Stochastic model for protein flexibility analysis

Xia, Kelin; Wei, Guo-Wei

2013

Xia, K., & Wei, G.-W. (2013). Stochastic model for protein flexibility analysis. Physical Review E, 88, 062709-.

https://hdl.handle.net/10356/82116

https://doi.org/10.1103/PhysRevE.88.062709

Stochastic model for protein flexibility analysis

Kelin Xia[1] and Guo-Wei Wei[1,2,3,*]

[1]*Department of Mathematics, Michigan State University, Michigan 48824, USA*

[2]*Department of Electrical and Computer Engineering, Michigan State University, Michigan 48824, USA*

[3]*Department of Biochemistry and Molecular Biology, Michigan State University, Michigan 48824, USA*

(Received 6 August 2013; revised manuscript received 3 October 2013; published 10 December 2013)

Protein flexibility is an intrinsic property and plays a fundamental role in protein functions. Computational analysis of protein flexibility is crucial to protein function prediction, macromolecular flexible docking, and rational drug design. Most current approaches for protein flexibility analysis are based on Hamiltonian mechanics. We introduce a stochastic model to study protein flexibility. The essential idea is to analyze the free induction decay of a perturbed protein structural probability, which satisfies the master equation. The transition probability matrix is constructed by using probability density estimators including monotonically decreasing radial basis functions. We show that the proposed stochastic model gives rise to some of the best predictions of Debye-Waller factors or B factors for three sets of protein data introduced in the literature.

DOI: 10.1103/PhysRevE.88.062709 PACS number(s): 87.15.−v, 87.14.et

I. INTRODUCTION

The importance of proteins to life and living organisms cannot be overemphasized. Apart from providing structural support in terms of tubulin, collagen, elastin, and keratin, proteins also catalyze all of the reactions of metabolism, regulate transcription and cell cycle, participate in signal transduction, and work as immune agents. For a long time, protein functions were believed to be determined by their uniquely folded structures, which in turn, are determined by their amino acid residue sequences [1]. This dogma has been seriously challenged due to the discovery that partially unfolded and intrinsically unfolded proteins are functional as well [2,3]. However, protein structure, either in its folded or unfolded form, still determines its function. Fortunately, the rapid progress in molecular biology in the past two decades has accumulated near a hundred thousand of protein structures in the protein data bank (PDB). Unfortunately, the prediction of protein functions from known protein structures remains a formidable task. What is needed is the in-depth understanding of the protein structure and function relation [4,5].

For a given protein structure, its geometric shape, electrostatic potential, and flexibility are some of the most relevant structural properties that determine its functions. The importance of protein geometry and electrostatics to protein functions is well known. However, the role of protein flexibility in protein functions is often elusive. It was argued that protein flexibility, not disorder, is intrinsic to molecular recognition [6]. Protein flexibility is the ability to deform from the equilibrium state under external forces. Under physiological condition, proteins experience constant bombardment by the fast-moving solvent molecules, ions, ligands, and cofactors—the so called Brownian dynamics. In solid-state or crystallized phase, proteins constantly interact with phonons generated by the lattice dynamics. In response, protein spontaneous fluctuations orchestrate with the Brownian dynamics or lattice dynamics. The degree of protein fluctuations is determined by both the strength of external stimuli and protein flexibility.

Protein flexibility can be investigated by using a number of experimental tools, including x-ray crystallography, nuclear magnetic resonance (NMR), atomic force microscopy, and optical tweezers. Apart from experimental means, protein flexibility is frequently studied by theoretical and computational approaches. Although molecular dynamics (MD) simulations are important for nonequlibrium processes and are able to deliver snapshots to analyze protein flexibility directly, they cannot yet be used to predict protein collective motions at biologically relevant time scales with quantitative accuracy [7]. In contrast, normal mode analysis (NMA) [8–11], elastic network model (ENM) [12], Gaussian network model (GNM) [13,14], and anisotropic network model (ANM) [15] are capable of accessing the long-time stability of proteins beyond the reach of molecular dynamics simulations [9–12,14]. Parallel to the relation between time-dependent and time-independent quantum dynamics, NMA, ENM, GNM, and ANM can be regarded as time-independent molecular mechanics approaches as they can be derived from their corresponding time-dependent molecular mechanics by using the time-harmonic approximation [16]. In the past few decades, these methods have been employed to analyze protein flexibility [17,18], protein thermal stability [19,20], enzyme site activities [21–23], side-chain mobilities [24,25], protein disordered regions [26,27], and binding [28]. Due to their reduced representation of protein structures, they are capable of investigating macroproteins and protein complexes, such as hemoglobin [29], F1 ATPase [30,31], chaperonin GroEL [32,33], viral capsids [34,35], and ribosome [36,37]. These approaches are often calibrated with experimental data—Debye-Waller factors or B factors. Physically, the B factor is a measure of the mean-squared atomic displacement due to thermal motion and possible experimental uncertainties. Typically, a flexible atom or particle has a large B factor while a rigid atom or particle has a small B factor. The analysis of B factors sheds light on the large-scale and long-time functional behaviors of biomolecules. This analysis is complementary to atomic detail simulations. Over years, flexibility analysis methods have been improved, including the consideration of cofactors and the periodicity of crystal structures [38–41]. The reader is referred to review papers [7,42–44] for the status and state of the art.

*Corresponding author: wei@math.msu.edu

A common feature of the aforementioned models is that they depend on the matrix diagonalization and mode decomposition. Motions of a few slowest modes are interpreted as collective and global behavior or functionality of the biomolecule. Another common feature of the above mentioned methods is that they utilize deterministic Hamiltonian dynamics or its time-independent representations. While, in the microscopic world, deterministic Hamiltonian systems typically have statistical or stochastic complements. However, the stochastic nature of protein equilibrium structures is not accounted for in the above mentioned models.

