Study of the Variability of the Native Protein Structure

Xusi Han, Woong-Hee Shin, Charles W Christoffer, Genki Terashi, Lyman Monroe, and Daisuke Kihara, Purdue University, West Lafayette, IN, United States

Introduction

Proteins are flexible molecules. After being translated from a messenger RNA by a ribosome, a protein folds into its native structure (the structure of lowest free energy), which is suitable for carrying out its biological function. Although the native structure of a protein is stabilized by physical interactions of atoms including hydrogen bonds, disulfide bonds between cysteine residues, van der Waals interactions, electrostatic interactions, and solvation (interactions with solvent), the structure still admits flexible motions. Motions include those of side-chains, and some parts of main-chains, especially regions that do not form the secondary structures, which are often called loop regions. In many cases, the flexibility of proteins plays an important or essential role in the biological functions of the proteins. For example, for some enzymes, such as triosephosphate isomerase (Derreumaux and Schlick, 1998), loop regions that exist in vicinity of active (i.e., enzymatic reaction) sites, takes part in binding and holding a ligand molecule. Transporters, such as maltose transporter (Chen, 2013), are known to make large open-close motions to transfer ligand molecules across the cellular membrane. For many motor proteins, such as myosin V that "walk" along actin filament as observed in muscle contraction (Kodera and Ando, 2014), flexibility is the central for their functions.

Reflecting such intrinsic flexibility of protein structures, differences are observed in protein tertiary structures determined by experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR) when they are solved under different conditions. In this article, we start by introducing two studies that surveyed such structural differences of the same proteins found in the public repository of protein structures, Protein Data Bank (PDB) (Berman *et al.*, 2000), which contains over 136,000 entries at the time of writing this article (December 2017). Thus, structural variability of a protein can be observed by comparing static structures determined under different conditions. Experimentally, conformational variability can be measured by NMR and other spectroscopic techniques (Greenleaf *et al.*, 2007; Parak, 2003; Hinterdorfer and Dufrene, 2006). Alternatively, structural changes, i.e., flexibility, can be observed for a single protein structure by performing computational simulations or predictions of dynamics of the protein structure. Computational methods for elucidating protein structure flexibility are also useful for estimating the free energy of molecular interactions as well as structure refinement needed when determining protein tertiary structures. In this article we overview such computational tools with some examples of application.

Discussion in this article is focused on protein structures that have overall stable fixed structures (with some flexibility in side-chains and parts of the main-chain, e.g., loops). However, note that there is a different class of proteins that do not form stable structures at all under physiological conditions. Such proteins are called intrinsically disordered proteins (IDPs) (Dunker *et al.*, 2008). IDPs were not paid much attention for a long time since their structures cannot be determined by regular experimental structural biology methods, such as X-ray crystallography or NMR, due to their intrinsic flexibility. But from around 2000, IDPs attracted large attention as long-neglected protein structures, and have been studied extensively since then. IDPs have characteristic amino acid sequence patterns, from which IDPs can be predicted from their amino acid sequences (Ferron *et al.*, 2006). It is estimated that about 5%–30% of amino acid sequences in an organism's proteome are intrinsically disordered (Oates *et al.*, 2013). It was found that disordered regions are often responsible for establishing protein-protein interactions, especially for proteins that interact with multiple proteins. The D2P2 database (Oates *et al.*, 2013) provides predicted IDPs in over 1700 genomes by many existing IDP prediction methods.

Fundamentals

Conformational Transitions Observed in Experimentally Determined Structures

Structural variability of proteins can be observed by comparing structures that are experimentally determined in different conditions. Kosloff and Kolodny (2008) performed a systematic study on protein structure variability relative to the sequence identity including cases that two structures are 100% identical in their sequences. The dataset was comprised of 19,295 protein chains taken from the April 2005 PDB, which are longer than 35 residues and were determined at a resolution of 2.5 Å or higher using X-ray crystallography. To eliminate redundant protein chains in the dataset, no pairs have 100% sequence identity and less than 1 Å root-mean square deviation (RMSD) between each other. Out of the 1941 chain pairs with 100% sequence identity in this dataset, there were 444 (22.9%) pairs with an RMSD of 3 Å or larger and 158 (8.1%) pairs with an RMSD over 6 Å. Such cases included chain pairs that have very different structures of over a 10 Å RMSD. They classified the sources of structural dissimilarity, which included different quaternary protein-protein interactions, protein-ligand and protein-DNA/RNA interactions, different crystallization conditions such as pH or salt conditions, and alternative crystallographic conformations of the same proteins. Thus, as is also discussed in other studies (Goh *et al.*, 2004; Boehr *et al.*, 2009), one of the main reasons for the existence of alternative conformations is due to physical interactions with other molecules. Most large conformational changes observed in protein structures are inherent to their biological functions, as Gan *et al.* (2002) pointed out in their work.

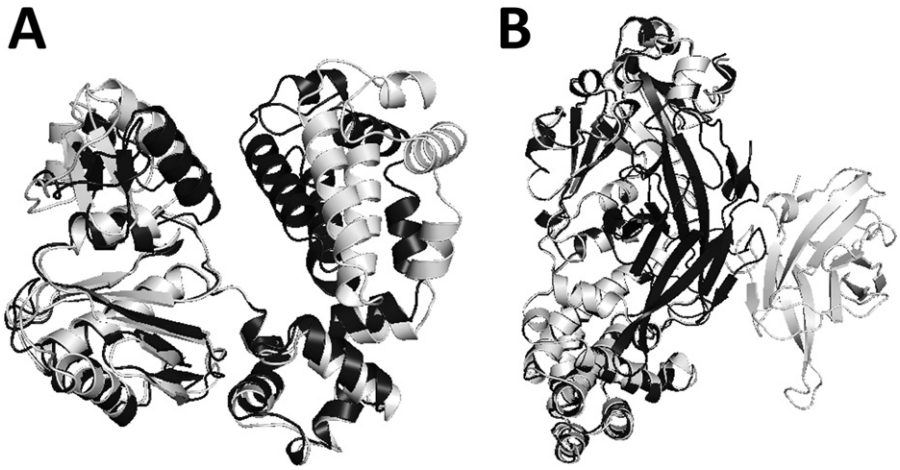


Fig. 1 Superimpositions of protein structures with a large conformational change. (A) Structure of DNA polymerase β in open (PDB ID: 9ICX, chain A, colored in light grey) and closed (PDB: 1BPY, chain A, colored in black) conformations. DNA in the crystal structures are omitted in the figure. (B) Structure of diphtheria toxin in open (PDB: 1DDT, chain A, colored in light grey) and closed (PDB: 1MDT, chain A, colored in black) conformations.

We show several examples of large conformational change of proteins. The first example is human DNA polymerase β in open and closed conformations (**Fig. 1(A)**) (Pelletier *et al.*, 1996; Sawaya *et al.*, 1997). The RMSD between the two structures is 5.3 Å. This conformational change is needed for its biological function, catalytic polymerization of nucleotides as well as binding and releasing DNA. The second example is diphtheria toxin in open and closed conformations (**Fig. 1(B)**). The RMSD is 11.0 Å (Bennett and Eisenberg, 1994). This protein has three domains, the catalytic domain, the transmembrane domain, and the C-terminus receptor binding domain. In the two structures superimposed in **Fig. 1(B)**, the arrangement of these three domains is very different. Particularly, the receptor domain shown on the right side of the figure is distant from the rest of the structure in the open conformation. This drastic conformational change is speculated to be essential for penetrating the host cell membrane, a critical step for intoxication.

