Optimal modeling of atomic fluctuations in protein crystal structures for weak crystal contact interactions

Jeffrey Hafner and Wenjun Zheng[a)]
Department of Physics, University at Buffalo, Buffalo, New York 14260, USA

(Received 22 October 2009; accepted 15 December 2009; published online 6 January 2010)

The accurate modeling of protein dynamics in crystalline states holds keys to the understanding of protein dynamics relevant to functions. In this study, we used coarse-grained elastic network models (ENMs) to explore the atomic fluctuations of a protein structure that interacts with its crystalline environment, and evaluated the modeling results using the anisotropic displacement parameters (ADPs) obtained from x-ray crystallography. To ensure the robustness of modeling results, we used three ENM schemes for assigning force constant combined with three boundary conditions for treating the crystalline environment. To explore the role of crystal contact interactions in the modeling of ADPs, we varied the strength of interactions between a protein structure and its environment. For a list of 83 high-resolution crystal structures, we found that the optimal modeling of ADPs, as assessed by a variety of metrics, is achieved for weak protein-environment interactions (compared to the interactions within a protein structure). As a result, the ADPs are dominated by contributions from rigid-body motions of the entire protein structure, and the internal protein dynamics is only weakly perturbed by crystal packing. Our finding of weak crystal contact interactions is also corroborated by the calculations of residue-residue contact energy within a protein structure and between protein molecules using a statistical potential. © *2010 American Institute of Physics*. [doi:10.1063/1.3288503]

I. INTRODUCTION

Protein structural dynamics, including both structural fluctuations in an equilibrium state and large conformational changes between equilibrium states, is increasingly recognized as a key linkage between protein structures and functions.[1] With the fast progressing of structural probing techniques such as nuclear magnetic resonance (NMR),[2] high-resolution information of protein dynamics can be obtained via the analysis of various dynamics data such as NMR order parameter.[3] The time-honored x-ray crystallography remains the primary source of protein dynamics data at atomic or near-atomic resolution. Atomic fluctuations in protein crystal structures have been traditionally quantified by the isotropic temperature factors (or B factors), which use an isotropic Gaussian distribution to characterize the spread of electron density of each atom.[4] Recently, a growing number of high-resolution protein crystal structures have been refined using anisotropic Gaussian distributions, which characterize atomic fluctuations by a symmetric tensor with six independent elements named anisotropic displacement parameters (ADPs).[5,6] Unlike the B factors, the ADPs describe not only the magnitude but also the direction of mean-squared atomic displacements. Therefore, they offer much richer information of protein structural dynamics in crystalline states.

To explore the fine details of protein structural dynamics, an array of computational methods ranging from all-atom molecular dynamics simulation[7] to various coarse-grained modeling techniques[8] have been developed. Several models have been employed to quantitatively describe the dynamic contributions to the crystallographic B factors and ADPs. One popular model (TLS model) describes a protein as an assembly of rigid subunits, and it fits the B factors with an optimized combination of translations, librations, and screwing motions.[9–11] Alternatively, low-frequency normal modes, which are known to capture the large-amplitude collective motions of protein structures, have been used to fit B factors[12,13] and refine ADPs.[14,15] Initially, the normal modes were solved from all-atom potential functions following energy minimization.[16–18] This procedure is computationally expensive for large protein structures, and it is susceptible to structural distortions caused by energy minimization. More recently, elastic network models (ENMs), including anisotropic network model (ANM)[19–21] and its isotropic variation—Gaussian network model (GNM),[22,23] have been developed to model protein dynamics with coarse-grained resolution. The ENM is usually constructed based on a C_α-only representation of protein structures, where neighboring C_α atoms are connected by harmonic springs with a uniform[19] or distance-dependent[24,25] force constant. Such dramatic simplification allows the coarse-grained normal modes to be calculated efficiently without energy minimization. The collective motions described by low-frequency modes remain unchanged despite the use of coarse-grained representation and harmonic potential function.[26,27] The crystallographic B factors have been modeled by GNM[28,29] and ANM[30] with remarkable success. GNM was found to achieve a better performance than a simplified TLS model without using any fitting parameters,[31] although the full TLS

