Flexibility–Rigidity Index for Protein–Nucleic Acid Flexibility and Fluctuation Analysis

Kristopher Opron,[a] Kelin Xia,[b] Zach Burton,[a] and Guo-Wei Wei*[c]†

Protein–nucleic acid complexes are important for many cellular processes including the most essential functions such as transcription and translation. For many protein–nucleic acid complexes, flexibility of both macromolecules has been shown to be critical for specificity and/or function. The flexibility-rigidity index (FRI) has been proposed as an accurate and efficient approach for protein flexibility analysis. In this article, we introduce FRI for the flexibility analysis of protein–nucleic acid complexes. We demonstrate that a multiscale strategy, which incorporates multiple kernels to capture various length scales in biomolecular collective motions, is able to significantly improve the state of art in the flexibility analysis of protein–nucleic acid complexes. We take the advantage of the high accuracy and $O(N)$ computational complexity of our multiscale FRI method to investigate the flexibility of ribosomal subunits, which are difficult to analyze by alternative approaches. An anisotropic FRI approach, which involves localized Hessian matrices, is utilized to study the translocation dynamics in an RNA polymerase. © 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24320

Introduction

Proteins and nucleic acids, which include deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), are among the most essential biomolecules for all known forms of life. In cells, proteins have a wide variety of important functions, including supporting organism structure, catalyzing reactions involved in transcription and translation participating in signal transduction, and working as immune agents. Nucleic acids typically function in association with proteins and play a crucial role in encoding, transmitting, and expressing genetic information. Genetic information is stored through the nucleic acid sequence, i.e., the order of nucleotides within a DNA or a RNA molecule and transmitted via transcription and translation processes. Protein rigidity, flexibility, and electrostatics strongly correlate to protein structure and function.[2] The impact of biomolecular electrostatics on their structure, function, and dynamics has been a subject of intensive study. However, the importance of biomolecular flexibility and rigidity in determining their structure and function has been overlooked. In general, protein rigidity is responsible for protein three-dimensional (3D) equilibrium geometric shapes and structural function in forms of tubulin, collagen, elastin, and keratin, while protein flexibility is an important factor in all other protein functions.[18] DNA flexibility is also an important effect in DNA packing. Although the flexibility of biomolecules is often associated with their motion and dynamics, which are their response to the external stimuli and die out at the absolute zero temperature, flexibility is an intrinsic property.

Biomolecular flexibility and rigidity can be measured directly or indirectly by many experimental approaches, such as X-ray crystallography, nuclear magnetic resonance (NMR), and single-molecule force experiments.[11] In single-molecule force experiments, including optical tweezers and nanopore force spectroscopy, the intrinsic rupture rate can be a direct measure of the flexibility and rigidity. In the X-ray structure, Debye–Waller factors, also known as B-factors or temperature factors, are computed as the uncertainty for each atom in the least square fitting between the X-ray diffraction data and the theoretical model. Debye–Waller factors are interpreted as atomic mean-square-fluctuations at the given experimental temperature, and are associated with biomolecular flexibility and rigidity. NMR is known for its ability to analyze biomolecular flexibility and rigidity under physiological conditions, and at various timescales.

The availability of experimental data makes the theoretical study of biomolecular flexibility and rigidity an interesting and important topic in which quantitative models can be calibrated and validated. Molecular dynamics (MD)[38] can be used to elucidate biomolecular collective motion and fluctuation. MD is a powerful technique for the understanding of the conformational landscapes of biomolecules. However, biomolecular flexibility and rigidity are intrinsic properties that can be measured at the motionless and fluctuation-free state. Therefore, MD might not be efficient for biomolecular flexibility and

[a] K. Opron, Z. Burton
Department of Biochemistry and Molecular Biology, Michigan State University, Michigan 48824
[b] K. Xia
Department of Mathematics, Michigan State University, Michigan 48824
[c] Guo-Wei Wei
Mathematical Biosciences Institute, The Ohio State University, Columbus, Ohio 43210
E-mail: wei@math.msu.edu
†On leave from the Department of Mathematics, Michigan State University.
Contract grant sponsor: NSF; Contract grant number: IIS-1302285 and DMS-1160352; Contract grant sponsor: NIH; Contract grant number: R01GM-090208; Contract grant sponsor: Michigan State University Center for Mathematical Molecular Biosciences Initiative

rigidity analysis. Alternative approaches including normal mode analysis (NMA),[6,20,32,50] graph theory,[26] and elastic network model (ENM)[3–5,22,34,48] have become the main workhorses for biomolecular flexibility and rigidity analysis during the past two decades. Analagous to the time-dependent and time-independent Schrödinger's equations, these approaches are designed as time-independent counterparts of the corresponding MD methods.[41] Consequently, diagonalization of the interaction matrix or Hamiltonian of a biomolecule is a required procedure to obtain biomolecular eigenmodes and associated eigenvalues, which are further organized to predict biomolecular temperature factors. The low-order eigenmodes computed from diagonalizing the Kirchhoff matrix or the Hessian matrix can be interpreted as the slow motions of the biomolecule around its equilibrium state and thus shed light on the long-time behavior of biomolecular dynamics near the equilibrium state beyond the reach of MD simulations.[6,32,50] Tirion argued that the potential used in NMA can be simplified to retain only the harmonic potential for elasticity, which is the dominant term in the MD Hamiltonian.[51] Network theory[17] has also had considerable impact in flexibility analysis. The combination of elasticity and coarse-grained network gives rise to elastic network model (ENM).[22] Many other network-based approaches, including Gaussian network model (GNM)[4,5] and anisotropic network model (ANM),[3] have been developed for biomolecular flexibility analysis.

It has been demonstrated by Yang et al.[67] that GNM is about one order more efficient than most other flexibility approaches. GNM is also typically more accurate than ANM in B-factor prediction.[39,41] Applications have been demonstrated in stability[36] analysis, docking simulation,[19] viral capsids,[42,46] and domain motions of hemoglobin,[65] F1 ATPase,[9,70] chaperonin GroEL,[28,69] and the ribosome.[49,56] More details can be found in a few recent reviews.[8,37,44,67] A common feature of the aforementioned time-independent methods is that they resort to a matrix diagonalization procedure. The computational complexity of matrix diagonalization is typically of the order of $\mathbf{O}(N^3)$, where N is the number of elements in the matrix. Such a computational complexity calls for new efficient strategies for the flexibility analysis of large biomolecules.

It is well known that NMA and GNM offer poor flexibility analysis for many macromolecules.[23,29,31,45] Park et al. had studied the performance of NMA and GNM methods for three sets of structures.[41] They found that both methods fail to work and deliver negative correlation coefficients (CCs) for many structures.[41] They have shown that mean correlation coefficients (MCCs) for the NMA B-factor prediction of small-sized, medium-sized, and large-sized sets of structures are about 0.480, 0.482, and 0.494, respectively.[39,41] GNM is considerably more accurate and delivers MCCs of 0.541, 0.550, and 0.529 for the above test sets.[39,41] Indeed, various improvements, including crystal environment, solvent type, and cofactors, are proposed.[23,29,31,45] Additionally, density—cluster rotational—translational blocking has been considered.[10,47] Alternative approaches have been proposed for the flexibility analysis of hinges in proteins using bioinformatics,[16] graph theory,[12,27,43] and energetics.[15] Moreover, low-quality experimental data due to collection conditions and structural refinement procedures may also contribute to poor flexibility predictions.

From observation of the relationship between flexibility and local packing density, Halle [21] proposed a much simplified model called local density model (LDM), and bypassed the whole eigenmode analysis for protein B-factor prediction. In this method, the inverse of contact density, defined as the number of noncovalent neighbor atoms within a local region, is found to be proportional to atomic mean-square displacements, thus can be directly used to predict the experimental B-factors. Another interesting method is the local contact model (LCM) proposed by Zhang et al. [68]. In this approach, the generalized order parameter of the atom is approximated by the summation of a series of exponential functions of atomic distances. Both LDM and LCM demonstrate great potential for protein flexibility prediction. Based on these approaches, many modifications have been proposed in the literature.[24,33,35] Among them, the weighted contact number (WCN) is able to deliver a better accuracy than GNM using an inverse square distance function.[35]

Recently, we have proposed a few matrix-decomposition-free methods for flexibility analysis, including molecular nonlinear dynamics,[62] stochastic dynamics,[61] and the flexibility–rigidity index (FRI).[39,60] Among them, FRI has been introduced to evaluate protein flexibility and rigidity that are required in a multiscale formalism called continuum elasticity with atomic rigidity (CEWAR) for macromolecular elasticity analysis.[60] The FRI method may appear to be akin to the "flexibility index" proposed independently by von der Lieth et al.[54] and Jacobs et al.[26] to describe bond strengths. However, these flexibility indices have little in common with FRI, which does not resort to any protein interaction Hamiltonian for predicting protein flexibility and rigidity. Instead, FRI is a structure-based approach. 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 protein structure is in turn determined by the relevant interactions. Consequently, FRI bypasses the $\mathbf{O}(N^3)$ matrix diagonalization. In fact, FRI does not even require the 3D geometric information of the protein structure. It assesses topological connectivity of the protein distance geometry and analyzes the geometric compactness of the protein structure. It can be regarded as a kernel generalization of the local density model.[21] Our initial FRI[60] has the computational complexity of $\mathbf{O}(N^2)$ and our fast FRI (fFRI)[39] based on a cell lists algorithm[1] is of $\mathbf{O}(N)$. FRI and fFRI have been extensively validated on a set of 365 proteins for parametrization, accuracy, and reliability. A 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 an HIV virus capsid (313,236 residues) in $<$30 s on a single core of a typical desktop processor, which would require GNM more than 120 years to accomplish if the computer memory is not a problem.[39]

