Q224(**1**)	11	1	0.27	8.6 / 8.5 / 5.4
1a0i	6.5.1.1	332	K34(**1**)	12	1	0.08	8.2 / 8.1/ 6.7
						Avg.	**7.3 / 7.0 / 5.6**

The 4th column lists the experimentally determined catalytic residues obtained from CSA except for those of 1a5v and 1br6 (denoted by * in the 1st column) where the catalytic site information is taken from the literature. The bold-face numbers in parentheses indicate the rank of the minimum the catalytic residue belongs to, predicted by *COMPACT*. (/) refers to the catalytic residues that are correctly identified to lie in a minimum, but cannot be captured by COMPACT. (x) is the case where the residue is not in any of the minima. NMP is the number of minima predicted. Note that the two selected enzymes from each E.C. functionally differ by at least the third digit of their E.C. number. Sensitivity and specificity are defined as TP/(TP+FN) and TP/(TP+FP), respectively where TP, FN and FP are the numbers of true positive, false negative and false positive results, respectively. Conservation score is obtained from ConSurf server (Glaser et al., 2003) ranging from 1 (most phylogenetically diversed) to 9 (most conserved). The last column designates the average conservation scores over all the predictions (**TP+FP**), for the false positives (**FP**), and for all residues (**All**) in a given protein respectively.

4. DATABASE (iGNM) AND ONLINE CALCULATION SERVER (oGNM) FOR PROTEIN MOTIONS BASED ON GAUSSIAN NETWORK MODEL[§]

4.1. ABSTRACT

The 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.

Here, 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. In additional to the database, we provide computations for newly deposited structures in PDB or customized structures submitted by the users. The later part of the efforts is summarized in a new online calculation web-server (oGNM) that computes

[§]Reprinted, added and modified from *Bioinformatics*, (2005; **21**(13):2978-2987), with permission by OXFORD UNIVERSITY PRESS

GNM dynamics for proteins or DNA/RNA/protein complexes that comprise single or multiple chains with selectable cutoffs for the constitutive nodes. Comprehensive graphic tools and visualization engines are powered by Chime plug-in, Jmol and Java applets. Most importantly, a quick eigensolver, BLZPACK, and the PowerB algorithm for solving theoretical B-factors (B_{theo}) were implemented in oGNM serving as a tool for fast computations and high throughput analyses of collective motions in biocomplexes. The results have shown that the BLZPACK can generate 20 low frequency normal modes within 20 seconds (excluding the time for file upload) and the residue fluctuations of all modes within 6 minutes for a protein with 5808 residues. Our new engine is 10^4-10^5 faster for essential normal mode analysis and 10^2-10^3 faster for B_{theo} (sum of all modes) calculation than conventional subroutines. A case study here suggests a C^{α}-C^{α} cutoff in the range 14-18 Å, a distance that covers second coordination shell, achieves the highest correlation between B_{theo} and experimental B-factor (B_{exp}) for 6 protein/DNA complexes. Database *i*GNM and the online calculation web-server *o*GNM are available at http://ignm.ccbb.pitt.edu/ and http://ignm.ccbb.pitt.edu/GNM_Online_Calculation.htm respectively.

An online service is provided for newly deposited structures:

http://ignm.ccbb.pitt.edu/gnmwebserver/index2.html

4.2. 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., 1997a; Doruker et al., 2000; 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, 2003a; Ma, 2004).

Among the EN models of different complexities, the simplest is the Gaussian Network Model (GNM) (Bahar et al., 1997a; 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 Chapter 1). 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. 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).

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 small proteins having < 300 residues in view of the time cost of energy minimization.

The continuation of Gerstein's work is the use of a simplified NMA, an implementation from Hinsen's MMTK (The Molecular Modelling Toolkit) package (Hinsen, 2000), to interpret the evolution of biomolecular motions in the low-frequency modes, especially the lowest global mode, which best characterizes the motion undertaken by residues that enjoy the largest fluctuations (Alexandrov et al., 2005). A similar application, developed by Reuter et al. (Hollup et al., 2005), is WEBnm@ (a web application for normal mode analyses of proteins) that provides protein motions of the first 6 modes for detailed analysis and reports deformation energy for first 14 modes using MMTK.

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).

Molecular Vibrations Evaluation Server (MoViES; http://ang.cz3.nus.edu.sg/cgi-bin/prog/norm.pl), developed by Chen's group, distinguishes itself as the only available 'true' NMA server based on full-atomic AMBER force field, which derives thermal vibrations for proteins and DNA/RNA up to 4000 heavy atoms (Cao et al. 2004). The results can be obtained in 7 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 chapter, 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.

As for NMA-based online calculation web-servers, in general, give a short response time (except for MoViES) for small proteins (< 1000 residues) due to the efficiency of matrix diagonalization tool for Hessian. For instance, MMTK uses sparse-matrix eigenvalue solver from the ARPACK library (Hinsen, 2000). It provides a subset of mode vectors and saves the running time substantially. However, for large structures over 2000 residues, the computation takes hours to days if the process is not hung up in the middle. Moreover, hardly any of the servers provide computations for protein/DNA/RNA complexes such as ribosome, DNA polymerases or nucleosomes. Also, the computed B_{theo} is either missing or under-represented by the sum of a subset of modes (Suhre and Sanejouand, 2004a). The quality of the agreement of B_{theo} and B_{exp} is an important assessment for good characterization of molecular motions using NMA-based models. The calculation of B_{theo}, which is the sum of all the eigenvalue-weighted normal modes, is nevertheless a computationally expensive task given that all the modes have to be obtained.

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, with various size (Figure 4-1), deposited in the PDB prior to September 15, 2003. The goal of constructing the DB 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 Chime plug-in (PC users), Jmol (users of any the platform) for 3D visualization or Java applets for 2-D mobility plots. 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., 1998a; Rader and Bahar, 2004).

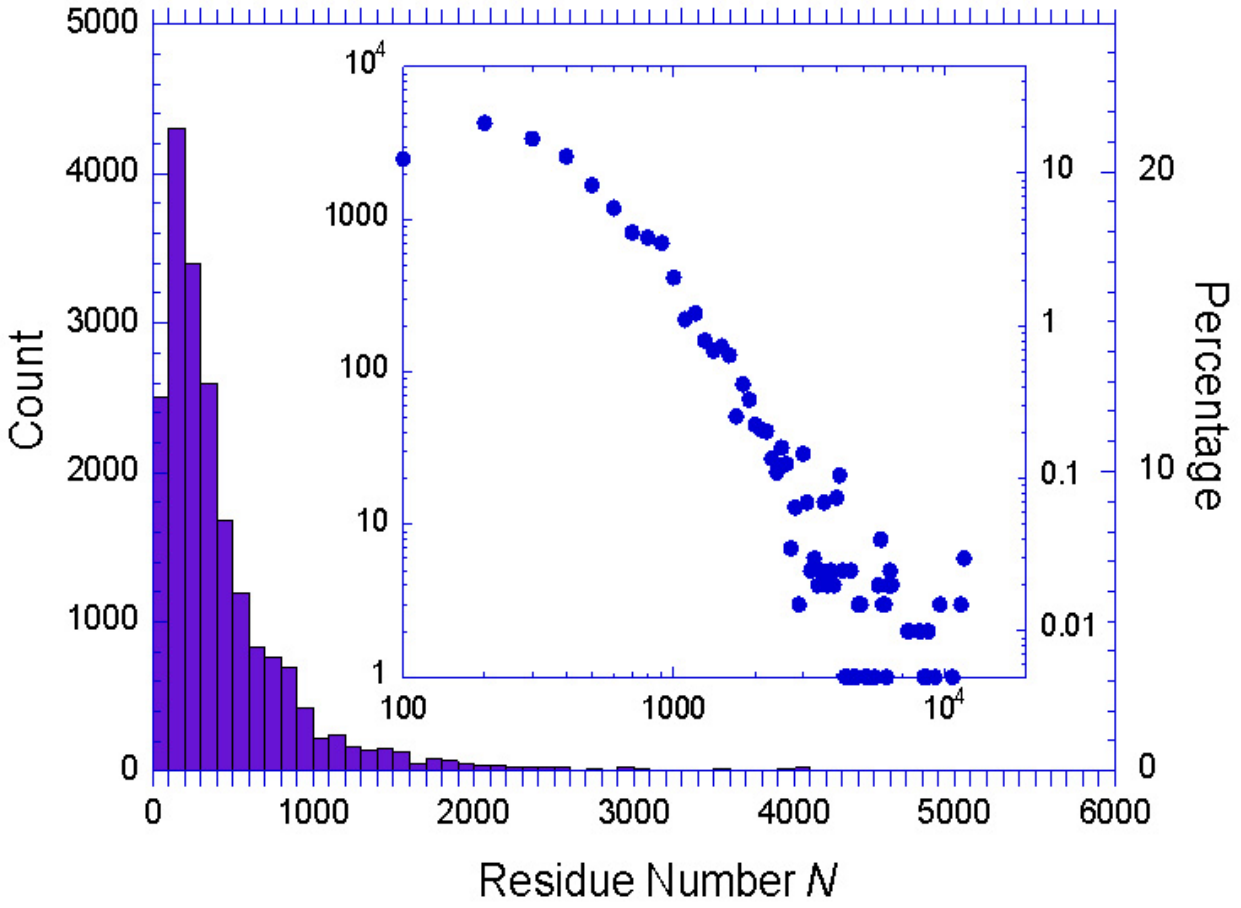

Figure 4-1 Distribution of the sizes of PDB structures compiled in the *i*GNM DB.
The number N of residues includes the number of amino acids contained in the examined PDB structures. 8.4% (1,701 out of 20,058) of the structures contain more than 10^3 residues. The inset displays the same distribution 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 $\Delta N = 50$.

On the other hand, to address the insufficiency of current computational tools, oGNM is developed to render fast computations for large protein/DNA/RNA complexes with the implementation of the BLZPACK package (Marques, 1995) based on the Lanczos algorithm, granting an efficient extraction of a subset of modes at either end of vibrational spectrum. A new algorithm, PowerB, based on power method (Mendelsohn, 1957), is proposed to expedite the computation for B_{theo} as well as spring constant by 10^2-10^3 folds compared to conventional

approaches. GNM results for 20 essential modes and summation of all modes (B_{theo}) can be obtained within seconds (excluding the file uploading time) and minutes respectively, for a protein of about 6,000 residues. This consequentially removes the need for the recruitment of any automated queuing system. Users can access the result of normal modes in text format as well as the multimedia presentations adopted in *i*GNM, allowing the browsing in an interactive fashion across the platforms. The results computed are purged monthly when allowing time for users' return to access.

The new engines is 10^4-10^5 and 10^2-10^3 faster for essential normal mode extraction and for B_{theo} (sum of all modes) calculation, respectively, than conventional engines. Here, we also present a case study to examine the agreement of B_{theo} and B_{exp} as a function of cutoff distances in the model for 6 proteins/DNA complexes. It is suggested that a C^{α}-C^{α} cutoff in the range 14-18 Å achieves the best agreement with experimental data.

*i*GNM is accessible at http://ignm.ccbb.pitt.edu/ and

oGNM is available at http://ignm.ccbb.pitt.edu/GNM_Online_Calculation.htm

4.3. METHOD

4.3.1. 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 or inaccurate coordinates deposited in the PDB. Figure 4-2 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 *4-1*.

4.3.2. The eigensolver

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.