The objective of the present work is to introduce a stochastic model for macromolecular flexibility analysis. Due to thermal motions, macromolecules constantly fluctuate around their equilibrium state. In our model, stochastic deviations are described by a probability function. We assume that transition probabilities between different nonequilibrium states are determined only by the structure of the protein. Therefore, configuration changes of the macromolecule can be formulated as a stationary Markov process. In the rest of this paper, we discuss our theory and model in Sec. II. Section III is devoted to the numerical validation of our model. Our prediction of B factors is validated with experimental data. The performance of the present stochastic model (SM) is compared that of the state of the art methods in the field. This paper ends with a conclusion.

II. THEORY AND MODEL

In probability theory, the Markov process is often used to describe a time-dependent or ordered stochastic process when the next state is determined only by the current state and is irrelevant of all the events preceded it. This characteristic is widely known as "memoryless". Some stochastic systems also satisfy a stationary property such that the joint probability distribution stays unchanged when there is a shift in time or space for all related terms. The combination of the stationary property with the Markov process gives us the stationary Markov process, which is directly related to the master equation used in physics, chemistry, and biology. Although these models are well developed and long established, their potential applications in protein flexibility analysis have never been fully explored, to our best knowledge.

In this work, we introduce a special stochastic model designed for the analysis of macromolecular flexibility. Let us consider a protein of N particles with the equilibrium reference configuration specified by a $3N$-dimensional position vector $\mathbf{r}^0 = (\mathbf{r}_1^0, \mathbf{r}_2^0, \ldots, \mathbf{r}_N^0)$. Due to its intrinsic motion, the protein configuration at time t is described by $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \in \mathbb{R}^{3N}$. Let us denote $d_j = \|\mathbf{r}_j - \mathbf{r}_j^0\|_2$ the deviation from the equilibrium in the jth particle and $\mathbf{d} = (d_1, d_2, \ldots, d_N) \in \mathbb{R}^{3N}$ a $3N$-dimensional vector of derivation. We denote $\mathbf{P}(t)$ the probability of finding the protein derivation $\mathbf{d}$ at time t. Here $\mathbf{P}(t) = (\mathbf{P}_1(t), \mathbf{P}_2(t), \ldots, \mathbf{P}_N(t))^T$ is a column vector. Assume that protein configurational dynamics is a stationary Markov process, one can derive the following master equation

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{A}\mathbf{P}(t). \tag{1}$$

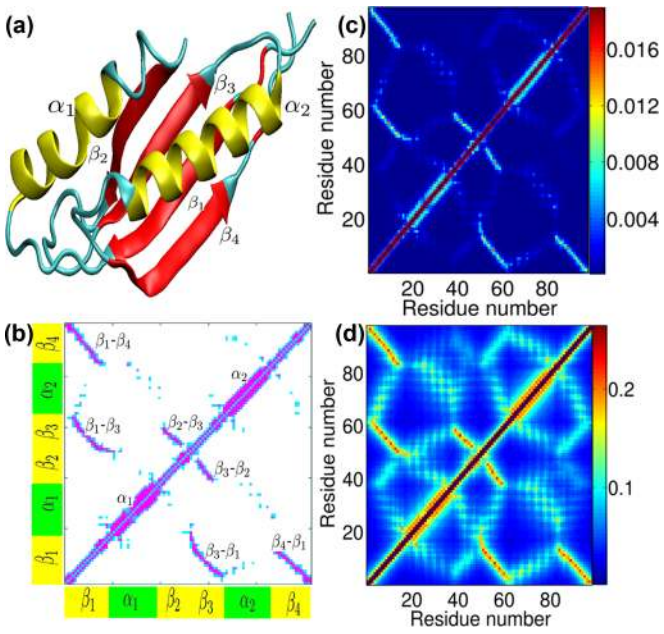

FIG. 1. (Color online) Structure and transition probability matrix (TPM) for protein 1J27. (a) The structure of protein 1J27. The α helices and β sheets are labeled and numbered. (b) TPM generated by using Eq. (3) with $\sigma = 3$ Å and $k = 2$. Diagonal elements have been excluded to emphasize nondiagonal transition probabilities. All α helices and β sheets and their interactions are identified. (c) TPM generated by using Eq. (4) with $\upsilon = 3$. (d) TPM generated by using Eq. (4) with $\upsilon = 1$. The interactions between α helices and β sheets can be easily identified.

where $\mathbf{A}$ is a transition matrix, which describes the transition probability between different protein configurations.

To estimate the element of $\mathbf{A}$, consider a configuration $\mathbf{d}$ generated by an infinitesimally small perturbation (Δ) on the ith particle

$$d_j = \Delta\delta_{ij}, \quad j = 1,2,\ldots,N, \tag{2}$$

where δ_{ij} is a Kronecker δ function. Denote the distance between the perturbed particle and any other particle as $d_{ij} \approx \|\mathbf{r}_i^0 - \mathbf{r}_j^0\|_2$. We assume that the transition probability for such an infinitesimally small derivation to propagate to any neighboring particle decays monotonically with respect to the distance. The transition probability matrix (TPM) can be constructed by using probability density estimators [45,46], including radial basis functions of exponential type

$$\mathbf{A}_{ij} = \begin{cases} e^{-(d_{ij}/\sigma_{ij})^k}, & \forall i \neq j, \quad k > 0; \\ -\sum_{j\neq i} \mathbf{A}_{ij}, & \forall i = j, \end{cases} \tag{3}$$

and power-law type

$$\mathbf{A}_{ij} = \begin{cases} \left(\frac{1}{d_{ij}}\right)^{\upsilon}, & \forall i \neq j, \quad \upsilon > 1; \\ -\sum_{j\neq i} \mathbf{A}_{ij}, & \forall i = j, \end{cases} \tag{4}$$

where σ_{ij} are characteristic distances between particles. In Eqs. (3) and (4), diagonal terms $\mathbf{A}_{ii}$ are chosen so that the detailed balance is maintained for the master equation (1). For simplicity, we utilize a coarse-grained representation of protein structures in terms of C_αs and set $\sigma_{ij} = \sigma$ in the present work.