However, for those protein structures that fail the traditional NMA and GNM methods, our earlier single-kernel FRI method is most likely not able to deliver good predictions either. In addition to aforementioned problems, neglecting of multiple characteristic length scales in protein structures is another drawback of present flexibility analysis. Indeed, biomolecules have many characteristic length scales, such as covalent bond scale, hydrogen bond scale, Van der Waals bond scale, intraresidue scale, inter-residue scale, alpha helix and beta sheet scale, domain scale, and protein interaction scale. When GNM or FRI is parametrized at a given cutoff or scale parameter, it captures only a subset of the characteristic length scales and inevitably misses other characteristic length scales of the protein. Consequently, these methods fail to offer accurate B-factor prediction for many biomolecules. A multiscale strategy has been proposed to resolve this problem by introducing two or three kernels that are parametrized at relatively small, medium, and/or large length scales in the FRI formulation.[40] We demonstrate that the resulting multiscale FRI (mFRI) works extremely well for many proteins for which the GNM method fails to offer accurate flexibility analysis.[40] Based on a set of 364 proteins, mFRI is 20% more accurate than GNM. It is interesting to note that there is no obvious way to incorporate multiple length scales in the aforementioned matrix diagonalization-based approaches.

Our FRI methods are similar to LDM, LCM, and WCN in terms of matrix-diagonalization-free. However, our FRI methods differ from LDM, LCM, and WCN in the following aspects. First, our original FRI methods were motivated from continuum mechanics, the CEWAR.[39,60] As a result, our FRI methods offer not only discrete flexibility index and discrete rigidity index but also continuous flexibility function and continuous rigidity function. The latter is equivalent to the volumetric biomolecular density distribution. Consequently, one can use the FRI rigidity function to fit electron microscope (EM) density maps.[52,58,63] A by-product of our FRI rigidity function is the smooth biomolecular surface extracted by setting an isosurface value, including the Gaussian surface as a special case.[30,64] In contrast, LDM, LCD, and WCN do not admit any continuum representation. Additionally, the discrete FRI formulations differ from those of LDM, LCD, and WCN by admitting the diagonal term in the summation. Moreover, we have considered the multiscale effects in biomolecules. Our mFRI captures biomolcular thermal fluctuations at various length scales and thus substantially improves the accuracy of the original FRI method. Finally, we have proposed an anisotropic FRI (aFRI) method to describe biomolecular collective motions. A unique feature of our aFRI method is that it allows adaptive Hessian matrices, from a completely global $3N \times 3N$ matrix to completely local 3×3 matrices. Therefore, one can use aFRI to pinpoint flexibility analysis to a given domain or region.

The objective of this work is to develop FRI methods for the flexibility analysis of protein–nucleic acid complexes. Proteins and nucleic acids are dramatically different biomolecules. Amino acid residues and nucleotides have different length scales and interaction characteristics. Therefore, a good model should not only allow residues and/or nucleotides to be treated with different length scales but also adapt a multiscale description of each residue and/or nucleotide. Unlike elastic network models that are parametrized at only one length scale for each particle, the mFRI provides a simultaneous multiscale description. Therefore, the present mFRI is able to better capture multiscale collective motions of protein–nucleic acid complexes. Additionally, many protein–nucleic acid complexes are very large biomolecules and pose difficulty to conventional mode decomposition-based methods. The **O**(N) scaling FRI methods provide an efficient approach to the flexibility analysis of large protein–nucleic acid complexes.

The rest of this article is organized as follows. Section II is devoted to methods and algorithms. To establish notation and facilitate further discussion, the basic FRI approach is briefly discussed. We then present the multikernel-based mFRI method to improve the accuracy of biomolecular flexibility analysis. The basic formulation of the aFRI is discussed. In Section III, we first analyze the benefit of adding an additional kernel with an appropriate length scale by comparing the performance of the B-factor prediction for a set of 64 protein–nucleic acid complexes introduced by Yang et al.[66] and a larger database of 203 high-resolution protein–nucleic acid structures. Three different coarse-grain representations of protein–nucleic acid complexes introduced by Yang et al.[66] are examined. Section IV is devoted to the application of the mFRI and aFRI methods. We consider a ribosomal structure to explore the utility and demonstrate the performance of the proposed mFRI. Further, we explore the use of aFRI for the prediction of collective motions of bridge helix, trigger loop, and nucleic acids in an RNA polymerase.

Methods and Algorithms

Flexibility–rigidity index

In FRI, the topological connectivity of a biomolecule is measured by rigidity index and flexibility index. In particular, the rigidity index represents the protein density profile. Consider an N atom representation of a biomolecule. The coordinates of these atoms are given as $\{\mathbf{r}_j | \mathbf{r}_j \in \mathbb{R}^3, j=1,2,\cdots,N\}$. We denote $||\mathbf{r}_i - \mathbf{r}_j||$ the Euclidean space distance between the ith atom and the jth atom. A general correlation kernel, $\Phi(||\mathbf{r}-\mathbf{r}_j||;\eta_j)$, is a real-valued monotonically decreasing radial basis function satisfying

$$\Phi(||\mathbf{r}-\mathbf{r}_j||;\eta_j)=1 \quad \text{as} \quad ||\mathbf{r}-\mathbf{r}_j|| \to 0 \tag{1}$$

$$\Phi(||\mathbf{r}-\mathbf{r}_j||;\eta_j)=0 \quad \text{as} \quad ||\mathbf{r}-\mathbf{r}_j|| \to \infty, \tag{2}$$

where η_j is an atomic type-dependent scale parameter. The correlation between the ith and jth particles is given by

$$C_{ij}=\Phi(||\mathbf{r}_i-\mathbf{r}_j||;\eta_j). \tag{3}$$

The correlation matrix $\{C_{ij}\}$ can be computed to visualize the connectivity among protein particles.

We define a position ($\mathbf{r}$)-dependent (continuous) rigidity function or density function[39,60]

$$\mu(\mathbf{r})=\sum_{j=1}^{N} w_j\Phi(||\mathbf{r}-\mathbf{r}_j||;\eta_j), \tag{4}$$

where w_j is an atomic type-dependent weight. For example, carbon, nitrogen, and phosphorus atoms can have different weights. Equation (4) can be understood as a discrete to continuum mapping. It maps a set of discrete values $\{w_j\}$ at $\{\mathbf{r}_j\}$ to the continuum domain. Although Delta sequences of the positive type discussed in an earlier work[57] are all good choices, generalized exponential functions

$$\Phi(||\mathbf{r}-\mathbf{r}_j||;\eta_j)=e^{-\left(||\mathbf{r}-\mathbf{r}_j||/\eta_j\right)^{\kappa}}, \quad \kappa>0 \tag{5}$$

and generalized Lorentz functions

$$\Phi(||\mathbf{r}-\mathbf{r}_j||;\eta_j)=\frac{1}{1+\left(||\mathbf{r}-\mathbf{r}_j||/\eta_j\right)^{\upsilon}}, \quad \upsilon>0 \tag{6}$$

have been commonly used in our recent work.[39,40,60] Since the rigidity function can be directly interpreted as a density distribution, it can be used to define the rigidity surface of a biomolecule by taking an isovalue. By taking $\kappa=2$ in eq. (5), we result in a formula for a Gaussian surface from eq. (4).

Similarly, we define a position ($\mathbf{r}$)-dependent (continuous) flexibility function[39,60]

$$F(\mathbf{r})=\frac{1}{\sum_{j=1}^{N} w_j\Phi(||\mathbf{r}-\mathbf{r}_j||;\eta_j)}. \tag{7}$$

This function is well defined in the computational domain containing the biomolecule. The flexibility function can be visualized by its projection on a given surface, such as the solvent-excluded surface of a biomolecule. The (discrete) rigidity index for the ith particle is obtained by restricting $\mathbf{r}$ to a given atomic position $\mathbf{r}_i$

$$\mu_i=\sum_{j=1}^{N} w_j\Phi(||\mathbf{r}_i-\mathbf{r}_j||;\eta_j). \tag{8}$$

Here, μ_i measures the total density or rigidity at the ith particle. In a similar manner, we define a set of (discrete) flexibility indices by

$$f_i=\frac{1}{\sum_{j=1}^{N} w_j\Phi(||\mathbf{r}_i-\mathbf{r}_j||;\eta_j)}. \tag{9}$$

The flexibility index f_i is directly associated with the B-factor of ith particle

$$B_i^t=af_i+b, \quad \forall i=1,2,\cdots,N \tag{10}$$

where $\{B_i^t\}$ are theoretically predicted B-factors, and a and b are two constants to be determined by a simple linear regression. This allows us to use experimental data to validate the FRI method. In our earlier work,[39,40,60] we set $w_j=1$ for the coarse-grained C_α representation of proteins. We have also developed parameter free FRI (pfFRI), such as $(\kappa=1,\eta=3)$ and $(\upsilon=3,\eta=3)$, to make our FRI robust for protein C_α B-factor prediction.