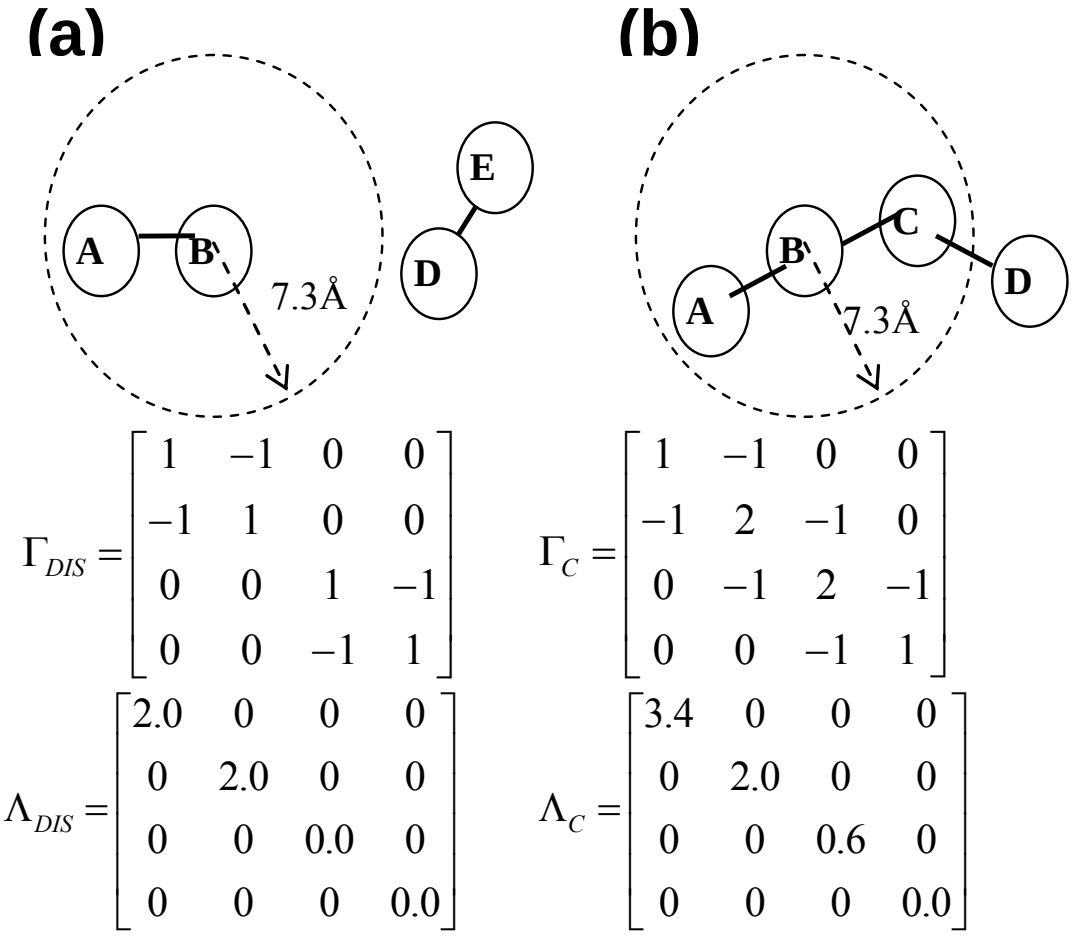

Figure 4-2 A schematic diagram to explain the cause of more than one eigenvalues.
Here, we illustrate how a discontinuity in the PDB sequence/coordinates may lead to more than one zero eigenvalue. In panel ***a***, the coordinates of residue C belonging to the A-B-C-D-E are missing. The distance between residues B and D is larger than the cutoff 7.3 Å, which leads to two independent blocks in the corresponding Kirchhoff matrix $\mathbf{\Gamma}_{\mathrm{DIS}}$ and more than one zero eigenvalue in the associated diagonal matrix of eigenvalues $\boldsymbol{\Lambda}_{\mathrm{DIS}}$. In contrast, the continuous tetrapeptide (no gap) in panel ***b*** accurately gives one zero eigenvalue, despite the possibly missing terminal residue E.

4.3.3. File parsing in oGNM

Any file created in PDB format, having size less than 10 MB, can be submitted to the oGNM website. The 'nodes' selected to construct the EN are the C^{α} atoms for amino acids, including

non-standard amino acid, and P atoms for nucleotides, including non-A, G, T, C, U nucleotides. C^{α}, P and C^{α}-P pairs in the network are considered to be coupled if they are located within a cutoff distance of 10 Å(r_c), 19 Å(r_p) and the average of r_c and r_p respectively. The values of 10 Å for r_c and 19 Å for r_p are chosen as default, yet can be interactively changed by the user. Currently, oGNM supports structures represented by ENs up 6200 nodes. For NMR structures, only the first model deposited in the PDB is used in the calculations.

4.4. RESULTS

4.4.1. Output files

Eleven output files can be accessed for each query structure (Fig 4-3a). 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 that can be accessed is presented below.

4.4.1.1. Contact topology (".ca" or ".nodes", ".cont", ".eigen" and ".kdat")

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 the last line of the file. In oGNM, this information is recorded in the ".nodes" file where the constitutive nodes include both the α-carbons and the nucleotide phosphor atoms. The number of nodes taken into the network along with the cutoffs used in the calculation is shown in the oGNM result page.

(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 4-3 The query enigines of *i*GNM

(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.

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 spatial 'discontinuity' that yielded more than one zero eigenvalues, which were captured in the corresponding '.eigen' files, will be re-submitted to an online service (http://ignm.ccbb.pitt.edu/gnmwebserver/index2.html) for GNM re-calculation with a larger cutoff. In oGNM, the users will be given warnings as the problematic structure being uploaded. In oGNM, since only a subset of modes is extracted by BLZPACK, only the eigenvalues of 21 lowest modes are listed. Also, the ".kdat" file in oGNM lists the non-zero contacts between residue i and j (i≠j) defined in the upper triangular Kirchoff matrix. One can see the contact elements are generally sparsely distributed, which rationalizes the employment of Lanczos algorithm.

4.4.1.2. Time average properties (".bfactor", ".cc", ".gamma" and ".corr" files)

The theoretical temperature factor ($B_{theo,i}$) predicted by the GNM is proportional to the inverse Kirchhoff matrix and also to the summation of all modes as

$$B_{theo,i} = (8\pi^2 k_B T/\gamma)\left[\Gamma^{-1}\right]_{ii} = (8\pi^2 k_B T/\gamma)\sum_{k=1}^{N-1}\lambda_k^{-1}\left[\boldsymbol{u}_k\right]_i\left[\boldsymbol{u}_k\right]_i \qquad (4\text{-}1)$$

Eq 4-1 follows from Eq 1-19 in Chapter 1 and the definition $B_{theo,i} = (8\pi^2/3) <(\Delta R_i)^2>$. The term $[\boldsymbol{u}_k]_i$ 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-*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) whose mobilities in the lowest mode are color-coded in panel ***a***. oGNM provides two more output files related to B-factors. A comparison of the quality of prediction – the correlation coefficient value of $B_{exp,i}$ and $B_{theo,i}$, is recorded in the ".corr" file. For example, the correlation coefficient value (R_{corr}) for this Dnaa/DNA complex (1J1V; Fujikawa et al., 2003), shown in Figure 4-4***e***, is 0.642 at cutoffs of 10 Å and 19 Å for Cα-Cα and P-P contacts respectively with an uniformed spring constant of $0.217 \times k_B T$ (kcal/mol.Å^2), derived from the equation:

$$\gamma = 8\pi^2 \frac{\sum_{i=1}^{n} \left[\Gamma^{-1}\right]_{ii}}{\sum_{i=1}^{n} B_{exp,i}} (k_B T) \qquad (4\text{-}2)$$

This value is reported in the ".gamma" file.

The predicted cross-correlations $<\Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j>$ between the fluctuations of residues *i* and *j* are listed in the '.cc' files. These are reported for small size proteins (N≤290 for iGNM; N≤500 for oGNM) due to the memory constraints. The data in these files are used to construct the color-coded correlation maps (called CCplot) (Figure *4-4****d***). $<\Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j>$ values are normalized between -1 and 1, by dividing them by $[<(\Delta R_i)^2><(\Delta R_j)^2>]^{1/2}$. 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).

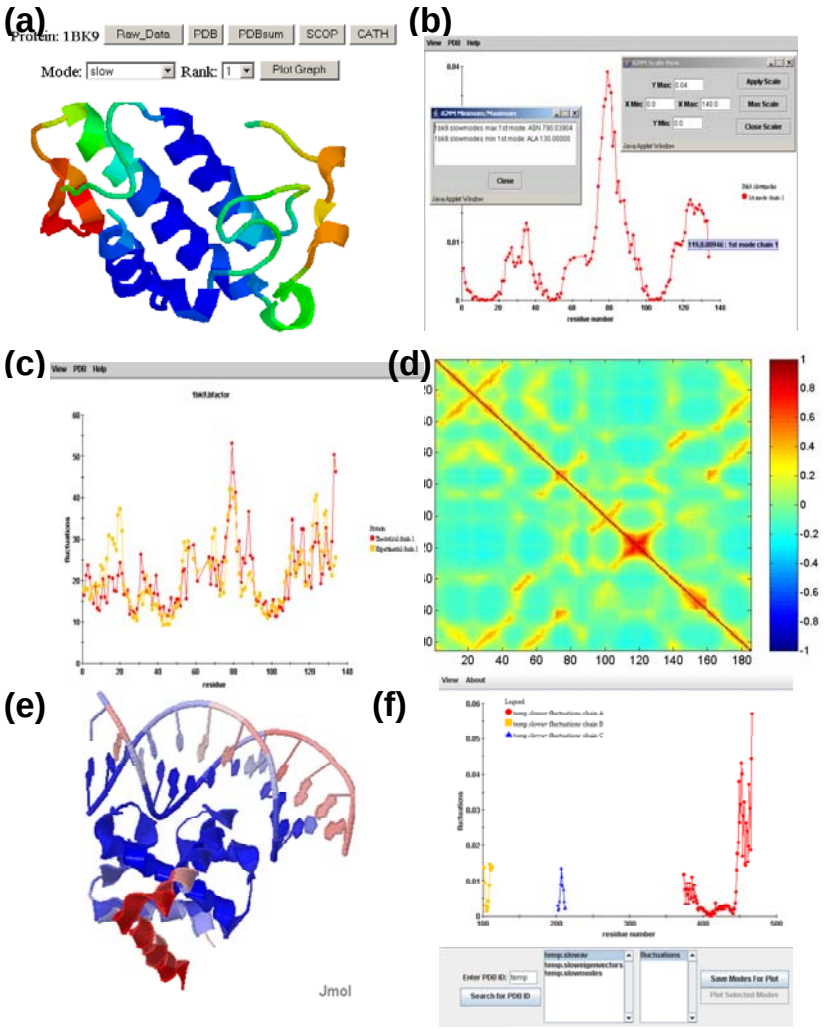

Figure 4-4 Visualization of GNM dynamics for phospholipase A2 (PDB ID: 1BK9).
(a) Color-coded ribbon diagram (Chime) that illustrates the mobilities in the slowest GNM mode (slow1). The structure is colored from dark blue, green, orange to red in the order of increasing mobility in the slow mode ***(b)*** 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 $\langle \Delta\mathbf{R}_i \bullet \Delta\mathbf{R}_j \rangle$ 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. The visualization of GNM dynamics for Domain IV of Chromosomal Replication Initiator Protein Dnaa Complexed with Dnaabox DNA (PDB ID**:** 1J1V) is shown in (e) and (f). ***(e)*** Color-coded ribbon diagram in *Jmol,* that illustrates the mobilities $[(\Delta R_i)^2]_{1\text{-}2}$ induced in the slowest two modes. The structure is colored from blue, white to red in the order of increasing mobility. ***(f)*** Mobilities $[(\Delta R_i)^2]_{1\text{-}2}$ of 1J1V, plotted against residue index. One protein chain (chain A) and DNA chains (chain B and C) are well separated on the plot. Note that the residue index of mobility plots such as *(b)* and *(e)* reflects the 'real' reported sequential numbers in the PDB file where the discontinuities in sequence numbers are commonly seen. However, the axes values in the cross-correlation map, shown in *(d),* give serial numbers for residues, which are created only for the plotting purposes.