[a)]Electronic mail: wjzheng@buffalo.edu.

model (with ten fitting parameters) seemed to perform better.[32] Recently, ANM has been used to model ADPs with reasonable success.[33–35] The major difference between TLS and ANM is that TLS only considers rigid-body movements (rotations and translations) and ignores internal motions, while ANM only accounts for internal motions (captured by normal modes with nonzero eigenvalue) and ignores global rigid-body motions. Attempts to include rigid-body motions in ENM-based fitting of B factors and ADPs supported the importance of their contributions,[32,35,36] although there are concerns about overfitting with many parameters.[35] It is thus desirable to develop a physically based model to incorporate both rigid-body motions and internal motions without relying on multiparameter fitting.

As testified by the great success of x-ray crystallography, it is generally agreed that a protein's crystal structures are relevant to its physiological states in solution. Therefore, crystal packing is unlikely to significantly alter protein native conformations, although one of several functionally relevant conformations may be favored by a particular crystal packing symmetry. However, it is still not clear how much the protein dynamics in a crystalline state correlates with the protein dynamics in solution. It is conceivable that crystal packing may affect the fluctuations of those atoms involved in crystal contacts. Indeed, previous studies revealed large effects of crystal packing on ADPs[6] and B factors.[37] Recent studies found better agreement between GNM/ANM-based predictions and NMR-based structural fluctuation data as compared to crystallographic B factors,[38,39] which was attributed to the absence/presence of crystal packing for NMR/x-ray structures. Therefore, to accurately model protein dynamics in crystalline states, it is important to properly consider the effects of crystal packing. To this end, a previous study[31] found that the GNM-based modeling of B factors was indeed improved by including neighboring protein molecules in a crystal. In a recent study, Phillips, Jr., and coworkers performed a systematic modeling of ADPs for 83 high-resolution protein crystal structures by using ENM coupled with rigorous treatments of boundary conditions and lattice vibrations.[40] Similar study was done for various crystal structures of lysozyme.[41]

To meet the challenge of modeling protein dynamics in crystalline states accurately and efficiently, the previous studies adopted the strategy of treating protein molecules of the entire crystal equally (under different boundary conditions) and summing up the contributions of lattice vibrations.[40,41] This approach is theoretically sound but computationally expensive (because normal modes have to be solved for many q values under the Born–von Ka´rma´n boundary condition, see Ref. 40). In this study, we will adopt a more efficient alternative strategy, which selects a single protein molecule as our main protein structure while treating the rest of crystal as its environment (also see Ref. 32). To reduce computing cost, the environment is truncated based on the distance to the main protein structure or by keeping the nearest and next nearest neighbors of the main protein structure (see Sec. II). The entire system is modeled using three ENM schemes, including ANM,[21] distance network model (DNM)[33] and HCA model[24] under three different boundary conditions (fixed, free and buffered environment, see Sec. II). The dynamic effects of crystal packing are systematically explored by varying the strength of protein-environment interactions. We performed ADP modeling for a list of 83 high-resolution crystal structures previously studied in Refs. 33 and 40. We found that the optimal modeling of ADPs, as assessed by a variety of metrics (see Sec. II), is achieved for weak protein-environment interactions (compared to the interactions within the main protein structure). As a result, the crystallographic ADPs and B factors are dominated by contributions from rigid-body motions of the main protein structure, and the internal protein dynamics is only weakly perturbed by crystal packing. The above results support the importance of explicit consideration of crystal packing to the correct modeling of ADPs[40] and parametrization of ENM. Our finding of weak crystal contact interactions is also corroborated by the calculations of residue-residue contact energy between neighboring proteins and within the main protein structure using a statistical potential.