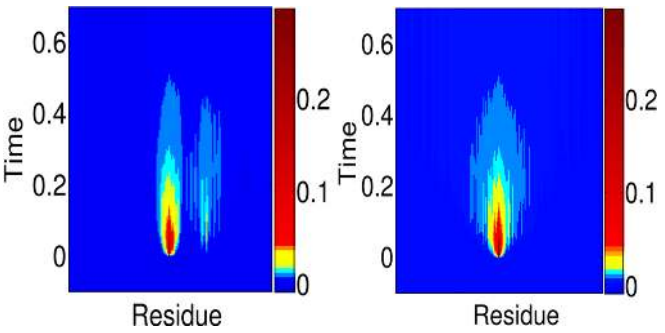

FIG. 2. (Color online) The relaxation process of the probability after the impulse perturbation of the equilibrium configuration of protein 1L11 at a given amino acid residue (the 82nd C_α). Left chart: The residues in horizontal axis are in their native order. Right chart: The residues in horizontal axis are listed in the descending order according to their distances with respect to the perturbed residue.

To understand the behavior of the transition probability matrix **A**, we consider protein 1J27, which has two α helices and four β sheets as shown in Fig. 1(a). Both the exponential type (3) and power-law type (4) of probability density estimators are used to construct the matrix. For the exponential type of radial basis functions, we choose $k = 2$ and set $\sigma = 10$ Å in our test. The resulting transition probability matrix is presented in Fig. 1(b). We examine the transition probability matrix without the diagonal terms so as to emphasize the interactions among protein residues. Clearly, large transition probabilities occur for the nearest neighboring particles, which means a perturbation introduced at a given particle will almost certainly create derivations from the equilibrium at nearest neighboring particles. Additional, nonzero transition probabilities exist for particles that are fairly close to the perturbed particle. Obviously, the transition probability matrix itself reflects protein structural information of connectivity and network topology, as shown in Fig. 1(b). It can be seen that the connectivity inside an α helix is usually represented by a thick and fat diagonal stripe, while β-sheet interactions are characterized by stripes orthogonal to the diagonal line. To be more specific, when two antiparallel β sheets are close to each other, their interactions are portrayed by stripes orthogonal to the diagonal line. While, when two parallel β sheets are close to each other, they create stripes parallel to the diagonal line in the transition probability matrix.

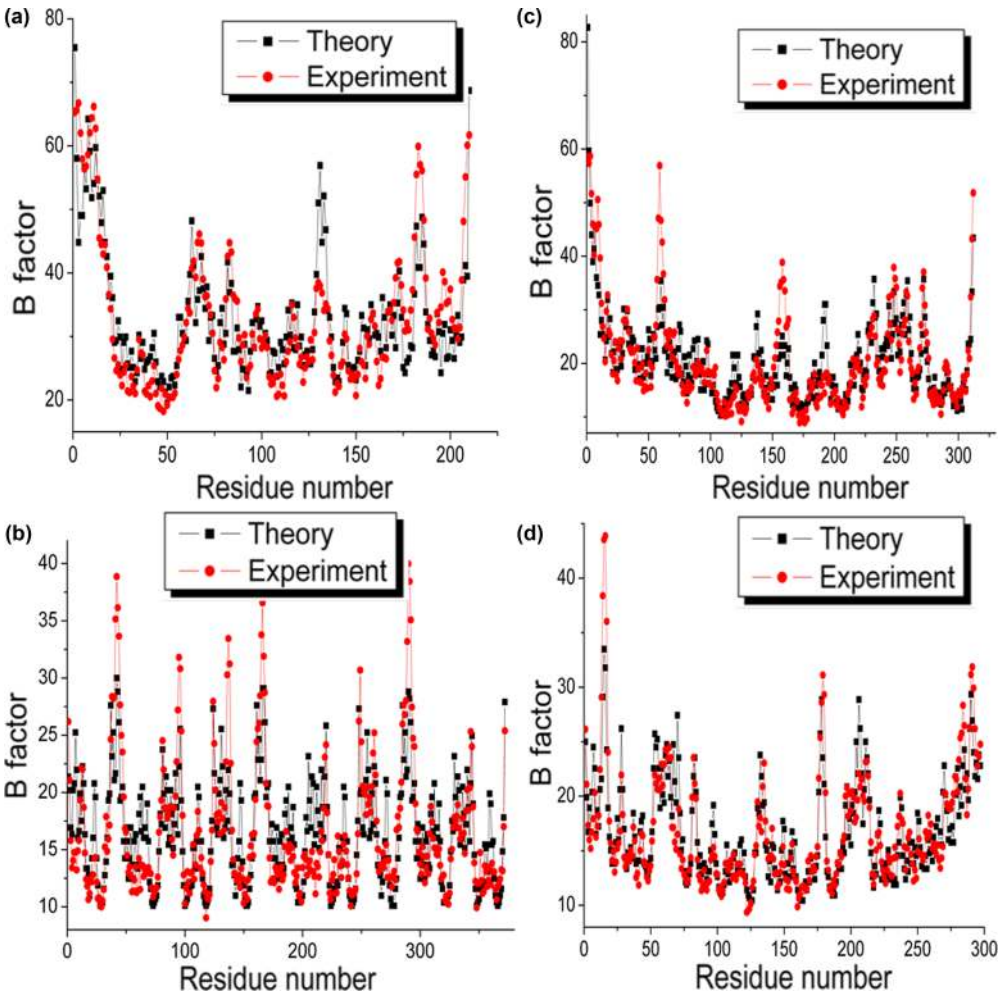

FIG. 3. (Color online) Comparison of B factors obtained from the proposed stochastic model and experiment. Power law probability density estimators are used. (a) B factors for protein 1MIZ predicted with $\upsilon = 3.0$ at correlation coefficient 0.849. (b) B factors for protein 1QD9 predicted with $\upsilon = 2.0$ at correlation coefficient 0.782. (c) B factors for protein 1QUS predicted with $\upsilon = 2.0$ at correlation coefficient 0.845. (d) B factors for protein 1RWR predicted with $\upsilon = 2.0$ at correlation coefficient 0.859.