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

The shapes of the 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 on mobilities can be found in the ".slowav" file. The entries therein refer to the weighted average

$$[(\Delta R_i)^2]_{1\text{-}2} = (\lambda_1^{-1} + \lambda_2^{-1})^{-1} (\lambda_1^{-1} [\boldsymbol{u}_1]_i [\boldsymbol{u}_1]_i + \lambda_2^{-1} [\boldsymbol{u}_2]_i [\boldsymbol{u}_2]_i) \qquad (4\text{-}3)$$

4.4.1.4. Global hinge residues

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.

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

These files are currently unavailable in oGNM. In *i*GNM, 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).

4.4.2. Query and Visualization

*i*GNM allows users to conveniently query and visualize GNM output files. All the search start with a single entry : http://ignm.ccbb.pitt.edu/FileDownload.htm. Upon the PDB ID is submitted, people can access the output files instantly. The results can be viewed, saved or re-directed to 3D Visualization Modules for normal modes or B-factors (B_{theo} and B_{exp}). These modules provide structure ribbon diagram that is colored according to the residue (or the nodes) mobilities of a single normal mode or of all modes (B-factors). The mobility plots can then be accessed from the buttons anchored in the top of these diagrams. Queries for accessing 3D Visualization Modules for normal modes and B-factors are also provided at http://ignm.ccbb.pitt.edu/3D_GNM.htm and http://ignm.ccbb.pitt.edu/BFactors.htm respectively.

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 4-3b).

Two major visualization engines are implemented in *i*GNM DB and oGNM are (i) the ribbon diagrams that are color-coded according to residue mobilities, and (ii) the mobility plots.

Ribbon diagrams are visualized with Jmol (http://jmol.sourceforge.net/), an open source molecule viewer written in Java. JmolApplet (http://jmol.sourceforge.net/applet/) is Jmol's Web browser applet version that can be integrated into web pages. Since JmolApplet is a cross-platform and runs with Java Virtual Machine (JVM) 1.1 which is included in most popular browsers, it is deployed easily without additional software installation by the end users. JmolApplet adopts Chime/Rasmol scripting language and allows users to manipulate color-coded structures in a way similar to the Chime plug-in (MDL Information Systems, Inc. www.mdlchime.com). Chime plug-in (for PC users only) and Jmol (cross-platform)are currently available in oGNM as shown in the illustrative Figures 4-4 panels ***a*** and ***e*** respectively. Note that Chime gives a broader color spectrum, while Jmol renders better representations for nucleotide-containing structures.

The interactive mobility plot viewer grants the user the ability to visually inspect the mode fluctuations of GNM outputs. The viewer is constructed as a Java applet using the Java Virtual Machine 1.5.0_02 and requires a Java Runtime Environment (Sun Microsystems, Inc. http://java.sun.com) of version 1.4.1 at a minimum. This applet is invoked from HTML tags with parameters including the data source, the desired PDB ID, the desired GNM output and desired normal modes from that output. The graphing ability is rendered through a product entitled JClass from Quest Technologies (http://www.quest.com/jclass/). The interface's controls are created from Java swing classes and provide a means to interact with the graph. The applet can also be run as a stand-alone application outside of a web page but still requires an Internet connection to connect to the data source. The current features allow users to load selected modes

of computed assemblies for comparison, scale the view of the plot and view the raw GNM output associated the selected modes. A demonstration is shown in Fig 4-4f.

Links to the raw *i*GNM data, PDB, PDBsum, SCOP and CATH are also included in *i*GNM for access by the users.

4.4.3. Online calculations

Currently (May 17, 2005), the PDB contains 30,963 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 the GNM files corresponding to this structure. If the GNM results are found, the results are displayed to the user through the Visualization Engine. For those PDB structures that are not included in *i*GNM, an online service (http://ignm.ccbb.pitt.edu/gnmwebserver/index2.html) is provided for automated file retrieving and subsequent GNM calculations (Liu et al., 2004). oGNM can also be used for the same purpose if the PDB file of the structure is available.

4.4.3.1. Database architecture

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 (Figure 4-5). 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.

4.4.3.2. The eigensolver - BLZPACK

BLZPACK was implemented for 20 lowest normal mode calculation in oGNM. The original implementation of GNM utilized the singular value decomposition (SVD) routine from Numerical Recipes (Press et al., 1992). Although sufficiently accurate and robust for small protein structures, this method became prohibitive for large protein structures because its computational time scales as $O(N^3)$ where N is the number of residues. Comparisons between SVD and several alternative eigenvalue solvers are shown in Figure 4-6a.

Two important aspects regarding the data are not considered by these complete eigensolvers. First, since the slow modes have been shown to correlate with functional motions, typically only a small fraction of the entire spectrum of eigenvalues and eigenvectors are of interest. Second, the Kirchhoff matrix is very sparse. We have taken advantage of these two characteristics by using a blocked Lanczos algorithm as implemented in the BLZPACK software (Marques, 1995). The lowest two curves in Figure 4-6 indicate an enormous speedup over previous methods for large structures. This allows for efficient online calculation (approximately 3 minutes for protein complexes with 6200 residues).

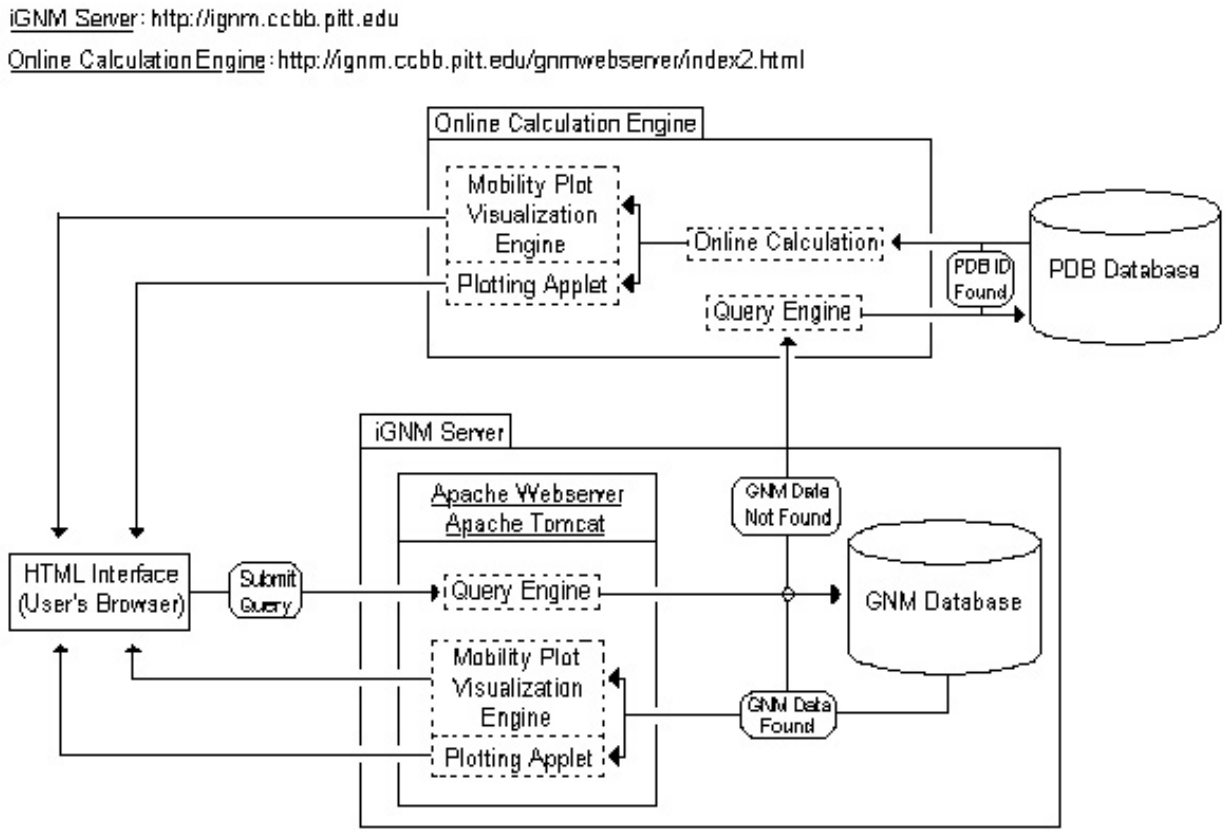

Figure 4-5 *i*GNM architecture.

*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 on the user's browser window. If they are not found, the user is offered the option to use an online service (http://ignm.ccbb.pitt.edu/gnmwebserver/index2.html) that invokes the search of PDB for the structure file, retrieves it to the server, and runs GNM calculations for it. 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 data with PDB.

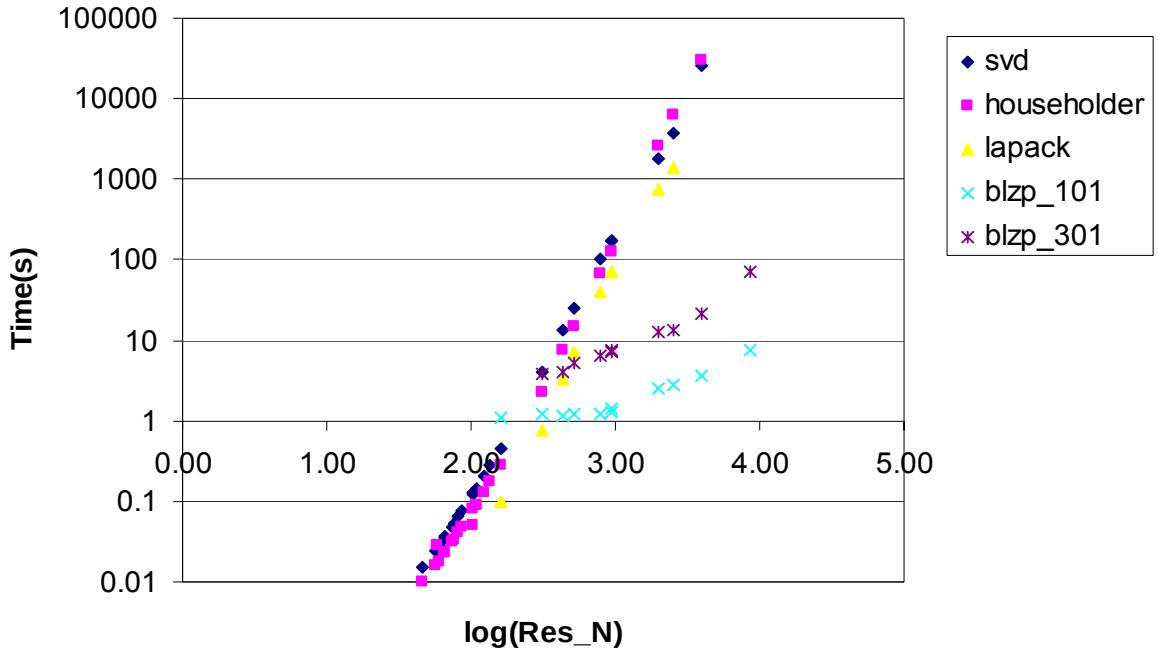

Figure 4-6 Improvement in computing time for calculating the slow modes by BLZPACK.
The execution times for several eigensolvers are shown by the different symbols (labeled on the right). These curves reflect the fact that the complete eigensolution using any of the top three methods, singular value decomposition (SVD), householder, or the dsyev subroutine of LAPACK), scales as $O(N^3)$. Exploiting both the sparsity of the Kirchhoff matrix and the utility of extracting the slowest modes as a first approach, allows for a significant decrease in calculation time for large structures. The final two methods, blzp_101 and blzp_301, return the slowest 101 and 301 modes respectively using the BLZPACK routine. Both of these methods are much faster, scaling rather with O(N).

4.4.3.3. B_{theo} computation

PowerB (described in Chapter 2) is encoded and implemented in oGNM. The B_{theo} calculation is computed optionally upon users' demand. It can only be started after the sparse form of Kirchoff matrix is computed and stored. The related output files such as the spring constant of the network, R_{corr} of theoretical and experimental B-factors and CCplot are shown soon after the B-factor

calculation is done. Currently, it takes around 10 minutes to compute the B_{theo} for the largest protein (6200 nodes) allowed in oGNM.