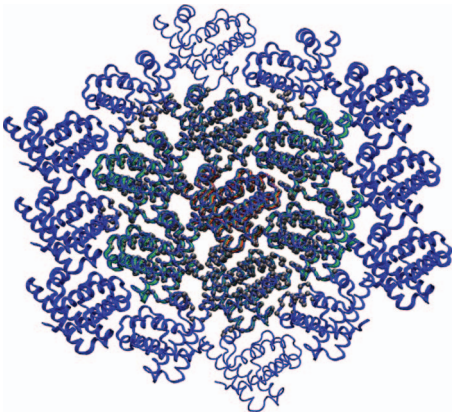

FIG. 1. Representation of crystalline environment for an oxymyoglobin crystal structure (PDB: 1A6M). The main protein structure, its nearest neighbor molecules (named buffer), and the nearest neighbor molecules of buffer are colored red, green, and blue, respectively. The C_α atoms of the environment residues within 25 Å from the main structure are shown as spheres colored in gray.

II. METHODS

A. Elastic network modeling of a protein structure embedded in crystalline environment

To explicitly model the effects of crystal packing, we construct a C_α-only ENM that consists of two components: first, the C_α atoms of a main protein structure (corresponding to an asymmetric unit of a crystal); second, an environment that includes the C_α atoms of other protein molecules in a crystal. To reduce computing cost, the environment is truncated in two different ways (see Fig. 1):

(1) Keeping the C_α atoms of environment within 25 Å from the C_α atoms of the main protein structure. We verified that the modeling results are insensitive to the choice of cutoff distance between 15 and 25 Å. The C_α coordinates of such truncated environment are generated using the What If webserver (http://swift.cmbi.ru.nl/servers/html/index.html).

(2) Keeping the nearest neighbor molecules of the main protein structure in a protein crystal (named buffer), together with the nearest neighbor molecules of the buffer (or the next nearest neighbors of the main protein structure). Here two protein molecules are said to be nearest neighbors if the minimal C_α–C_α atomic distance between them is <10 Å. To construct a protein crystal, we first build a unit cell from the main protein structure using the crystallographic symmetry transformations from REMARK 290 of a PDB file. Then, a protein crystal is built from the unit cell using the three translational vectors derived from the SCALEn records of a PDB file.

The potential energy of the two-component ENM is

$$E = \frac{1}{2}\sum_{i<j} C_{ij}(d_{ij} - d_{ij,0})^2 + \frac{1}{2}f_{\text{env}}\sum_{i,I} C_{iI}(d_{iI} - d_{iI,0})^2 + \frac{1}{2}\sum_{I<J} g_{IJ}C_{IJ}(d_{IJ} - d_{IJ,0})^2, \quad (1)$$

where i and j (I and J) are indices for C_α atoms in the main structure (environment). C_{ij}, C_{IJ}, or C_{iI} represents spring force constant whose assignments vary between ENM schemes (see below). d_{ij}, d_{IJ}, and d_{iI} are C_α–C_α atomic distances. $d_{ij,0}$, $d_{IJ,0}$, and $d_{iI,0}$ are the values of d_{ij}, d_{IJ}, and d_{iI} given by the crystal structure. A new model parameter f_{env} within the range [0, 1] is introduced to tune the strength of interactions between different protein molecules (including all protein-environment interactions and some interactions within the environment). $f_{\text{env}}=0$ corresponds to the case of an isolated protein structure. $f_{\text{env}}=1$ if we assume equal strength of interprotein and intraprotein interactions. Another parameter $g_{IJ}=1$ if the C_α atoms I and J belong to the same protein molecule, and $g_{IJ}=f_{\text{env}}$ if they belong to two different protein molecules.

The following three schemes of force constant assignments are considered:

(1) ANM:

$$C_{\text{ij}} = \begin{cases} 1 & \text{if } d_{ij,0} < R_c \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where the unit for C_{ij} is arbitrary, and R_c is a cutoff distance. In agreement with Refs. 33 and 40, the default value of R_c is set to 10 Å because it gives the optimal modeling of ADPs (see below).