Multiscale flexibility–rigidity index

The basic idea of multiscale FRI or multikernel FRI (mFRI) is quite simple. Since macromolecules are inherently multiscale in nature, we utilize multiple kernels that are parametrized at multiple length scales to characterize the multiscale thermal fluctuations of macromolecules

$$f_i^n=\frac{1}{\sum_{j=1}^{N} w_j^n\Phi^n(||\mathbf{r}_i-\mathbf{r}_j||;\eta_j^n)}, \tag{11}$$

where w_j^n, $\Phi^n(||\mathbf{r}_i-\mathbf{r}_j||;\eta_j^n)$ and η_j^n are the corresponding quantities associated with the nth kernel. We seek the minimization of the form

$$\mathrm{Min}_{a^n,b}\left\{\sum_i\left|\sum_n a^n f_i^n+b-B_i^e\right|^2\right\} \tag{12}$$

where $\{B_i^e\}$ are the experimental B-factors. In principle, all parameters can be optimized. For simplicity and computational efficiency, we only determine $\{a^n\}$ and b in the above minimization process. For each kernel, Φ^n, w_j^n, and η_j^n will be selected according to the type of particles.

Specifically, for a simple C_α network, we can set $w_j^n=1$, $\eta_j^n=\eta^n$ and choose a single-kernel function parametrized at different scales. The predicted B-factors can be expressed as

$$B_i^{\mathrm{mFRI}}=b+\sum_{n=1}\frac{a^n}{\sum_{j=1}^{N}\Phi(||\mathbf{r}_i-\mathbf{r}_j||;\eta^n)}. \tag{13}$$

The difference between eqs. (11) and (13) is that, in eqs. (11), both the kernel and the scale can be changed for different n. In contrast, in eq. (13), only the scale is changed. One can use a given kernel, such as

$$\Phi(||\mathbf{r}-\mathbf{r}_j||;\eta^n)=\frac{1}{1+\left(||\mathbf{r}-\mathbf{r}_j||/\eta^n\right)^3}, \tag{14}$$

to achieve good multiscale predictions [40].

Anisotropic flexibility–rigidity index

The anisotropic flexibility–rigidity index (aFRI) model was built in a very unique manner. Different from the existing normal mode analysis or anisotropic elastic network models, in which the Hessian matrix is always global, our aFRI model delivers a local and adaptive Hessian matrix. This means that for a molecule with N particles, the Hessian matrix is always $3N\times 3N$ for ANM, whereas, our Hessian matrix may vary from a set of N 3

$\times$ 3 matrices for a completely local aFRI to $3N \times 3N$ for a complete global aFRI, depending on the need of a physical problem. For instance, if one is particularly interested in certain structures like alpha helices, domains, or binding sites of a protein, or certain subunits of a biomolecular complex, one design an aFRI for these portions of the molecule. We partition all the N particles in a molecule or a biomolecular complex into a total of M clusters $\{c_1, c_2, \cdots, c_k, \cdots, c_M\}$. Cluster c_k has N_k particles or atoms, so that $N=\sum_{k=1}^{M} N_k$. We choose clusters based on our physical interest as aforementioned. In this way, two very special situations can be found. The first one corresponds to the completely local situation, i.e., N clusters and each cluster has only one atom. The other situation contains only one cluster, which is then completely global. It is straightforward to construct a $3N \times 3N$ Hessian matrix and analyze the collective motion. The problem arises when we consider the global motion of a selected cluster, at the same time include the influence from the rest clusters. The essential idea is to construct a cluster Hessian matrix for each cluster individually and then incorporate the information from nearby clusters into its diagonal terms.

For example, if we want to know the thermal fluctuation of a particular cluster c_k with N_k particles or atoms, we need to find $3N_k$ eigenvectors for the cluster. Let us keep in mind that each position vector in $\mathbb{R}^3$ has three components, i.e., $\mathbf{r}=(x,y,z)$. For each given pair of particles i and j, we can define a local anisotropic matrix $\Phi^{ij}=\left(\Phi^{ij}_{uv}\right)$ as

$$\Phi^{ij}=\begin{pmatrix} \Phi^{ij}_{xx} & \Phi^{ij}_{xy} & \Phi^{ij}_{xz} \\ \Phi^{ij}_{yx} & \Phi^{ij}_{yy} & \Phi^{ij}_{yz} \\ \Phi^{ij}_{zx} & \Phi^{ij}_{zy} & \Phi^{ij}_{zz} \end{pmatrix}, \tag{15}$$

where Φ^{ij}_{uv} are defined as

$$\Phi^{ij}_{uv}=\frac{\partial}{\partial u_i}\frac{\partial}{\partial v_j}\Phi(||\mathbf{r}_i-\mathbf{r}_j||;\eta_j), \quad u,v=x,y,z; i,j=1,2,\cdots,N. \tag{16}$$

Due to the inner connection between rigidity and flexibility, we have two different aFRI algorithms. The specially designed cluster Hessian matrix with a smaller size can incorporate nonlocal geometric impact and predict collective thermal motions of the cluster. The details are presented below.

Anisotropic rigidity. In anisotropic rigidity-based aFRI, a rigidity Hessian matrix is needed. For a cluster c_k, if we denote its rigidity Hessian matrix as $\left(\mu^{ij}_{uv}(c_k)\right)$ with elements,

$$\mu^{ij}_{uv}(c_k)=-w_j\Phi^{ij}_{uv}, \quad i,j\in c_k; i\neq j; u,v=x,y,z \tag{17}$$

$$\mu^{ii}_{uv}(c_k)=\sum_{j=1}^{N} w_j\Phi^{ij}_{uv}, \quad i\in c_k; u,v=x,y,z \tag{18}$$

$$\mu^{ij}_{uv}(c_k)=0, \quad i,j\notin c_k; u,v=x,y,z. \tag{19}$$

In this way, the rigidity Hessian matrix is of $3N_k\times 3N_k$ dimensions. More importantly, the information from all other clusters are built into diagonal terms, even if the cluster itself is completely localized, i.e., $N_k=1, \ \forall k$.

For B-factor prediction, we define a set of anisotropic rigidity (AR)-based flexibility indices by

$$f_i^{\mathrm{AR}}=\frac{1}{\mu^i_{\mathrm{diag}}}, \tag{20}$$

where the ith diagonal term μ^i_{diag} is of the form,

$$\mu^i_{\mathrm{diag}}=\mathrm{Tr}\left(\mu^i_{uv}\right) \tag{21}$$

$$=\sum_{j=1}^{N} w_j\left[\Phi^{ij}_{xx}+\Phi^{ij}_{yy}+\Phi^{ij}_{zz}\right]. \tag{22}$$

Here, f_i^{AR} is employed in the linear regression to determine B-factors.

Anisotropic flexibility. The other way to construct aFRI is to construct a flexibility Hessian matrix, which is denoted as $\mathbf{F}(c_k)$ for cluster c_k with elements,

$$\mathbf{F}^{ij}(c_k)=-\frac{1}{w_j}(\Phi^{ij})^{-1}, \quad i,j\in c_k; i\neq j; u,v=x,y,z \tag{23}$$

$$\mathbf{F}^{ii}(c_k)=\sum_{j=1}^{N}\frac{1}{w_j}(\Phi^{ij})^{-1}, \quad i\in c_k; u,v=x,y,z \tag{24}$$

$$\mathbf{F}^{ij}(c_k)=0, \quad i,j\notin c_k; u,v=x,y,z. \tag{25}$$

Note that $(\Phi^{ij})^{-1}$ represents the unscaled inverse of matrix Φ^{ij} such that $\Phi^{ij}(\Phi^{ij})^{-1}=|\Phi^{ij}|$. The diagonalization of $\mathbf{F}(c_k)$ gives rise to eigenmodes, which represent the cluster motions. Additionally, the diagonal part $\mathbf{F}^{ii}(c_k)$ has built in information from all particles in the system. In this way, we deliver a cluster Hessian matrix. By diagonalizing $\mathbf{F}(c_k)$, we obtain $3N_k$ eigenvectors for the N_k particles in the cluster c_k of interest. Furthermore, instead of predicting the B-factors via the eigenvalues and eigenmodes, we directly predict the B-factors by using our anisotropic flexibility (AF)-based flexibility indices defined as,

$$f_i^{\mathrm{AF}}=\mathrm{Tr}(\mathbf{F}(c_k))^{ii}, \tag{26}$$

$$=(\mathbf{F}(c_k))^{ii}_{xx}+(\mathbf{F}(c_k))^{ii}_{yy}+(\mathbf{F}(c_k))^{ii}_{zz}. \tag{27}$$

Finally, we employ f_i^{AF} to predict B-factors.

Implementation and Validation

In this section, we parametrize and test the previously described mFRI on protein–nucleic acid structures. A immediate concern is whether the proposed mFRI is as efficient on protein–nucleic structures as it is on protein-only structures as shown in a previous study.[59] The accuracy of the mFRI method is tested by the B-factor prediction of two sets of protein–nucleic acid structures, including a set of 64 molecules used in a recent GNM study[66] and a set of 203 molecules for more accurate parametrization of mFRI.