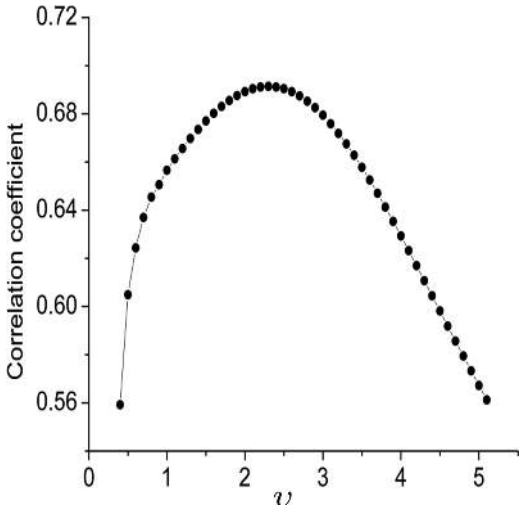

FIG. 4. The optimal parameter search for power-law type of probability density estimators in Eq. (4). A set of 60 proteins is used and the correlation coefficients averaged over the set plotted with respect to the change of parameter υ. The maximal correlation coefficient is reached when υ is around 2.5.

To further explore the transition probability matrix and examine the performance of the power-law functions, we select $\upsilon = 3$ and 1 in Eq. (4). The resulting transition probability matrices are plotted in Figs. 1(c) and 1(d). It is interesting to note that Fig. 1(d) is able to illustrate some transition probabilities due to long-distance interactions, such as the interactions between two α helices. As shown in Fig. 1(a), β_1 and β_2 are quite far apart and parallel to each other. However, their interaction can be clearly seen in Fig. 1(d).

Obviously, transition probability matrix $\mathbf{A}$ is a Hermitian matrix and its eigenvalues are real. Mathematically, the solution of Eq. (1) can be expressed as

$$\mathbf{P}(t) = \sum_{l=1}^{N} c_l \xi^{(l)} e^{r_l t}, \tag{5}$$

where $\xi^{(l)}$ and r_l are eigenvectors and eigenvalues of $\mathbf{A}$, respectively. Here, c_l are coefficients and can be determined by initial values of $\mathbf{P}$. Instead of finding out the global solution of Eq. (1), it is more useful to sample the response of a well-defined perturbation to a given particle described by Eq. (2). Since the diagonal elements of $\mathbf{A}$ are all real and negative, the probability $\mathbf{P}(t)$ must decay toward the equilibrium, just like the free induction decay of the spin dynamics in NMR experiments. Therefore, we define the relaxation time (τ_i) as the time used for $\mathbf{P}(t)$ to recover to a factor of $1/e$ after the perturbation given in Eq. (2). Interestingly, the relaxation time τ_i is a measure of the strength of ith particle's interactions with its environment. A strong interaction leads to a short relaxation time and a low structural flexibility. Therefore, we are able to establish the correlation between the structural flexibility and the relaxation time τ_i. As a result, the relaxation time computed from the proposed master equation must directly correlate with B factors.

To illustrate our ideas, we solve Eq. (1) numerically with the second-order forward Euler scheme. We consider Eq. (3) with $k = 2$ and $\sigma = 6$ Å. We set the initial value as an impulse perturbation on the ith particle with $i = 82$, i.e., $\mathbf{P}_j(0) = \delta_{ij}, j = 1,2,\ldots,N$. Figure 2 demonstrates the relaxation process of the probability after the perturbation of the equilibrium configuration of protein 1L11 on the given particle. We have plotted residues in the horizontal axis in two ways, namely, in their original order and in the descending order according to their distances with respect to the perturbed residue. Amazingly, the impulse relaxation pattern with residues in the original order shown in Fig. 2 (left chart) highlights the connectivity and interaction strength of the perturbed residue. While the other pattern shown in Fig. 2 indicates that the perturbation gradually propagates from the nearest neighboring amino acid residues to a large set of nearby residues over a short time period before it diminishes finally.

III. RESULTS AND DISCUSSIONS

It remains to be proven that the proposed probability decay carries detailed structural information as the free induction decay in NMR. To this end, we utilize the relaxation time to predict protein B factors and compare our prediction with

TABLE I. Comparison of B-factor prediction by parameter-free stochastic model (pfSM), Gaussian normal mode (GNM), and normal mode analysis (NMA) for small-sized data set [16]. The asterisk sign indicates improved prediction with modified protein data.