(2) DNM:

$$C_{ij} = \sum_{i_1 j_1 : d_{i_1 j_1,0} < 9} \frac{1}{d_{i_1 j_1,0}^2}, \quad (3)$$

where the summation is over all pairs of heavy atoms of residues i and j within a cutoff distance (chosen to be 9 Å, following Ref. 33). Here i_1 and j_1 are indices for the heavy atoms of residues i and j. The DNM was proposed in Ref. 33, which sets force constants for several distance ranges to the reciprocal of the total number of atomic contacts in each range. Because the number of atomic contacts grows quadratically with distance (assuming the atomic density is constant), our DNM formulation is essentially a continuous counterpart of the original DNM.[33] The use of $1/d^2$ distance dependence also agrees with the new parameter-free ENM proposed in Ref. 25.

(3) HCA:

$$C_{ij} = \begin{cases} 205.5 \times d_{ij,0} - 571.2 & \text{if } d_{ij,0} \le 4 \text{ Å} \\ 305.9 \times 10^3 \times {d_{ij,0}}^{-6} & \text{if } 4 \text{ Å} < d_{ij,0} \le 25 \text{ Å} \\ 0 & \text{if } d_{ij,0} > 25 \text{ Å} \end{cases}, \quad (4)$$

where the unit for C_{ij} is kcal/mol Å^2. The HCA scheme[24] was developed by a best fit to the all-atom normal modes calculated based on the AMBER force field.[42]

The Hessian matrix H is calculated as the second derivatives of potential energy E [see Eq. (1)] with respective to C_α coordinates. For a protein structure with N residues, H contains $N \times N$ superelements (named H_{ij}) with size 3×3 given as follows:

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 E}{\partial x_i \partial x_j} & \frac{\partial^2 E}{\partial x_i \partial y_j} & \frac{\partial^2 E}{\partial x_i \partial z_j} \\ \frac{\partial^2 E}{\partial y_i \partial x_j} & \frac{\partial^2 E}{\partial y_i \partial y_j} & \frac{\partial^2 E}{\partial y_i \partial z_j} \\ \frac{\partial^2 E}{\partial z_i \partial x_j} & \frac{\partial^2 E}{\partial z_i \partial y_j} & \frac{\partial^2 E}{\partial z_i \partial z_j} \end{bmatrix}, \quad (5)$$

where x_i, y_i, and z_i are the Cartesian coordinates of the C_α atom i.

B. Calculation of ADP and B factors

The Hessian matrix H can be partitioned into four submatrices as follows (P denotes main protein structure, E denotes environment):

$$H = \begin{bmatrix} H_{PP} & H_{PE} \\ H_{EP} & H_{EE} \end{bmatrix}. \quad (6)$$

We consider the following three boundary conditions for treating the crystalline environment (assuming $f_{\text{env}}>0$):

(1) *Fixed environment*. The C_α atoms of environment are fixed in space, so the covariance matrix of C_α atoms in the main protein structure is

$$\langle u_P {u_P}^T \rangle = k_B T {H_{PP}}^{-1}, \quad (7)$$

where u_P is the displacement vector of C_α atoms in the main protein structure, k_B is the Boltzmann constant, T is temperature, ${H_{PP}}^{-1}$ is the inverse of the H_{PP} submatrix.

(2) *Free environment*. The C_α atoms of environment are free to move, so the covariance matrix of C_α atoms in the main protein structure is

$$\langle u_P {u_P}^T \rangle = k_B T {H^{-1}}_{PP}, \quad (8)$$

where ${H^{-1}}_{PP}$ is the PP-submatrix of the pseudoinverse H^{-1}, which is calculated after projecting out six trans-

lational and rotational zero modes of the entire two-component system.

(3) *Buffered environment*. The C_α atoms of buffer (the nearest neighbors of the main protein structure) are free to move, while the rest of environment (the next nearest neighbors of the main protein structure) is fixed, so the covariance matrix of C_α atoms in the main protein structure is

$$\langle u_P u_P{}^T \rangle = k_B T (H_{PB,PB})^{-1}{}_{PP}, \tag{9}$$

where $(H_{PB,PB})^{-1}{}_{PP}$ is the PP-submatrix of the inverse of the submatrix of H that involves C_α coordinates in the main protein structure and buffer (named $H_{PB,PB}$).

