High-throughput modeling and analysis of protein structural dynamics

Xiong Liu and Hassan A. Karimi

Received: 12th October 2006

Abstract
Protein function is a dynamic property closely related to the conformational mechanisms of protein structure in its physiological environment. To understand and control the function of target proteins, it becomes increasingly important to develop methods and tools for predicting collective motions at the molecular level. In this article, we review computational methods for predicting conformational dynamics and discuss software tools for data analysis. In particular, we discuss a high-throughput, web-based system called iGNM for protein structural dynamics. iGNM contains a database of protein motions for more than 20,000 PDB structures and supports online calculations for newly deposited PDB structures or user-modified structures. iGNM allows dynamics analysis of protein structures ranging from enzymes to large complexes and assemblies, and enables the exploration of protein sequence-structure-dynamics-function relations.

Keywords: protein dynamics; protein function; normal mode analysis; elastic network models; databases; web-based systems

INTRODUCTION
The detailed understanding of protein functions is a goal of bioinformatics in the post-genomic era. A function is a property closely related to the conformational mechanics of the structure in its physiological environment. Protein catalytic activity, folding, binding and molecular recognition all involve protein motions or fluctuations [1]. Protein motions or conformational changes relevant to biological events are called functional motions or functional dynamics.

The description of functional dynamics is a challenging scientific problem, since proteins do not work in a deterministic way as macro machines do. X-ray crystallography provides useful structural information about thermal and other fluctuations of the atoms in a protein. For example, B-factor, or thermal factor, is used to measure the positional uncertainty associated with each atom in the thermal fluctuations. However, experimentally observed B-factors cannot explain molecular collective motions relevant to biological functions [2]. Instead, protein dynamics modeling provides insights into the mechanisms of motion of large proteins by dissecting observed motions into a collection of normal modes. The slowest modes usually provide information on the collective motions relevant to biological function, as demonstrated in many applications [3-5].

Traditionally, theoretical and experimental normal modes have only been compared to actual motions on a case-by-case basis [6]. With the rapid accumulation of biomolecular structures in the Protein Data Bank (PDB) [7], it has become evident that structural and dynamic information per se is not sufficient for gaining insights into the mechanism of function. For further exploration and establishment of biomolecular structure-dynamics-function relations, efficient methods and tools for predicting and managing collective motions at the molecular level are becoming increasingly important.
Bahar and co-workers [8, 9] have proposed the Gaussian Network Model (GNM) and further simplified high-resolution models by only considering inter-residue contact topology in the folded state with single parameter potentials. The accumulating evidence that supports the utility of GNM as an efficient tool for a first estimation of the machinery or proteins and their complexes has led to iGNM (Internet-based GNM) (http://ignm.ccbb.pitt.edu) [10, 11], a high-throughput, web-based system of GNM results compiled for protein structures ranging from enzymes to large complexes and assemblies. iGNM contains a database of calculated dynamics for more than 20,000 PDB structures and supports online calculations for customized structures or newly deposited PDB structures. iGNM’s features include: (i) capability of analyzing relatively large structures or multiple domains; (ii) tools for retrieving calculated results or releasing online calculation results with high performance (in seconds)—this is a significant improvement over similar systems that require minutes, hours, or days for predicting protein motions; and (iii) tools for providing an integrated environment for querying, visualizing, and comparing protein collective motions.

iGNM allows large-scaled dynamics analysis of protein structures and enables the exploration of protein structure—dynamics—function relations. Due to its efficiency and applicability to large structures, iGNM is gaining attention of researchers in the biological community. The rest of the article is organized as follows. In the section titled Computational Prediction of Protein Dynamics, computational models on protein dynamics are reviewed. In the section titled Web-based Systems for Protein Dynamics, related work on web-based systems for protein dynamics is discussed. In the section titled iGNM Architecture, the architecture of iGNM is described. In the section titled iGNM Utility, three scenarios are presented to show the utility of iGNM. Finally, conclusions and future research are discussed in the section titled Conclusions and Future Research.