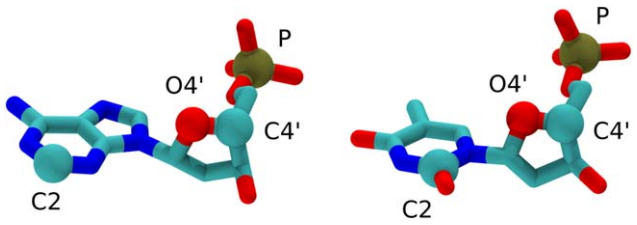


Figure 1. Illustration highlighting atoms used for coarse-grained representations in protein–nucleic acid complexes for FRI and GNM. In addition to protein Cα atoms, Model M1 considers the backbone P atoms for nucleotides. Model M2 includes M1 atoms and adds the sugar O4′ atoms for nucleotides. Model M3 includes M1 atoms and adds the sugar C4′ atoms and the base C2 atoms for nucleotides.

Coarse-grained representations of protein–nucleic acid complexes

In this section, we consider flexibility analysis of protein–nucleic acid complexes. To this end, we need coarse-grained representations. We consider three coarse-grained representation of nucleic acids to be used in conjugation with the Cα-only representation used for proteins. These three models are identical to those used by Yang et al.[66] and are named M1, M2, and M3. Model M1 consists of the backbone P atoms and protein Cα atoms. Model M2 contains the same atoms as M1 but also includes sugar O4′ atoms. Model M3 includes atoms from M1 and adds the sugar C4′ atoms and base C2 atoms (Fig. 1).

Model M1 is similar to protein Cα representations because they are both backbone-only representations. The atoms in M1 are 6 bonds apart while Cα atoms are 3 bonds apart. Model M2 includes P atoms and adds the O4′ atoms located on the ribose portion of the nucleotide. Finally, model M3 includes atoms of P, C4′, and C2, a carbon from the base portion of the nucleotide (Fig. 1). As pointed out by Yang et al.,[66] nucleotides are approximately three times more massive than amino acids and so model M3 with three nodes per nucleotide is consistent in this sense with using Cα atoms for the protein representation.

Multiscale/multikernel FRI

To parametrize and test the accuracy of multikernel fFRI on protein–nucleic acid structures, we use a dataset from Yang et al.[66] containing 64 structures. In addition, we construct a larger database of 203 high-resolution structures. This expanded protein–nucleic structure set was obtained by searching the Protein Data Bank (PDB) for structures that contain both Protein and DNA and have an X-ray resolution between 0.0 and 1.75 Å. All PDB files are processed by removing low occupancy atomic coordinates for structures having residues with multiple possible coordinates. The PDB IDs of the 64 and 203 structures can be found in Tables 1 and 2, respectively.

To quantitatively assess the performance of the proposed multikernel FRI method, we consider the correlation coefficient (CC)

$$\mathrm{CC}=\frac{\sum_{i=1}^{N}\left(B_i^e-\bar{B}^e\right)\left(B_i^t-\bar{B}^t\right)}{\left[\sum_{i=1}^{N}\left(B_i^e-\bar{B}^e\right)^2\sum_{i=1}^{N}\left(B_i^t-\bar{B}^t\right)^2\right]^{1/2}}, \tag{28}$$

where $\{B_i^t, i=1,2,\cdots,N\}$ are a set of predicted B-factors by using the proposed method and $\{B_i^e, i=1,2,\cdots,N\}$ are a set of experimental B-factors read from the PDB file. Here, $\bar{B}^t$ and $\bar{B}^e$ are the statistical averages of theoretical and experimental B-factors, respectively.

Multikernel FRI testing on protein–nucleic acid structures. Previous tests of single-kernel FRI indicate that the Lorentz type and exponential type correlation kernels are the two most accurate kernel types. This leads us to try the combination of these two types of kernels. The resulting multikernel FRI method requires four parameters, namely, κ and η, for the exponential kernel and υ and η for the Lorentz kernel.

Single-kernel FRI testing. To compare FRI and GNM methods for protein–nucleic acid structures, we test our single-kernel FRI at a range of η values. For this test, we use the Lorentz kernel with $\upsilon=3$ for B-factor prediction on both structure sets and all three representations (M1, M2, and M3). The results are shown in Figure 2. For the 64 structure set, single-kernel FRI has a maximum mean correlation coefficient (MCC) to experimental B-factors for M1, M2, and M3 representations of 0.620, 0.612, and 0.555. Comparatively, GNM had an MCC of approximately 0.59, 0.58, and 0.55 for M1, M2, and M3 for the same data set.[66] The maximum MCCs for FRI on the larger data set for M1, M2, and M3 are 0.613, 0.625, and 0.586, respectively. The M1 and M2 representations perform better than the M3 representation.

Parameter-free multikernel FRI. As with protein-only structures, we develop multikernel FRIs with multiple kernels to improve accuracy of prediction on protein–nucleic acid structures. To simplify the FRI method, we try to develop an accurate parameter-free version for a two-kernel mFRI. We use a combination of one Lorentz and one exponential kernel. Values for parameters υ and κ are set to 3.0 and 1.0, respectively, based on the results of previous FRI studies.[39] The optimal values for η in both kernels are determined by testing a range of possible values from 2 to 20 Å. All three representations (M1, M2, and M3) described previously are considered. The results of these tests on the set of 203 protein–nucleic acid structures are shown in Figure 3.

As expected, the addition of another kernel results in an overall increase in accuracy for the 203 complex set. For two-kernel mFRI, the MCCs increase up to 0.68 for M1, 0.67 for M2, and 0.63 for M3. The choice of η turns out to be very robust based on results shown in Figure 3.

We have also carried out a similar test of two-kernel mFRI ($\upsilon=3.0$ and $\kappa=1.0$) for the set of 64 protein–nucleic acid structures. Note that this has many large complexes. The MCCs for M1, M2, and M3 models are 0.668, 0.666, and 0.620, respectively, which are similar to what we have found for the

Table 1. Correlation coefficients (CCs) between predicted and experimental B-factors for the set of 64 protein–nucleic acid structures [66].

	M1		M2		M3	
PDB ID	CC	N1	CC	N2	CC	N3
1asy	0.647	1114	0.645	1248	0.631	1382
1b23	0.751	471	0.774	537	0.714	603
1c0a	0.763	653	0.704	721	0.598	789
1CX0	0.821	162	0.763	234	0.627	306
1drz	0.846	162	0.754	234	0.585	306
1efw	0.537	1286	0.647	1412	0.660	1538
1egk*	0.273	104	0.298	212	0.267	320
1ehz*	0.623	62	0.706	124	0.722	186
1evv*	0.710	62	0.769	124	0.770	186
1f7u	0.577	670	0.588	734	0.603	798
1ffk	0.759	6482	0.793	9310	0.809	12138
1ffy	0.520	991	0.549	1066	0.568	1141
1fg0*	0.720	498	0.723	996	0.721	1494
1fir*	0.687	61	0.576	122	0.439	183
1fjg	0.461	3915	0.585	5428	0.600	6941
1gid*	0.649	316	0.643	632	0.583	948
1gtr	0.724	603	0.747	677	0.645	751
1h3e	0.717	507	0.724	586	0.645	663
1h4s	0.671	1011	0.704	1076	0.626	1141
1hr2*	0.599	313	0.589	628	0.585	943
1i94	0.489	3923	0.615	5437	0.652	6951
1i9v*	0.615	73	0.631	147	0.642	220
1j1u	0.730	372	0.671	446	0.456	520
1j2b	0.686	1300	0.712	1448	0.672	1596
1j5a	0.532	3158	0.548	5932	0.510	8706
1j5e	0.427	3909	0.546	5422	0.553	6935
1jj2	0.799	6567	0.839	9443	0.836	12319
1jzx	0.586	3158	0.600	5932	0.561	8706
1l8v*	0.700	312	0.688	626	0.672	940
1l9a	0.849	211	0.789	336	0.675	461
1lng	0.780	183	0.595	280	0.405	377
1m5k	0.904	402	0.841	622	0.760	842
1m5o	0.921	405	0.872	629	0.810	853
1mfq	0.773	341	0.688	468	0.543	595
1mms	0.507	317	0.548	433	0.646	549
1n32	0.388	3916	0.494	5447	0.517	6978
1nbs*	0.547	270	0.566	540	0.573	810
1o0c	0.766	602	0.758	676	0.636	750
1qf6	0.608	710	0.578	779	0.540	848
1qrs	0.671	603	0.672	677	0.586	751
1qtq	0.620	602	0.640	676	0.596	750
1qu2	0.520	991	0.549	1066	0.568	1141
1qu3	0.579	954	0.599	1029	0.613	1104
1rc7	0.599	256	0.566	296	0.470	336
1s72	0.823	6636	0.839	9507	0.831	12378
1ser	0.748	855	0.743	917	0.657	978
1sj3	0.880	167	0.805	240	0.614	313
1tn2*	0.686	62	0.712	124	0.676	186
1tra*	0.624	62	0.670	124	0.660	186
1ttt	0.578	1401	0.564	1587	0.515	1773
1u0b	0.757	535	0.754	609	0.621	683
1u6b	0.476	312	0.490	531	0.506	750
1u9s*	0.446	155	0.432	310	0.419	465
1vby	0.877	167	0.792	240	0.587	313
1vc0	0.878	167	0.804	240	0.611	313
1vc5	0.861	164	0.840	234	0.685	304
1y0q*	0.491	230	0.484	463	0.472	696
1y26*	0.677	70	0.697	141	0.709	212
1yfg*	0.565	64	0.600	128	0.623	192
1yhq	0.835	6636	0.840	9507	0.831	12378
1yij	0.836	6636	0.851	9507	0.842	12378
2tra*	0.614	65	0.614	130	0.613	195
3tra*	0.645	64	0.615	128	0.620	192
4tra*	0.679	62	0.715	124	0.694	186

Here N1, N2, and N3 values represent the number of atoms used for the M1, M2, and M3 representations, respectively, for each structure. We use the parameter-free two-kernel mFRI model, i.e., one exponential kernel ($\kappa = 1$ and $\eta = 18$ Å) and one Lorentz kernels ($\upsilon = 3$, $\eta = 18$ Å). PDB IDs marked with an asterisk (*) indicate structure containing only nucleic acid residues.