PDB ID	N	pfSM	GNM [16]	pfSM-GNM	NMA [16]	pfSM-NMA
1AIE	31	0.390	0.155	0.235	0.712	−0.322
1AKG	16	0.192	0.185	0.007	−0.229	0.421
1BX7	51	0.651	0.706	−0.055	0.868	−0.217
1ETL	12	0.939	0.628	0.311	0.355	0.584
1ETM	12	0.768	0.432	0.336	0.027	0.741
1ETN	12	0.075	−0.274	0.349	−0.537	0.612
1FF4	65	0.631	0.674	−0.043	0.555	0.076
1GK7	39	0.684	0.821	−0.137	0.822	−0.138
1GVD	52	0.629	0.591	0.038	0.570	0.059
1HJE	13	0.721	0.616	0.105	0.562	0.159
1KYC	15	0.841	0.754	0.087	0.784	0.057
1NOT	13	0.842	0.523	0.319	0.567	0.275
1O06	20	0.873	0.844	0.029	0.900	−0.027
1OB4	16	0.769	0.750*	0.019	0.930	−0.161
1OB7	16	0.556	0.652*	−0.096	0.952	−0.396
1P9I	29	0.555	0.625	0.070	0.603	−0.048
1PEF	18	0.877	0.808	−0.069	0.888	−0.011
1PEN	16	0.291	0.270	0.021	0.056	0.235
1Q9B	43	0.767	0.656	0.111	0.646	0.121
1RJU	36	0.329	0.431	−0.102	0.235	0.094
1U06	55	0.386	0.434	−0.048	0.377	0.009
1UOY	64	0.653	0.671	−0.018	0.628	0.025
1USE	40	0.138	−0.142	0.280	−0.399	0.537
1VRZ	21	0.548	0.677*	−0.129	−0.203	0.751
1XY2	8	0.118	0.562	−0.444	0.458	−0.340
1YJO	6	0.322	0.434	−0.112	0.445	−0.123
1YZM	46	0.847	0.901	−0.054	0.939	−0.092
2DSX	52	0.329	0.127	0.202	0.433	−0.104
2JKU	35	0.837	0.656	0.181	0.850	−0.013
2NLS	36	0.613	0.530	0.083	0.088	0.525
2OL9	6	0.529	0.689	−0.160	0.886	−0.357
2OLX	4	0.795	0.885	−0.090	0.776	0.119
6RXN	45	0.577	0.594	−0.017	0.304	0.273

experimental data from x-ray crystallography. We compute the relaxation time for a controlled impulse response process of a given particle. This procedure is repeated over all particles of interest in the molecule. The set of relaxation times $\{\tau_i\}$ is then converted to B factors by using a standard linear regression. Figures 3(a)–3(d) provide such comparisons for four protein structures, namely 1MIZ, 1QD9, 1QUS, and 1RWR. Surprisingly, our new approach gives rise to very good prediction of B factors for these proteins.

It is important to quantitatively assess the performance of the proposed stochastic model for the B-factor prediction. For this purpose, we make use of the correlation coefficient C_c

$$C_c = \frac{\Sigma_{i=1}^{N}\left(B_i^e - \bar{B}^e\right)\left(B_i^t - \bar{B}^t\right)}{\left[\Sigma_{i=1}^{N}\left(B_i^e - \bar{B}^e\right)^2 \Sigma_{i=1}^{N}\left(B_i^t - \bar{B}^t\right)^2\right]^{1/2}}, \tag{6}$$

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

To systematically validate our SM, we consider a set of 60 protein structures downloaded from the PDB. All structures of these proteins are obtained by the x-ray diffraction with resolution about 2.0 Å. No multiple conformations exist in these proteins, which means that for each protein, all the occupancy terms equal 1.0. The power-law type of probability density estimators in Eq. (4) is considered. We search the optimal value of parameter υ by calculating correlation coefficients averaged over 60 proteins in the range of [0.4, 5.1]. It is seen from Figure 4 that the best average correlation coefficient is achieved around $\upsilon = 2.5$. This result enables us

TABLE II. Comparison of B-factor prediction in terms of correlation coefficients by parameter-free stochastic model (pfSM), Gaussian normal mode (GNM), and normal mode analysis (NMA) for medium-sized data set [16]. The asterisk sign indicates improved prediction with modified protein data.

PDB ID	N	pfSM	GNM [16]	pfSM-GNM	NMA [16]	pfSM-NMA
1ABA	87	0.737	0.613	0.124	0.057	0.680
1CYO	88	0.736	0.741	−0.005	0.774	−0.038
1FK5	93	0.586	0.485	0.101	0.362	0.224
1GXU	88	0.681	0.421	0.260	0.581	0.100
1I71	83	0.375	0.549	−0.174	0.380	−0.005
1LR7	73	0.685	0.620	0.065	0.795	−0.110
1N7E	95	0.507	0.497	0.010	0.385	0.122
1NNX	93	0.773	0.631	0.142	0.517	0.256
1NOA	113	0.596	0.615	−0.019	0.485	0.111
1OPD	85	0.385	0.398	−0.013	0.796	−0.411
1QAU	112	0.678	0.620	0.058	0.533	0.145
1R7J	90	0.435	0.368	0.067	0.078	0.357
1UHA	83	0.675	0.638*	0.092	0.308	0.367
1ULR	87	0.596	0.495	0.101	0.223	0.373
1USM	77	0.833	0.798	0.035	0.780	0.053
1V05	96	0.592	0.632	−0.040	0.389	0.203
1W2L	97	0.612	0.397	0.215	0.432	0.180
1X3O	80	0.526	0.654	−0.128	0.453	0.073
1Z21	96	0.609	0.433	0.176	0.289	0.320
1ZVA	75	0.537	0.690	−0.153	0.579	−0.042
2BF9	36	0.517	0.680*	−0.163	0.521	−0.004
2BRF	100	0.757	0.710	0.047	0.535	0.222
2CE0	99	0.588	0.529	0.059	0.628	−0.040
2E3H	81	0.690	0.605	0.085	0.632	0.058
2EAQ	89	0.740	0.695	0.045	0.688	0.052
2EHS	75	0.718	0.747	−0.029	0.565	0.153
2FQ3	85	0.748	0.348	0.400	0.508	0.24
2IP6	87	0.595	0.572	0.023	0.826	−0.231
2MCM	112	0.782	0.820	−0.038	0.643	0.139
2NUH	104	0.762	0.771	−0.009	0.685	0.077
2PKT	93	0.180	−0.193*	0.373	−0.165	0.345
2PLT	99	0.444	0.509*	−0.065	0.187	0.257
2QJL	99	0.574	0.594	−0.020	0.497	0.077
2RB8	93	0.603	0.517	0.086	0.485	0.118
3BZQ	99	0.514	0.466	0.048	0.351	0.163
5CYT	103	0.420	0.331	0.089	0.102	0.318

TABLE III. Comparison of B-factor prediction in terms of correlation coefficients by parameter-free stochastic model (pfSM), Gaussian normal mode (GNM), and normal mode analysis (NMA) for large-sized data set [16]. The asterisk sign indicates improved prediction with modified protein data.

