Communication: Capturing protein multiscale thermal fluctuations

Kristopher Opron,[1] Kelin Xia,[2] and Guo-Wei Wei[1,2,3,a)]
[1]*Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, USA*
[2]*Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, USA*
[3]*Department of Electrical and Computer Engineering Michigan State University, East Lansing, Michigan 48824, USA*

(Received 21 April 2015; accepted 21 May 2015; published online 2 June 2015)

Existing elastic network models are typically parametrized at a given cutoff distance and often fail to properly predict the thermal fluctuation of many macromolecules that involve multiple characteristic length scales. We introduce a multiscale flexibility-rigidity index (mFRI) method to resolve this problem. The proposed mFRI utilizes two or three correlation kernels parametrized at different length scales to capture protein interactions at corresponding scales. It is about 20% more accurate than the Gaussian network model (GNM) in the B-factor prediction of a set of 364 proteins. Additionally, the present method is able to deliver accurate predictions for some large macromolecules on which GNM fails to produce accurate predictions. Finally, for a protein of N residues, mFRI is of linear scaling ($O(N)$) in computational complexity, in contrast to the order of $O(N^3)$ for GNM. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4922045]

Proteins are among the most essential biomolecules for life. Many protein functions, such as structural support, catalyzing chemical reactions, and allosteric regulation are strongly correlated to protein flexibility.[14] Protein flexibility is an intrinsic property of proteins and can be measured directly or indirectly by many experimental approaches, such as X-ray crystallography, nuclear magnetic resonance (NMR), and single-molecule force experiments.[10] Theoretically, protein flexibility can be computed by normal mode analysis (NMA),[7,15,23,33] graph theory,[19] rotation translation blocks (RTB) method,[9,31] and elastic network models (ENM),[4–6,16,24,32] including Gaussian network model (GNM)[5,6] and anisotropic network model (ANM).[4] A common feature of the above mentioned time-independent methods is that they resort to matrix diagonalization procedure. The computational complexity of matrix diagonalization is typically on the order of $O(N^3)$, where N is the number of elements in the matrix. Such a computational complexity calls for new more efficient strategies for the flexibility analysis of large biomolecules.

It is well known that NMA and GNM do not work well for many macromolecules. Park *et al.* had collected three sets of structures to test performance of NMA and GNM methods.[27] It was found that both methods fail to work and deliver negative correlation coefficients for many structures.[27] The mean correlation coefficients (MCCs) for the B-factor prediction of small-sized, medium-sized, and large-sized sets of structures are about 0.480, 0.482, and 0.494 for NMA, respectively.[26,27] The GNM preforms slightly better, with the mean correlation coefficients of 0.541, 0.550, and 0.529 for the above test sets.[26,27] Obviously, there is a pressing need to develop innovative approaches for biomolecular flexibility analysis.

Recently, we have proposed a few matrix-decomposition-free methods for flexibility analysis, including molecular nonlinear dynamics,[36] stochastic dynamics,[35] and flexibility-rigidity index (FRI).[26,34] Among them, FRI has been introduced to evaluate protein flexibility and rigidity. The fundamental assumptions of the FRI method are as follows. Protein functions, such as flexibility, rigidity, and energy, are fully determined by the structure of the protein and its environment, and the protein structure is in turn determined by the relevant interactions. Therefore, whenever the protein structure is available, there is no need to analyze protein flexibility and rigidity by tracing back to the protein interaction Hamiltonian. Consequently, the FRI bypasses the $O(N^3)$ matrix diagonalization. Our initial FRI[34] has the computational complexity of $O(N^2)$ and our fast FRI (fFRI)[26] based on a cell lists algorithm[3] is of $O(N)$. The FRI and the fFRI have been extensively validated by a set of 365 proteins for parametrization, accuracy, and reliability. The parameter free fFRI is about 10% more accurate than the GNM on the 365 protein test set and is orders of magnitude faster than GNM on a set of 44 proteins. FRI is able to predict the B-factors of a HIV virus capsid (313 236 residues) in less than 30 s on a single desktop CPU (AMD Phenom II X6 1100T), which would require GNM more than 120 yr to accomplish if the computer memory is not a problem.[26] See the supplementary material for details.[1]