COMPUTATIONAL PREDICTION OF PROTEIN DYNAMICS

Overview

Experimental methods, such as X-ray crystallography, nuclear magnetic resonance (NMR) and hydrogen/deuterium (H/D) exchange, reveal atomic-level information on protein internal motions. Not surprisingly, a major endeavor in recent years has been devoted to developing computational models and methods for simulating protein dynamics using structural data and relating the observed behavior to other experimental data. Molecular dynamics (MD) simulations have proven to be a useful approach for generating conformational trajectories of macromolecules in order to visualize the correlation of their dynamics to the biological functions [12]. However, MD simulations are expensive in terms of CPU time and memory. An efficient method for identifying function-related conformational changes is normal mode analysis (NMA), a method widely used for characterizing molecular fluctuations near a given equilibrium state using vibrational modes. The utility of NMA for protein dynamics has been recognized for the last 20 years [12, 13], but has been revitalized in recent years with the success of elastic network models used in NMA. In these models, atoms or groups of atoms (e.g., residues or groups of residues) are modeled as point sites (network nodes) connected by springs, which account for the force field that stabilizes the native structure. The utility of these models in NMA was first pointed out by Tirion [14]. Given the insensitivity of the most cooperative modes to the detailed structure, a large majority of recent analyses have been performed using lower-resolution elastic network models. Among the elastic network models of different complexities, the simplest is GNM [8, 9]. GNM is a one-dimensional model that does not convey information on the directions of fluctuations, but their size. Also, GNM, with its present form, is based on alpha carbons only, although it is possible to base it on an all-atom representation of the structure.

GNM

The roots of GNM are well founded in fundamental statistical mechanical theories of polymer networks where the junctions of the network undergo Gaussian-distributed fluctuations under the potential of the pendant chains [15, 16].

In GNM, the alpha carbon atoms of residues are identified as the junctions or nodes of the network, and the pairs of nodes closer than a cutoff distance are connected by harmonic potentials with a uniform spring constant $\gamma$ (Figure 1). In addition to nonbonded interactions, the effect of chain connectivity is also considered, as the model automatically includes the constraints imposed by the first neighboring alpha carbon atoms along the backbone. Thus, the residues fluctuate under the
The connectivity (or Kirchhoff) matrix of contacts, $\Gamma_{ij}$, is used to describe the inter-residue contact topology. The definition of $\Gamma_{ij}$ is given in Equation (1), where $i$ and $j$ are residue indices, $R_{ij}$ is the distance between residue $i$ and residue $j$, $r_c$ is the distance cutoff typically in the range between 5 and 7 Å and $\sum \Gamma_{ik}$ is the number of coordination residues within the cutoff. The off-diagonal elements of $\Gamma$ are defined as $\Gamma_{ij} = \frac{1}{C_{0}}$ if $R_{ij}$ is shorter than $r_c$, and zero otherwise; and the $i$th diagonal term is the degree of node $i$, or the coordination number of residue $i$.

$$
\Gamma_{ij} = \begin{cases} 
-1 & \text{if } i \neq j \text{ and } R_{ij} \leq r_c \\
0 & \text{if } i \neq j \text{ and } R_{ij} > r_c \\
\sum_{k \neq i} \Gamma_{ik} & \text{if } i = j
\end{cases}
$$ (1)

Residue 1 2 3 4... 18 19 20 21... N

$$
\Gamma(R) = \begin{pmatrix}
-1 & -1 & 0 & \cdots & 0 \\
-1 & 9 & -1 & \cdots & 0 \\
0 & -1 & -1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0 
\end{pmatrix}
$$ (2)

Figure 2 shows an example of calculating $\Gamma$, where residue '3' is at the center of a cutoff sphere. The third row of $\Gamma$, which is the connectivity between residue '3' and other residues, is given in Equation (2).

The statistical thermodynamics of the network are controlled by the Hamiltonian [17]:

$$
H = \frac{\gamma}{2} \left[ \Delta X \Gamma \Delta X^T + \Delta Y \Gamma \Delta Y^T + \Delta Z \Gamma \Delta Z^T \right]
$$ (3)
where $\gamma$ is the spring constant, $\Delta X$, $\Delta Y$ and $\Delta Z$ are the $N$-dimensional vectors of the $X$, $Y$- and $Z$- components of the fluctuation vectors $\{\Delta R_1, \Delta R_2, \ldots, \Delta R_N\}$ of the $N$ residues in the examined protein. The mean-square fluctuations of residue $i$ scale with the $i$th diagonal element of the inverse of $\Gamma$ [8, 9] as:

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

where $k$ is constant and $T$ is temperature. The cross-correlations $\{\Delta R_i \cdot \Delta R_j\}$ scale with the $ij$-th off-diagonal elements of $\Gamma^{-1}$.