PDB ID	N	pfSM	GNM [16]	pfSM-GNM	NMA [16]	pfSM-NMA
1AHO	64	0.598	0.562	0.036	0.339	0.259
1ATG	231	0.605	0.497	0.108	0.154	0.451
1BYI	224	0.468	0.552	−0.084	0.133	0.335
1CCR	111	0.532	0.351	0.181	0.530	−0.002
1E5K	188	0.742	0.859	−0.117	0.620	0.122
1EW4	106	0.603	0.547	0.056	0.447	0.156
1IFR	113	0.706	0.637	0.069	0.330	0.376
1NKO	122	0.518	0.368	0.150	0.322	0.196
1NLS	238	0.575	0.523*	0.052	0.385	0.190
1O08	221	0.410	0.309	0.101	0.616	−0.206
1PMY	123	0.654	0.685	−0.031	0.702	−0.048
1PZ4	113	0.843	0.843	0.000	0.844	−0.001
1QTO	122	0.421	0.334	0.087	0.725	−0.304
1RRO	108	0.399	0.529	−0.130	0.546	−0.147
1UKU	102	0.648	0.742	−0.094	0.720	−0.072
1V70	105	0.431	0.162	0.269	0.285	0.146
1WBE	204	0.558	0.549	0.009	0.574	−0.016
1WHI	122	0.479	0.270	0.209	0.414	0.165
1WPA	107	0.528	0.417	0.111	0.380	0.148
2AGK	233	0.683	0.512	0.171	0.514	0.169
2C71	205	0.677	0.560	0.117	0.584	0.093
2CG7	90	0.494	0.379	0.115	0.308	0.186
2CWS	227	0.648	0.696	−0.048	0.524	0.124
2HQK	213	0.810	0.365	0.445	0.743	0.067
2HYK	237	0.586	0.515	0.071	0.593	−0.007
2I24	113	0.430	0.494	−0.064	0.441	−0.011
2IMF	203	0.611	0.514	0.097	0.401	0.210
2PPN	107	0.640	0.668	−0.028	0.468	0.172
2R16	176	0.474	0.618*	−0.144	0.411	0.063
2V9V	135	0.599	0.528	0.071	0.594	0.005
2VIM	104	0.376	0.282	0.164	0.273	0.155
2VPA	204	0.772	0.576	0.196	0.594	0.178
2VYO	206	0.693	0.761	−0.068	0.739	−0.046
3SEB	238	0.768	0.826	−0.058	0.720	0.048
3VUB	101	0.641	0.607	0.034	0.365	0.276

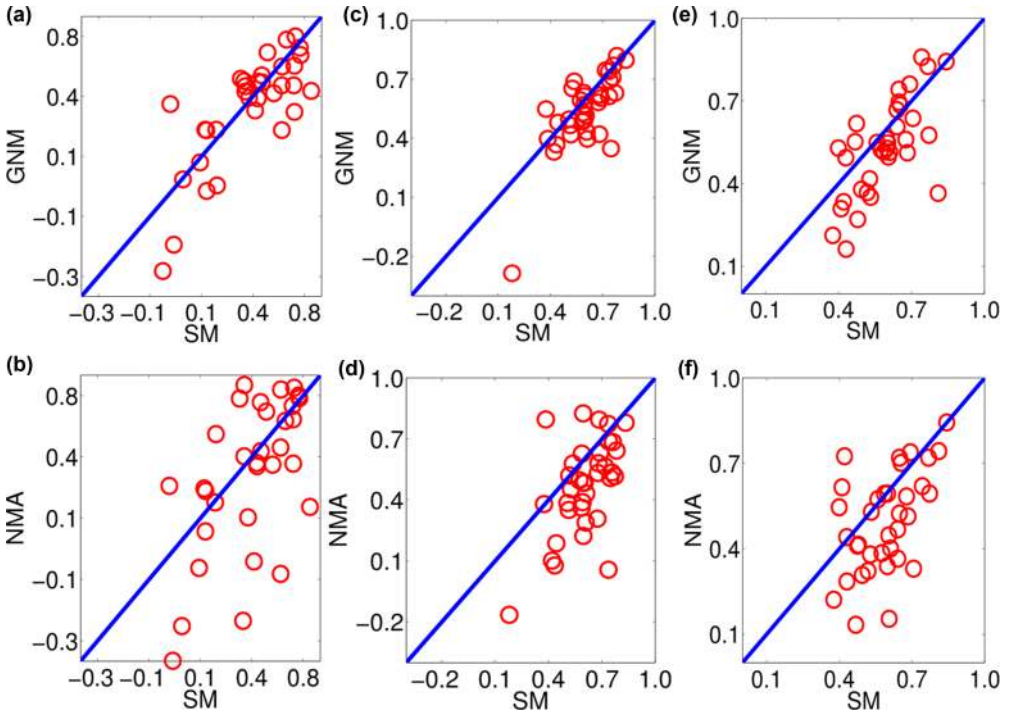

FIG. 5. (Color online) The comparison of correlation coefficients predicted by the GNM, NMA, and pfSM ($\upsilon = 2.5$) for three sets of proteins. (a) GNM vs SM for the small-sized protein set; (b) NMA vs SM for the small-sized protein set; (c) GNM vs SM for the medium-sized protein set; (d) NMA vs SM for the medium-sized protein set; (e) GNM vs SM for the large-sized protein set; (f) NMA vs SM for the large-sized protein set.

to obtain a parameter free stochastic model (pfSM) by setting $\upsilon = 2.5$.