Nevertheless, there are structures for which FRI does not work either. In fact, structures that fail NMA and GNM are likely to be difficult for the original FRI method as well. One such structure is pictured in Figure 2 where the GNM method fails to predict the high flexibility of a hinge region in calmodulin with any cutoff distance. There are a number of reasons for this and other types of failure. Crystal environment, solvent type, co-factors, data collection conditions, and structural refinement procedures are well-known causes.[17,21,22,30]

a)Author to whom correspondence should be addressed. Electronic mail: wei@math.msu.edu

However, there is one more important cause that has not been discussed in the literature to our best knowledge, namely, multiple characteristic length scales in a single protein structure. Indeed, contrary to small molecules, macromolecular interactions have a wide variety of characteristic length scales, ranging from covalent bond, hydrogen bond, van der Waals bond, residue, alpha helix and beta sheet, domain, and protein scales. Protein flexibility is intrinsically associated with protein interactions and thus must have a multiscale trait as well. When the GNM or FRI method is parametrized at a given cutoff or scale parameter, it captures only a subset of the characteristic length scales but inevitably misses the other characteristic length scales of the protein. Consequently, none of them is able to provide an accurate B-factor prediction.

The multiscale flexibility-rigidity index (mFRI) is constructed to capture the multiscale collective motions of macromolecules. We utilize multiple correlation kernels, with each kernel being parametrized at specific scale to characterize the multiscale flexibility of macromolecules. The nth flexibility index of the ith (coarse-grained) particle is given by

$$f_i^n = \frac{1}{\sum_{j=1}^{N} w_j^n \Phi^n(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_j^n)}, \tag{1}$$

where w_j^n is an atomic type dependent parameter, $\Phi^n(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_j^n)$ is a correlation kernel, and η_j^n is a scale parameter. Here, $\mathbf{r}_i$ and $\mathbf{r}_j$ are the coordinates for ith and jth particles, respectively. We seek the minimization of the form

$$\text{Min}_{a^n,b}\left\{\sum_i \left|\sum_n a^n f_i^n + b - B_i^e\right|^2\right\}, \tag{2}$$

where $\{B_i^e\}$ are the experimental B-factors. We use generalized exponential kernels[26,34]

$$\Phi^n(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j^n) = e^{-\left(\|\mathbf{r}-\mathbf{r}_j\|/\eta_j^n\right)^\kappa}, \quad \kappa > 0 \tag{3}$$

and generalized Lorentz kernels

$$\Phi^n(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j^n) = \frac{1}{1 + \left(\|\mathbf{r} - \mathbf{r}_j\|/\eta_j^n\right)^\upsilon}, \quad \upsilon > 0. \tag{4}$$

In principle, all parameters can be optimized. For simplicity and computational efficiency, we only determine $\{a^n\}$ and b in the above minimization process. In this work, we limit the number of kernels to at most three and set $w_j^n = 1$. Both generalized exponential kernels and generalized Lorentz kernels are employed. More detailed description of the mFRI is given in the supplementary material.[1]

To understand the multiscale behavior of flexibility analysis, we consider a test set containing 364 protein structures whose Protein Data Bank (PDB) identities are listed in the literature[26] and it contains test sets used in GNM studies.[27] This test set omits one structure present in previous FRI studies (PDB ID: 1AGN) due to unrealistic B-factor data. Our goal is to examine how an additional kernel with a large length scale impacts the flexibility analysis. To this end, we consider two smooth Lorentz type of kernels with $\upsilon = 3$. We explore a number of scale combinations as shown in Fig. 1, which plots the MCC values for B-factor prediction on the set of 364 structures. The low MCC values on the diagonal line