The fluctuation dynamics of the structure results from $N-1$ superposed GNM modes. The modes can be extracted by the eigenvalue decomposition of $\Gamma$. The decomposition is calculated as follows: $\Gamma = U \Lambda U^T$, where $U$ is an orthogonal matrix whose columns $u_i$ ($2 \leq i \leq N$) are the eigenvectors of $\Gamma$ and $\Lambda$ is the diagonal matrix of the eigenvalues $\lambda_i$, usually organized in ascending order. The first eigenvalue, identically equal to zero, is not included. The $i$-th eigenvector reflects the shape of the $i$-th mode as a function of the residue index. The $i$-th eigenvalue is proportional to the $i$-th mode’s frequency [9].

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

$$
B_i = (8\pi^2 kT/\gamma)[\Gamma^{-1}]_{ii} = (8\pi^2 kT/\gamma) \sum_{k=1}^{N-1} \lambda_k^{-1}[u_k][u_k].
$$

Equation (5) follows from Equation (4) and the definition $B_i = (8\pi^2 / 3)(\langle \Delta R_i \rangle^2)$. The term $[u_k]$ designates the $i$-th element (corresponding to $i$-th residue) of the $k$-th eigenvector.

From the computational time standpoint, the eigenvalue decomposition of the connectivity matrix $\Gamma$ is the most expensive task in GNM calculations. The singular value decomposition (SVD) method [18], whose computation timescales with $N^3$ for a network of $N$ residues, is often used. When $N<1500$, the computations are performed within minutes, while for $N>1500$, the computations may be performed in 15 days.

An alternative decomposition algorithm that utilizes the BLZPACK software [19] is based on Block Lanczos Method for large structures. This method evaluates a subset ($1 \leq k \leq 100$) of dominant (slowest) modes, within a timescale of $N^2$, i.e. the computation time is more than three orders of magnitude faster than the routine SVD, when structures of more than $10^3$ residues are analyzed.

### Functional implication of GNM outputs

Despite its simplicity, GNM has proven to yield results in good quantitative and qualitative agreement with experimental data and MD simulations [8, 20–22]. Previous studies have shown that GNM can satisfactorily reproduce experimentally observed fluctuations and functional motions of proteins complexed with RNA or DNA [17], including supramolecular structures like ribosomal complexes [23] or viral capsids [24].

The GNM slow modes provide insights about the mechanisms of the cooperative molecular motions relevant to function. Previous studies [17, 25–29] show that the minima in the slowest-mode shapes coincide with the hinge sites of the molecule, whereas the maxima usually correspond to substrate recognition sites.

The fast modes usually contain white-noises that need to be filtered out to obtain physically meaningful information. The GNM results differ from those extracted from conventional simulations in that they are devoid of random noise effects. The fastest modes indicate the most strongly constrained sites in the presence of the intricate coupling between all residues. The peaks in the fastest modes are referred to as ‘kinetically hot residues’ in view of their high frequencies [21, 27]. These sites are usually involved in the folding nuclei, or in the key tertiary contacts stabilizing the overall fold.

Correlations between residue fluctuations describe those regions of the structure that move collectively and how these regions move with respect to one another. The positive and negative limits of cross-correlation are 1 and $-1$. The limits correspond to pairs of residues exhibiting fully correlated (same direction and same sense) and fully anti-correlated (same direction and opposite sense) motions, respectively. Zero correlation represents uncorrelated or orthogonal motions. Correlation maps are useful for identifying functional regions. For example, three highly correlated regions are observed for Tubulin dynamics [30]. The partitioning of the structure into three such regions is consistent with the equilibrium structure found by Nogales et al. [31].
WEB-BASED SYSTEMS FOR PROTEIN DYNAMICS

Overview

Traditionally, biologists manually extract interested protein structures from PDB and plug them into dynamics models to obtain dynamics information. This manual approach is not efficient, and is limited to specific case studies. To enable dynamics modeling and analysis for a large number of protein structures, the automatic construction of a biological workflow is required. The workflow should contain a series of tasks including acquisition of protein structures, dynamics modeling on acquired structures and management of predicted dynamics information for efficient query and visualization. It is now widely recognized that for the prompt dissemination of research results such a workflow must be implemented through web-based approaches.

Web-based systems on protein dynamics are based on computational models (see the section titled Computational Prediction of Protein Dynamics). These models have two roles in web-based systems: (i) they are used as online computational engines to infer the functional motions of the protein structures submitted by users and (ii) they are used as offline engines to process a large number of PDB structures for building web-enabled protein motion databases. While online calculation engines provide the dynamics information for given protein structures, protein motion databases provide dynamics information in a much larger scale and therefore allow researchers to perform statistical analysis or data mining to answer deeper questions.