To exploit the sparseness of Hessian matrix, we use a sparse linear-equation solver CHOLMOD[43] to calculate matrix inversion in Eqs. (7)–(9). To eliminate the overflow due to zero modes, a small positive number $\varepsilon=0.0001$ is added to the diagonal matrix elements of H before its inversion, then the six translational and rotational zero modes are projected out from $(H+\varepsilon)^{-1}$. The use of sparse linear equation solver is computationally more efficient and accurate than the calculation of H^{-1} using a subset of low-frequency modes.[33]

The ith 3×3 diagonal block C_{ii} of the covariance matrix $\langle u_P u_P{}^T \rangle$ gives the theoretical prediction of ADP tensor for the C_α atom i:

$$C_{ii} = \begin{bmatrix} \langle \delta x_i^2 \rangle & \langle \delta x_i \delta y_i \rangle & \langle \delta x_i \delta z_i \rangle \\ \langle \delta x_i \delta y_i \rangle & \langle \delta y_i^2 \rangle & \langle \delta y_i \delta z_i \rangle \\ \langle \delta x_i \delta z_i \rangle & \langle \delta y_i \delta z_i \rangle & \langle \delta z_i^2 \rangle \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ U_{12} & U_{22} & U_{23} \\ U_{13} & U_{23} & U_{33} \end{bmatrix}, \tag{10}$$

where the diagonal elements U_{11}, U_{22}, and U_{33} give the mean-squared fluctuations of the C_α atom i along the x, y, and z directions, and the off-diagonal elements U_{12}, U_{13}, and U_{23} describe the covariance among the displacements of the C_α atom i along the x, y, and z directions. Together, the six ADP elements determine a three-dimensional Gaussian distribution function which describes both the direction and magnitude of the atomic fluctuations.[6] For fixed probability value, the distribution is ellipsoidal with a directional preference along the long axis, which is given by the eigenvector of ADP tensor with the largest eigenvalue. The anisotropy of the Gaussian distribution is defined as the ratio of the smallest to the largest eigenvalue of ADP tensor.

The B factor is related to the trace of ADP tensor as follows:

$$B = 8\pi^2 (U_{11} + U_{22} + U_{33})/3. \tag{11}$$

C. Comparison between theoretical and experimental ADP

We use the following metrics to assess the similarity between experimental and theoretical ADP tensors (represented as U and V):

1. Real-space correlation coefficient

The following real-space correlation coefficient is calculated to evaluate the overlap integral of two three-dimensional Gaussian distributions given by U and V:[6]

$$\mathrm{cc}(U,V) = \frac{(\det U^{-1} \det V^{-1})^{1/4}}{[\det(U^{-1}+V^{-1})/8]^{1/2}}. \tag{12}$$

Based on the real-space correlation coefficient, the following two metrics have been introduced to evaluate the directional similarity of two ADPs:

a. Normalized correlation coefficient

$$\mathrm{ncc}(U,V) = \frac{\mathrm{cc}(U,V)}{\mathrm{cc}(U,U_{\mathrm{iso}})\mathrm{cc}(V,V_{\mathrm{iso}})}, \tag{13}$$

where $U_{\mathrm{iso}}=V_{\mathrm{iso}}=I_3/3$, I_3 is a 3×3 identity matrix, and U and V have been normalized by their trace. The normalized correlation coefficient (ncc) measures the similarity between U and V relative to their similarities to an isotropic tensor.[6] Following Ref. 35, we use a simple ratio of the number of ADPs with $\mathrm{ncc}>1$ and the total number of ADPs (named f_{ncc}) to measure the overall similarity between two sets of ADPs.

b. Modified correlation coefficient ($\mathrm{cc}_{\mathrm{mod}}$)

$$\mathrm{cc}_{\mathrm{mod}}(U,V) = \frac{\mathrm{cc}(U,V) - \mathrm{cc}(U,V^*)}{1 - \mathrm{cc}(U,V^*)}, \tag{14}$$

where V^* is a 3×3 tensor generated by taking the eigenvectors of U and using the eigenvalues of V, with the largest and smallest switched, to define the two ellipsoids with perfect misalignment.[33,40] $\mathrm{cc}_{\mathrm{mod}}$ is 1.0 (0) if the two ellipsoids are perfectly aligned (misaligned).