4.4.3.4. Prediction of B_{theo} for six protein/DNA complexes

Six protein/DNA complexes belonging to distinctive CATH topology groups were selected from the PDB to investigate how R_{corr} of B_{theo} and B_{exp} change with r_c as r_p is fixed at 19 Å (Table 4-1). As shown in Figure 4-7, the complexes have an optimal cutoff between 12-18 Å at different X-ray diffraction temperatures. Since the number of amino acids way outnumbers that of the nucleotides, we expect a similar optimal cutoff to be observed for protein/DNA complexes when the number of amino acids dominates the entire structure. We realize that a larger number of DNA/protein complexes should be investigated before any solid conclusions can be drawn. The analysis demonstrates, however, that such studies can be carried out using the oGNM.

Table 4-1 Attributes of Six Protein/DNA complexes for R_{corr} vs. cutoff distance study

PDB ID	CATH	Nodes	Protein/**DNA**	Resolution(Å)	XDT (K)
1J1V	3.40.50.2000	118	A:94,**B**:12,**C**:12	2.1	100
1QTM	A1:3.30.420.10 A2:1.20.1060.10 A3:3.30.70.370 A4:1.10.150.20	563	A:539,**B**:11,**C**:13	2.3	100
1AAY	A1,A2,A3: 3.30.160.60	105	A:85,**B**:10,**C**:10	1.6	295
1AOI	A0,B0,C0,D0,E0,F0, G0,H0: 1.10.20.10	1095	A:98,B:83,C:115,D:99 , E:116,F:87,G:108, H:99, **I**:145,**J**:145	2.8	110
1ECR	A1:3.50.14.10 A2:3.30.54.10	333	A:305,**B**:14,**C**:14	2.7	288
1OTC	A1,A2,A3: 2.40.50.140 B0:2.40.200.10	683	A:459,B:213, **D**:11	2.8	288

CATH number describes the structural characteristics of the domains in the 6 complexes. Note that the domains list here differ from each other at the level of structural sub-classes 'T'. 'Protein/**DNA**' lists the number of residues in the protein and DNA chains, the DNA chains being written in bold face, followed by the number of nodes in the chain. XDT is the temperature at which the X-ray diffraction is measured.

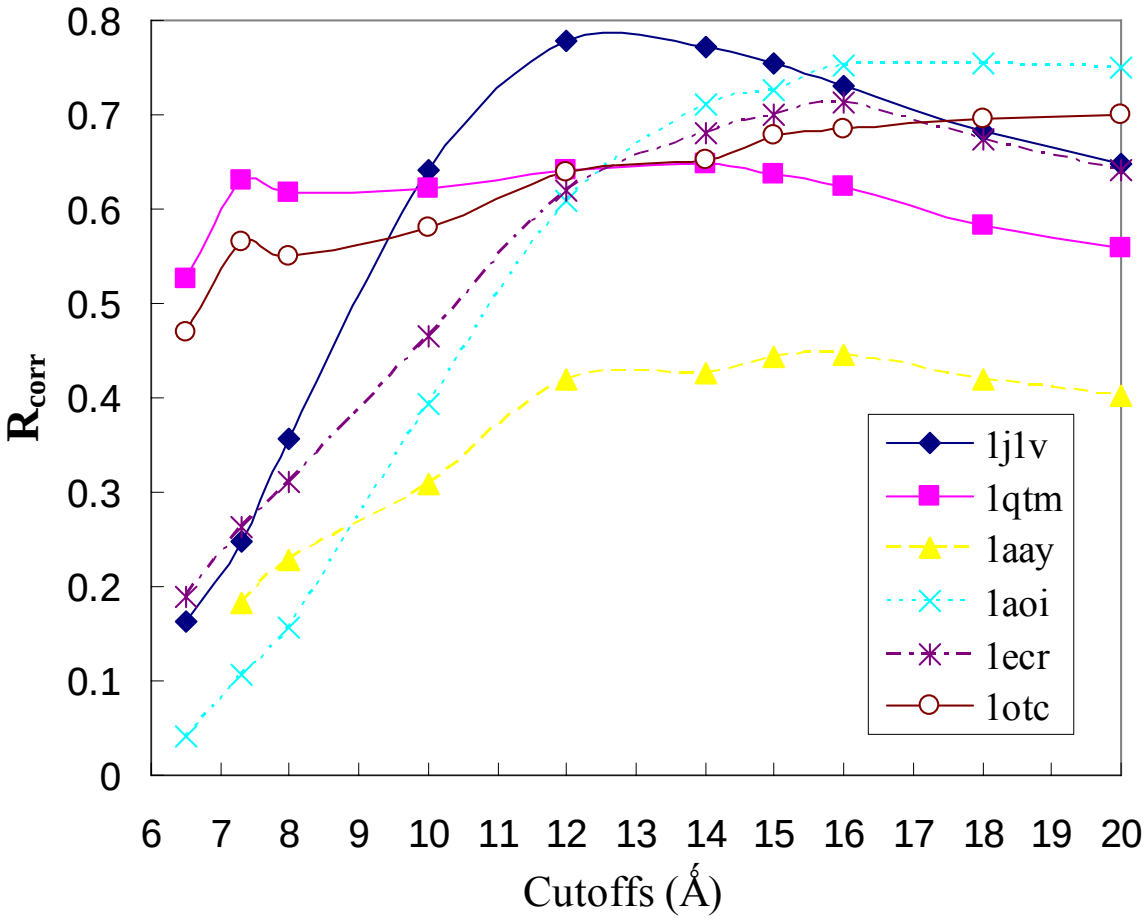

Figure 4-7 Correlation coefficient (R_{corr}) between B_{theo} and B_{theo} as a function of cutoff distance for six protein/DNA complexes.
The attributes of the six complexes are listed in Table 4-1.

4.5. DISCUSSION

We generated information on the equilibrium dynamics of 20,058 structures in the reach of covering the entire PDB. The variation of the spring constant γ and the correlation R_{corr} between B_{theo} and B_{exp} as a function of cutoff distances can also be examined in oGNM for various

biocomplexes as shown in the above illustrative example for 6 protein/DNA complexes. The utility of such a protein dynamics DB and computing tool is to give us insights about the relation between the structure, functions and collective motions of proteins.

The current file-parsing strategy in oGNM is to include all the C^{α}s and Ps in the EN. This includes short peptide or nucleotide inhibitors. Users can manually delete the coordinates of those in the uploaded file if those are deemed insignificant to the overall dynamics. The default r_c, 10 Å, for C^{α}-C^{α} is set to cover wider spatial discontinuity in structures due to the incomplete or inaccurate report of atom positions defined by X-ray diffraction. This enlarged cutoff, compared with the cutoff 7.3 Å in *i*GNM, is expected to generate less eigenvalue- problematic dynamics results. A default r_p, 19 Å, for P-P is proposed to cover the distance, across DNA strands, of the P atoms in two based-paired nucleotides.

The size of the uploaded PDB file currently determines the responding time of oGNM website for mode calculation. The real time taken for mode computation is trivial for complexes with a size > 6,000 nodes. The B-factors calculation is initiated upon users' request and completed within 10 minutes for the largest structure allowed on the server.

The eigenvalue decomposition of the connectivity matrix $\mathbf{\Gamma}$ 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 for *i*GNM, 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 increase up to 15 days in the case of the largest structure (11730 residues), 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, a much faster algorithm has been implemented in oGNM. The BLZPACK software (Marques, 1995) based on Block Lanczos Method for large structures is used to evaluate a subset ($1 \le k \le 20$) of dominant (slowest) modes, within a time scale of *N* (Figure 4-6a), i.e. the computing times is more than 6 orders of magnitude shorter than the subroutine SVD, when structures of $>10^3$ residues are analyzed. The same algorithm will be particularly useful for generating the ANM (anisotropic network model) (Doruker et al., 2000; Atilgan et al., 2001) data that we plan to incorporate in the near future in the *i*GNM DB.

In theory, PowerB can be applied to any type of Hessians. The extension of this method to ANM or NMA is currently under study. One can subtract 6 zero eigenmodes from the pseudo-inversed Hessian subject to small perturbation in order to obtain the summation of all modes. The future implementation includes computing fast modes with power method. The convergence rate of power method is highly related to the distribution of eigenvalues. In general, Γ gives a sparse distribution in lower modes, contributing a higher $\left|\frac{\lambda_k}{\lambda_{k-1}}\right|$ value and shorter computation time than that in higher modes. Currently, the bottleneck in PowerB is the matrix inversion process. A faster routine is currently being tested to boost up the speed.

Iteration loops k = 1000, threshold t = 0.01 are good enough for an instant extraction of the zero mode. The convergence reaches as $\left|\frac{\boldsymbol{A}^{k+1}\boldsymbol{X}_o}{\left|\boldsymbol{A}^{k+1}\boldsymbol{X}_o\right|} - \frac{\boldsymbol{A}^{k}\boldsymbol{X}_o}{\left|\boldsymbol{A}^{k}\boldsymbol{X}_o\right|}\right| \le |t|$.

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, 1998b; Bahar et al., 1999b; 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 have been adopted in these studies 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. However, given that in some PDB files, only the phosphorus atoms in nucleotides are provided, we adopted single node representation for the nucleotides in DNA and RNA with an enlarged cutoff distance to 19 Å. This by no means implies that 1-node-representation outperforms, in terms of the agreement with B_{exp} or NMR results, the 2- or 3-node representations. More detailed studies are needed to reach conclusive results. Yet, oGNM here provides a comprehensive and physically tangible model for researchers to explore the collective dynamics of complexes of interest. On the other hand, one should note that the current *i*GNM DB does *not* contain the results for such complexes or assemblies containing RNA/DNA components. An updated version of iGNM DB that incorporates the DNA/RNA/protein complexes plus biological units (see the next paragraph) is in progress.

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 the PDB, and (ii) the method is applicable to fluctuations near the native structure. 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 conditions. 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 (the physiologically functional structures) 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 (< 30 nodes) may not be always justifiable. First, small structures are amenable to analysis using more detailed full atomic models that take account of their specific interactions. Second, the Gaussian approximation for residue fluctuations becomes more accurate with increasing size of the network, as follows from the 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.

5. CONCLUSION AND FUTURE WORK

5.1. ASSESSMENT OF THE ELASTIC NETWORK MODELS

In this dissertation, we showed that the Gaussian Network Model (GNM) agrees with crystallographic results slightly better than Normal Mode Analysis (NMA) does. This does not necessarily imply that the introduction of residue specificities into simplified EN-models is unnecessary. The findings redirected our thoughts to two facets. First, we need multidimensional experimental results to assess and confirm our predictions. These benchmarks should directly correlate with the protein function. For example, we would like to know how accurate the statement 'catalytic sites are distinguished by their low mobilities in the global modes of motion, compared to their neighbors along the sequence' is, when we resort to NMA instead of GNM to examine this phenomenon. This observation should hold true across all the dynamics models, probably more so for NMA given its 'detailed' nature at atomic scale. Second, more importantly, we have to understand why GNM performs better than NMA, despite its simplicity and underlying isotropic deformations assumptions. We currently invite attention to the difference between the potential functions implicitly adopted in the GNM and ANM. An intuitive thinking would lead to the proposal of creating a modified ANM model that adopts a potential that penalizes the orientational deformation of residue pairs, a topic that can be explored in future work.

Since the GNM considers no solvent damping effect nor are other constrains besides topological features of the molecule imposed, we expect our predictions to match experimental data better as the proteins move in a less restricted environments. It is indeed encouraging to observe that the GNM yields better agreement with NMR MSD than those with the X-ray temperature factors, and also that the GNM performs better when applied to structural diffraction data collected at higher XDT. To summarize the results, GNM appears to give better agreement with experimental data if the measurements are taken at higher temperature or in solution.