Results from NMA of proteins can now be accessed in a number of web-based systems. MolMovDB (http://molmovdb.org/) is a collection of data and software tools pertaining to flexibility in protein and RNA structures [32]. About 17,000 movies (morphs) of transitions between known structures are available in the database. MolMovDB also offers a simplified NMA to display the molecular motions in the low-frequency modes [33]. ProMode (http://cube.socs.waseda.ac.jp/pages/jsp/index.jsp) is a database containing normal modes for single-chain proteins from PDB [34]. About 1442 structures are available in the database. ElNe´mo (http://www.igs.cnrs-mrs.fr/elnemo/) is an online calculation server for large structures using an all-atom model [35]. Similar calculation servers include MoViES (http://ang.cz3.nus.edu.sg/cgi-bin/prog/norm.pl) [36], WEBnm (http://www.bioinfo.no/tools/normalmodes) [37], NOMAD-Ref (http://lorentz.immstr.pasteur.fr/nomad-ref.php) [38] and AD-ENM/DC-ENM (http://enm.lobos.nih.gov/) [39]. More recently, the Orozco group has been working on a consensus view of protein dynamics using all-atom MD simulations [40]. The results for 359 PDB structures are now available in the MODEL database (http://mnb.pcb.ub.es/MODEL/).

Compared to these related works, iGNM (http://ignm.ccbb.pitt.edu/) [11] contains a database of GNM calculations for up to 22,549 structures (as of 15 September, 2003), and supports online calculations based on GNM for submitted structures. Unlike other similar systems that require minutes, hours or days for predicting protein motions, given the same computing platform, iGNM has a much better performance (seconds for database query; seconds for small structures online calculation; and minutes for very large structures online calculation). Also, iGNM can handle various sizes of structures, or single domains of proteins. It allows researchers to examine the essential dynamics of the complete set of PDB structures and customized structures.

Comparison of computational methods

Table 1 outlines the main characteristics (components, computational methods and outputs) of the systems discussed earlier. As for computational methods, MolMovDB uses interpolation between known conformations to generate morphs. The simplified NMA of MolMovDB uses the Molecular Modelling Toolkit (MMTK) package [41] and alpha carbon (CA) atom representation of structures. Although MolMovDB provides the option of generating and downloading movies of motions, it is restricted to the analysis of single-domain or single-chain proteins.

ProMode generates normal modes using the ECEPP/2 force field [42]. The structures are pre-equilibrated prior to NMA computation. The simplified NMA of MolMovDB uses the Molecular Modelling Toolkit (MMTK) package [41] and alpha carbon (CA) atom representation of structures. Although MolMovDB provides the option of generating and downloading movies of motions, it is restricted to the analysis of single-domain or single-chain proteins.

ProMode generates normal modes using the ECEPP/2 force field [42]. The structures are pre-equilibrated prior to NMA computation. The simplified NMA of MolMovDB uses the Molecular Modelling Toolkit (MMTK) package [41] and alpha carbon (CA) atom representation of structures. Although MolMovDB provides the option of generating and downloading movies of motions, it is restricted to the analysis of single-domain or single-chain proteins.

ProMode generates normal modes using the ECEPP/2 force field [42]. The structures are pre-equilibrated prior to NMA computation. The simplified NMA of MolMovDB uses the Molecular Modelling Toolkit (MMTK) package [41] and alpha carbon (CA) atom representation of structures. Although MolMovDB provides the option of generating and downloading movies of motions, it is restricted to the analysis of single-domain or single-chain proteins.
residues into a single super-residue. Due to this approximation, it is possible to model very large proteins using an all-atom model in a reasonable time. Tama et al. [44] have shown that this approximation has very little influence on the low-frequency modes.

MoViES uses an NMA model based on full-atomic AMBER force field, which derives thermal vibrations for proteins and DNA/RNA up to 4000 heavy atoms [36]. WEBnm uses the MMTK package and CA atom representation of structures for dynamics analysis [37]. NOMAD-Ref uses an all-atom elastic network model to calculate normal modes, which in turn are used to refine models against experimental data such as X-ray diffraction [38]. AD-ENM also uses elastic network model to perform analysis of macromolecular dynamics but with a residue (or CA atom) representation of structures [39]. Recently, a new modeling tool called DC-ENM has been added that uses low-frequency normal modes and a few pairs of distance constraints to build protein structural models.