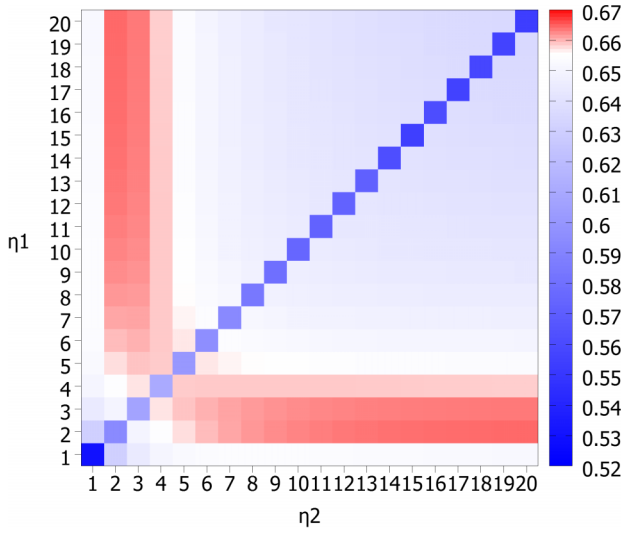

FIG. 1. An illustration of multiscale behavior in protein flexibility analysis. Two Lorentz kernels ($\upsilon = 3$) are used. Their scale values, η values, are listed along the horizontal and vertical axes. The mean correlation coefficient value for B-factor prediction on a set of 364 proteins is shown in each cell of the matrix and color coded for convenience with red representing the highest correlation coefficients and green the lowest. Obvious, the combination of a relatively small-scale kernel and a relatively large-scale kernel delivers best prediction, which indicates the importance of incorporating multiscale in protein flexibility analysis.

indicate that two-scale methods are always better than a single scale one. The best results are achieved at the combination of a relatively small-scale kernel and a relatively large-scale kernel. This behavior proves the importance of incorporating multiscale in the biomolecular flexibility analysis. The best MCC for the test set is 0.67, which is about 20% better than the best GNM prediction and about 6% improvement over our single scale FRI approach.

The improvement in the MCC for B-factor prediction on a set of 364 proteins discussed above obscures the fact that the proposed multiscale method is able to capture the multiscale behaviors in many structures that fail the original FRI and GNM. In the rest of this paper, we demonstrate utility of the proposed multiscale method by a few case studies. A three-scale FRI is employed.

Protein hinge regions have been shown to be correlated with active sites and catalysis in enzymes. Flexibility has a major role in specificity of binding of a protein to other proteins, nucleic acids, or other molecules. An active site or docking region that is more flexible will accommodate more varied substrates or partners while more rigid domains are more specific. Protein hinges are also found separating large domains of proteins. In this context, the hinges can be very important for protein conformational changes. The protein featured in this section, calmodulin, is a good example of a hinge that affects both structure and function.

The central region of calmodulin shown in Figure 2 is a long α-helix which is unwound or kinked at the middle when no calcium is bound to the two distal metal coordinating domains. In both forms, with or without calcium bound, this helix retains a large degree of flexibility based on B-factor values from the PDB files (1CLL and 1CFD).