Fig. 2 shows three more examples where binding with other molecule is involved in the conformational change. **Fig. 2(A)** shows structure of export chaperone FliS in *Aquifex aeolicus* (Evdokimov *et al.*, 2003). FliS acts as a flagellar export chaperone, which binds specifically to FliC to regulate its export and assembly process. The structure of unbound FliS is an antiparallel four-helix bundle (**Fig. 2(A)**, Left). The N-terminal cap blocks the hydrophobic binding site in FliS when FliC is absent. Upon complex formation with FliC, there is a substantial conformation change in FliS (**Fig. 2(A)**, Right), where the N-terminus in FliS displaces to form a short helix on one side of the bundle (on the right side of the panel), interacting with one helical segment in FliC. The RMSD between the two conformation is 6.4 Å. The next one is a textbook example, calmodulin (**Fig. 2(B)**). Calmodulin undergoes a large conformational change between the apo (i.e., non-ligand binding) state and the calcium-binding state (Yamniuk and Vogel, 2004). In calcium-free calmodulin, each globular domain is made up of four helices running in parallel/antiparallel orientations to each other (**Fig. 2(B)**, Left). In the calcium-bound form, the interdomain region forms a long α-helix (the right panel) instead of two short α-helices in the apo form. This helical rearrangement exposes hydrophobic binding patches on the surface of each domain, which binds to peptide sequences in target enzymes. The RMSD between the two structures is 12.9 Å. The last example in **Fig. 2** is RfaH, a member of a universally conserved family of transcription factors (**Fig. 2(C)**). In its closed form, RfaH C-terminal domain (CTD) forms an α hairpin that masks an RNA Polymerase binding site in RfaH N-terminal domain (NTD) (Belogurov *et al.*, 2007) (**Fig. 2(C)**, Left, the light grey region). But upon release from RfaH-NTD, RfaH-CTD refolds into a β barrel that binds to the ribosome to activate translation (the right panel) (Burmann *et al.*, 2012). The RMSD between the two structures of the RfaH-CTD is 14.0 Å. The dramatic switch from α helix to β barrel transforms the function of RfaH from transcription factor to translation factor.

Molecular Dynamics

From this subsection, we introduce several computational methods that can simulate or model structural variability of proteins.

Molecular dynamics (MD) simulation is a computer simulation technique to investigate the movement of atoms and molecules based on the principles of physics. Since its first application for biomolecules published 40 years ago (Mccammon *et al.*, 1977), it has become a popular and standard tool for studying biomolecular motion. The basic concepts of MD are (1) to divide time into discrete time steps (1 or 2 fs, generally) and (2) to solve Newton's equations of motion (Eq. (1)) at every time step:

$$F(\mathbf{x}) = -\nabla U(\mathbf{x}) = m\frac{d^2\mathbf{x}}{dt^2} \tag{1}$$

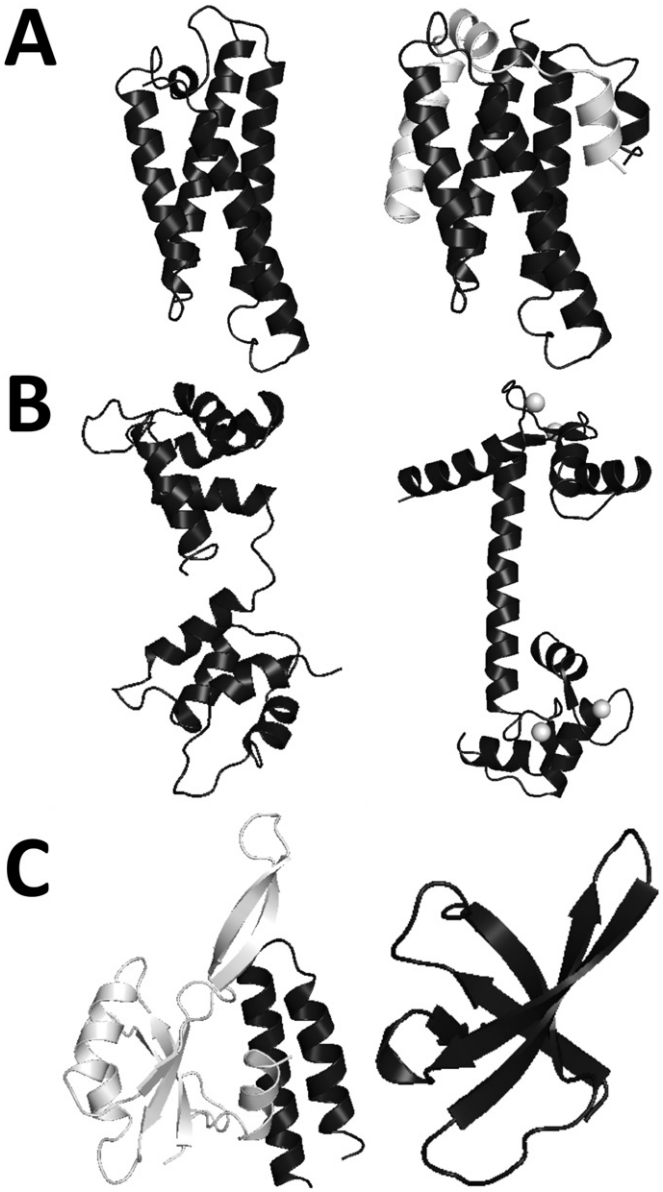


Fig. 2 Conformational transitions upon protein-ligand interactions. (A) Structure of FliS in unbound state (PDB: 1ORJ, chain A, left panel) and bound state (PDB :1ORY, right panel). FliC is colored in light grey in the FliS-FliC complex. (B) Structure of calmodulin in the unbound state (PDB: 1DMO, chain A) and the calcium-binding state (PDB: 1CLL, chain A). Calcium ions in the crystal structure are shown as spheres colored in light grey on the right panel. (C) Structure of full-length RfaH in closed form (PDB: 2OUG, chain A) and RfaH C terminal domain in open form (PDB 2LCL, chain A). RfaH N terminal domain is colored in light grey in the closed state (left).

$\mathbf{F}(\mathbf{x})$, $\mathbf{U}(\mathbf{x})$, $\mathbf{x}$, and m are a force, a potential energy, a coordinate of an atom, and a mass of an atom, respectively. Since the system is composed of a number of atoms, Eq. (1) cannot be solved analytically. Therefore, the trajectory of an atom can be obtained by integrating the potential energy at every time step numerically. The most famous algorithm for the numerical integration is Verlet algorithm (Verlet, 1967). A potential energy acting on an atom, $\mathbf{U}(\mathbf{x})$, is calculated by a force field, which is composed of a functional form representing each force terms. Basically, the force field is composed of bonded and nonbonded terms (Eq. (2)).

$$U_{total} = U_{Bonded} + U_{Nonbonded} \tag{2}$$

A bonded term is a pairwise interaction energy of atoms that are connected by covalent bonds. It is composed of four terms; bond, angle, torsion, and improper torsion (**Fig. 3**, Eq. (3)). Different from the previous three terms that work on all atoms, improper torsion term only acts on special sets of atoms such as a peptide bond and benzene and prevents a deviation from planarity of the atom sets.