MODEL is based on MD simulations using state-of-the-art supercomputers (~50 years of CPU) and four most widely used force fields (OPLS, CHARMM, AMBER and GROMOS) [40]. Equilibrated structures are used as starting points for 10ns production trajectories. Those trajectories are analyzed to obtain structural and dynamic properties such as B-factors and backbone root mean square deviation (rmsd) values.

iGNM uses GNM, which models CA atoms with a single-potential parameter. Because GNM is one-dimensional, it only predicts the size (amplitude) of fluctuations without information on the directions of fluctuations. However, due to the simplicity and efficiency of GNM (see the section titled GNM), iGNM contains a much larger number of protein structures than any other system.

**Comparison of outputs**

MolMovDB offers movies and classification of molecular motions according to their size and mechanism [32]. The NMA calculation server of

<table>
<thead>
<tr>
<th>System</th>
<th>Software components</th>
<th>Computational methods</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>MolMovDB (USA)</td>
<td>Database of 17000 morphs of transitions between PDB structures; NMA calculation server</td>
<td>Interpolation between known conformations; NMA using the MMTK package and CA atom representation of structures</td>
<td>Movies (morphs) of molecular motions; classification of motions; low-frequency modes for submitted structures</td>
</tr>
<tr>
<td>ProMode (Japan)</td>
<td>Database of NMA results for 1442 structures</td>
<td>All-atom NMA; energy minimization prior to NMA</td>
<td>20 slowest modes</td>
</tr>
<tr>
<td>ElNémo (France)</td>
<td>NMA calculation server</td>
<td>NMA combined with RTB algorithm</td>
<td>Up to 100 slowest modes</td>
</tr>
<tr>
<td>MoViES (Singapore)</td>
<td>NMA calculation server</td>
<td>All-atom NMA;</td>
<td>Normal modes; mode distribution; thermodynamics</td>
</tr>
<tr>
<td>WEBnm (Norway)</td>
<td>NMA calculation server</td>
<td>NMA using the MMTK package and CA atom representation of structures</td>
<td>Motions for the first six modes; deformation energy for first 14 modes</td>
</tr>
<tr>
<td>NOMAD-Ref (France)</td>
<td>NMA calculation server; tools for deformation and refinement of structures</td>
<td>All-atom elastic network model</td>
<td>Normal modes; Refinement of lowest frequency amplitudes</td>
</tr>
<tr>
<td>AD-ENM/DC-ENM</td>
<td>NMA calculation server</td>
<td>Elastic network model using CA atom representation of protein structures and P atom representation of DNA/RNA</td>
<td>Up to 20 slowest modes from AD-ENM; protein structural models from DC-ENM</td>
</tr>
<tr>
<td>MODEL (Spain)</td>
<td>Database of MD simulation results for 359 PDB structures</td>
<td>MD simulation using state-of-the-art supercomputer and four most widely used force fields</td>
<td>B-factors; rmsd (Å) values; pair-cross rmsd (Å) values; frequencies (cm⁻¹) for the first 100 modes</td>
</tr>
<tr>
<td>iGNM (USA)</td>
<td>Database of GNM results for 22 549 structures; GNM calculation server</td>
<td>GNM using CA atom representation of protein structures and P atom representation of DNA/RNA</td>
<td>20 slowest modes and 20 fastest modes; average modes (B-factors); correlation between the fluctuations</td>
</tr>
</tbody>
</table>

**Table 1: Web-based systems on protein dynamics**
MolMovDB calculates the five lowest frequency modes for submitted structures [33].

The ProMode outputs include 20 slowest modes and average overall normal modes and time. The fluctuations in normal modes include fluctuation of positions, fluctuation of torsion angles and correlations between CA movements [34].

ElNémo calculates normal modes up to 100 slowest modes for submitted structures. The different properties of these modes (i.e. frequency, degree of collectivity of movement and mean square displacement) are displayed using 3D animations or 2D plots [35].

The MoViES outputs include the normal modes, the distribution of normal modes with respect to frequencies, the thermal fluctuational bond disruption for all of the hydrogen bonds and the vibrational thermodynamic quantities such as vibrational free energy and entropy [36].

The WEBnm outputs include motions for the first six modes, deformation energy for the first 14 modes, vector field representation of the displacement associated with the low-frequency modes and transconformation between two unknown structures [37].

The NOMAD-Ref outputs include normal modes, refinement of lowest-frequency amplitudes and refinement of docking solutions [38].

The AD-ENM outputs include up to 20 slowest modes, and the DC-ENM outputs include protein structural models [39].

MODEL provides MD simulation results such as B-factors, average backbone rmsds and pair-cross rmsd [40].