Table 2. The PDB IDs of the 203 high-resolution protein–nucleic acid structures used in our single-kernel FRI parameter test.

PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID
1A1H	1A1I	1AAY	1AZP	1BF4	1C8C	1D02	1D2I	1DC1	1DFM
1DP7	1DSZ	1EGW	1EON	1F0V	1FIU	1H6F	1I3W	1JK2	1JX4
1K3W	1K3X	1L1Z	1L3L	1L3S	1L3T	1L3V	1LLM	1MNN	1NJX
1NK0	1NK4	1OJ8	1ORN	1PFE	1QUM	1R2Z	1RFF	1RH6	1SX5
1T9I	1U4B	1VTG	1WTO	1WTQ	1WTV	1XJV	1XVK	1XVN	1XVR
1XYI	1ZS4	2ADW	2AXY	2BCQ	2BCR	2BOP	2C62	2C7P	2EA0
2ETW	2EUW	2EUX	2EUZ	2EVF	2EVG	2FMP	2GB7	2HAX	2HEO
2HHV	2IBT	2IH2	2ITL	2NQ9	2O4A	2OAA	2ODI	2P2R	2PY5
2Q10	2R1J	2VLA	2VOA	2WBS	2XHI	2Z70	2ZKD	3BIE	3BKZ
3BM3	3BS1	3D2W	3EY1	3EYI	3FC3	3FDE	3FDQ	3FSI	3FYL
3G00	3G9M	3G9O	3G9P	3GO3	3GOX	3GPU	3GQ4	3HPO	3HT3
3HTS	3I0W	3I2O	3I3M	3I49	3I8D	3IGK	3JR5	3JX7	3JXB
3JXY	3JXZ	3KDE	3KXT	3M4A	3MR3	3MXM	3NDH	3O1M	3O1P
3O1S	3O1T	3O1U	3OQG	3PV8	3PVI	3PX0	3PX4	3PX6	3PY8
3QEX	3RKQ	3RZG	3S57	3S5A	3SAU	3SJM	3TAN	3TAP	3TAQ
3TAR	3THV	3TI0	3U6E	3U6P	3V9W	3ZDA	3ZDB	3ZDC	3ZDD
4A75	4B21	4B9S	4DFK	4DQI	4DQP	4DQQ	4DS4	4DS5	4DSE
4DSF	4E0D	4ECQ	4ECV	4ECX	4ED0	4ED2	4ED7	4ED8	4EZ6
4F1H	4F2R	4F2S	4F3O	4F4K	4F8R	4FPV	4GZ1	4GZN	4HC9
4HIK	4HIM	4HLY	4HTU	4HUE	4HUF	4HUG	4IBU	4IX7	4KLG
4KLI	4KLM	4KMF							

IDs marked with an asterisk indicate those containing only nucleic acids residues.

set of 203 structures. The set of 64 structures includes 19 structures composed of nucleic acids and no amino acids. The MCCs for this nucleic acid-only subset are 0.608, 0.617, and 0.603 for M1, M2, and M3 models, respectively. The correlation coefficients for all 64 individual molecular complexes are listed in Table 1.

To summarize the performance of Gaussian network model, single-kernel FRI, and two-kernel mFRI, we list their MCCs for the 64 protein–nucleic acid structures in Table 3. It can be seen that the FRI outperforms GNM in all three representations, and two-kernel mFRI further significantly improves the accuracy of our method and achieves up to 15% improvement compared with GNM.[66] Based on our earlier test,[40] we believe that our three-kernel mFRI can deliver a better prediction.

Applications

In this section, we briefly explore the applications of the mFRI and aFRI methods to large protein–nucleic acid complexes. We highlight a few particular examples where mFRI improves upon previous FRI methods, in particular, for the flexibility prediction of ribosomes. Further, we show how aFRI is well suited for the study of the dynamics of large macromolecular complexes using the bacterial RNA polymerase active site as an example.

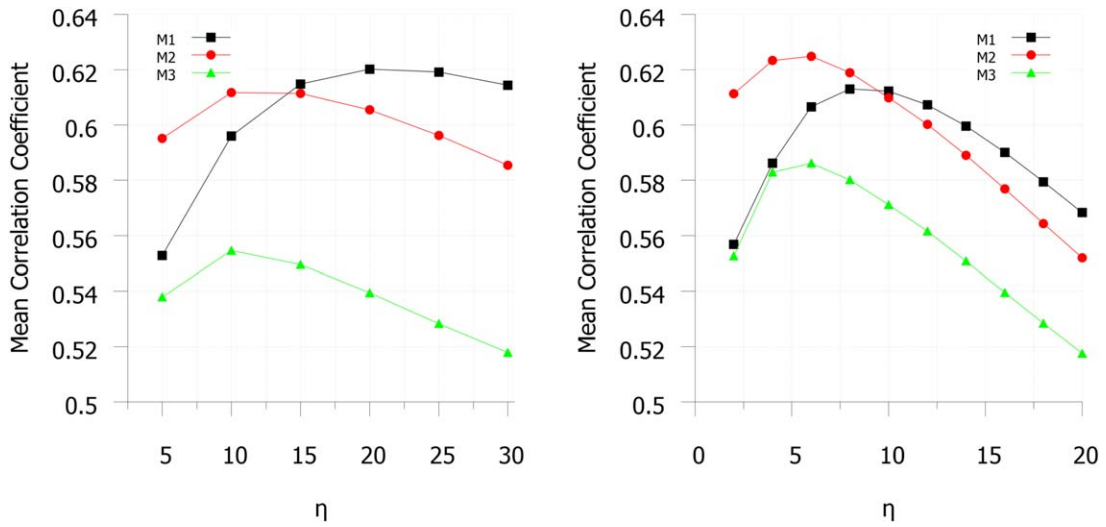


Figure 2. MCCs for single kernel parameter test using the M1 (squares), M2 (circles), and M3 (triangles) representations. Lorentz kernel with $\upsilon = 3$ is used. The parameter η is varied to find the maximum MCC on the test set of structures. The results for a set of 64 protein–nucleic acid structures (PDB IDs listed in Table 1) are shown on the left, while results for a separate set of 203 structures (PDB IDs listed in Table 2) is shown on the right for more general selections.

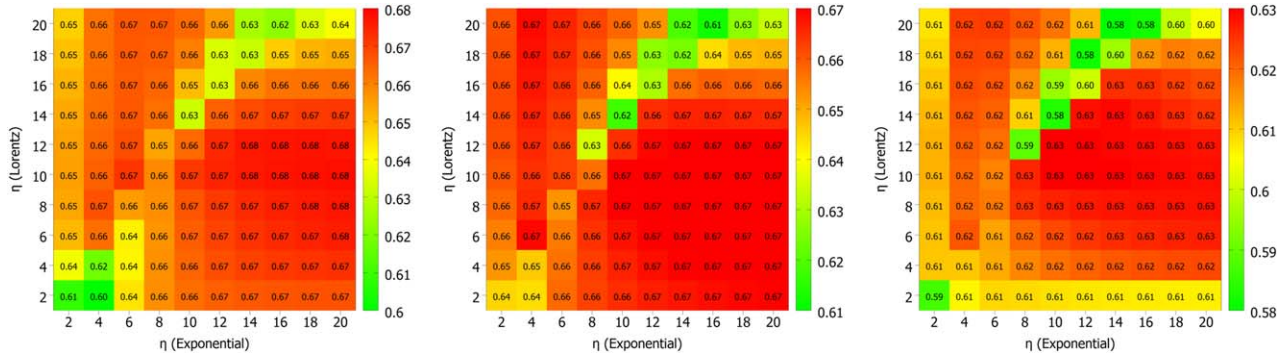


Figure 3. Mean correlation coefficients (MCCs) for two-kernel FRI models on a set of 203 protein–nucleic acid structures. From left to right, MCC values are shown for M1, M2, and M3 representations. We use one Lorentz kernel with $\upsilon = 3.0$ and one exponential kernel with $\kappa = 1.0$. The values of parameter η for both kernels are varied from 2 to 20 Å. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Multikernel FRI flexibility prediction for protein–nucleic acid structures—ribosomes

Some of the largest and most biologically important structures that contain both protein and nucleic acids are ribosomes. Ribosomes are the protein synthesizers of the cell and connect amino acid into polymer chains. In ribosomes, proteins and RNA interact through intermolecular effects, such as electrostatic interactions, hydrogen bonding, hydrophobic interactions, base stacking, and base pairing. RNA tertiary structures can significantly influence protein–RNA interactions. Ribosomes are primarily composed of RNA with many smaller associated proteins as shown in Figure 4. The top of Figure 4 shows the 50S subunit of the ribosome (PDB ID: 1YIJ) with the nucleic acids in a smooth surface representation with the protein subunits bound and shown in a secondary structure representation. The set of 64 structures used in our tests contains a number of ribosomal subunits. Due to their multiscale nature, these structures also happen to be among those that benefit the most from using multikernel FRI over single-kernel FRI or GNM. For example, in the case of ribosome 50S subunit structure (PDB ID:1YIJ), B-factor prediction with three-kernel FRI yields a CC value of 0.85, while that of single-kernel FRI is only around 0.3. GNM does not provide a good B-factor prediction for this structure either. The three-kernel mFRI model we used is one exponential kernel ($\kappa = 1$ and $\eta = 15$ Å) and two Lorentz kernels ($\upsilon = 3$, $\eta = 3$ Å, and $\upsilon = 3$, $\eta = 7$ Å). The comparison between mFRI-predicted and experimental B-factors for ribosome 50S subunit structure is demonstrated in Figure 4.