2. Kullback–Leibler distance

The Kullback–Leibler (KL) distance[44] evaluates the difference between the three-dimensional Gaussian distributions a and b as defined by U and V.[34] The KL distance can be expressed in terms of the eigenvalues (d_{ak} and d_{bk}, $k=1,2,3$) and eigenvectors (v_{ak} and v_{bk}, $k=1,2,3$) of U and V as follows:

$$D_{ab} = -\frac{3}{2} + \frac{1}{2}\sum_{k=1}^{3} \ln\frac{d_{bk}}{d_{ak}} + \frac{1}{2}\sum_{k=1}^{3}\sum_{l=1}^{3} \frac{d_{ak}}{d_{bl}} |\boldsymbol{v}_{ak}^T \boldsymbol{v}_{bl}|^2. \tag{15}$$

Since the KL distance is asymmetric ($D_{ab} \neq D_{ba}$), the arithmetic average $(D_{ab}+D_{ba})/2$ was calculated in Ref. 34. We notice that D_{ab} diverges if the distribution b is highly anisotropic (with a near-zero eigenvalue). To avoid such divergence, we use $min\{D_{ab},D_{ba}\}$ instead of $(D_{ab}+D_{ba})/2$ as our KL distance metric.

3. Dot product

It is defined as the absolute value of the dot product between the two eigenvectors of U and V with the largest eigenvalue.[33] The dot product is 1 if the long axes of U and V are perfectly aligned, and 0 if their long axes are perpendicular to each other.

4. Pearson correlations

The above metrics only evaluate the directional similarity of two ADPs. To include the magnitude of ADPs into comparison, we compute the Pearson correlation (termed pc_{all}) between two sets of ADPs as two $6N'$-dimensional vectors $\tilde{U}$ and $\tilde{V}$ (Ref. 34) (N' is the number of ADPs):

$$pc = \frac{\Sigma_{j=1}^{6N'}(\tilde{U}_j - \langle\tilde{U}\rangle)(\tilde{V}_j - \langle\tilde{V}\rangle)}{\sqrt{\Sigma_{j=1}^{6N'}(\tilde{U}_j - \langle\tilde{U}\rangle)^2 \Sigma_{j=1}^{6N'}(\tilde{V}_j - \langle\tilde{V}\rangle)^2}}. \quad (16)$$

We also calculate the Pearson correlations for $3N'$ diagonal and $3N'$ off-diagonal ADP elements separately (termed $pc_{diagonal}$ and $pc_{offdiagonal}$, respectively), and the Pearson correlation between theoretical and experimental B factors (termed pc_{trace}).[34]

D. Crystallographic data set for model evaluation

We evaluate our modeling of ADPs using a set of 83 ultrahigh-resolution crystal structures (with resolution at or beyond 1 Å) collected and studied in Refs. 33 and 40. From the ANISOU records of these PDB structures, we collect 16 852 usable ADPs for those C_α atoms with occupancy of 1.0 (though all C_α atoms are included in the construction of ENM).

Following Refs. 33 and 40, for the evaluation of Pearson correlations of all, diagonal, off-diagonal ADP elements and B factors, we use all 16 852 ADPs; for the evaluation of directional metrics (f_{ncc}, cc_{mod}, KL distance and dot product), we use a subset of 6784 ADPs with anisotropy of ≤ 0.5.

III. RESULTS AND DISCUSSION

To explore how crystal packing affects the modeling of ADPs, we performed ENM-based modeling of a protein structure embedded in crystalline environment for a list of 83 high-resolution crystal structures.[33,40] We will address the following questions based on the modeling results.