The GNM outputs are mainly composed of the fluctuations in a set of slowest and fastest modes, as individual modes and as average modes, and the correlation between the fluctuations. While the normal modes predicted by other systems contain the size and direction of fluctuations, GNM normal modes only contain the size of fluctuations. The functional implications of the GNM outputs were discussed in the section titled Functional Implication of GNM Outputs.

### Comparison of technologies employed

Tables 2–4 summarize the technologies (database, web interface and visualization) employed by the existing systems. Table 2 shows that MolMovDB uses relational database technology to store movies and statistics of protein motions. It provides PDB ID search and browsing of protein motions; ProMode stores NMA results and animations of these results, but with an unknown database approach. ProMode provides only browsing of protein motions; MODEL allows browsing of the MD simulations for a set of PDB structures. The underlying database approach is unknown. iGNM is based on a customized database approach. iGNM contains GNM outputs indexed by PDB ID and allows for both PDB ID search and keyword-based search. Both MolMovDB and ProMode contain information about residue motion directions, and therefore provide animation/movies of protein motions. However, iGNM only contains information about amplitudes of residue fluctuations, and therefore only static pictures are available.

Table 3 shows that all systems are based on the client–server architecture. For access methods, MolMovDB, ElNémo and MoViES use CGI; ProMode and MODEL use JSP; WEBnm uses DTML script; NOMAD-Ref and AD-ENM/
DC-ENM use PHP; iGNM uses Java Servlet. For visualization, MolMovDB uses PyMOL and MolScript; ProMode uses Chime and Java applets; ElNémo uses MolScript; WEBnm uses VMD and Image Magick; NOMAD-Ref uses PyMOL and VMD; AD-ENM/DC-ENM uses Jmol and VMD; MODEL uses Jmol and iGNM uses Chime, Jmol and Java applets.

Table 4 shows the different visualization systems adopted. All these tools allow 3D displaying and animation of molecular structures. Chime and JmolApplet provide 3D rendering, while MolScript and PyMOL provide image files such as GIF and JPEG. All tools, except Chime, are open source. Since JmolApplet is written in Java, it is cross-platform and runs with Java Virtual Machine (JVM) 1.1 included in most popular browsers; this is in contrast to other tools that have different versions, depending on the platform where the source code is compiled.

**iGNM Architecture**

iGNM is based on a client–server architecture for protein motion query, visualization and calculation (Figure 3). The client is based on standard web browsers, and there are two components: GNM and PDB. iGNM consists of two servers: online calculation server and database server. The details of iGNM servers are discussed in the following subsections.
The goal of constructing the database server 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. Currently, the database in iGNM contains visual and quantitative information on the collective modes predicted by GNM for 20,058 structures deposited in PDB prior to 15 September, 2003.

There are five major database entities in the database server: (i) GNM entity that stores both structural and dynamics information for each protein structure; (ii) B-factors or temperature factors entity that stores equilibrium fluctuations for each residue of a protein structure; (iii) slow-modes entity that stores the slowest (lowest frequency) modes; (iv) fast-modes entity that stores the fastest (highest frequency) modes; and (v) crosscorr entity that stores the correlation of fluctuations between different residues. GNM entity has a one-to-one relationship with the rest of the entities. Table 5 shows the database schema used in iGNM.

The GNM entity contains nine attributes, including PDBID, which is a key to a protein record. Attributes such as protein name, protein class, structure and descriptions are PDB objects, while attributes such as B-factors, slow-modes, fast-modes and crosscorr are iGNM objects to be retrieved for any given query.

The B-factors entity contains three attributes: residue index, GNM-calculated theoretical B-factors, and X-ray crystallographic B-factors taken from PDB.

The slow-modes entity contains eleven attributes: residue index and slow-mode shapes (attributes 2–11) associated with the 10 slowest modes, starting from the slowest (first) mode. The dimension of attributes 2–11 in each row is in Angstrom square, giving the fluctuations resulting from these independent modes. Figure 4 shows a data instance example of the slow-modes entity.

There are also 11 attributes in fast-modes entity: residue index and fast-mode shapes (attributes 2–11) associated with the 10 fastest modes, starting from the highest mode. Since the last modes reflect localized fast motions in the protein, these modes have few non-zero elements.

The crosscorr entity contains three attributes: two residue indices and one correlation value.

The iGNM database is a customized database design based on GNM output files. The database has one instance of the GNM table schema for each protein in PDB. The attributes that are iGNM objects (e.g. B-factors) are linked through PDB ID to each corresponding table. As the database schema shows, there is only one-to-one relationship between the GNM table and other tables. Since there are no complicated join operations between the tables, each GNM instance is implemented as a folder, and the object attributes (e.g. B-factors, slow-modes) associated with the instance are stored as text files.