To further validate the proposed SM for flexibility analysis, we carry out a comparison study. We employ the coarse-grain GNM, one of the cutting edge approaches in the field to calibrate the present SM. The computer code for the GNM is downloaded from the Jernigan Laboratory [47] with some minor modifications to improve its performance. The cutoff distance of 7 Å, which is near the optimal for the set of proteins, is used in all the GNM calculations. Three data sets, proposed by Park, Jernigan, and Wu [16], including relatively small-, medium-, and large-sized proteins are utilized. It is found that some data in these sets have multiple conformations and missing residues. To use these data directly without appropriate modifications would underscore theoretical methods. Therefore, we carefully add in the missing residues and remove the repeated atoms with lower occupancy values. As a result, the prediction of GNM for modified proteins is significantly improved from that reported in the literature [16]. Results involving modified data are marked with an asterisk in Tables I, II, and III. We also compare our results with those obtained by using the coarse-grain normal mode analysis (NMA). The related data are directly taken from Park *et al*. [16]. We use our pfSM with the optimal value $\upsilon = 2.5$. As shown in Figure 5, our pfSM outperforms the GNM and the NMA in most cases. The same conclusion can be reached by examining the detailed

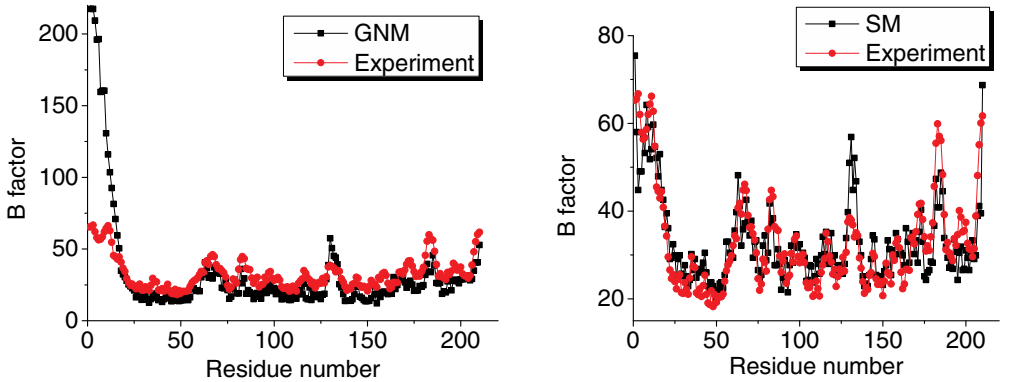

FIG. 6. (Color online) The comparison of B-factor prediction between the proposed SM model and GNM for protein 1MIZ. The result in the left chart is the prediction by GNM with the correlation coefficient of 0.761. The result in right chart is from the proposed SM with the correlation coefficient of 0.849.

TABLE IV. The comparison of average correlation coefficients calculated by pfSM, GNM, and NMA over three data sets.

PDB set	pfSM	GNM	NMA	Improvement with respect to GNM (%)	Improvement with respect to NMA (%)
Small	0.578	0.541	0.480	6.80	20.4
Medium	0.605	0.555	0.469	9.01	29.0
Large	0.589	0.530	0.494	11.1	19.2

data in Tables I, II, and III. Figure 6 gives a comparison of the B-factor prediction between the proposed SM model and the GNM for protein 1MIZ. In this case, our method gives a better overall prediction. However, both methods over predict B factors around residue 130.

To further analyze the performance of pfSM, GNM, and NMA, we compute the average correlation coefficient predicted by each method for each data set. Based on the details listed in Tables I, II, and III. The results of our analysis are presented in Table IV for a comparison. It is seen that our pfSM gives better results. Comparing with the GNM, pfSM has about 11% increase in its B-factor prediction over the set of large-sized proteins. The pfSM's improvement over the set of small-sized proteins is 6.8%. Therefore, the pfSM seems to work relatively better on large-sized proteins. Apparently, there is a huge improvement in the average correlation coefficient over that predicted by the NMA for all the protein structures considered. Although part of the improvement is due to the fact that modified protein data sets were employed in the pfSM calculations, pfSM still outperforms NMA when a few modified protein data are not counted. The conclusion that pfSM performs better than the NMA can also be drawn from the well-known fact that the GNM outperforms the NMA in B-factor prediction [16].

Although our method demonstrates its ability for the accurate analysis of macromolecular flexibility, there exists much room for its further improvement. It is well known that cofactors and nearby structures in a protein crystal significantly influence flexibility properties. Some ions, such as zinc ions, play a central role in the stabilization of the protein. The absence of this type of ion will dramatically change flexibility properties. For instance, in Table II, the crystal structure of protein 2PKT has cofactors $C_2H_3O_3$ and Ca^{2+} and Cl^{2+}. Because this effect has not been considered in the present work, the SM prediction is inaccurate with the correlation coefficient being only about 0.180. Another limitation of the present model is the use of the coarse-grain representation. In our method, all amino acids are denoted by their C_α atoms and treated equally. However, the size and the characteristic of amino acids in a protein can vary significantly from each other and lead to different flexibility behavior. The incorporation of these properties in our model will definitely yield a better method.