A. How does the quality of ADP modeling depend on the strength of protein-environment interactions?

In previous modeling of protein structures in crystalline states,[40,41] it was assumed that the intraprotein interactions are of the same strength as the interactions between neighboring protein molecules. Although this assumption is chemically sound (same types of atomic forces are involved in both intraprotein and interprotein interactions), the following three lines of reasoning suggest that the latter may be significantly weaker than the former:

(1) The intraprotein interactions among densely packed residues are dominated by favorable hydrophobic interactions, hydrogen bonds, and weakly screened electrostatic interactions. However, the crystal contact interactions often involve loosely packed surface residues exposed to solvents, so they are energetically less favorable (in particular, the electrostatic interactions are subject to strong solvent screening).
(2) The intraprotein interactions, thanks to their key roles in stabilizing protein native conformations, tend to be evolutionally conserved, while the crystal contact interactions are not under evolutionary pressure.
(3) Weak crystal contact interactions are also consistent with the general observation that protein native conformations are not significantly altered by crystal packing, and proteins can remain functionally active in crystalline states (see Ref. 45).

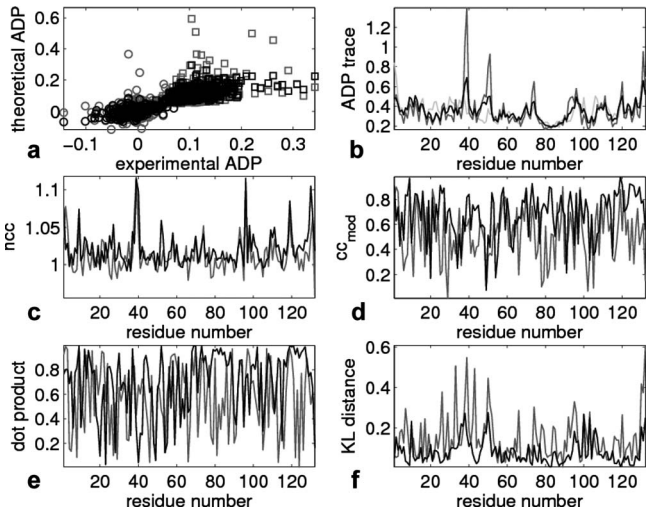

FIG. 2. The results of ADP modeling for a zinc protease crystal structure (PDB: 1C7K): (a) Scatter plot of experimental ADPs vs theoretical ADPs (in unit of Å^2). The theoretical ADPs are calculated for isolated structure ($f_{env}=0$, colored gray) and weak protein-environment interactions ($f_{env}=0.02$, colored black). The diagonal and off-diagonal ADP elements are shown as squares (□) and circles (○), respectively. Panels (b)–(f) show the ADP traces and several directional metrics (ncc, cc_{mod}, dot product, KL distance) as a function of residue number for isolated structure ($f_{env}=0$, colored gray) and weak protein-environment interactions ($f_{env}=0.02$, colored black). In panel (b), the experimental B factors (rescaled by $3/8\pi^2$) are colored in light gray. The ANM scheme with fixed-environment boundary condition is used here (see Sec. II). The theoretical ADPs are normalized so that the sum of theoretical ADP traces is equal to the sum of experimental ADP traces.

To further explore the above issue, we performed ENM-based modeling of a two-component system comprised of a protein structure and its crystalline environment. A new parameter f_{env} is introduced to describe the strength of interprotein interactions relative to that of intraprotein interactions (see Sec. II). To offer a glimpse to the modeling results, we have shown the results of ADP modeling for a zinc protease crystal structure (PDB: 1C7K, see Fig. 2) under two modeling conditions: isolated protein structure ($f_{env}=0$) and weak protein-environment interactions ($f_{env}=0.02$). The fixed-environment boundary condition (see Sec. II) is used here. A better agreement between theoretical and experimental ADPs [see Fig. 2(a)] and B factors [see Fig. 2(b)] is found at $f_{env}=0.02$; the Pearson correlations for all, diagonal, off-diagonal ADP elements and B factors (see Sec. II) increase from 0.72, 0.32, 0.24, and 0.38 at $f_{env}=0$ to 0.89, 0.54, 0.53, and 0.65 at $f_{env}=0.02$. For the directional comparison of experimental and theoretical ADPs, we focus on 66 out of 132 experimental ADPs with anisotropy ≤ 0.5 (following Ref. 40). We calculate four metrics for directional similarity between theoretical and experimental ADPs (f_{ncc}, cc_{mod}, KL distance and dot product, see Sec. II), which all indicate improvement from $f_{env}=0$ to $f_{env}=0.02-f_{ncc}$ increases from