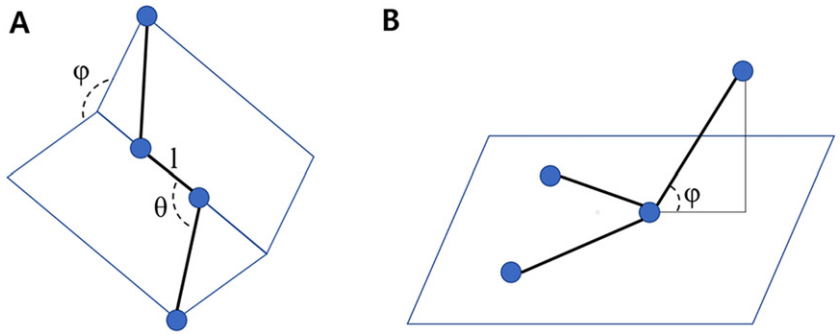


Fig. 3 Definitions of angles in the force field. (A) The bond length *l*, bond angle θ and the torsion angle, (φ). (B) Improper torsion.

$$U_{Bonded} = U_{Bond} + U_{angle} + U_{Torsion} + U_{Improper}$$
$$= \sum_{Bonds} \frac{k_{bi}}{2}(l - l_{i,eq})^2 + \sum_{Angles} \frac{k_{\theta i}}{2}(\theta_i - \theta_{i,eq})^2 + \sum_{Torsions} \sum_{n} \frac{V_{n,i}}{2}\left[1 + cos\left(n(\varphi)_i - (\varphi)_{i,0}\right)\right] + \sum_{Improper} \frac{k_{(\varphi)i}}{2}\left((\varphi)_i - (\varphi)_{i,eq}\right)^2 \quad (3)$$

The bond and the angle terms reflect vibration of bond lengths and angles along with covalent bonds. l and θ are the bond length and bond angle at the current state (**Fig. 3(A)**). The parameters, k_b, l_{eq}, k_θ, and θ_{eq}, are force constants and values of a bond in the equilibrium state. They are derived from gas phase spectroscopy and/or crystal structures of small molecules. The functional forms of the terms are approximated as harmonic oscillators. The torsion term calculates bond energies associated with proper torsion angles, (φ) in **Fig. 3(A)**. Since it is periodic, it has a cosine functional form, and the parameters, V_n and $(\varphi)_0$ are obtained from quantum mechanics calculation. The improper torsion has a functional form of a harmonic oscillator (**Fig. 3(B)**).

A nonbonded term describes the long-range interaction of atoms, which consists of van der Waals and Coulomb potentials:

$$U_{Nonbonded} = U_{van\ der\ Waals} + U_{Coulomb} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left\{ 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \right\} \quad (4)$$

In the van der Waals term, σ_{ij} and r_{ij} are the radius and well depth, respectively. They are calculated by taking the arithmetic mean ($\sigma_{ij} = (\sigma_{ii} + \sigma_{jj})/2$) and the geometric mean ($\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\ \varepsilon_{jj}}$) of parameters of the di-atomic system. For the Coulomb term, q_i is an atomic charge of atom i, and ε is a dielectric constant that reflects a solvent polarity (e.g., $\varepsilon = 1$ for vacuum, $\varepsilon = 4$ for protein, and $\varepsilon = 80$ for water in implicit solvent model).

Most biological reactions take place in a cell, surrounded by water molecules; thus it is important to consider the effects of water. In MD and other simulation methods, water molecules can be treated in two ways; an implicit solvation and an explicit solvation model. An implicit solvation model (**Fig. 4(A)**), which is also called continuum solvent, considers solvent as a continuous isotropic medium. It approximates solvent molecules as a homogeneously polarizable, averaged medium (Skyner *et al.*, 2015). The main parameter for the continuum model is a dielectric constant ε, and supplementary parameters are used to describe solvent properties such as surface tension. Atoms in solution are represented as spherical cavities with atomic charges embedded in homogeneously polarizable solvent (**Fig. 4(A)**). An explicit solvation model (**Fig. 4(B)**) treats water molecules explicitly. This is a more realistic model than the implicit solvation model, since physical interaction between a solute and solvent molecules can be directly calculated from the same functional form of the force field. However, the number of atoms in the system is increased drastically in an explicit solvation model; therefore, it takes much more time to calculate potential energy of the solute atoms and trajectory of them. The most famous model of explicit water is TIP3P (Jorgensen *et al.*, 1983).

Coarse-Grained Protein Structure Models

If the size of a protein to simulate is very large, a MD simulation for a biologically relevant time span with an atomistic representation may be computationally infeasible. One way to solve this issue is to represent a protein structure with a simplified (coarse-grained, CG) model. In a coarse-grained model, the complexity of the system is reduced by grouping atoms together into pseudo particles (Barnoud and Monticelli, 2015). Reducing the number of particles decreases the number of interactions to calculate and makes the energy landscape smoother, both of which contribute to shortening the computational time for simulation compared to an atomistic model. The functional form of a CG force field is similar to that for an atomistic force field. The parameters of a CG force field are determined to be able to produce similar trajectories as atomistic MD simulation.

One of the well-known CG force fields is MARTINI (Marrink *et al.*, 2007). Generally, the MARTINI force field maps a group of four heavy atoms to one interaction site (bead). The beads of amino acids are mainly classified into four; polar, nonpolar, apolar, and charged. These are further categorized into subclasses based on hydrogen-bonding capabilities and polarity, resulting in 18 types of beads. Comparing to an atomistic simulation, the MARTINI force field yields a speed-up of 2–3 orders of magnitude. Other CG protein models include UNRES (United RESidues) (Liwo *et al.*, 2004, 2005) and CABS (C-Alpha, c-Beta, Side-chain)

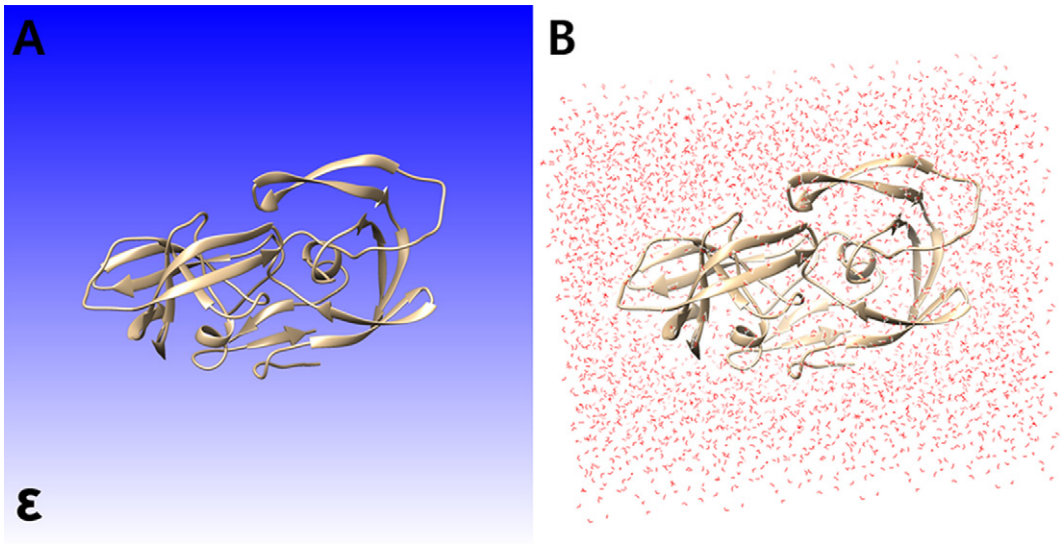


Fig. 4 Two solvent models. (A) The implicit solvent model, where solvent is represented as a dielectric constant of ε. (B) The explicit solvent model. A water molecule is represented as a bent line with oxygen and hydrogen atoms colored in red and white, respectively. Interactions between atoms in water molecules and proteins are computed in the same way using the equations of the force field used.

(Kolinski, 2004). In the UNRES model, each amino acid residue is represented by only two interaction sites, an ellipsoid at the side-chain centroid and another one at the center of Cα-Cα bond. The force field of UNRES is derived as a restricted free energy function, which considers averaging the all-atom energy over the degrees of freedom that are neglected in the UNRES model. Thus, UNIRES is a physics-based united residue force field. In the CABS model, each amino acid residue is represented by four interaction centers; the Cα atom, the center of Cα-Cα bond, the Cβ atom, and the center of mass of the side-group. To facilitate faster simulations, Cα atoms in the CABS model are projected onto a cubic lattice with 0.61 Å spacing. The force field of CABS is based on statistical potentials that mimic averaged interactions observed in globular protein structures in PDB.