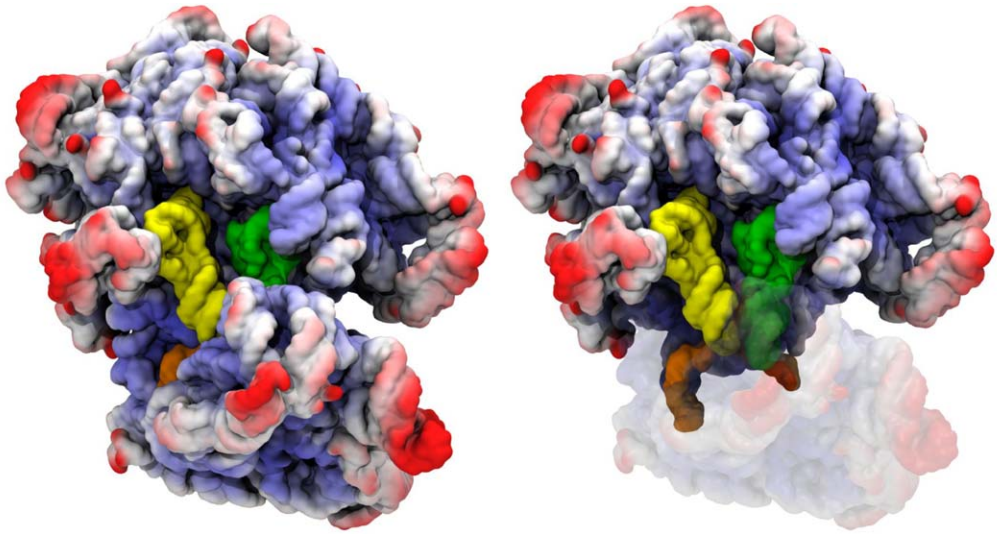
By using the fitting coefficients from the above 50S subunit (1YIJ) flexibility analysis, we have obtained flexibility predictions for the entire ribosome (PDB ID:4V4J) as well as many protein subunits and other RNAs that associate with it (Fig. 4). To avoid confusion, the B-factors for 4V4J are uniquely determined by using not only the same three-kernel mFRI model from the case 1YIJ, and also its fitting parameters, i.e., $a^1 =, a^2 =, a^3$, and b. Again, the FRI values are mapped by color to the smooth surface of the nucleic acids; however, in these bottom figures, the protein subunits are omitted to draw attention instead to the various types of RNA involved in this structure.

Anisotropic FRI for conformational motion prediction of an RNA polymerase

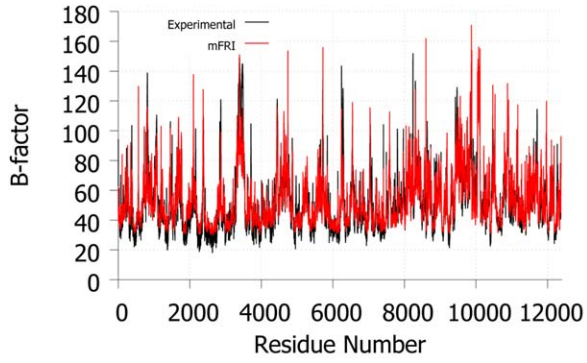
RNA polymerase is one of the essential enzymes for all life on Earth as we know it today and possibly from the very beginning of life.[7,25] Despite this importance, the mechanisms for many of the polymerase's functions are still not well understood on the atomic level. Considerable effort has been spent both experimentally and computationally to understand RNAP polymerase function in more detail but many questions remain. The study of RNA polymerase experimentally or computationally is difficult and often expensive due to the size of the system and variety of molecules involved. The minimal required elements for a bacterial or eukaryotic RNA polymerase include multiple protein subunits, a double-stranded DNA molecule, a single-stranded RNA molecule, free nucleotides, various ions (Mg^{2+}, Zn^{2+}, Na^{+}, etc.), and solvent. A typical setup for this system in all-atom molecular dynamics includes 300,000 atoms when solvated. With this number of atoms and current computer power, it is often not feasible to simulate these molecules on biologically relevant timescales using MD. Perhaps the most popular tool for studying long-time dynamics of biomolecules is normal mode analysis (NMA) and its related methods such as the anisotropic network model (ANM). These methods have been successfully used to study protein dynamics for many proteins; however, at their maximum accuracy, their computational complexity is of $\mathbf{O}(N^3)$, where N is the number of atoms. This is a problem because many cellular functions involve a large number of macromolecules with many thousands to millions of residues to consider. Therefore, future computational studies of biomolecules

Table 3. MCCs of Gaussian network model (GNM) [66], single-kernel flexibility–rigidity index (FRI), and two-kernel mFRI for three coarse-grained representations (M1, M2, and M3). A set of 64 protein–nucleic acid structures [66] is used.

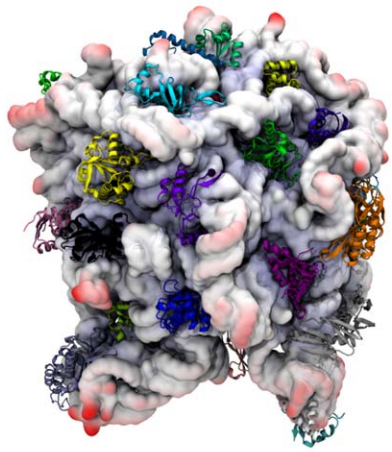
	GNM [66]	FRI	Two-kernel mFRI
M1	0.59	0.620	0.666
M2	0.58	0.612	0.668
M3	0.55	0.555	0.620



(a) Complete ribosome with bound tRNAs PDB ID: 4V4J.



(b) Ribosome 50S subunit PDB ID: 1YIJ B factors



(c) Ribosome 50S subunit PDB ID: 1YIJ

Figure 4. Complete ribosome with bound tRNAs (yellow (A site) and green (P site)) and mRNA Shine-Delgarno sequence (orange) PDB ID: 4V4J. The same correlation coefficients and fitting parameters from mFRI model of protein 1YIJ are used. A comparison of predicted and experimental B-factor data for Ribosome 50S subunit PDB ID: 1YIJ. The CC value is 0.85 using the parameter-free three-kernel mFRI model. Nucleic acids are shown as a smooth surface colored by FRI flexibility values (red for more flexible regions) while bound protein subunits are colored randomly and shown in a secondary structure representation. We achieve a CC value up to 0.85 using parameter-free three-kernel mFRI model, i.e., one exponential kernel ($\kappa = 1$ and $\eta = 15$ Å) and two Lorentz kernels ($\upsilon = 3$, $\eta = 3$ Å, and $\upsilon = 3$, $\eta = 7$ Å). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

beyond the protein scale will require methods with better scaling properties such as FRI and aFRI.

In this example, we use completely local anisotropic FRI to examine correlated motions in regions near the active site of bacterial RNA polymerase, including the bridge helix, trigger loop, and nucleic acid chains. We examine the relationship between these components' motions and their contributions to critical functions such as catalysis and translocation. We use the anisotropic rigidity form in the section "Anisotropic rigidity" with the Lorentz kernel ($\upsilon = 2$ and $\eta = 3$ Å). Figure 5a is a simplified representation of RNA polymerase (PDB ID 2PPB) that shows these important features which are buried in the core of the largest protein subunits, β and β'. The bridge helix and trigger loop, shown in green and blue, respectively, are parts of the protein that have been implicated in most of the essential functions of the polymerase. Mutational studies of these regions result in modulation of the polymerase speed and accuracy, both positively and negatively, indicating the regions are important for normal functioning of the enzyme. How these regions aid these functions and how they interact remains an open question. With this demonstration of local aFRI analysis, we hope to shed some light on how these essential parts of RNA polymerase work together.