Although the GNM yields the same level of agreement with experimental data over a wide range of cutoff distances, R_c = 15Å could be a safer choice to employ in order to cover the possible spatial discontinuities in structure originated from unresolved atom coordinates, to eliminate any overestimation in the mobilities of hanging tails/ends. Other situations where a higher cutoff distance may be more appropriate are the 'stretched' structures observed under intensive crystal constraints at low temperature, or the loop that stick into a hollow catalytic pocket as illustrated for penicillopepsin (1BXO).

In assessing the performance of different harmonic models or the effects of parameters in a model, the comparison of the predictions with X-ray crystallographic B-factors seems straightforward and considers the physical properties in a good approximation. In fact, B-factor reports the uncertainty in the atom positions as a result of two effects – static and dynamic. Static effects come from the fact that some subset of atoms can orient themselves in more than one energy well bearing a similar depth. These atoms therefore distribute themselves in different spatial positions, corresponding to the different energy wells, with a similar possibility. Dynamic

effect, on the other hand, is what we analyze in this thesis, originating from the thermal fluctuations of atoms about their local energy minima.

However, the B-factors do not distinguish these two effects. Structures of high resolution are often accompanied with low atomic B-factors. It is not very clear that the decrease in B-factors results from the decrease in purely static effects or both effects. The crystallographic B-factors are not therefore the most accurate experimental measures of fluctuation behavior. In this respect, NMR measurements may provide a better measure of dynamics, not limited by crystal contacts or experimental temperatures. This feature is further confirmed by the fact that GNM predictions show a better agreement with NMR measurements than X-ray data. In this thesis, we have not conducted a systematic comparative study on the agreement levels of fluctuations revealed by X-ray B-factors and by NMR MSD between models. Such a future assessment should be valuable to clarify the suitability of different experimental approaches for benchmarking theoretical predictions.

In a systematic analysis of 235 proteins structurally characterized at different resolutions (from 1 to 2.4 Å) at XDT ≥ 273 K, a temperature considered to allow for 'vivid' protein dynamics and thereby decreased static effects on B-factors, the R_{corr} values show a very minor dependence on crystal resolution with a correlation coefficient factor less than 0.15. A regression line of a negative slope -0.086 is observed in the R_{corr} - Resolution plot, showing a very weak increase in R_{corr} as the resolution improves. An even smaller correlation of 0.12 between R_{corr} and crystal resolution has been observed for 831 proteins under XDT at 100K.

One should note that although the observed absolute residue fluctuations vary with the resolutions of a given structure, their relative fluctuation profiles (as a function of residue index) remain consistent. Moreover, the RMSD between two structures at different resolutions for a given protein are usually very small, which leads to the same B_{theo} profile. Hence, in terms of correlation with theoretical predictions, the influence of crystal resolutions on the R_{corr} is trivial, at least for a certain range of resolutions, say, 1 – 3 Å.

5.2. ACTIVE SITE PREDICTION USING *COMPACT* AND NEURAL NETWORK ALGORITHMS

Within the scope of this thesis, the potential utility of the GNM for elucidating structure-dynamics-function relations in enzymes of different EC classes has been explored systematically. It was found that catalytic residues have highly restricted mobilities in the global modes. Also, catalytic residues are subject to more restricted mobilities than their four flanking neighbors (two on each side) along the sequence, and this feature holds for more than 70% of the examined catalytic residues. A close interplay between chemical activity and molecular mechanics is suggested by these findings.

A new software, *COMPACT*, has been developed within the scope of this thesis. *COMPACT* takes advantage of the lower mobility concept observed for the active sites, and is found to observe amazing results. For the first time, we can 'roughly' predict the position of the active sites on a given enzyme structure without knowing the amino acid sequence. The results are observed to show high sensitivity, but low specificity. The low specificity (TP/(TP+FP)) of the

prediction should be overcome by employing more sequence and structure information, especially the knowledge of residue conservation, which has been observed to be most effective in previous algorithms (Gutterige et al., 2003). Our findings suggest that the active sites should be located at regions where dynamics requirements are met. The specific function is apparently achieved with the right spatial arrangement of atoms and local composition of amino acids, the specific function can be conducted.

We are currently in the process of incorporating the dynamics information along with other sequence/ structural features into a neural network algorithm for active site predictions. The preliminary results are encouraging and yield a high accuracy rate that is deemed the best amongst reported algorithms. The idea of *COMPACT* can be further utilized and validated in low-resolution structures such as those from cryo-EM. Inspired by Ma's work (Ming et al., 2002a, b), we are planning to predict the catalytic loci of low-resolution enzymes from the fuzzy electron density map provided by EM or X-ray. The 'nodes', connected by the springs, can be created by clustering and quantizing the electron density clouds. This idea will be pursued and tested in the near future.

5.3. DEVELOPMENT OF *i*GNM

GNM, as the simplest EN model, provides a means of performing a high-throughput analysis of protein dynamics. With the deposition of ever growing large number of structures in the PDB, as well as the determination of increasingly larger supramolecular structures by experimentalists, GNM demonstrate a unique ability to compute and generate dynamics information in a database

framework, *i*GNM. The database *i*GNM has been a pioneering effort to address this need; it was compiled and interfaced to allow biologists to access the dynamics of their molecules of interest with minimal computational hustle. Online calculation web servers, such as the current oGNM, allow us to upload and obtain the dynamics of proteins or protein/DNA/RNA complexes that belong to (1) part of the biological units, (2) a combined complex comprising two PDB files, or (3) the structure derived from homology modeling. We describe three recent applications assisted by the stored and computed GNM dynamics.

PolQ, a new sequence identified by Professor Richard Wood and coworkers at the University of Pittsburgh Cancer Institute has been shown to have DNA polymerase activities (Seki et al., 2004). This polymerase, however, was found to have low-fidelity in DNA replication, which facilitated the bypass of DNA damages, despite its evolutionary similarity to the high-fidelity polymerase family, Pol I, observed in the sequence alignment. It bypasses damaged (AP) sites, inserting a nucleotide A with 22% of the efficiency of a normal template, and then continuing the extension as avidly as with a normally-paired base. To gain insight into the mechanism by which PolQ bypasses DNA damage, we carried a homology modeling study using Pol I members as templates. A converged model verified by two potentials was obtained for PolQ. Insert 1, an insertion of 22 residues between the conserved motifs in its catalytic pocket, was found to insert into the tip of the polymerase thumb subdomain and dock into the minor groove of the double helix DNA. GNM results for PolQ were retrieved from the calculation engine currently implemented in oGNM. It was found that in the global modes this highly positively charged stretch, Insert 1, would exhibit outstanding mobilities. The considerable flexibility conferred by Insert 1 to the thumb domain of the polymerase could alter the cooperativity of PolQ processing

and offer a stronger interaction with DNA that helps bypass damages without dissociating from the DNA template.

Creatinase (Creatine amidinohydrolase) is a component enzyme in the biocomplex that was immobilized in a biosensor that monitors the creatine levels in patients' blood. It was found to be highly susceptible to the inhibition of silver ions from the amperometric electrodes that contain Ag/AgCl (Berberich et al., 2005). Preliminary studies have shown that silver would lead to deactivation provided that their concentration would identically match that of the enzyme in solution. This inhibition does not apparently occur at the active site, given that the addition of substrate would not protect the enzyme from losing its activity. Addition of silver scavenging molecules, thiol containing molecules in particular, were extremely effective in preventing the loss of enzymatic activity due to the decreased free silver ions in chelating reagents. To identify remote sites that were impacted by silver ion binding, we used GNM results from *i*GNM. We proposed the critical cysteines in creatinase that allosterically alter the catalytic motions upon silver binding. We were able to show that two critical thiol residues, Cys297 and Cys60, exist, that if modified/bound by silver ions, may perturb the enzyme function and lead to inactivation. Among these, Cys297 is surrounded by negatively charged residues which may attract positively charged silver ions. We also concluded that the enzyme motions mediated by Cys297, which act as a hinge center in the global modes, could be impeded by silver ions and this could be the mechanism that leads to creatinase inactivation.

The future extensions of *i*GNM are several folds. First, we would like to compute and store the dynamics data for structures in which the biological unit is different from the structure in the

PDB file. For instance, in our recent study on HK97 bacteriophage capsid (Rader et al., 2005), the initial step before examining the capsid dynamics was to assemble the 420 asymmetric units (PDB ID:1ifo and 1ohg), each of which comprises seven monomers of 385 residues each, into a shell (Prohead I) of 60 hexamers and 12 pentamers, forming an icosahedrally symmetric structure. Complexes as such can be derived from the combination of repeated subunits. Usually the coordinates of repeat units are available in the PDB, as well as the transformation (rotation and translation) matrices for constructing the entire structure of the multimer or assembly. The complex (entire HK97 bacteriophage capsid) rather than the constituent heptamers, is the biologically functional unit. This type of reconstruction of structure draws our attention to the necessity of generating and storing the dynamics of biological units.

Second, EN-models that provide information on the directionalities of molecular motions, such as ANM, will be incorporated into the *i*GNM server. The challenge lies in the computer memory allocation given that the Hessian is 9 times larger than the Kirchhoff matrix in the GNM. However, coarse-grained approaches, fast algorithms such as BLZPACK and PowerB that extract a subset of modes and compute B_{theo} respectively, and parallel computing environment can always help us push up the limit of computation to describe functional dynamics of supramolecules. During the course of present doctoral studies, it has been for the first time possible to compute GNM dynamics for proteins over 10^5 residues (Rader et al., 2005).

Third, an automated system is needed to download and compute the dynamics for the newly deposited structures. The mirror site of PDB has been created in our center to synchronize this attempt. Although *i*GNM is a continuing, active projects, we would like to coordinate our

activities with those on-going for PDB update to provide an one-on-one structure-to-dynamics mapping as the former three goals are met.

*i*GNM datamining is probably the most interesting issue to pursue in the near future. We have unprecedented opportunities to explore protein dynamics in a systematical fashion. In addition to characterizing the conserved dynamics pattern of given protein families, we are going to ask more fundamental questions. What is the average dynamics for individual amino acids in general? How do the dynamics vary in different structural classes? Are there specific dynamic patterns preferred by secondary structural elements, or structural motifs, or by particular sequences (triplets or higher n-grams) of residues? As we unravel more and more such patterns, we will be able to start to infer dynamics from sequence compositions. This effort should nevertheless go hand-in-hand with other techniques such as comparative modeling, machine learning technique and clustering algorithms, and most importantly, with experimental studies that can provide feedback for improving or testing our computational models and results.

Overall, this dissertation supports the view that protein structures have been designed to undergo conformational changes that dictate their biological functions. Many applications based on EN-models seem feasible with promising results. 'Dynamics database', *i*GNM, opens the possibility storing pre-computed dynamics information in a database framework for various uses and implementations that can potentially shed light into the conformational mechanisms of protein function.

Bibliography

Agarwal, P., Geist, A., and Gorin, A. (2004) Protein dynamics and enzymatic catalysis: investigating the peptidyl-prolyl cis-trans isomerization activity of cyclophilin A. *Biochemistry* **43,** 10605-10618.

Alexandrov, V., Lehnart, U., Echols, N., Milburn, D., Gerstein, M. and Engelman, D. (2005) Normal modes for predicting protein motions: A comprehensive database assessment and associated web tool. *Protein Sci* **14**, 633-643.

Amadei, A., Linssen, A.B.M., Berendsen, H.J.C. (1993) Essential dynamics of proteins. *Proteins*, **17**, 412–425.

Archontis, G., Simonson, T. and Karplus, M. (2001) Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase. *J. Mol. Biol.* **306**, 307-327.

Argiriadi, M.A., Morisseau, C., Hammock, B.D., and Christianson, D.W. (1999) Detoxification of Environmental Mutagens and Carcinogens: Structure-Based Mechanism and Evolution of Liver Epoxide Hydrolase. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 10637-10642.