Elastic Network Models

As discussed in Section "Coarse-Grained Protein Structure Models", coarse-grained models of protein dynamics are used to reduce the computational cost of analysis. Some of the simplest such models are elastic network models (ENMs). In an ENM, pairwise interactions in a structure are modelled as Hookean springs at equilibrium, i.e., a model where beads representing interaction centers (e.g., amino acid residues) are connected by springs. In other words, the potential energy function has the form $\sum \frac{1}{2}\gamma_{ij}\left(r_{ij}-r_{ij}^{0}\right)$, where γ_{ij} are positive constants characterizing the stiffness of the interaction and r_{ij}^{0} is the initial distance between atoms i and j; in contrast to the general form of MD potentials given in Section "Coarse-Grained Protein Structure Models", ENMs effectively model all interactions as bonds. Tirion considered an all-atom elastic network model with all γ_{ij} identical and atom pairs interacting if their distance was less than the sum of their van der Waals radii and an arbitrary cutoff (Tirion, 1996). That study established that normal mode analysis (NMA), which decomposes the general motion of a system into linearly independent modes of harmonic motion, of such potentials could reproduce the density of slow normal modes of more complicated and costly potentials. Hinsen showed that the slow modes of ENMs can in fact reproduce large-scale deformation, even when only α carbons were considered (Hinsen, 1998). Such coarse ENMs with a single point mass per residue have become popular; in the following sections, we discuss two such models.

Anisotropic network model

Although not chronologically the first residue-level ENM, the anisotropic network model (ANM) is the more general of the two we discuss here. Introduced by Doruker *et al.* (2000) and expounded by Atilgan *et al.* (2001) and Eyal *et al.* (2006), the ANM potential has all γ_{ij} identical and all interaction cutoffs identical (e.g., 7 Å). Here, the NMA is performed with respect to the Cartesian coordinates of each α carbon; i.e., the Hessian of the potential has the block matrix form

$$H=\begin{bmatrix} H_{ii} & \cdots & H_{ij} \\ \vdots & \ddots & \vdots \\ H_{ji} & \cdots & H_{jj} \end{bmatrix} \tag{5}$$

where each block expands to

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 U_{ij}}{\partial x_i \partial x_j} & \frac{\partial^2 U_{ij}}{\partial x_i \partial y_j} & \frac{\partial^2 U_{ij}}{\partial x_i \partial z_j} \\ \frac{\partial^2 U_{ij}}{\partial y_i \partial x_j} & \frac{\partial^2 U_{ij}}{\partial y_i \partial y_j} & \frac{\partial^2 U_{ij}}{\partial y_i \partial z_j} \\ \frac{\partial^2 U_{ij}}{\partial z_i \partial x_j} & \frac{\partial^2 U_{ij}}{\partial z_i \partial y_j} & \frac{\partial^2 U_{ij}}{\partial z_i \partial z_j} \end{bmatrix} \tag{6}$$

Computing the eigenpairs of H yields $3N-6$ eigenpairs corresponding to internal modes. The eigenvalues λ_k and frequencies ω_k of the modes are related by

$$\omega_k^2 = \gamma \lambda_k \tag{7}$$

Although H is not invertible, we can construct a pseudo-inverse

$$H^{-1} = \sum_{k=7}^{3N} \frac{u_k u_k^T}{\lambda_k} \tag{8}$$

where u_k are eigenvectors and eigenpairs are in increasing order by eigenvalue magnitude. Cross-correlations between the equilibrium fluctuations of residues i and j are then given by

$$C_{ij} = \frac{tr\left(H_{ij}^{-1}\right)}{\sqrt{tr\left(H_{ii}^{-1}\right) tr\left(H_{jj}^{-1}\right)}} \tag{9}$$

where H_{ij}^{-1} is the (i,j)th 3×3 block of H^{-1}. The B-factor of atom i, defined as $8\pi^2$ times the mean squared displacement of atom i,

$$B_i = \frac{8\pi^2 k_B T}{3\gamma} tr\left(H_{ii}^{-1}\right) \tag{10}$$

where k_B is the Boltzmann constant and T is the temperature.

ANM has been used to study the dynamics of biomolecules that are related to their biological functions (Navizet *et al.*, 2004; Keskin *et al.*, 2002).

Gaussian network model

Compared to ANM, which can evaluate directional preferences of vibrational dynamics of proteins, Gaussian Network Model (GNM) only provides displacements and correlations between displacements of amino acid residues (Bahar *et al.*, 1997). Mathematically, any information obtained from GNM can also be obtained from ANM; however, if only isotropic information is needed, GNM is computationally cheaper. To calculate the isotropic cross-correlations and B-factors under a GNM, we start from the graph Laplacian Γ of the graph where vertices correspond to atoms and edges correspond to interactions. Then, the cross-correlations are given by

$$C_{ij} = \frac{\Gamma_{ij}^{-1}}{\sqrt{\Gamma_{ii}^{-1} \Gamma_{jj}^{-1}}} \tag{10}$$

where Γ^{-1} is a pseudoinverse as in Section "Anisotropic network model", and the B-factors are given by

$$B_i = \frac{8\pi^2 k_B T}{3\gamma} \Gamma_{ii}^{-1} \tag{11}$$

Residue fluctuations computed from GNM have been shown to agree with the X-ray crystallographic B-factor, which reflects fluctuation of each atom at its position in the crystal structure (Bahar *et al.*, 1998). GNM has also been used for characterizing functional motions of proteins (Yang and Bahar, 2005) and macromolecules, such as ribosome (Wang *et al.*, 2004).

Flexpred

Variability of positions of residues in a protein structure can be also predicted from structural features of amino acids. Jamroz *et al.* (2012) developed FlexPred, which uses support vector regression (SVR) to predict residue fluctuations observed in 10 ns MD simulations, with a mean error in tests of 1.1 Å. Fluctuation is defined as

$$\sqrt{\langle (\Delta R_i)^2 \rangle}^{\text{ref}} = \sqrt{\frac{1}{T} \sum_{t_j=1}^{T} \left(x_i(t_j) - x_i^{\text{ref}}\right)^2} \tag{12}$$

where T is the number of frames in the MD trajectory, $x_i(t_j)$ is the position of the α carbon of residue i at time t_j, and x_i^{ref} is the position of the α carbon of residue i in the protein structure. Structural features for each α carbon/residue that were found to be useful as input to SVR include distance to the structure's center of mass, the number of surrounding residues that are in contact, the

number of hydrophobic/hydrophilic contacts, solvent-accessible surface area, residue depth (the distance of the residue from protein surface), and secondary structure type.

Application and Case Studies

In this section we provide available software for analysing protein structure variability with some example applications.

Examples of Structural Variability in Protein Families

Structural variation of a protein can be quantified by superimposing the structures. Structure superimposition can help us reveal conserved regions among the structures which may be of functional importance as well as structure divergence and flexibility.

There have been a variety of multiple structure alignment methods developed over the years, which employ different scoring functions and different heuristics. Two examples are MAMMOTH-mult (Lupyan *et al.*, 2005) (see "Relevant Websites section") and POSA (Ye and Godzik, 2005) (see "Relevant Websites section"). Both methods first conducts all pairwise structure alignments and constructs a dendrogram from pairwise similarity scores. Then, following the dendrogram, pairwise structure alignment is performed from leaf nodes, and finally all the structures are superimposed. This superimposition then undergoes an iterative refinement process (progressive alignment).