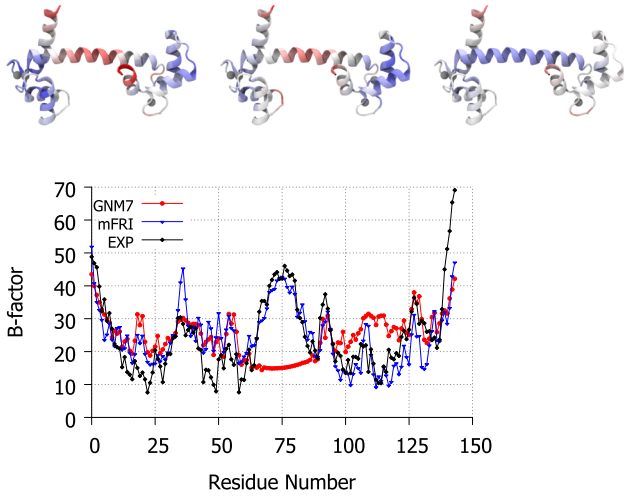

FIG. 2. Top, the structure of calmodulin (PDB ID: 1CLL) visualized in Visual Molecular Dynamics (VMD)[18] and colored by experimental B-factors (left), mFRI predicted B-factors (middle), and GNM predicted B-factors (right) with red representing the most flexible regions. Bottom, the experimental and predicted B-factor values plotted per residue. The GNM7 is for the GNM method with a cutoff distance of 7 Å. Clearly, GNM misses the flexible hinge region. The mFRI is parametrized at $\upsilon^1 = 3$, $\eta^1 = 3$ Å, $\upsilon^2 = 3$, $\eta^2 = 7$ Å, $\kappa^3 = 1$, and $\eta^3 = 15$ Å.

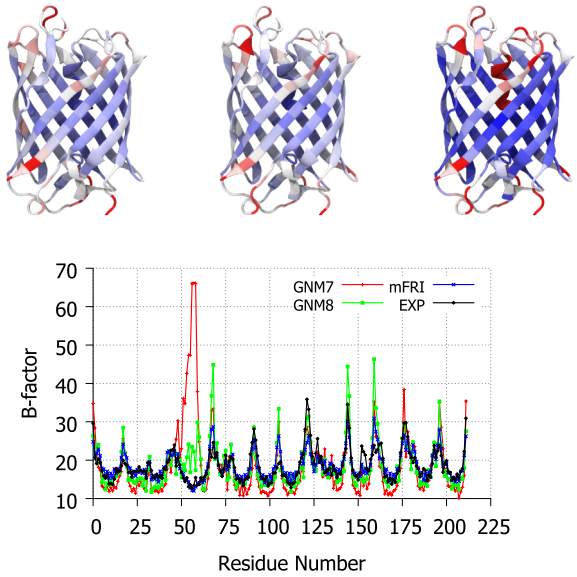

FIG. 3. Top, a visual comparison of experimental B-factors (left), mFRI predicted B-factors (middle), and GNM predicted B-factors (right) for the engineered cyan fluorescent protein, mTFP1 (PDB ID:2HQK). Bottom, the experimental and predicted B-factor values plotted per residue. The GNM naming convention indicates the cutoff used for the GNM method in Å, i.e., GNM7 is the GNM method with a cutoff distance of 7 Å.

Many tools exist for the prediction and analysis of hinges in proteins using bioinformatics,[13] graph theory,[11,20,28] and energetics.[12] The proposed mFRI has capabilities similar to those in these tools. The mFRI can be used to predict hinge regions by high FRI values or predicted B-values.

A comparison of mFRI method and GNM for the B-factor prediction of calcium-bound calmodulin is displayed in Figure 2. B-factor prediction by single kernel FRI and GNM is unable to accurately predict the hinge region in the middle of the protein with any parameter. Two- and three-kernel based mFRI methods, on the other hand, are much more accurate in the hinge region. As more kernels are added, the accuracy can be seen to grow but sufficient accuracy is achieved at three kernels.

We have shown in our supplementary material[1] that a similarly good B-factor prediction for calmodulin type of structures can be achieved by the original FRI method if the crystal effect is taken into consideration. This result suggests that the proposed mFRI method may be able to take care some crystal effects.

Cyan fluorescent protein (CFP), shown in Figure 3, is a homolog of the famous green fluorescent protein (GFP). Isolated from the crystal jellyfish in the 1990s,[29] GFP enabled a revolution in biochemistry by allowing the tagging and tracking of a wide range of molecules. CFP was found later in Anthozoa coral species which have turned out to be a good source of fluorescent proteins with varied emission spectra.[25] In this example, we examine the flexibility of an engineered CFP from Clavularia coral[2] (PDB ID: 2HQK), mTFP1. It is clear in Figure 3 that GNM B-factor predictions contain a large error around residues 50-60 which is very pronounced at the recommended cutoff of 7 Å and is still somewhat problematic when the cutoff is changed to 8 Å, the best alternate parameter found by searching incrementally outward from 7 Å in either direction. mFRI on the other hand has no issue with this particular region. Upon further inspection, it is clear that the offending region is the small, alpha-helical region suspended in the center of the beta-barrel. It is not surprising that this sort of configuration would be highly cutoff parameter dependent in a scheme such as GNM, which has hard cutoffs for connectivity. It would appear that this structure is dominated by short range interaction but the region of residues 50-60 is affected to a large degree by mid-range interactions, i.e., there are at least two important scales of interaction in this case. It follows then that mFRI, which has kernels to capture short- and mid-range interactions, would perform better than GNM7 or GNM8 parameterizations alone in B-factor predictions, Figure 3, which is exactly what we see from the results.