IV. CONCLUSION

Flexibility is an intrinsic property of proteins and is essential for protein functions. Conventional flexibility analysis relies on the Hamiltonian mechanics and matrix decomposition. We introduce a stochastic model for macromolecular flexibility analysis. An NMR-free induction-decay-like perturbation process is designed to stimulate the probability transfer from the nonequilibrium to equilibrium after an impulse perturbation. We show that the speed of the probability transfer at each residue correlates with its flexibility. As a result, we develop a stochastic model for protein B-factor prediction. The proposed method bypasses the construction of any Hamiltonian and does not require the matrix diagonalization. Comparison with experimental data and established methods validates the present approach for flexibility analysis. A further comparison with a newly developed flexibility-rigidity index (FRI) approach [46] is a topic for future research.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grants No. DMS-1160352 and No. IIS-1302285, and NIH Grant No. R01GM-090208. The authors acknowledge the Mathematical Biosciences Institute for hosting valuable workshops.

[1] C. B. Anfinsen, Science **181**, 223 (1973).
[2] M. Schroder and R. J. Kaufman, Annu. Rev. Biochem. **74**, 739 (2005).
[3] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).
[4] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[5] S. H. White and W. C. Wimley, Annu. Rev. Biophys. Biomol. Struct. **28**, 319 (1999).
[6] J. Janin and M. J. Sternberg, F1000 Biology Reports **5**, 1 (2013).
[7] C. P. Chng and L. W. Yang, Bioinformatics and Biology Insights **2**, 171 (2008).
[8] N. Go, T. Noguti, and T. Nishikawa, Proc. Natl. Acad. Sci. USA **80**, 3696 (1983).
[9] M. Tasumi, H. Takenchi, S. Ataka, A. M. Dwidedi, and S. Krimm, Biopolymers **21**, 711 (1982).
[10] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[11] M. Levitt, C. Sander, and P. S. Stern, J. Mol. Biol. **181**, 423 (1985).
[12] M. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).
[13] P. J. Flory, Proc. R. Soc. London A **351**, 351 (1976).
[14] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[15] A. R. Atilgan, S. R. Durrell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).
[16] J. K. Park, R. Jernigan, and Z. Wu, Bull. Math. Biol. **75**, 124 (2013).
[17] P. A. Karplus and G. E. Schulz, Naturwissenschaften **72**, 212 (1985).
[18] M. Vihinen, E. Torkkila, and P. Riikonen, Proteins **19**, 141 (1994).
[19] M. Vihinen, Protein Eng. **1**, 477 (1987).
[20] S. Parthasarathy and M. R. N. Murthy, Protein Eng. **13**, 9 (2000).
[21] O. Carugo and P. Argos, Proteins **31**, 201 (1998).

[22] Z. Yuan, J. Zhao, and Z. X. Wang, Protein Eng. **16**, 109 (2003).
[23] S. Mohan, N. Sinha, and J. Smith-Gill, Biophys. J. **85**, 3221 (2003).
[24] O. Carugo and P. Argos, Protein Eng. **10**, 777 (1997).
[25] E. Eyal, R. Najmanovich, M. Edelman, and V. Sobolev, Proteins **50**, 272 (2003).
[26] R. B. Altman, C. Hughes, D. Zhao, and O. Jardetzky, Prot. Pept. Lett. **1**, 120 (1994).
[27] P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. D. Lawson, and A. K. Dunker, Protein Sci. **13**, 71 (2004).
[28] J. A. Marsh, S. A. Teichmann, and J. D. Forman-Kay, Curr. Opin. Struct. Biol. **22**, 643 (2012).
[29] C. Xu, D. Tobi, and I. Bahar, J. Mol. Biol. **333**, 153 (2003).
[30] W. J. Zheng and S. Doniach, Proc. Natl. Acad. Sci. USA **100**, 13253 (2003).
[31] Q. Cui, G. J. Li, J. Ma, and M. Karplus, J. Mol. Biol. **340**, 345 (2004).
[32] O. Keskin, I. Bahar, D. Flatow, D. G. Covell, and R. L. Jernigan, Biochem. **41**, 491 (2002).
[33] W. Zheng, B. R. Brooks, and D. Thirumalai, Biophys. J. **93**, 2289 (2007).
[34] A. J. Rader, D. H. Vlad, and I. Bahar, Structure **13**, 413 (2005).
[35] F. Tama and C. K. Brooks III, J. Mol. Biol. **345**, 299 (2005).
[36] F. Tama, M. Valle, J. Frank, and C. K. Brooks III, Proc. Natl. Acad. Sci. USA **100**, 9319 (2003).
[37] Y. Wang, A. J. Rader, I. Bahar, and R. L. Jernigan, J. Struct. Biol. **147**, 302 (2004).
[38] S. Kundu, J. S. Melton, D. C. Sorensen, and G. N. Phillips, Jr., Biophys. J. **83**, 723 (2002).
[39] D. A. Kondrashov, A. W. Van Wynsberghe, R. M. Bannen, Q. Cui, and G. N. Phillips, Jr., Structure **15**, 169 (2007).
[40] K. Hinsen, Bioinformatics **24**, 521 (2008).
[41] G. Song and R. L. Jernigan, J. Mol. Biol. **369**, 880 (2007).
[42] J. Ma, Structure **13**, 373 (2005).
[43] L. Skjaerven, S. M. Hollup, and N. Reuter, J. Mol. Struct (THEOCHEM) **898**, 42 (2009).
[44] Q. Cui and I. Bahar, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems* (Chapman and Hall, London, 2010).
[45] G. W. Wei, J. Phys. A **33**, 8577 (2000).
[46] K. L. Xia, K. Opron, and G. W. Wei, J. Chem. Phys. **139**, 194109 (2013).
[47] http://ribosome.bb.iastate.edu/software.html