Multiple structure alignment is useful for understanding structural variability of single protein as well as structures of a protein family. **Fig. 5** shows such an example, an alignment of three lectins, which have the legume lectin-like fold. Lectins are a group of carbohydrate binding proteins with large variation in size and structure. Characteristics of the lectin fold is the presence of a two

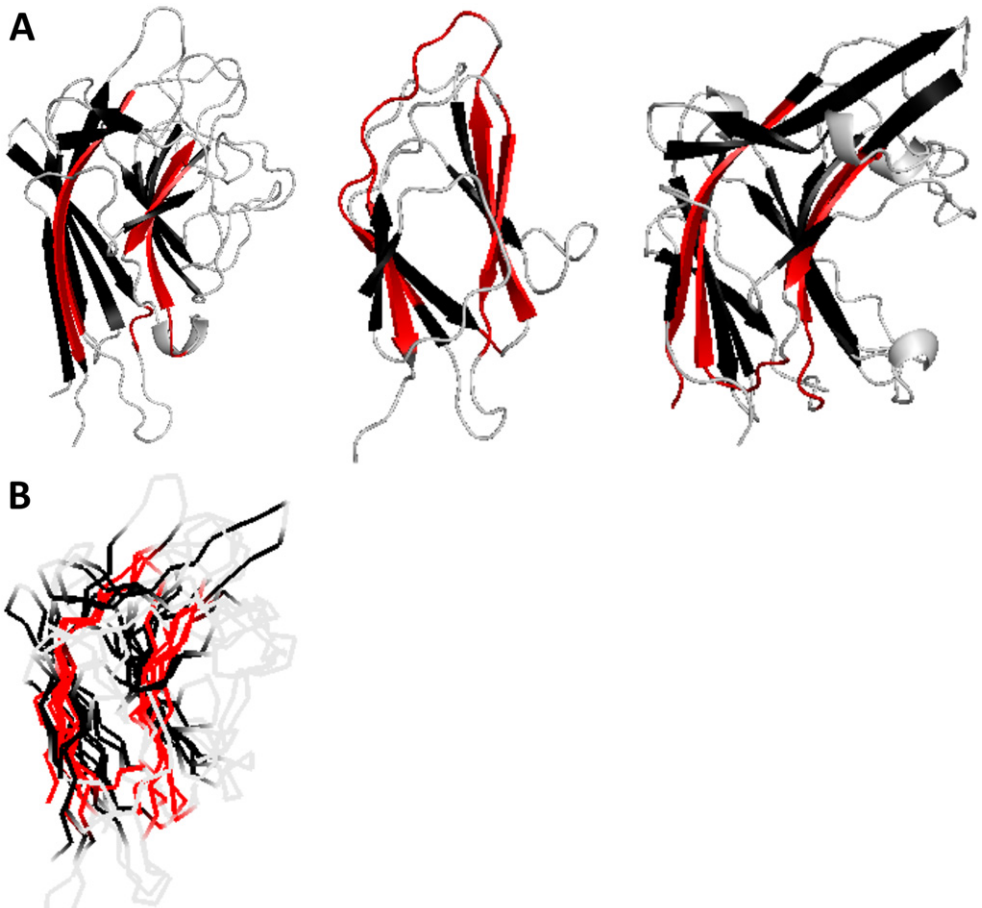


Fig. 5 Structure variation of proteins containing lectin fold. (A) Ribbon diagrams of three structures containing legume lectin-like fold. From left to right: Peanut lectin (PDB: 2PEL, chain A), Spermadhesin (PDB: 1SPP, chain A), Glucanase (PDB: 2AYH, chain A). All structures all shown in the same orientation. Common core identified in the POSA alignment are colored in red. Residues in common core are: residues 44–48, 51–53, 64–69, 200–209, 217–228 in 2pel-A; residues 28–35, 44–49, 85, 87–95, 97–108 in 1spp-A; residues 62–66, 69–77, 175–184, 203–214 in 2ayh-A. Beta sheets are colored in black. Helices and loops are colored in grey. (B) Superimposition of all structures. Structure alignments were produced using POSA.

Table 1 Popular molecular dynamics software

Name	*Website*
Amber	ambermd.org
CHARMM	www.charmm.org
GROMACS	gromacs.org
NAMD	www.ks.uiuc.edu/Research/namd
Desmond	www.deshawresearch.com/resources_desmond.html

β-sheets positioned almost in parallel (Chandra *et al.*, 2001). Although all structures with lectin fold possess the β-sheets with the particular geometry, curvature of a sheet and the presence of binding site loops and hydrophobic cores vary among those structures (**Fig. 5(A)**). The multiple structure alignment by POSA overlays β-sheets of the three proteins and particularly reveals that all three proteins possess a common core of 36 amino acids, three β strands shown in red in **Fig. 5(B)**. The average RMSD between three structure pairs is 3.0 Å. Compared with peanut lectin (Left in **Fig. 5(A)**), spermadhesin (Middle) and glucanase (Right) have an RMSD of 2.88 Å (for aligned 80 residues) and an RMSD of 3.02 Å (for aligned 172 residues), respectively.

Software for Molecular Dynamics Simulations

MD can be performed using several program packages to simulate protein flexibility. **Table 1** lists popular MD simulation programs. All of the programs are free for academic users. The first MD package, Amber (Assisted Model Building with Energy Refinement) (Case *et al.*, 2005), was originally developed for refining NMR structures. The name 'Amber' refers to both force fields for the simulation of biomolecules and the MD program package. The most widely used force field versions are Amber94, Amber99SB, and Amber03. The next one, CHARMM (Chemistry at HARvard Macromolecular Mechanics) (Brooks *et al.*, 2009), also refers to both force fields and the program package. It was originally developed by Martin Karplus of Harvard University, USA. It is the oldest biomolecular MD package listed in **Table 1**. CHARMM-GUI (see "Relevant Website section") provides a web-based graphical tools for setting up input files for the simulation. GROMACS (GROningen MAchine for Chemical Simulations) (Pronk *et al.*, 2013) typically runs 3–10 times faster than other MD programs. Unlike Amber and CHARMM, GROMACS does not have its own force field. Instead, it can import Amber, CHARMM, GROMOS, and the OPLS force field to run MD simulation. The unique feature of the program is that it is an open-source software released under the GPL license. NAMD (NAnoscale Molecular Dynamics) is developed by the Theoretical and Computational Biophysics Group at the University of Illinois, Urbana-Champaign, USA (Phillips *et al.*, 2005). It is designed to efficiently run on parallel machines for simulating large molecules. The program has high compatibility with other MD programs such as CHARMM, since NAMD has same input, output, and force field formats as CHARMM.

Desmond is developed by D.E. Shaw Research (Bowers *et al.*, 2006). The program uses its novel parallel algorithms and numerical methods to achieve high computing performance. Desmond can import AMBER, CHARMM, and the OPLS force field to run MD simulation.

MD-Based Protein Structure Model Refinement

From Sections "MD-Based Protein Structure Model Refinement", "Refinement of Structures from Electron Microscopy Data", "Protein Folding" and "Free Energy Calculation" we will discuss notable applications of MD simulation for addressing protein structure variability. The first one is structure refinement for computationally modelled protein structure models. Protein structures can be computationally modelled with a reasonable accuracy if a structure of a related protein is already solved and available in PDB. The protein structure modelling technique that uses known structures as a template is called homology modelling or comparative modelling (Fiser, 2010). However, structure models usually still have deviation from the correct (native) structure of the protein, which may be as subtle as side-chain orientations to as large as a domain or loop motion relative to the native structure. The basic assumption of using MD for refining a structure model is that the conformational ensemble generated by MD simulation from the model contains a better structure that is closer to the native structure. The approach can also be applied to structure models built from low-resolution experimental data such as low-resolution X-ray crystallography (Mcgreevy *et al.*, 2014) and cryo-electron microscopy (cryo-EM) density maps (Singharoy *et al.*, 2016).

In the field of protein structure prediction, It has been recognized that running MD simulations for a structure model does not improve a model consistently, but rather drifts the structure away from the native structure because the conformational space is very large. However, the situation has changed in 2012, when MD-based methods were developed that refined models consistently in a protein structure prediction contest, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Nugent *et al.*, 2014). Since then several MD-based structure refinement methods have been developed. These methods typically run MD with constraints so that the overall structure does not change drastically. Also the methods combine a structural averaging step and cross-check structures with additional structure evaluation scores (Mirjalili and Feig, 2013; Lee *et al.*, 2018; Terashi and Kihara, 2018). A drawback of the current methods is that the structure improvement they can achieve is small due to the constraints applied in the simulation, which do not allow large conformational changes in the model.