A similar situation exists with the structure 1V70, a probable antibiotic synthesis protein, which is shown in Figure 4. As in the last example, the problematic portion for B-factor prediction comes at the end of a protein chain. In this case, there is an overestimation of flexibility for residues 1-10 when using GNM. Again, varying parameters from the recommended 7 Å results in marginally better results; however, no parametrization is able to reach the accuracy of mFRI.

The final example is a biologically important molecule, ribosomal protein L14, a component of the 60S ribosomal subunit.[8] Depicted in Figure 5, L14 is a structurally diverse protein containing regions of alpha helix, beta-barrel, parallel beta strands, and a beta-hairpin motif. The pattern of flexibility predicted by GNM for this structure is shown to be over-exaggerated, i.e., rigid areas are predicted to be more rigid than they actually are and vice versa. This pattern exists in most GNM results due to the use of a hard cutoff in the Kirchhoff matrix. Such a hard cutoff will inevitably lead to the overestimation of bond importance near the edge of the cutoff; therefore, if a large number of interactions exist for a particular atom near the cutoff point, there is likely to be a large error in the estimation of flexibility for that atom. This is likely what is happening with the errors in GNM calculation of the proteins

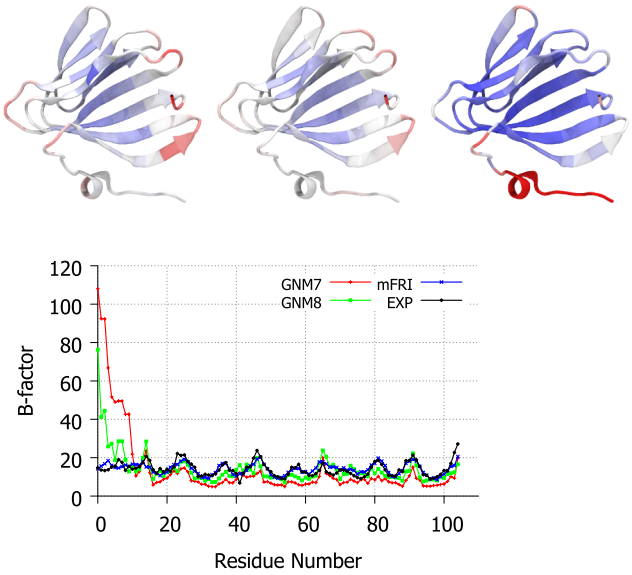

FIG. 4. Top, a visual comparison of experimental B-factors (left), mFRI predicted B-factors (middle), and GNM predicted B-factors (right) for a probable antibiotic synthesis protein (PDB ID:1V70). Bottom, the experimental and predicted B-factor values plotted per residue.