Atilgan, A.R., Durrell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O. and Bahar, I. (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model, *Biophys. J.*, **80**, 505-515

Baker, N. A., Sept, D., Joseph, S., Holst, M. J. and McCammon, J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A* **98**, 10037-10041.

Bahar, I., Atilgan, A.R. and Erman, B. (1997a) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, *Folding Des.*, **2**, 173-181.

Bahar, I. and Jernigan, R.L. (1997b) Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation, *J. Mol. Biol.*, **266**, 195-214.

Bahar, I., Atilgan, A.R., Demirel, M.C., and Erman, B. (1998a) Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability, *Phys. Rev. Lett.*, **80**, 2733 -2736.

Bahar, I. and Jernigan, R.L. (1998b) Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms, *J. Mol. Biol.*, **281**, 871-884.

Bahar, I. Wallqvist, A., Covell, D.G., and Jernigan, R.L. (1998c) Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model, *Biochem.*, **37**, 1067-1075.

Bahar, I. (1999a) Dynamics of proteins and biomolecular complexes: inferring functional motions from structure, *Rev. Chem. Eng.*, **15**, 319-349.

Bahar, I., Erman, B., Jernigan, R.L., Atilgan, A.R., and Covell, D. (1999b) Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function, *J. Mol. Biol.*, **285**, 1023-1037.

Bahar, I. and Jernigan, R.L. (1999c) Cooperative fluctuations and subunit communication in tryptophan synthase. *Biochemistry*, **38**, 3478–3490.

Banks, R.D., Blake, C.C.F., Evans, P.R., Haser, R., Rice, D.W., Hardy, G.W., Merret, M. and Phillips, A.W. (1979) Sequence, structure and activity of phosphoglycerate kinase: a possible hinge-bending enzyme. *Nature* **279,** 773-777.

Bartlett,G., Porter,C., Borkakoti,N., and Thornton, J.M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324,** 105-121.

Berberich, J., Yang, L.-W., Bahar, I., Russell, A.J. (2005) Analysis of the impact of silver ions on creatine amidinohydrolase. *ACTA Biomaterialia*, **1**, 183-191.

Berendsen, H. J. and Hayward, S. (2000) Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.* **10**, 165-169.

Benkovic,S.J., and Hammes-Schiffer,S. (2003) A perspective on enzyme catalysis. *Science*, **301,** 1196-1202.

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

Brooks, B. and Karplus, M. (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor, *Proc. Natl. Acad. Sci. USA,* **80**, 6571-6575.

Brünger,A.T. (1993) X-PLOR 3.1: A system for X-Ray crystallography and NMR. Yale University Press, New Haven, USA.

Burioni, R. et al. (2004) Topological thermal instability and length of proteins, *Proteins*, **55**, 529-535.

Cao,Z.W. *et al*. (2004) MoViES: molecular vibrations evaluation server for analysis of fluctuational dynamics of proteins and nucleic acids. *Nucleic Acids Res.*, **32**, W679–W685.

Caves LSD, Evanseck JD, Karplus M. (1998) Locally accessible conformations of proteins: multiple molecular dynamics simulations of proteins. *Protein Sci*; **7**, 649–666.

Chan HS and Dill KA, (1998) Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics, *Proteins*, **30**, 2-33

Chen,C.C., and Herzberg,O. (1992) Inhibition of beta-lactamase by clavulanate. Trapped intermediates in cryocrystallographic studies. *J. Mol. Biol.* **224,** 1103-1113.

Chen, S.C., Bahar, I. (2004) Mining frequent patterns in protein structures: a study of protease families. *Bioinformatics*, **20**, i77-i85.

Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips JGN. (1995) A sampling problem in molecular dynamics simulations of macromolecules. *Proc Natl Acad Sci USA,* **92**, 3288–3292.

Clark,DS. (2004) Characteristics of nearly dry enzymes in organic solvents: implications for biocatalysis in the absence of water. *Philos Trans R Soc Lond B Biol Sci.* **359**, 1299-1307.

Cregut,D., Drin,G., Liautard,J.P., and Chiche,L. (1998) Hinge-bending motions in annexins: molecular dynamics and essential dynamics of apo-annexin V and of calcium bound annexin V and I. *Protein Eng.* **11 ,** 891-900.

Cui, Q. et al. (2004) A normal mode analysis of structural plasticity in the biomolecular motor F-1-ATPase, *J. Mol. Biol.*, **340**, 345-372.

Daniel,R.M., Dunn,R.V., Finney,J.L., and Smith,J.C. (2002) The role of dynamics in enzyme activity. *Annu Rev Biophys Biomol Struct.* **32,** 69-92.

de Groot BL, Hayward S, van Aalten DMF, Amadei A. (1998) Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. *Proteins*, **31**, 116–127.

Delarue, M. and Sanejouand, Y. H. (2002) Simplified normal mode analysis of conformational transitions in DNA- dependent polymerases: the elastic network model. *J. Mol. Biol.* **320**, 1011-1024.

Delarue M and Dumas P (2004) On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc Natl Acad Sci USA*, **101**, 6957-6962.

Demirel, M.C. et al. (1998) Identification of kinetically hot residues in proteins, *Protein Sci.,* **7**, 2522-2532.

Diaz,N., Sordo,T.L., Merz,Jr.K.M., and Suarez,D. (2003) Insights into the acylation mechanism of class A beta-lactamases from molecular dynamics simulations of the TEM-1 enzyme complexed with benzylpenicillin. *J. Am. Chem. Soc.* **125**, 672-684.

Doruker et al. (2000) Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor. *Proteins* **40**, 512-524.

Doruker P, Jernigan RL, Bahar I (2002) Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J Comput Chem*, **23**, 119-127.

Duan, Y. and Kollman, P. A. (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740-744.

Echols, N., Milburn, D. and Gerstein, M. (2003) MolMovDB: analysis and visualization of conformational change and structural flexibility, *Nucl. Acids Res.,* **31**, 478-482.

Eisenmesser,EZ., Bosco,DA., Akke,M., and Kern,D. (2002) Enzyme dynamics during catalysis. *Science* **295,** 1480-1481.

Erkip,A. and Erman,B. (2004) Dynamics of large-scale fluctuations in native proteins. Analysis based on harmonic inter-residue potentials and random external noise. *Polymer*, **45**, 641–648.

Eschenburg,S., Genov,N., Peters,K., Fittkau,S., Stoeva,S., Wilson,K.S., and Betzel,C. (1998) Crystal structure of subtilisin DY, a random mutant of subtilisin Carlsberg. *Eur J Biochem*, **257,** 309-318.

Falcon,C.M., and Matthews,K.S. (1999) Glycine insertion in the hinge region of lactose repressor protein alters DNA binding. *J. Biol. Chem.* **274,** 30849-30857.

Fauman,E.B., Yuvaniyama, C., Schubert,H.L., Stuckey, J.A. and Saper, M.A. (1996) The X-ray crystal structures of *Yersinia* tyrosine phosphatase with bound tungstate and nitrate. Mechanistic Implications. *J. Biol. Chem.*, **271**, 18780–18788.

Flory, P.J. (1976) Statistical thermodynamics of random networks, *Proc. Roy. Soc. Lond. A,* **351**, 351-378.

Frank, J. and Agrawal, R.K. (2000) A ratchet-like inter-subunit reorganization of the ribosome during translocation, *Nature,* **406**, 318-322.

Frauenfelder, H. and McMahon, B. (1998) Dynamics and function of proteins: The search for general concepts *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4795-4797.

Freire,E. (1999) The propagation of binding interactions to remote sites in proteins: analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 10118-10122.

Fujikawa, N., Kurumizaka, H., Nureki, O., Terada, T., Shirouzu, M., Katayama, T., Yokoyama, S. (2003) Structural basis of replication origin recognition by the Dnaa protein. *Nucleic Acids Res.*, **31**, 2077-2086.

Garcia AE, Harman JG. (1996) Simulations of CRP:(cAMP)2 in noncrystalline environments show a subunit transition from the open to the closed conformation. *Protein Sci.* **5**, 62–71.

Gerstein, M. and Krebs, W. (1998) A database of macromolecular motions, *Nucl. Acids Res.,* **26**, 4280-4290.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**, 163-164.

Go, N., Noguti, T. and Nishikawa, T. (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes, *Proc. Natl. Acad. Sci. USA*, **80**, 3696-3700.

Goodman, J. L., Pagel, M. D. and Stone, M. J. (2000) Relationships between protein structure and dynamics from a database of NMR-derived backbone order parameters *J. Mol. Biol.* **295**, 963-978.

Gutteridge,A., Bartlett G.J. and Thornton J.M. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **330**, 719-734

Haliloglu, T., Bahar, I. and Erman, B. (1997) Gaussian dynamics of folded proteins, *Phys. Rev. Lett.*, **79**, 3090-3093.

Haliloglu,T. and Bahar,I. (1999) Structure-based analysis of protein dynamics. Comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. *Proteins*, **37**, 654–667.

Halle, B. (2002) Flexibility and packing in proteins, *Proc. Natl. Acad. Sci. U.S.A*., **99**, 1274-1279.

He, J., Zhang, Z., Shi, Y., Liu, H. (2003) Efficiently explore the energy landscape of proteins in molecular dynamics simulations by amplifying collective motions. *J Chem Phys*, **119**, 4005-4017.

Hinsen,K. (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417–429.

Hinsen, K. and Kneller, G.R. (1999) A simplified force field for describing vibrational protein dynamics over the whole frequency range, *J. Chem. Phys.*, **111**, 10766-10769.

Hinsen K, Reuter N, Navaza J, Stokes DL, Lacapère JJ (2005) Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase. *Biophys J,* **88**, 818-827.

Hirano,M., and Hirano,T. (2002). Hinge-mediated dimerization of SMC protein is essential for its dynamic interaction with DNA. *EMBO J.* **21,** 5733-5744.

Hollup SM, Salensminde G, Reuter N (2005) WEBnm@: a web application for normal mode analyses of proteins. *BMC Bioinformatics*, **6**, 1-8.

Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: visual molecular dynamics, *J. Mol. Graph.*, **14**, 33-38.

Ichiye T, Karplus M. (1991) Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins,* **11**, 205–217

Isin, B., Doruker, P. and Bahar, I. (2002) Functional motions of influenza virus hemagglutinin: a structure-based analytical approach, *Biophys. J.,* **82**, 569-581.

Itoh K, Sasai M (2004) Dynamical transition and proteinquake in photoactive yellow protein. *Proc Natl Acad Sci USA*, **101**, 14736-14741.

Jaravine, V.A., Rathgeb-Szabo, K. and Alexandrescu, A.T. (2000) Microscopic stability of cold shock protein A examined by NMR native state hydrogen exchange as a function of urea and trimethylamine N-oxide, *Prot. Sci.*, **9**, 290-301.

Jernigan, R. L. and Bahar, I. (1996). Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* **6**, 195-209.

Jernigan, R.L., Demirel, M.C. and Bahar, I. (1999) Relating structure to function through the dominant slow modes of motion of DNA topoisomerase II, *Intl. J. Quant. Chem.*, **75**, 301-312.

Jernigan, R.L., Bahar, I., Covell, D.G., Atilgan, A.R., Erman, B. and Flatow, D.T. (2000) Relating the structure of HIV-1 reverse transcriptase to its processing step, *J. Biomol. Struct. Dyn.*, **11**, 49-55.

Jing,H., Babu,Y.S., Moore,D., Kilpatrick,J.M., Liu,X.Y., Volanakis,J.E., and Narayana,S.V. (1998) Structures of native and complexed complement factor d: implications of the atypical his57 conformation and self-inhibitory loop in the regulation of specific serine protease activity. *J. Mol. Biol.* **282,** 1061-1081.

Kaledin M, Brown A, Kaledin AL, Bowman JM (2004) Normal mode analysis using the driven molecular dynamics method. II. An application to biological macromolecules. *J Chem Phys*, **121**, 5646-5653.