Refinement of Structures From Electron Microscopy Data

MD is also used for refining protein structures derived from low-resolution experimental data. The experimental data referred to here is the three-dimensional reconstruction of electron microscopy (EM) data, commonly referred to as EM maps. The fundamental methodology for combining MD approaches and experimental data is to perform the MD with an additional metric meant to measure the quality of fit between the atomic model and the EM map.

One method for measuring quality of fit is to convert the EM map data into a potential energy landscape. In this methodology, high density regions of the EM map will have low energy penalties for atoms to occupy that space, while low density regions will have higher energy penalties for atoms to occupy that space. The energy term is then added to the rest of the terms in an MD simulation (Eq. (3)). The most popular method which applies a quality-of-fit metric in this way is Molecular Dynamics Flexible Fitting (MDFF) (Singharoy *et al.*, 2016). An alternative metric for fit quality is cross correlation. The cross correlation between an atomic model and an EM map is determined by generating a simulated density map from the atomic model and determining the similarity of the simulated and experimental maps. A popular method which implements this kind of quality of fit metric is ROSETTA (Dimaio *et al.*, 2015). It is important when implementing these kinds of refinement techniques to remember that the refinement can only be as good as the experimental data allows. As the resolution of EM maps decreases, the extent to which multiple refinement techniques agree with each other will tend to decrease (Monroe *et al.*, 2017).

Protein Folding

Many proteins fold over a time scale on the order of millisecond or longer. Some proteins in the low millisecond range are Trp cage and Villin headpiece with sizes of 20 and 35 residues, respectively. These two α-helical proteins have successfully been simulated from unfolded states to folded states in simulations lasting 10 ns to 1 ms (Simmerling *et al.*, 2002; Duan and Kollman, 1998), and achieving Cα RMSDs of 0.97 and 3.0 Å, respectively. On an even larger scale, the folding and unfolding of ubiquitin, a 76 residue $\alpha\beta$ protein, has been studied through high temperature, 1 ms simulations to an RMSD of 0.5 Å to the crystal structure of the protein (Piana *et al.*, 2013).

Free Energy Calculation

One of the most interesting applications of MD simulation is free energy calculation of a system. From a statistical thermodynamics point of view, free energy calculation is based on the ergodicity hypothesis, which claims that all possible microstates in the phase space can be covered by a trajectory of a system in a long time. With this hypothesis, we can assume that the average of an observable, such as enthalpy, over time is the same as its statistical ensemble. One of the classical example of this application is a protein folding energy landscape (Lei *et al.*, 2007).

Another useful application of free energy calculation is protein-ligand binding free energy (ΔG_{bind}) using The Molecular Mechanics energies combined with the Poisson–Boltzmann and Surface Area continuum solvation (MM/PBSA) method (Genheden and Ryde, 2015). MM/PBSA starts by generating MD trajectories of a protein-ligand complex, a protein, and a ligand. From the trajectories, snapshots of the system are picked up with a certain time interval. The Gibbs free energy of a molecule at a snapshot is calculated as:

$$G = E_{Bond} + E_{vdW} + E_{Coul} + G_{pol} + G_{np} - TS \tag{13}$$

E_{Bond}, E_{vdW}, and E_{Coul} are bonded term, van der Waals, and Coulomb force of a molecule, respectively. The three terms are calculated by the molecular mechanics force field (MM). G_{pol} and G_{np} are solvation energies of polar and nonpolar atoms of a molecule, obtained by solving the Poisson-Boltzmann equation for polar atoms and is proportional to surface area for nonpolar atoms. Although the MD trajectory is obtained from explicit solvent condition, the solvation energy is calculated with implicit solvation (PBSA). The last term, TS in Eq. (13) is entropy, calculated from normal mode analysis, multiplied by absolute temperature.

By taking average of Gibbs free energy across snapshots, binding free energy is calculated as follows:

$$\Delta G_{bind} = \langle G_{Protein-Ligand} \rangle - \langle G_{Protein} \rangle - \langle G_{Ligand} \rangle \tag{14}$$

$\langle Gi \rangle$ is an average of Gibbs free energy of a molecule *i*. The MM/PBSA has been successfully applied to estimating binding free energies of drugs (Zoete *et al.*, 2003; Pearlman, 2005).

Protein Flexibility Prediction With Flexpred

As discussed in Section "FlexPred", flexibility of a protein structure can be predicted from structural features of amino acids in the structure. The method, FlexPred is available as a web server (see "Relevant Websites section") and downloadable software (Peterson *et al.*, 2017). **Fig. 6** shows an example of prediction for bovine angiogenin (PDB: 1AGI, 125 residues long). The plot shows flexibility, how far in Euclidean distance (Å) each amino acid moves on average (in 10 ns MD simulation) along the protein chain (Eq. (12)). At the bottom of the page, there are links to download the fluctuation predictions in CSV form and in the B-factor field of a PDB file. In this example, the predicted fluctuations and crystallographic B-factors correlate with correlation coefficient 0.82. **Fig. 6(B)** shows a side-by-side comparison of the predicted fluctuations (Left) and the B-factors (Right).

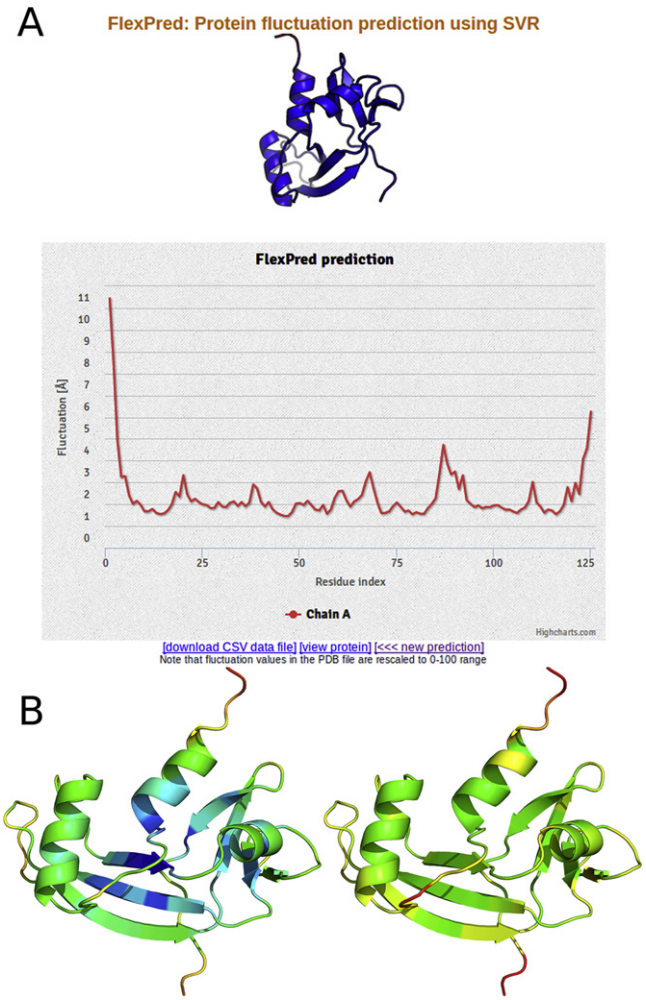


Fig. 6 Example of protein structure flexibility prediction. FlexPred was used. (A) Predicted fluctuation of residues of bovine angiogenin (PDB ID: 1AGI). The y-axis shows the fluctuation (Eq. (12)) of each residue. (B) Predicted fluctuation mapped on residues in the PDB file. Values are stored in the B-factor field of the PDB file. Left, prediction; Right, structure colored by crystallographic B-factor of 1AGI.

Conclusion

Proteins are flexible molecules. Some motions are important for biological function of proteins. How the structure of a protein varies can be obtained if the protein structure is solved under different conditions by experimental methods. Variability of a structure can be also experimentally measured by using NMR or spectroscopic techniques. Computationally, structure variability can be simulated by atomistic MD or coarse-grained models, and can also be predicted from the structure. Applications of MDs, multiple structure alignment, and flexibility prediction are shown.