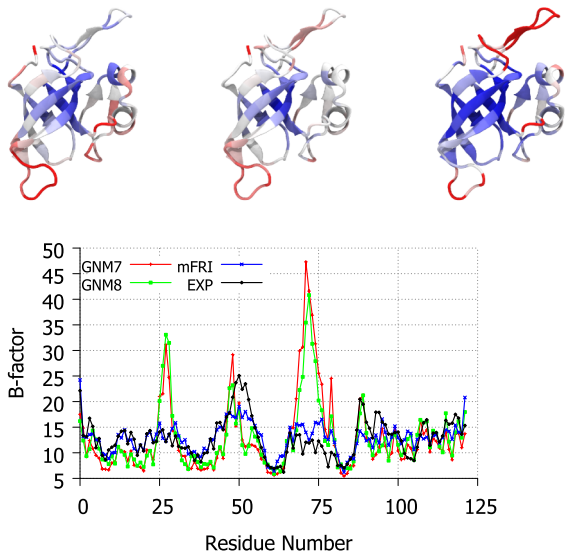

FIG. 5. Top, a visual comparison of experimental B-factors (left), mFRI predicted B-factors (middle), and GNM predicted B-factors (right) for the ribosomal protein L14 (PDB ID:1WHI). Bottom, the experimental and predicted B-factor values plotted per residue.

in Figures 3–5; the protein at the end of the chain may be near the edge of the cutoff distance for many interactions with the bulk of the proteins. While adjusting GNM's cutoff distance may temper the error being introduced, it cannot eliminate it completely unless they change to a soft-decaying kernel method such as FRI. Nevertheless, soft-decaying kernel based methods can only alleviate the problem. They do not deliver satisfactory B-factor predictions unless a multiscale strategy is employed. We note that it is not obvious how to incorporate a multiscale strategy in matrix diagonalization based methods.

This work was supported in part by NSF Grant Nos. IIS-1302285 and DMS-1160352, NIH Grant No. R01GM-090208, and MSU Center for Mathematical Molecular Biosciences Initiative.

[1]See supplementary material at http://dx.doi.org/10.1063/1.4922045 for theoretical formulation, parametrization, efficiency test, additional examples, and crystal packing effects.
[2]H. W. Ai, J. Henderson, S. Remington, and R. Campbell, "Directed evolution of a monomeric, bright and photostable version of clavularia cyan fluorescent protein: Structural characterization and applications in fluorescence imaging," Biochem. J **400**, 531–540 (2006).
[3]M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987).
[4]A. R. Atilgan, S. R. Durrell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," Biophys. J. **80**, 505 – 515 (2001).
[5]I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, "Vibrational dynamics of proteins: Significance of slow and fast modes in relation to function and stability," Phys. Rev. Lett. **80**, 2733 – 2736 (1998).
[6]I. Bahar, A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential," Folding Des. **2**, 173 – 181 (1997).
[7]B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. States, S. Swaminathan, and M. Karplus, "Charmm: A program for macromolecular energy, minimization, and dynamics calculations," J. Comput. Chem. **4**, 187–217 (1983).
[8]C. Davies, S. W. White, and V. Ramakrishnan, "The crystal structure of ribosomal protein l14 reveals an important organizational component of the translational apparatus," Structure **4**(1), 55–66 (1996).
[9]O. N. A. Demerdash and J. C. Mitchell, "Density-cluster NMA: A new protein decomposition technique for coarse-grained normal mode analysis," Proteins: Struct., Funct., Bioinf. **80**(7), 1766–1779 (2012).
[10]O. K. Dudko, G. Hummer, and A. Szabo, "Intrinsic rates and activation free energies from single-molecule pulling experiments," Phys. Rev. Lett. **96**, 108101 (2006).
[11]U. Emekli, S. Dina, H. Wolfson, R. Nussinov, and T. Haliloglu, "HingeProt: Automated prediction of hinges in protein structures," Proteins **70**(4), 1219–1227 (2008).
[12]S. Flores and M. Gerstein, "FlexOracle: Predicting flexible hinges by identification of stable domains," BMC Bioinf. **8**(1), 215 (2007).
[13]S. Flores, L. Lu, J. Yang, N. Carriero, and M. Gerstein, "Hinge atlas: Relating protein sequence to sites of structural flexibility," BMC Bioinf. **8**, 167 (2007).
[14]H. Frauenfelder, S. G. Slihar, and P. G. Wolynes, "The energy landsapes and motion of proteins," Science **254**(5038), 1598–1603 (1991).
[15]N. Go, T. Noguti, and T. Nishikawa, "Dynamics of a small globular protein in terms of low-frequency vibrational modes," Proc. Natl. Acad. Sci. U. S. A. **80**, 3696 – 3700 (1983).
[16]K. Hinsen, "Analysis of domain motions by approximate normal mode calculations," Proteins **33**, 417 – 429 (1998).
[17]K. Hinsen, "Structural flexibility in proteins: Impact of the crystal environment," Bioinformatics **24**, 521 – 528 (2008).
[18]W. Humphrey, A. Dalke, and K. Schulten, "VMD – visual molecular dynamics," J. Mol. Graphics **14**(1), 33–38 (1996).
[19]D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, "Protein flexibility predictions using graph theory," Proteins: Struct., Funct., Genet. **44**(2), 150–165 (2001).
[20]K. S. Keating, S. C. Flores, M. B. Gerstein, and L. A. Kuhn, "StoneHinge: Hinge prediction by network analysis of individual protein structures," Protein Sci. **18**(2), 359–371 (2009).
[21]D. A. Kondrashov, A. W. Van Wynsberghe, R. M. Bannen, Q. Cui, and J. G. N. Phillips, "Protein structural variation in computational models and crystallographic data," Structure **15**, 169 – 177 (2007).
[22]S. Kundu, J. S. Melton, D. C. Sorensen, and J. G. N. Phillips, "Dynamics of proteins in crystals: Comparison of experiment with simple models," Biophys. J. **83**, 723 – 732 (2002).
[23]M. Levitt, C. Sander, and P. S. Stern, "Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme," J. Mol. Biol. **181**(3), 423 – 447 (1985).
[24]G. H. Li and Q. Cui, "A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca(2+)-ATPase," Bipohys. J. **83**, 2457 – 2474 (2002).
[25]M. V. Matz, A. F. Fradkov, Y. A. Labas, A. P. Savitsky, A. G. Zaraisky, M. L. Markelov, and S. A. Lukyanov, "Fluorescent proteins from nonbioluminescent anthozoa species," Nat. biotechnol. **17**(10), 969–973 (1999).