Two types of queries are possible in iGNM version 1.2: (i) retrieving dynamics information for a given PDB structure from a list of all PDB structures and (ii) retrieving residues’ dynamic information for each structure.

iGNM allows users to retrieve information through a simple search engine by entering the PDB identifier of the protein structure of interest.
For example, ‘2hmg’ is the PDB code for influenza virus hemagglutinin A (HA). The output includes: (i) the sizes of residue motions in different collective modes; (ii) the equilibrium fluctuations of residues and comparison with X-ray crystallographic B-factors; (iii) the cross-correlations between residue fluctuations, or domain motions in the collective modes; and (iv) the identity of residues that assume a key mechanical role (e.g. hinge) in the global dynamics of the molecule, as well as those potentially participating in folding nuclei/cores [20, 22].

In addition to queries using PDB ID, iGNM is integrated with PDB SearchLite query interface for keyword-based queries [10]. By typing keywords related to the biological macromolecules of interest, users can browse PDB records and GNM output files for a given protein family in an integrated environment.

Retrieving dynamic information for each residue is achieved through visual queries. After a protein structure is retrieved, the fluctuations of each residue are displayed in both 2D mobility graphs and 3D ribbon diagrams (see the section titled iGNM Utility). iGNM allows the visual query of each residue’s fluctuation by either interactively clicking a residue’s position in the 2D graph or using embedded menus to select residues with desired features in the 3D diagrams.

### iGNM UTILITY

**Retrieving and exploring protein dynamics**

iGNM uses several visualization tools to provide both 2D mobility plots and 3D mobility ribbon diagrams of GNM outputs. By typing PDB ID in the 3D Visualization Module (http://ignm.ccb.pitt.edu/3D_GNM.htm), users can view and compare the ribbon diagrams of the query structures color-coded according to the mobilities of residues in the slowest or fastest 20 modes. Likewise, the B-factors Visualization Module (http://ignm.ccb.pitt.edu/BFactors.htm) provides access to ribbon diagrams colored by the mean-square fluctuations predicted and observed for all modes (Figure 5).
iGNM uses Chime and Jmol to visualize 3D protein mobility ribbon diagrams for each mode. Figure 6(a) shows a color-coded ribbon diagram (Chime) that illustrates the mobilities in the slowest GNM mode for carbonic anhydrase (PDB ID: 1ca2). The structure is colored from dark blue, green, orange, to red in the order of increasing mobility in the slowest mode. A Java applet is used to plot 2D mobility graphs [Figure 6(b)]. The user can point the cursor to the positions of interest (minima or maxima) on the graphs to view the corresponding residue index and relative fluctuations. Comparison of experimental and theoretical B factor is also provided [Figure 6(c)]. The cross-correlation map, i.e. the dot products of a pair of residue fluctuations, is generated using Matlab [Figure 6(d)].

**Customized online calculation**

oGNM can routinely update the iGNM database engine by calculating newly deposited PDB structures. Furthermore, it allows users to perform GNM computations for customized structures (e.g. site mutagenesis) or structures not deposited in the PDB (Figure 7).

The user can choose a different cutoff distance $r_c$ of GNM. According to a recent study on the effect of cutoff distance on the coefficient between experimental and theoretical B factors [45], GNM results are rather insensitive to the cutoff distance provided in the range $7.3 \, \text{Å} \leq r_c \leq 15 \, \text{Å}$.

Although pure oligonucleotide (DNA/RNA) structures and protein–oligonucleotide complexes account for only a small fraction ($\sim 0.5\%$) of all the structures in PDB, these structures are involved in

---

**Figure 6:** iGNM visualization examples. (a) Ribbon diagram in cartoon display; (b) 2D mobility diagram; (c) theoretical fluctuations and experimental fluctuations; (d) cross-correlations between residue fluctuations. The highly correlated regions are colored red, and the moderately correlated or anti-correlated regions are blue.
some of the most important subcellular functions, such as gene replication, storage and repair, and it is important to properly model the DNA and RNA components' dynamics in elastic network (EN) calculations. Few of web-based systems on NMA (Section 'Web-based systems for protein dynamics') include a representation for oligonucleotides in their underlying EN model. This makes such systems unable to completely characterize the motions of structures such as the ribosome, DNA polymerases and tRNA synthetases.