Local aFRI, as described in earlier work, is much less computationally costly than global aFRI or NMA and has been shown to have qualitatively similar results for small to large size single proteins. To further validate the local aFRI method, we compare the conclusions from a local aFRI study of RNAP to those

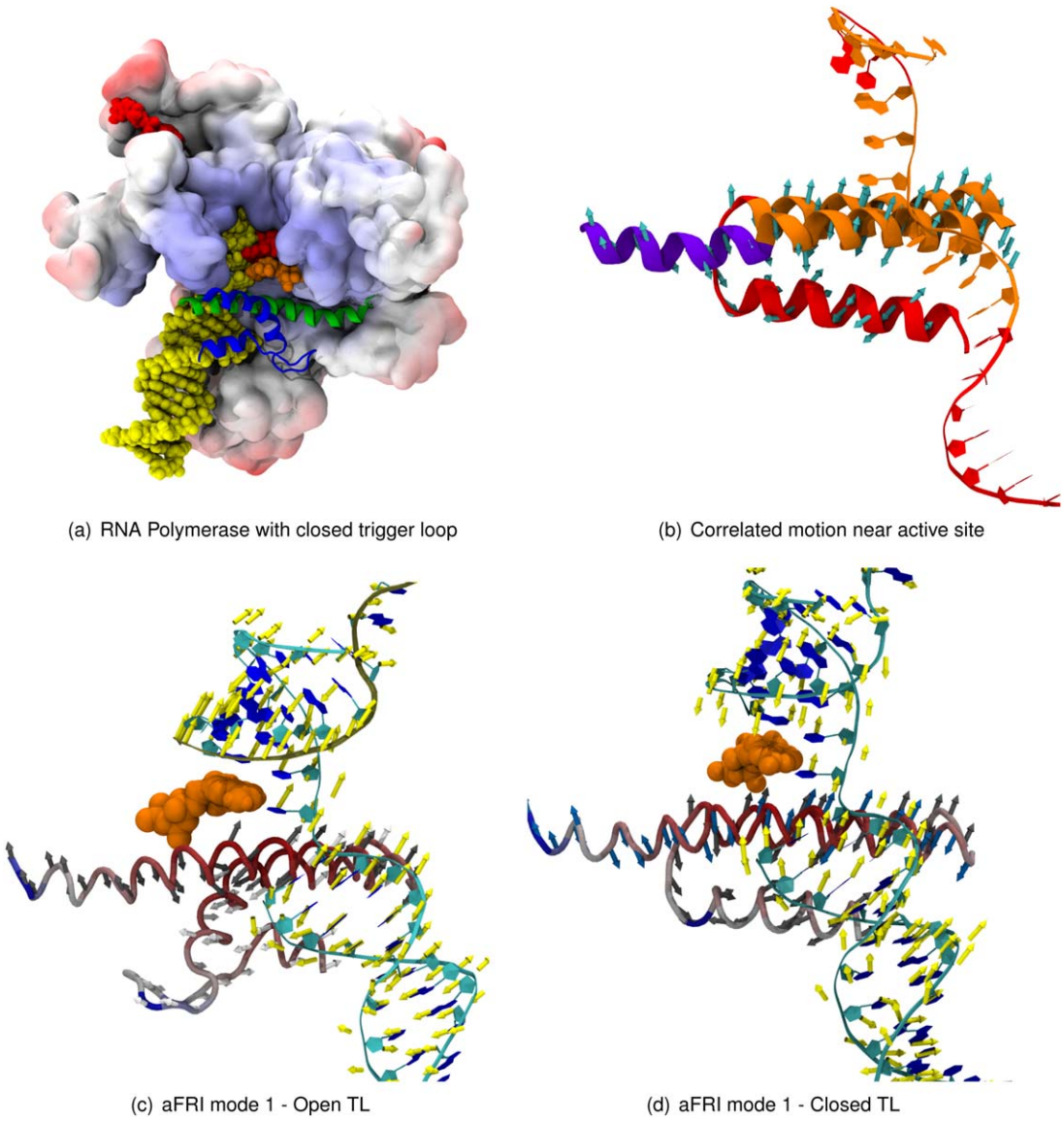


Figure 5. The first RNAP local FRI mode for the bridge helix, trigger loop, and nucleic acids from both open (PDB ID: 2PPB) and close (PDB ID: 2O5J) configurations. Arrows represent the direction and relative magnitude of atomic fluctuations. Arrows for the bridge helix, trigger loop, and nucleic acids are pictured as blue, white, and yellow, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of NMA based studies. The RNA polymerase elongation complex is a relatively large system but it is still tenable for NMA methods. NMA has been applied to both bacterial and eukaryotic RNA polymerase in the past[13,53] which provides us with a point of comparison for our results.

Local aFRI produces three modes of motion sorted from lowest to highest frequency vibration according to eigenvalue as in NMA. In Figure 5, we present findings from the lowest frequency mode effectively focusing on the most dominant motion of each conformation. Two major conformations of RNA polymerase are considered, those with open and closed trigger loop regions (Figs. 5c and 5d). A closed trigger loop is one that is completely folded into two parallel alpha helices while an open trigger loop has a region of disordered loop between two shorter helices and is slightly bent away from the bridge helix. The closing or folding of the trigger loop into the closed conformation is assumed to follow binding of an NTP in the active site and to precede catalysis. After catalysis, it is suspected that the trigger loop opens or unfolds to facilitate translocation and permit new NTPs to enter the active site.

The results of aFRI analysis on the effect of trigger loop closing reveal a distinct change in correlated motions in open and closed trigger loop conformations. These changes involve interactions between the bridge helix, the trigger loop, and the nucleic acid regions. In Figure 5b, regions of high correlation are color coded which reveals that the bridge helix is composed of two highly self-correlated portions suggesting the presence of a hinge in the bridge helix. In fact, the central portion of the bridge helix has been observed as a kinked or bent helix in a yeast RNAP structure.[55] Additionally, it is observed that a portion of the bridge helix and the N-terminal

helix of the trigger loop are highly correlated in the closed trigger loop structure only. This set of two helices is situated directly next to the active site and could provide stability to aid catalysis after trigger loop closing.

Additionally, correlation between nucleic acids and protein shows marked differences from the open trigger loop to closed trigger loop structures. The motions indicated in Figures 5c and 5d show that the open trigger loop structure is primed to translocate based on the direction of highly correlated motions of the upstream and downstream nucleic acids. By contrast, the closed trigger loop nucleic acid motions are considerably less correlated and not in the direction of translocation. This is the expected relationship as it matches the results from previous biological and NMA studies of RNA polymerase.[13]

These differences between a closed trigger loop and open trigger loop structure reveal potentially important structural changes that arise as the RNA polymerase switches between open and closed trigger loop conformations during the transition between translocation and catalysis. Specifically, the results for the closed trigger loop conformation suggest the presence of a stabilized catalytic area which is made of the N-terminal helix of the trigger loop and the bridge helix. The results for the open trigger loop conformation show no such coordination of the active site helices and instead indicate a less defined hinge and coordinated motion in the direction of translocation. Taken together these results provide a potential explanation for how trigger loop opening and closing is correlated with translocation and catalysis, respectively.

Concluding Remarks

Protein–nucleic acid complexes are essential to all living organisms. The function of these complexes depends crucially on their flexibility, an intrinsic property of a macromolecule. However, for many large protein–nucleic acid complexes, such as ribosomes and RNA polymerases, the present flexibility analysis approaches can be problematic due to their computational complexity of $\mathbf{O}(N^3)$ and neglecting multiscale effects.

This work introduces the FRI methods[39,40,60] for the flexibility analysis of protein–nucleic acid structures. We show that a multiscale FRI (mFRI) realized by multiple kernels parametrized at multiple length scales is able to significantly outperform the Gaussian network model (GNM) for the B-factor prediction of a set of 64 protein–nucleic acid complexes.[66] The FRI methods are not only accurate but also efficient, as their computational complexity is $\mathbf{O}(N)$. Additionally, anisotropic FRI (aFRI), which has cluster Hessian matrices, offers collective motion analysis for any given cluster, i.e., subunit or domain in a biomolecular complex.

We apply FRI methods to a ribosomal subunit (1YIJ) with multiple subunits. We note that both original single-scale FRI and GNM do not work well for this structure. It is found that the multiscale strategy is crucial for the flexibility analysis of multisubunit structures. The correlation coefficients between FRI predictions and experimental B-factors for 1YIJ improve from 0.3 for single-scale FRI to 0.85 for multiscale FRI. We further use the fitting coefficients obtained from 1YIJ to predict the flexibility of an entire ribosome, 4V4J. We find that mFRI has an advantage for analyzing large biomolecular complexes due to both higher speeds and accuracy.

We also demonstrate the utility of the anisotropic FRI (aFRI) for analyzing the translocation of an RNA polymerase, which involves protein, DNA, RNA, nucleotide substrates, and various ions. Both experimental and computational studies of RNA polymerases are difficult and expensive due to the size and complexity of the biomolecular complex. The molecular mechanism of RNA polymerase translocation is an interesting, open-research topic. This work makes use of localized aFRI to elucidate the synergistic local motions of a bacterial RNA polymerase. Our findings are consistent with those from much more expensive molecular dynamics simulations and normal mode analysis.[13,14]

The study of hinges has been an important topic and much research has been done in the past.[12,15,16,27,43] Identification of hinge residues is useful for inferring motion and function when molecules are too large for MD simulation on relevant timescales. Other methods, such as GNM and NMA, have been utilized for this purpose. FRI-based methods could play a significant role in hinge analysis. This aspect will be carefully analyzed in our future work.