Karplus M, Kushick JN. (1981) Method for estimating the configurational entropy of macromolecules. *Macromolecules*. **14**, 325–332.

Karplus,M., and McCammon, J. (1983) Dynamics of proteins: Elements and functions. *Ann Rev Biochem*, **53,** 263-300.

Keskin, O., Jernigan, R.L. and Bahar, I. (2000) Proteins with similar architecture exhibit similar large-scale dynamic behavior, *Biophys. J.*, **78**, 2093-2016.

Keskin, O. Bahar, I., Flatow, D., Covell, D.G., Jernigan, R.L. (2002a) Molecular mechanisms of chaperonin GroEL-GroES function, *Biochem.*, **41**, 491-501.

Keskin, O., Durell, S.R., Bahar, I., Jernigan, R.L., Covell, D.G. (2002b) Relating molecular flexibility to function: a case study of tubulin, *Biophys. J.*, **83**, 663-680.

Khan,A.R., Parrish,J.C., Fraser,M.E., Smith,W.W., Bartlett,P.A., and James,M.N. (1998) Lowering the entropic barrier for binding conformationally flexible inhibitors to enzymes. *Biochemistry*, **37,** 16839-16845.

Kim MK, Jernigan RL, Chirikjian GS (2002) Efficient generation of feasible pathways for protein conformational transitions. *Biophys. J.*, **83**, 1620-1630.

Kitao A, Hirata F, Go N. (1991) The effects of solvent on the conformation and the collective motions of a protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem Phys.* **158**, 447–472.

Kitao A, Go N. (1999) Investigating protein dynamics in collective coordinate space. *Curr Opin Struct Biol.* **9**, 164–169

Kohen,A., Cannio,R., Bartolucci,S., and Klinman,JP. (1999) Enzyme dynamics and hydrogen tunnelling in a thermophilic alcohol dehydrogenase. *Nature*, **399,** 417-418.

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S. H., Chong, L., Lee, M., Lee, T. E., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A. and Cheatham, T. (2000) Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33**, 889-897.

Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu H, Gerstein M (2002) Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, **48**, 682-695.

Kundu, S. et al. (2002) Dynamics of proteins in crystals: comparison of experiment with simple models, *Biophys. J.*, **83**, 723-732.

Kundu, S. and Jernigan, R.L. (2004a) Molecular mechanism of domain swapping in proteins: an analysis of slower motions, *Bipohys. J.*, **86**, 3846-3854.

Kundu, S., Sorensen, D.C. and Phillips Jr., G.N. (2004b) Automatic domain decomposition of proteins by a Gaussian network model, *Proteins*, **57**, 725-733.

Kurkcuoglu O, Jernigan RL, Doruker P (2004) Mixed levels of coarse-graining of large proteins using elastic network model succeeds in extracting the slowest motions. *Polymers*, **45**, 649-657.

Kurt,N. *et al*. (2003) Cooperative fluctuations of unliganded and substrate-bound HIV-1 protease: a structure-based analysis on a variety of conformations from crystallography and molecular dynamics simulations. *Proteins*, **51**, 409–422.

Lattanzi,G. (2004) Application of coarse grained models to the analysis of macromolecular structures. *Comput. Mat. Sci.*, **30**, 163–171.

Leach, A. R. (2001). Molecular Modelling. *Principles and Applications.* 2nd, 353-404

Liao,J.L. and Beratan,D.N. (2004) How does protein architecture facilitate the transduction of ATP chemical-bond energy into mechanical work? The cases of nitrogenase and ATP binding-cassette proteins. *Biophys J.*, **87**, 1369–1377.

Liu, X., Karimi, H., Yang, L-.W. and Bahar, I. (2004) Protein Functional Motion Query and Visualization. *IEEE Proceedings*, 28th Annual International Computer Software and Applications Conference (COMPSAC'04)

Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR (2005) An analysis of core deformations in protein superfamilies. *Biophys J.* **88**, 1291-1299.

Levitt,M., Sander,C., and Stern,P.S. (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* **181,** 423-447.

Levy RM, Karplus M, Kushick J, Perahia D. (1984) Evaluation of the configurational entropy for proteins: application to molecular dynamics simulations of an a-helix. *Macromolecules,* **17**, 1370–1374.

Li GH, Cui Q (2002) A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca^{2+}-ATPase. *Biophys. J.*, **83**, 2457-2474.

Liu,S., Widom,J., Kemp,C.W., Crews,C.M., and Clardy,J. (1998) Structure of Human Methionine Aminopeptidase-2 Complexed with Fumagillin. *Science*, **282,** 1324-1327.

Loris, R. et al. (1999) Conserved water molecules in a large family of microbial ribonucleases, *Proteins,* **36**, 117-134.

Louis,J.M., Dyda,F., Nashed,N.T., Kimmel,A.R., and Davies,D.R. (1998) Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease. *Biochemistry,* **37,** 2105-2110.

Lubkowski,J., Yang,F., Alexandratos,J., Wlodawer,A., Zhao,H., Neamati,N., Pommier,Y., Merkel,G., and Skalka,A.M. (1998) Structure of the catalytic domain of avian sarcoma virus integrase with a bound HIV-1 integrase-targeted inhibitor. *Proc. Natl. Acad. Sci. USA.* **95,** 4831-

4836.

Luo, J., and Bruice,TC. (2004) Anticorrelated motions as a driving force in enzyme catalysis: the dehydrogenase reaction. *Proc. Natl. Acad. Sci. USA*. **101,** 13152-13156.

Ma, J., and Karplus, M. (1998) The allosteric mechanism of the chaperonin GroEL: A dynamic analysis. *Proc. Natl. Acad. Sci. USA*. **95,** 8502-8507.

Ma, J.P. (2004) New advances in normal mode analysis of supermolecular complexes and applications to structural refinement. *Curr. Protein. Pept. Sci.*, **5**, 119–123.

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

Matthews, D., Dragovich, P.S., Webber, S.E., Fuhrman, S.A., Patick, A.K., Zalman, L.S., Hendrickson,T.F., Love, R.A., Prins, T.J., Marakovits, J.T., Zhou, R., Tikhe, J., Ford, C.E., Meador, J.W., Ferre, R.A., Brown, E.L., Binford, S.L., Brothers, M.A., Delisle, D.M., and Worland, S.T. (1999) Structure-Assisted Design of Mechanism-Based Irreversible Inhibitors of Human Rhinovirus 3C Protease with Potent Antiviral Activity Against Multiple Rhinovirus Serotypes. *Proc. Natl. Acad. Sci. USA*. **96,** 11000-11007.

Mattice, W.L. and Suter, U.W. (1994) *Conformational theory of large molecules.* John Wiley and Sons, Inc., New York.

Maguid, S., Fernandez Alberti, S., Ferrelli, L. and Echave, J. (2005) Exploring the common dynamics of homologous proteins. Application to the globin family. *Biophys. J.* **89**, 3-13

McCallum, S.A., Hitchens,T.K., Torborg,C., Rule, G.S. (2000) Ligand-induced changes in the structure and dynamics of a human class Mu glutathione S-transferase. *Biochemistry*, **39**, 7343–7356.

McCammon, J.A., Gelin, B.R., Karplus, M., and Wolynes, P.G. (1976). Hinge-Bending Mode in Lysozyme. *Nature* **262,** 325-326.

McCarter, J. D., and Withers, S. G. (1994) Mechanisms of enzymatic glycoside hydrolysis *Curr. Opin. Struct. Biol.* **4**, 885-892.

Mendelsohn, N. S. (1957) An iterative method for the solution of linear equations based on the power method for proper vectors. MathSciNet. *Math. Tables Aids Comput.* **11**, 88–91.

Micheletti, C., Cecconi, F., Flammini, A. and Maritan, A. (2002a) Crucial stages of protein folding through a solvable model: predicting target sites for enzyme-inhibiting drugs, *Prot. Sci.*, **11**, 1878-1887.

Micheletti, C., Lattanzi, G. and Maritan, A. (2002b) Elastic properties of proteins: insight on the folding process and evolutionary selection of native structures, *J. Mol. Biol.*, **321**, 909-921.

Micheletti, C., Carloni, P. and Maritan, A. (2004) Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models, *Proteins*, **55**, 635-648.

Ming, D., Kong, Y.F., Lambert, M.A., Huang, Z. and Ma, J.P. (2002a) How to describe protein motion without amino acid sequence and atomic coordinates. *Proc Natl Acad Sci USA*, **99**, 8620-8625.

Ming, D., Kong, Y., Wakil, S.J., Brink, J., and Ma, J. (2002b) Domain movements in human fatty acid synthase by quantized elastic deformational model. *Proc. Natl. Acad. Sci. USA*. **99,** 7895-7899.

Ming, D., Kong, Y., Wu, Y. and Ma, J. (2003a) Simulation of F-Actin filaments of several Microns, *Bipohys. J.,* **85**, 27-35.

Ming, D., Kong, Y., Wu, Y. and Ma, J. (2003b) Substructure synthesis method for simulating large molecular complexes, *Proc. Natl. Acad. Sci. USA,* **100**, 104-109.

Mittl, P.R., Di Marco, S., Krebs, J.F., Bai, X., Karanewsky, D.S., Priestle, J.P., Tomaselli, K.J., and Grutter, M.G. (1997). Structure of recombinant human CPP32 in complex with the tetrapeptide acetyl-Asp-Val-Ala-Asp fluoromethyl ketone. *J. Biol. Chem.* **272,** 6539-6547.

Miyashita, O., Onuchic, J.N. and Wolynes, P.G. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci USA*, **100**, 12570-12575.

Miyashita, O., Wolynes, P.G. and Onuchic, J.N. (2005) Simple energy landscape model for the kinetics of functional transitions in proteins. *J Phys Chem B*, **109**, 1959-1969.

Miyazawa, S. and Jernigan, R. L. (1985) Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation, *Macromol.,* **18**, 534-552.

Moche, M., Schneider, G., Edwards, P., Dehesh, K., and Lindqvist,Y. (1999). Structure of the complex between the antibiotic cerulenin and its target, beta-ketoacyl-acyl carrier protein

synthase. *J. Biol. Chem.* **274,** 6031-6034.

Mullins, L. S., Pace, C. N. and Raushel, F. M. (1997) Conformational stability of ribonuclease T1 determined by hydrogen-deuterium exchange, *Prot. Sci.,* **6**, 1387-1395.

Nemethy, G ., Pottle, M.S., and Scheraga, H.A. (1983) Energy parameters in polypeptides. Updating of geometrical parameters, nonbonded interactions and hydrogen bond interactions for the naturally occurring amino acids. *J. Phys. Chem.*, **87**, 1883–1887.

Noguchi, T., Onizuka, K., Akiyama, Y. and Saito, M. (1997) PDB-REPRDB: A Database of Representative Protein Chains in PDB (Protein Data Bank). *In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, AAAI press, Menlo Park, CA.

Oakley, A.J., Bello, M.L., Battistoni, A., Ricci, G., Rossjohn, J., Villar, H.O., and Parker, M.W. (1997) The structures of human glutathione transferase P1-1 in complex with glutathione and various inhibitors at high resolution. *J. Mol. Biol.* **274,** 84-100.

Ortiz, A.R. and Skolnick, J. (2000) Sequence evolution and the mechanism of protein folding, *Bipohys. J.,* **79**, 1787-1799.

Pang, A., Arinaminpathy, Y., Sansom, M., and Biggin, P. (2003) Interdomain dynamics and ligand binding: molecular dynamics simulations of glutamine binding protein. *FEBS letters*, **550,** 168-174.

Papoulis, A. (1965) Sequences of Random Variables, in Probability, Random Variables, and Stochastic Processes, Kogakusha Company, Tokyo, p. 233

Porter, C.T., Bartlett, G.J., and Thornton, J.M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **32,** D129-D133.

Press,W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (1992) *Numerical Recipes in Fortran: 2nd Ed.* Cambridge University Press, **Chp 2.6**, 51–62.