Acknowledgement

This work was partly supported by the National Institutes of Health (R01GM123055) and the National Science Foundation (IIS1319551, IOS1127027, DMS1614777).

References

Atilgan, A.R., Durell, S.R., Jernigan, R.L., *et al.*, 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J. 80, 505–515.
Barnoud, J., Monticelli, L., 2015. Coarse-grained force fields for molecular simulations. Methods Mol. Biol. 1215, 125–149.
Bahar, I., Atilgan, A.R., Erman, B., 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold. Des. 2, 173–181.
Bahar, I., Wallqvist, A., Covell, D.G., Jernigan, R.L., 1998. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. Biochemistry 37, 1067–1075.

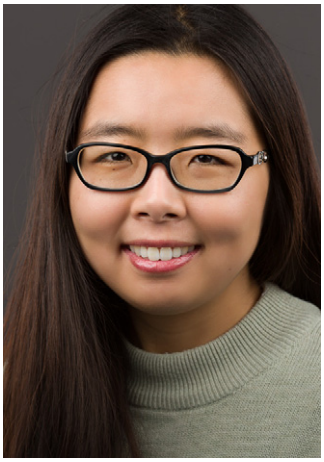
Belogurov, G.A., Vassylyeva, M.N., Svetlov, V., *et al*., 2007. Structural basis for converting a general transcription factor into an operon-specific virulence regulator. Mol. Cell 26, 117–129.
Bennett, M.J., Eisenberg, D., 1994. Refined structure of monomeric diphtheria toxin at 2.3 A resolution. Protein Sci. 3, 1464–1475.
Berman, H.M., Westbrook, J., Feng, Z., *et al*., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242.
Boehr, D.D., Nussinov, R., Wright, P.E., 2009. The role of dynamic conformational ensembles in biomolecular recognition. Nat. Chem. Biol. 5, 789–796.
Bowers, K.J., Chow, E., Xu, H., *et al*., 2006. Scalable algorithms for molecular dynamics simulations on commodity clusters. In: Proceedings of the ACM/IEEE Conference on Supercomputing (SC06). Tampa, Florida.
Brooks, B.R., Brooks, C.L., Mackerell 3rd, A.D., *et al*., 2009. CHARMM: The biomolecular simulation program. J. Comput. Chem. 30, 1545–1614.
Burmann, B.M., Knauer, S.H., Sevostyanova, A., *et al*., 2012. An alpha helix to beta barrel domain switch transforms the transcription factor RfaH into a translation factor. Cell 150, 291–303.
Case, D.A., Cheatham, T.E., Darden 3rd, T., *et al*., 2005. The Amber biomolecular simulation programs. J. Comput. Chem. 26, 1668–1688.
Chandra, N.R., Prabu, M.M., Suguna, K., Vijayan, M., 2001. Structural similarity and functional diversity in proteins containing the legume lectin fold. Protein Eng. 14, 857–866.
Chen, J., 2013. Molecular mechanism of the Escherichia coli maltose transporter. Curr. Opin. Struct. Biol. 23, 492–498.
Derreumaux, P., Schlick, T., 1998. The loop opening/closing motion of the enzyme triosephosphate isomerase. Biophys. J. 74, 72–81.
Dimaio, F., Song, Y., Li, X., *et al*., 2015. Atomic-accuracy models from 4.5-A cryo-electron microscopy data with density-guided iterative local refinement. Nat. Methods 12, 361–365.
Doruker, P., Atilgan, A.R., Bahar, I., 2000. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to alpha-amylase inhibitor. Proteins Struct. Funct. Genet. 40, 512–524.
Duan, Y., Kollman, P.A., 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 282, 740–744.
Dunker, A.K., Silman, I., Uversky, V.N., Sussman, J.L., 2008. Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. 18, 756–764.
Evdokimov, A.G., Phan, J., Tropea, J.E., *et al*., 2003. Similar modes of polypeptide recognition by export chaperones in flagellar biosynthesis and type III secretion. Nat. Struct. Biol. 10, 789–793.
Eyal, E., Yang, L.W., Bahar, I., 2006. Anisotropic network model: Systematic evaluation and a new web interface. Bioinformatics 22, 2619–2627.
Ferron, F., Longhi, S., Canard, B., Karlin, D., 2006. A practical overview of protein disorder prediction methods. Proteins 65, 1–14.
Fiser, A., 2010. Template-based protein structure modeling. Methods Mol. Biol. 673, 73–94.
Gan, H.H., Perlow, R.A., Roy, S., *et al*., 2002. Analysis of protein sequence/structure similarity relationships. Biophys. J. 83, 2781–2791.
Genheden, S., Ryde, U., 2015. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin. Drug Discov. 10, 449–461.
Goh, C.S., Milburn, D., Gerstein, M., 2004. Conformational changes associated with protein-protein interactions. Curr. Opin. Struct. Biol. 14, 104–109.
Greenleaf, W.J., Woodside, M.T., Block, S.M., 2007. High-resolution, single-molecule measurements of biomolecular motion. Annu. Rev. Biophys. Biomol. Struct. 36, 171–190.
Hinsen, K., 1998. Analysis of domain motions by approximate normal mode calculations. Proteins Struct. Funct. Genet. 33, 417–429.
Hinterdorfer, P., Dufrene, Y.F., 2006. Detection and localization of single molecular recognition events using atomic force microscopy. Nat. Methods 3, 347–355.
Jamroz, M., Kolinski, A., Kihara, D., 2012. Structural features that predict real-value fluctuations of globular proteins. Proteins 80, 1425–1435.
Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L., 1983. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. 79, 926–935.
Keskin, O., Durell, S.R., Bahar, I., Jernigan, R.L., Covell, D.G., 2002. Relating molecular flexibility to function: A case study of tubulin. Biophys. J. 83, 663–680.
Kodera, N., Ando, T., 2014. The path to visualization of walking myosin V by high-speed atomic force microscopy. Biophys. Rev. 6, 237–260.
Kolinski, A., 2004. Protein modeling and structure prediction with a reduced representation. Acta Biochim. Pol. 51, 349–371.
Kosloff, M., Kolodny, R., 2008. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins 71, 891–902.
Lee, G.R., Heo, L., Seok, C., 2017. Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. Proteins 86 (Suppl. 1), S168–S176.
Lei, H., Wu, C., Liu, H., Duan, Y., 2007. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proc. Natl. Acad. Sci. USA 104, 4925–4930.
Liwo, A., Khalili, M., Scheraga, H.A., 2005. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc. Natl. Acad. Sci. USA 102, 2362–2367.
Liwo, A., Oldziej, S., Czaplewski, C., Kozlowska, U., Scheraga, H.A., 2004. Parameterization of backbone-electrostatic and multibody contributions to the UNRES force field for protein-structure prediction from ab initio energy surfaces of model systems. J. Phys. Chem. B 108, 9421–9438.
Lupyan, D., Leo-Macias, A., Ortiz, A.R., 2005. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 21, 3255–3263.
Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., De Vries, A.H., 2007. The MARTINI force field: Coarse grained model for biomolecular simulations. J. Phys. Chem. B 111, 7812–7824.
Mccammon, J.A., Gelin, B.R., Karplus, M., 1977. Dynamics of folded proteins. Nature 267, 585–590.
Mcgreevy, R., Singharoy, A., Li, Q., *et al*., 2014. xMDFF: Molecular dynamics flexible fitting of low-resolution X-ray structures. Acta Crystallogr. D Biol. Crystallogr. 70, 2344–2355.
Mirjalili, V., Feig, M., 2013. Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. J. Chem. Theory Comput. 9, 1294–1303.
Monroe, L., Terashi, G., Kihara, D., 2017. Variability of protein structure models from electron microscopy. Structure 25, 592–602. e2.
Navizet, I., Lavery, R., Jernigan, R.L., 2004. Myosin flexibility: Structural domains and collective vibrations. Proteins 54, 384–393.
Nugent, T., Cozzetto, D., Jones, D.T., 2014. Evaluation of predictions in the CASP10 model refinement category. Proteins 82 (Suppl. 2), S98–S111.
Oates, M.E., Romero, P., Ishida, T., *et al*., 2013. D(2)P(2): Database of disordered protein predictions. Nucleic Acids Res. 41, D508–D516.
Parak, F.G., 2003. Proteins in action: The physics of structural fluctuations and conformational changes. Curr. Opin. Struct. Biol. 13, 552–557.
Pearlman, D.A., 2005. Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. J. Med. Chem. 48, 7796–7807.
Pelletier, H., Sawaya, M.R., Wolfle, W., Wilson, S.H., Kraut, J., 1996. Crystal structures of human DNA polymerase beta complexed with DNA: Implications for catalytic mechanism, processivity, and fidelity. Biochemistry 35, 12742–12761.
Peterson, L., Jamroz, M., Kolinski, A., Kihara, D., 2017. Predicting real-valued protein residue fluctuation using FlexPred. Methods Mol. Biol. 1484, 175–186.
Phillips, J.C., Braun, R., Wang, W., *et al*., 2005. Scalable molecular dynamics with NAMD. J. Comput. Chem. 26, 1781–1802.
Piana, S., Lindorff-Larsen, K., Shaw, D.E., 2013. Atomic-level description of ubiquitin folding. Proc. Natl. Acad. Sci. USA 110, 5915–5920.
Pronk, S., Pall, S., Schulz, R., *et al*., 2013. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29, 845–854.
Sawaya, M.R., Prasad, R., Wilson, S.H., Kraut, J., Pelletier, H., 1997. Crystal structures of human DNA polymerase beta complexed with gapped and nicked DNA: Evidence for an induced fit mechanism. Biochemistry 36, 11205–11215.
Simmerling, C., Strockbine, B., Roitberg, A.E., 2002. All-atom structure prediction and folding simulations of a stable protein. J. Am. Chem. Soc. 124, 11258–11259.
Singharoy, A., Teo, I., Mcgreevy, R., *et al*., 2016. Molecular dynamics-based refinement and validation for sub-5 A cryo-electron microscopy maps. eLife. 5.
Skyner, R.E., Mcdonagh, J.L., Groom, C.R., Van Mourik, T., Mitchell, J.B.O., 2015. A review of methods for the calculation of solution free energies and the modelling of systems in solution. Phys. Chem. Chem. Phys. 17, 6174–6191.
Terashi, G., Kihara, D., 2018. Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. Proteins 86 (Suppl. 1), S189–S201.
Tirion, M.M., 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. 77, 1905–1908.
Verlet, L., 1967. Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. Phys. Rev. 159, 98.