Acknowledgment

The authors acknowledge the Mathematical Biosciences Institute for hosting valuable workshops.

Keywords: thermal fluctuation · atomic flexibility · protein–nucleic acid complex · multiscale

How to cite this article: K. Opron, K. Xia, Z. Burton, G.-W. Wei. *J. Comput. Chem.* **2016**, *37*, 1283–1295. DOI: 10.1002/jcc.24320

[1] M. P. Allen, D. J. Tildesley. *Computer Simulation of Liquids*; Clarendon Press: Oxford, **1987**.
[2] C. B. Anfinsen, *Science* **1973**, *181*, 223.
[3] A. R. Atilgan, S. R. Durrell, R. L. Jernigan, M. C. Demirel, O. Keskin, I. Bahar, *Biophys J* **2001**, *80*, 505.
[4] I. Bahar, A. R. Atilgan, M. C. Demirel, B. Erman, *Phys Rev Lett* **1998**, *80*, 2733.
[5] I. Bahar, A. R. Atilgan, B. Erman, *Fold Des* **1997**, *2*, 173.
[6] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. States, S. Swaminathan, M. Karplus, *J Comput Chem* **1983**, *4*, 187.
[7] Z. F. Burton, *Transcription* **2014**, *5*, e28674.
[8] Q. Cui, I. Bahar. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*; Chapman and Hall/CRC, **2010**.
[9] Q. Cui, G. J. Li, J. Ma, M. Karplus, *J Mol Biol* **2004**, *340*, 345.
[10] O. N. A. Demerdash, J. C. Mitchell, *Protein Struct Funct Bioinform* JUL 2012, *80*, 1766.
[11] O. K. Dudko, G. Hummer, A. Szabo, *Phys Rev Lett* **2006**, *96*, 108101.
[12] U. Emekli, S. Dina, H. Wolfson, R. Nussinov, T. Haliloglu, *Proteins* **2008**, *70*, 1219.
[13] M. Feig, Z. F. Burton, *Protein Struct Funct Bioinform* **2010**, *78*, 434.
[14] M. Feig, Z. F. Burton, *Biophys J* **2010**, *99*, 2577.
[15] S. Flores, M. Gerstein, *BMC Bioinf* **2007**, *8*, 125.
[16] S. Flores, L. Lu, J. Yang, N. Carriero, M. Gerstein, *BMC Bioinf* **2007**, *8*, 167.
[17] P. J. Flory, *Proc R Soc Lond A* **1976**, *351*, 351.

[18] H. Frauenfelder, S. G. Slihar, P. G. Wolynes, *Science* DEC 13 1991, *254*, 1598.
[19] Z. N. Gerek, S. B. Ozkan, *Protein Sci* **2010**, *19*, 914.
[20] N. Go, T. Noguti, T. Nishikawa, *Proc Natl Acad Sci USA* **1983**, *80*, 3696.
[21] B. Halle, *Proc Natl Acad Sci USA* **2002**, *99*, 1274.
[22] K. Hinsen, *Proteins* **1998**, *33*, 417.
[23] K. Hinsen, *Bioinformatics* **2008**, *24*, 521.
[24] S. W. Huang, C. H. Shih, C. P. Lin, J. K. Hwang, *Theor Chem Acc* **2008**, *121*, 197.
[25] L. M. Iyer, E. V. Koonin, L. Aravind, *BMC Struct Biol* **2003**, *3*, 1):1.
[26] D. J. Jacobs, A. J. Rader, L. A. Kuhn, M. F. Thorpe, *Protein Struct Funct Genet* **2001**, *44*, 150.
[27] K. S. Keating, S. C. Flores, M. B. Gerstein, L. A. Kuhn, *Protein Sci* **2009**, *18*, 359.
[28] O. Keskin, I. Bahar, D. Flatow, D. G. Covell, R. L. Jernigan, *Biochemistry* **2002**, *41*, 491.
[29] D. A. Kondrashov, A. W. Van Wynsberghe, R. M. Bannen, Q. Cui, J. G. N. Phillips, *Structure* **2007**, *15*, 169.
[30] M. Krone, J. E. Stone, T. Ertl, K. Schulten, *EuroVis-Short Papers* **2012**, *2012*, 67.
[31] S. Kundu, J. S. Melton, D. C. Sorensen, J. G. N. Phillips, *Biophys J* **2002**, *83*, 723.
[32] M. Levitt, C. Sander, P. S. Stern, *J Mol Biol* **1985**, *181*, 423.
[33] D. W. Li, R. Brüschweiler, *Biophys J* **2009**, *96*, 3074.
[34] G. H. Li, Q. Cui, *Bipohys J* **2002**, *83*, 2457.
[35] C. P. Lin, S. W. Huang, Y. L. Lai, S. C. Yen, C. H. Shih, C. H. Lu, C. C. Huang, J. K. Hwang, *Protein Struct Funct Bioinform* **2008**, *72*, 929.
[36] D. R. Livesay, S. Dallakyan, G. G. Wood, D. J. Jacobs, *FEBS Lett* **2004**, *576*, 468.
[37] J. P. Ma, *Structure* **2005**, *13*, 373.
[38] J. A. McCammon, B. R. Gelin, M. Karplus, *Nature* **1977**, *267*, 585.
[39] K. Opron, K. L. Xia, G. W. Wei, *J Chem Phys* **2014**, *140*, 234105.
[40] K. Opron, K. L. Xia, G. W. Wei, *J Chem Phys* **2015**, *142*, 211101.
[41] J. K. Park, R. Jernigan, Z. Wu, *Bull Math Biol* **2013**, *75*, 124.
[42] A. J. Rader, D. H. Vlad, I. Bahar, *Structure* **2005**, *13*, 413.
[43] M. Shatsky, R. Nussinov, H. J. Wolfson, *J Comput Biol* **2004**, *11*, 83.
[44] L. Skjaerven, S. M. Hollup, N. Reuter, *J Mol Struct Theochem* **2009**, *898*, 42.,
[45] G. Song, R. L. Jernigan, *J Mol Biol* **2007**, *369*, 880.
[46] F. Tama, C. K. Brooks III, *J Mol Biol* **2005**, *345*, 299.
[47] F. Tama, F. X. Gadea, O. Marques, Y. H. Sanejouand, *Protein Struct Funct Bioinform* **2000**, *41*, 1.
[48] F. Tama, Y. H. Sanejouand, *Protein Eng* **2001**, *14*, 1.
[49] F. Tama, M. Valle, J. Frank, C. K. Brooks III, *Proc Natl Acad Sci USA* **2003**, *100*, 9319.
[50] M. Tasumi, H. Takenchi, S. Ataka, A. M. Dwidedi, S. Krimm, *Biopolymers* **1982**, *21*, 711.
[51] M. M. Tirion, *Phys Rev Lett* **1996**, *77*, 1905.
[52] M. Topf, K. Lasker, B. Webb, H. Wolfson, W. C. A. Sali, *Structure* **2008**, *16*, 295.
[53] A. Van Wynsberghe, G. Li, Q. Cui, *Biochemistry* **2004**, *43*, 13083.
[54] C. W. von der Lieth, K. Stumpf-Nothof, U. Prior, *J Chem Inform Comput Sci* **1996**, *36*, 711.
[55] D. Wang, D. A. Bushnell, K. D. Westover, C. D. Kaplan, R. D. Kornberg, *Cell* **2006**, *127*, 941.
[56] Y. Wang, A. J. Rader, I. Bahar, R. L. Jernigan, *J Struct Biol* **2004**, *147*, 302.
[57] G. W. Wei, *J Phys A Math Gen* **2000**, *33*, 8577.
[58] W. Wriggers, R. A. Milligan, J. A. McCammon, *J Struct Biol* **1999**, *125*, 185.
[59] K. L. Xia, X. Feng, Y. Y. Tong, G. W. Wei, *J Comput Chem* **2015**, *36*, 408.
[60] K. L. Xia, K. Opron, G. W. Wei, *J Chem Phys* **2013**, *139*, 194109.
[61] K. L. Xia, G. W. Wei, *Phys Rev E* **2013**, *88*, 062709.
[62] K. L. Xia, G. W. Wei, *Chaos* **2014**, *24*, 013103.
[63] K. L. Xia, G. W. Wei, *Int J Num Met Biomed Eng* **2015**, *31*, e02719.
[64] K. L. Xia, Z. X. Zhao, G. W. Wei, *J Comput Biol* **2015**, *22*, 1.
[65] C. Xu, D. Tobi, I. Bahar, *J Mol Biol* **2003**, *333*, 153.
[66] L. Yang, A. Rader, X. Liu, C. Jursa, S. Chen, H. Karimi, I. Bahar, *Nucleic Acids Res* **2006**, *34*, W24.
[67] L. W. Yang, C. P. Chng, *Bioinf Biol Insights* **2008**, *2*, 25.
[68] F. L. Zhang, R. Brüschweiler, *J Am Chem Soc* **2002**, *124*, 12654.
[69] W. Zheng, B. R. Brooks, D. Thirumalai, *Biophys J* **2007**, *93*, 2289.
[70] W. J. Zheng, S. Doniach, *Proc Natl Acad Sci USA* **2003**, *100*, 13253.,

Received: 18 June 2015
Revised: 2 December 2015
Accepted: 17 January 2016
Published online on 1 March 2016