[26]K. Opron, K. L. Xia, and G. W. Wei, "Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis," J. Chem. Phys. **140**, 234105 (2014).
[27]J. K. Park, R. Jernigan, and Z. Wu, "Coarse grained normal mode analysis vs. refined Gaussian network model for protein residue-level structural fluctuations," Bull. Math. Biol. **75**, 124 – 160 (2013).
[28]M. Shatsky, R. Nussinov, and H. J. Wolfson, "FlexProt: Alignment of flexible protein structures without a predefinition of hinge regions," J. Comput. Biol. **11**(1), 83–8106 (2004).
[29]O. Shimomura, F. H. Johnson, and Y. Saiga, "Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, aequorea," J. Cell. Comp. Physiol. **59**(3), 223–239 (1962).
[30]G. Song and R. L. Jernigan, "vgnm: A better model for understanding the dynamics of proteins in crystals," J. Mol. Biol. **369**(3), 880 – 893 (2007).
[31]F. Tama, F. X. Gadea, O. Marques, and Y. H. Sanejouand, "Building-block approach for determining low-frequency normal modes of macromolecules," Proteins: Struct., Funct., Bioinf. **41**(1), 1–7 (2000).
[32]F. Tama and Y. H. Sanejouand, "Conformational change of proteins arising from normal mode calculations," Protein Eng. **14**, 1 – 6 (2001).
[33]M. Tasumi, H. Takenchi, S. Ataka, A. M. Dwidedi, and S. Krimm, "Normal vibrations of proteins: Glucagon," Biopolymers **21**, 711 – 714 (1982).
[34]K. L. Xia, K. Opron, and G. W. Wei, "Multiscale multiphysics and multidomain models — Flexibility and rigidity," J. Chem. Phys. **139**, 194109 (2013).
[35]K. L. Xia and G. W. Wei, "A stochastic model for protein flexibility analysis," Phys. Rev. E **88**, 062709 (2013).
[36]K. L. Xia and G. W. Wei, "Molecular nonlinear dynamics and protein thermal uncertainty quantification," Chaos **24**, 013103 (2014).