In oGNM, the user can choose different models for different structures. A single-node model is used for proteins and a three-node model for nucleotides. A small value for nucleotide–nucleotide cutoff distance, $r_p$, is not an appropriate physical representation of the nucleotide since the distance between the P-atoms of base-paired bases is larger than this cutoff value. It is shown that the optimal correlation for all nodes in the examined structure occurs near a cutoff value comparable to that used for amino acids, $r_c = 7.3$ Å [45].

**Interplay between dynamics and function**

The data contained in iGNM allow users to explore the relationships between protein dynamics and function. For example, a systematic analysis was performed by Yang and Bahar [46] to delineate the coupling between catalysis and conformational mechanics. In their study, slow modes are extracted from iGNM and compared with experimental data for 98 enzymes of different structural subclasses. The results show that the global hinge centers in slow modes (evidenced by low mobility values) tend to be co-localized with the catalytic sites experimentally identified. Thus, chemically active residues are found to also participate in critical sites from a conformational mechanics point of view.

Figure 8 shows one example for β-lactamase class A (1BTL), where the weighted average mobilities for the two slowest modes $\langle (\Delta R)^2 \rangle_{-2}$ are correlated with catalytic residues (S70, K73, S130 and E166). The result shows that the catalytic sites occupy regions that are spatially constrained in general (indicated by lower $\langle (\Delta R)^2 \rangle_{-2}$ values).

Users can also explore the utility of fast modes and cross-correlations between fluctuations in identifying functionally important residues or regions (see the section titled Functional implication of GNM outputs). It is also useful to correlate iGNM normal modes with sequence-derived features such as conservation or hydrophobicity to explore the sequence–structure–function relationships.

**CONCLUSIONS AND FUTURE RESEARCH**

Traditional web-based systems on protein dynamics rely on the high-resolution approach to model protein dynamics. They provide detailed information about protein motions, but their low time performance prevents them from accumulating dynamics information for all known protein structures. GNM, as an efficient coarse-grained model, simplifies traditional normal mode analyses and elastic network models by only considering inter-residue contact
topology in the folded state with single parameter potentials. GNM allows high-throughput analysis of protein structures ranging from enzymes to large complexes and assemblies. iGNM is a web-based system that allows query of calculated GNM results for more than 20,000 protein structures as well as online dynamics calculation of newly deposited PDB structures or user-modified structures. iGNM is not only a data resource for querying single protein structure, but also a knowledge base for the exploration of protein sequence–structure–dynamics–function relations.

There are fundamental issues for developing an efficient and automatic web-based system for protein functional dynamics modeling and analysis. Future research will include relating GNM data to other types of protein data and integrative data mining. Integrating sequence and structure information with GNM data allows for further exploration and establishment of protein sequence–structure–dynamics–function relations. For example, sequential patterns based on N-gram analysis [47, 48] could be integrated with residue fluctuation patterns found in iGNM to detect possible relationships between protein sequences and dynamics. Another example is integrating protein classification databases, such as Pfam (http://www.sanger.ac.uk/Software/Pfam/) [49] and Enzyme Structures Database (http://www.ebi.ac.uk/thornton-srv/databases/enzymes/) with GNM data to study dynamics characteristics for a family of proteins.

Exploration and establishment of protein sequence–structure–dynamics–function relations call for an efficient and automatic architecture for integrating protein data resources, including databases, applications and tools. Currently, protein data resources are accessible to researchers through web application interfaces, e.g. through an HTTP form and a corresponding Java servlet. To integrate data through web applications (current approach), users have to perform such tedious tasks as screen scraping or writing scripts to parse HTML code (to extract data) while ignoring explanatory text and graphics. This approach is labor intensive and fragile, for minor changes in the HTML code of a given web application may cause failure [50]. The web services approach, on the other hand, offers an environment for flexible integration of various bioinformatics data resources [50, 51]. We plan to develop a web-services-based architecture for flexible and expandable integration of heterogeneous, geographically dispersed data resources. Furthermore, we plan to develop chaining algorithms to enable the creation of composite services.

Key Points

- Efficient computational methods and tools are needed for understanding the dynamics, and thereby controlling the functions of proteins.
- Time cost of molecular dynamics simulations has been a major drawback for a systematic characterization of protein dynamics.
- Lower-resolution elastic network models (e.g. the Gaussian Network Model) allow efficient prediction of protein dynamics.
- iGNM is an efficient web-based system for querying, computing and visualizing protein dynamics.

Acknowledgments

The authors would like to thank Dr. Lee-Wei Yang for his feedback on the first draft of this article. This research was funded by a grant from the National Institute of Health (grant no.1 R33 GM068400-01A2).

References