Rader, A.J., Anderson, G., Isin, B., Khorana, H.G., Bahar, I. and Klein-Seetharaman, J. (2004a) Identification of core amino acids stabilizing rhodopsin, *Proc. Natl. Acad. Sci. USA*, **101**, 7246-7251.

Rader, A.J. and Bahar, I. (2004b) Folding core predictions from network models of proteins, *Polymer*, **45**, 659-668.

Rader, A.J., Vlad, D.H. and Bahar, I. (2005) Maturation dynamics of bacteriophage HK97 capsid, *Structure,* **13**, 413-421.

Renatus, M., Stubbs, M.T., Huber, R., Bringmann, P., Donner, P., Schleuning, W.D., and Bode, W. (1997) Catalytic domain structure of vampire bat plasminogen activator: a molecular paradigm for proteolysis without activation cleavage. *Biochemistry,* **36,** 13483-13493.

Ringe, D., and Petsko, GA. (2004) The 'glass transition' in protein dynamics: what it is, why it occurs, and how to exploit it. *Biophys Chem.* **105,** 667-680.

Roe, S.M., Prodromou, C., O'Brien, R., Ladbury, J.E., Piper, P.W., and Pearl, L.H. (1999) Structural basis for inhibition of the Hsp90 molecular chaperone by the antitumor antibiotics radicicol and geldanamycin. *J. Med. Chem.* **42,** 260-266.

Scapin, G., Reddy, S.G., Zheng, R., and Blanchard, J.S. (1997) Three-dimensional structure of Escherichia coli dihydrodipicolinate reductase in complex with NADH and the inhibitor 2,6-pyridinedicarboxylate. *Biochemistry* **36,** 15081-15088.

Seki, M., Masutani, C., Yang, L,-W, Schuffert, A., Iwai, S., Bahar, I. and Wood, R.D. (2004) High efficiency bypass of DNA damage by a single human DNA polymerase. *EMBO J.* **23**, 4484-4494.

Service, R. F. (2000) Structural genomics offers high-speed look at proteins. *Science*, **287**, 1954-1956.

Sidhu,G., Withers, S.G., Nguyen, N.T., McIntosh, L.P., Ziser, L., and Brayer, G.D. (1999) Sugar ring distortion in the glycosyl-enzyme intermediate of a family G/11 xylanase. *Biochemistry,* **38,** 5346-5354.

Sinha, N., Kumar, S., and Nussinov, R. (2001) Inter-Domain Interactions in Hinge-Bending Transitions. *Structure*, **9,** 1165-1181

Sluis-Cremer, N., Temiz, N.A. and Bahar, I. (2004) Conformational changes in HIV-1 reverse transcriptase induced by nonnucleoside reverse transcriptase inhibitor binding, *Curr. HIV Res.,* **2**, 323-332.

Smith, L. J., Daura, X. and van Gunsteren, W. F. (2002) Assessing equilibration and convergence in biomolecular simulations. *Proteins*, **48**, 487-496.

Stams, T., Chen, Y., Boriack-Sjodin, P.A., Hurt, J.D., Liao, J., May, J.A., Dean, T., Laipis, P., Silverman, D.N., and Christianson, D.W. (1998) Structures of murine carbonic anhydrase IV and human carbonic anhydrase II complexed with brinzolamide: molecular basis of isozyme-drug discrimination. *Protein Sci.* **7**, 556-563.

Stenberg, K., and Lindqvist, Y. (1997). Three-dimensional structures of glycolate oxidase with bound active-site inhibitors. *Protein Sci.* **6,** 1009-1015.

Stock, A. (1999) Relating dynamics to function. *Nature*, **400**, 221-222.

Suhre, K. and Sanejouand, Y.-H. (2004a) ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, **32**, W610–W614.

Suhre, K. and Sanejouand, Y.-H. (2004b) On the potential of normal mode analysis for solving difficult molecular replacement problems. *Acta Crystallogr., Sect. D* **60**, 796–799.

Suguna, K., Padlan, E. A., Smith, C. W., Carlson, W. D., and Davies, D. R. (1987) Binding of a reduced peptide inhibitor to the aspartic proteinase from Rhizopus chinensis: implications for a mechanism of action. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7009-7013

Tama, F., Gadea, F.X., Marques, O. and Sanejouand, Y.H. (2000) Building-block approach for determining low-frequency normal modes of macromolecules, *Proteins,* **41**, 1-7.

Tama, F. and Sanejouand, Y.H. (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, **14**, 1-6.

Tama, F. and Brooks, C.L. III, (2002a) The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus, *J. Mol. Biol.*, **318**, 733-747.

Tama, F., Wriggers, W. and Brooks, C.L. III, (2002b) Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J. Mol. Biol.*, **321**, 297-305.

Tama, F. (2003a) Normal mode analysis with simplified models to investigate the global dynamics of biological systems. *Protein Peptide Lett.*, **10**, 119–132.

Tama, F., Valle, M., Frank, J. and Brooks CL III, (2003b) Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy, *Proc. Natl. Acad. Sci. USA,* **100**, 9319-9323.

Tama, F., Miyashita, O., Brooks, C.L. III (2004a), Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol*., **337**, 985-999.

Tama, F., Miyashita, O., Brooks, C.L. III (2004b), Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *J. Struct. Biol.*, **147**, 315-326.

Tama, F. and Brooks III, C.L. (2005) Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis. *J. Mol. Biol.,* **345**, 299-314.

Tatsumi, R., Fukunishi, Y., Nakamura, H. (2004) A hybrid method of molecular dynamics and harmonic dynamics for docking of flexible ligand to flexible receptor. *J. Comput. Chem*. **25**, 1995-2005.

Taylor, W. R. (1986) The classification of amino acid conservation. *J. Theo. Biol.* **119**, 205–218.

Temiz, N.A. and Bahar, I. (2002) Inhibitor binding alters the directions of domain motions in HIV-1 reverse transcriptase, *Proteins,* **49**, 61-70.

Temiz, N.A., Meirovitch, E., and Bahar, I. (2004). Escherichia coli adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling (15)N-NMR relaxation data. *Proteins*, **57,** 468-480.

Thomas, A., Hinsen, K., Field, M.J., Perahia, D. (1999) Tertiary and quaternary conformational changes in aspartate transcarbamylase: a normal mode study, *Proteins,* **34**, 96-112.

Tirion, M.M. (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis, *Phys. Rev. Lett.,* **77**, 1905-1908.

Tousignant, A., and Pelletier, J.N. (2004). Protein motions promote catalysis. *Chem Biol.* **11,** 1037-1042.

Towler, P., Staker, B., Prasad, S.G., Menon, S., Tang, J., Parsons, T., Ryan, D., Fisher, M., Williams, D., Dales, N.A., Patane, M.A., and Pantoliano, M.W. (2004). ACE2 X-Ray Structures Reveal a Large Hinge-bending Motion Important for Inhibitor Binding and Catalysis. *J. Biol. Chem.* **279,** 17996-18007.

Van Wynsberghe, A., Li, G. and Cui, Q. (2004) Normal-mode analysis suggests protein flexibility modulation throughout rna polyerase's functional cycle, *Biochem.,* **43**, 13083-13096.

Varughese, K.I., Su, Y., Cromwell, D., Hasnain, S., and Xuong, N.H. (1992). Crystal structure of an actinidin-E-64 complex. *Biochemistry*, **31,** 5172-5176.

Vila-Sanjurjo, A., Ridgeway , W.K., Seymaner, V., Zhang, W., Santoso, S., Yu, K. and Doudna Cate, J.H. (2003) X-ray crystal structures of the wt and a hyper-accurate ribosome from Escherichia coli, *Proc. Natl. Acad. Sci. USA,* **100**, 8682-8687.

Wand, A. J. (2001) Dynamic activation of protein function: A view emerging from NMR spectroscopy *Nat. Struct. Biol.* **8**, 926-931.

Wang, W. and Kollman, P. A. (2000). Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. *J. Mol. Biol.* **303**, 567-582.

Wang, Y., Rader, A.J., Bahar, I. and Jernigan, R.L. (2004) Global ribosome motions revealed with elastic network model, *J. Struct. Biol.,* **147**, 302-314.

Wako, H., Endo, S., Nagayama, K. and Go, N. (1995) FEDER/2: program for static and dynamic conformational energy analysis of macro-molecules in dihedral angle space. *Comp. Phys. Comm.*, **91**, 233–251.

Wako, H. and Endo, S. (2002) ProMode: a database of normal mode analysis of proteins. *Genome Informatics*, **13**, 519–520.

Wako, H., Kato, M. and Endo, S. (2003) Improvements in ProMode (a Database of Normal Mode Analyses of Proteins). *Genome Informatics*, **14**, 663–664.

Wako, H., Kato, M. and Endo, S. (2004) ProMode: a database of normal mode analyses on protein molecules with a full-atom model. *Bioinformatics*, **20**, 2035–2043.

Wilce, M.C., Bond, C.S., Dixon, N.E., Freeman,H.C., Guss,J.M., Lilley,P.E., and Wilce,J.A. (1998) Structure and mechanism of a proline-specific aminopeptidase from Escherichia coli. *Proc. Natl. Acad. Sci. USA.* **95,** 3472-3477.

Wilson, K.P., Malcolm, B.A. and Matthews, B.W. (1992) Structural and thermodynamic analysis of compensating mutations within the core of chicken egg white lysozyme, *J. Biol. Chem.,* **267**, 10842-10849.

Wind, R.D., Uitdehaag, J.C., Buitelaar, R.M., Dijkstra, B.W., and Dijkhuizen, L. (1998) Engineering of cyclodextrin product specificity and pH optima of the thermostable cyclodextrin glycosyltransferase from Thermoanaerobacterium thermosulfurigenes EM1. *J. Biol. Chem.* **273,** 5771-5779.

Wolf-watz, M., Thai, V., Henzler-Wiilman, K., Hadjipavlou, G., Eisenmesser, E.Z., and Kern, D. (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nature Struct & Mol Biol.* **11,** 945-949.

Wu, Y. Wu, Y., Yuan, X., Gao, X., Fang, H. and Zi, J. (2003) Universal behavior of localization of residue fluctuations in globular proteins, *Phys. Rev. E.,* **67**, 041909.

Xiang, J., Sun, J., and Sampson, N.S. (2001). The Importance of Hinge Sequence for Loop Function and Catalytic Activity in the Reaction Catalyzed by Triosephosphate Isomerase. *J. Mol. Biol.* **307,** 1103-1112.

Xu, C., Tobi, D. and Bahar, I. (2003) Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T<-> R2 transition, *J. Mol. Biol.,* **333**, 153-168.

Yan, X., Hollis, T., Svinth, M., Day, P., Monzingo, A.F., Milne, G.W., and Robertus, J.D. (1997) Structure-based identification of a ricin inhibitor. *J. Mol. Biol.* **266,** 1043-1049.

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

Yang, L.-W., Liu, X., Jursa, C.J., Holliman, M., Rader, A.J., Karimi, H.A., Bahar, I. (2005b) iGNM: a database of protein functional motions based on Gaussian Network Model. http://ignm.ccbb.pitt.edu/ *Bioinformatics*, **21**, 2978-2987.

Zdanov, A., Wu, S., DiMaio, J., Konishi, Y., Li, Y., Wu, X., Edwards, B.F., Martin, P.D., and Cygler, M. (1993) Crystal structure of the complex of human alpha-thrombin and nonhydrolyzable bifunctional inhibitors, hirutonin-2 and hirutonin-6. *Proteins,* **17,** 252-265.

Zhang, Z.Y., Shi, Y.Y., and Liu, H.Y. (2003) Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys J.* **84,** 3583-3593.

Zhao, H., Tang, L., Wang, X., Zhou, Y., and Lin, Z. (1998) Structure of a snake venom phospholipase A2 modified by p-bromo-phenacyl-bromide, *Toxicon*, **36**, 875-886.