Wang, Y., Rader, A.J., Bahar, I., Jernigan, R.L., 2004. Global ribosome motions revealed with elastic network model. J. Struct. Biol. 147, 302–314.
Yamniuk, A.P., Vogel, H.J., 2004. Calmodulin's flexibility allows for promiscuity in its interactions with target proteins and peptides. Mol. Biotechnol. 27, 33–57.
Yang, L.W., Bahar, I., 2005. Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. Structure 13, 893–904.
Ye, Y., Godzik, A., 2005. Multiple flexible structure alignment using partial order graphs. Bioinformatics 21, 2362–2369.
Zoete, V., Michielin, O., Karplus, M., 2003. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. Application to HIV-1 protease inhibitors. J. Comput. Aided Mol. Des. 17, 861–880.

Relevant Websites

http://www.charmm-gui.org
Charmm-Gui.
http://kiharalab.org/flexPred/
FlexPred Server - Kihara Lab.
https://ub.cbm.uam.es/software/online/mamothmult.php
MAMMOTH-Mult - Unidad de Bioinformatica CBMSO.
http://posa.sanfordburnham.org/
POSA Home - Sanford Burnham Prebys Medical Discovery Institute.

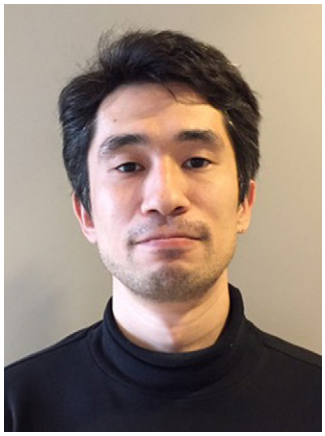
Biographical Sketch



Xusi Han is currently a PhD candidate in Department of Biological Sciences at Purdue University, West Lafayette, Indiana, USA. She is also pursuing toward the M.S. degree in Applied Statistics at Purdue University. She received the B.S. degree from Minzu University of China, Beijing, China, in 2009. Her current research interests include protein global structure comparison, protein structure classification, and graph analysis for electron microscopy density maps.

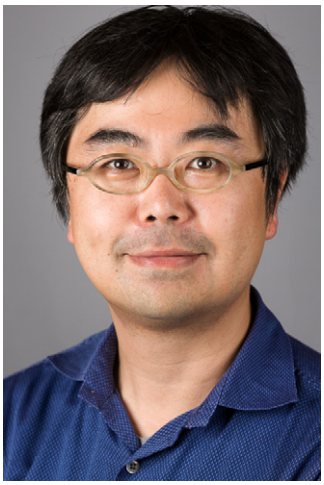
Woong-Hee Shin is currently a postdoctoral researcher of the Kihara Lab in the Department of Biological Sciences at Purdue University. He received the B.Sc. in chemistry from Seoul National University, South Korea in 2008. He also received the PhD degree in chemistry in 2014 from Seoul National University. During his PhD research, he developed a flexible-receptor ligand docking program, GalaxyDock, and performed virtual ligand library screening to find active molecules of nuclear receptors. After receiving his PhD, he joined the Kihara lab in August 2014. His current research interests are development and application of computational drug discovery programs.

Charles Christoffer is currently a graduate member of the Kihara Lab at Purdue University, West Lafayette, Indiana, USA. He received the B.S. degree in Computer Science and Applied Mathematics from Purdue University in 2015. He is currently pursuing the PhD degree with the Department of Computer Science at Purdue University. His research interests include geometric data structures, computer vision, and image processing, especially their application to structural bioinformatics. Within structural bioinformatics, he is particularly interested in protein docking and structure database search.



Genki Terashi is currently a postdoctoral researcher of the Kihara Lab in the Department of Biological Sciences at Purdue University, West Lafayette, Indiana, USA. He received the B.S., M.S., and PhD degrees in Pharmaceutical Sciences from Kitasato University, Tokyo, Japan, in 2002, 2004 and 2008, respectively. His current research interests include protein structure model refinement, protein structure modelling and comparison, and protein-protein docking. Application of Machine Learning method to the structural biology is also his research interest.

Lyman Monroe is a PhD candidate in the Department of Biological Sciences at Purdue University, West Lafayette, Indiana, USA. He received the B.S. degree in Physics from Purdue University in 2012, as well as a second B.S. degree in Chemistry from Purdue University in 2013. His projects include protein structure modelling and refinement using molecular dynamics simulation.



Daisuke Kihara is a full professor in the Department of Biological Sciences and the Department of Computer Science at Purdue University, West Lafayette, Indiana, USA. He received the B.S. degree in Biochemistry from the University of Tokyo, Japan in 1994, and the M.S. and PhD degrees from Kyoto University, Japan in 1996 and 1999, respectively. He was a postdoctoral researcher with Prof. Jeffrey Skolnick at the Danforth Plant Science Center, St. Louis and at SUNY Buffalo from 1999 to 2003. He joined Purdue University as an assistant professor in 2003 and was promoted to associate professor in 2009 and to full professor in 2014. His research field is bioinformatics. His research projects include protein docking, protein tertiary structure prediction, structure- and sequence-based protein function prediction, and computational drug design. He has published over 140 research papers and book chapters. His research projects have been supported by funding from the National Institutes of Health, the National Science Foundation, the Office of the Director of National Intelligence, and industry. In 2013, he was named a University Faculty Scholar by Purdue University.