0.77 to 0.94 [see Fig. 2(c)], the average cc_{mod} increases from 0.56 to 0.71 [see Fig. 2(d)], the average dot product increases from 0.63 to 0.69 [see Fig. 2(e)], and the average KL distance decreases from 0.17 to 0.10 [see Fig. 2(f)]. Notably, the introduction of weak protein-environment interactions has suppressed sharp peaks in the theoretical B factors [see Fig. 2(b)] and KL distances [see Fig. 2(f)]. For comparison, we also explored strong protein-environment interactions ($f_{env}=1$, data not shown). The modeling quality of $f_{env}=1$ lies between $f_{env}=0$ and 0.02; the Pearson correlations for all, diagonal, off-diagonal ADP elements and B factors are 0.84, 0.39, 0.32, and 0.48, f_{ncc} is 0.80, the average cc_{mod} is 0.61, the average dot product is 0.66, and the average KL distance is 0.14. Therefore, among the above three modeling conditions ($f_{env}=0, 0.02, 1$), the optimal ADP modeling for 1C7K is achieved at weak protein-environment interactions ($f_{env}=0.02$).

The above ADP modeling and evaluation have been conducted for 16 852 ADPs collected from 83 PDB structures.[33,40] To deduce the average performance of our ADP modeling, we average four Pearson correlations (for all, diagonal, off-diagonal ADP elements and B factors) over 83 structures, and three directional metrics (cc_{mod}, dot product and KL distance, see Sec. II) over a subset of 6784 ADPs with anisotropy of ≤ 0.5 (another directional metric f_{ncc} is calculated over this ADP subset, see Sec. II). To explore how the ADP modeling quality depends on the strength of protein-environment interactions, these average metrics are plotted as a function of f_{env} for three boundary conditions (free, fixed, and buffered environment) and three ENM schemes (ANM, DNM, and HCA) (see Sec. II).

For ANM combined with three boundary conditions, it is found that the bottom of average KL distance and the peaks of other metrics are roughly aligned near $f_{env} \sim 0.02$ (see Fig. 3). Therefore, the optimal ADP modeling is attained when the protein-environment interactions are much weaker than the intraprotein interactions. For fixed environment, the improvement from $f_{env}=0$ to 0.02 is significant for all metrics [see Fig. 3(a)]; the average Pearson correlations for all, diagonal, and off-diagonal ADP elements and B factors increase from 0.60, 0.40, 0.26, and 0.48 at $f_{env}=0$ to 0.83, 0.58, 0.46, and 0.64 at $f_{env}=0.02$; f_{ncc} increases from 0.79 to 0.94, the average cc_{mod} increases from 0.57 to 0.74, the average dot product increases from 0.64 to 0.74, and the average KL distance decreases from 0.23 to 0.10. The performance at $f_{env}=1$ is intermediate between $f_{env}=0$ and 0.02 [for example, the Pearson correlation of B factors is 0.51 and f_{ncc} is 0.81 at $f_{env}=1$, see Fig. 3(a)]. Similar results are found for the free-environment and buffered-environment boundary conditions [see Figs. 3(b) and 3(c)].

Similar results are obtained for DNM and HCA combined with three boundary conditions,[46] although the optimal f_{env} (~ 0.04 for DNM and ~ 0.06 for HCA under fixed-environment boundary condition) is higher than ANM. The optimal performance of HCA and DNM is slightly better than ANM—for HCA and fixed environment, the average Pearson correlation of B factors is 0.69, f_{ncc} is 0.95, the average cc_{mod} is 0.75, the average dot product is 0.75, and the average KL distance is 0.09 at $f_{env}=0.06$; for DNM and fixed environment, the average Pearson correlation of B factors is 0.68, f_{ncc} is 0.96, the average cc_{mod} is 0.75, the average dot product is 0.76, and the average KL distance is 0.09 at $f_{env}=0.04$.

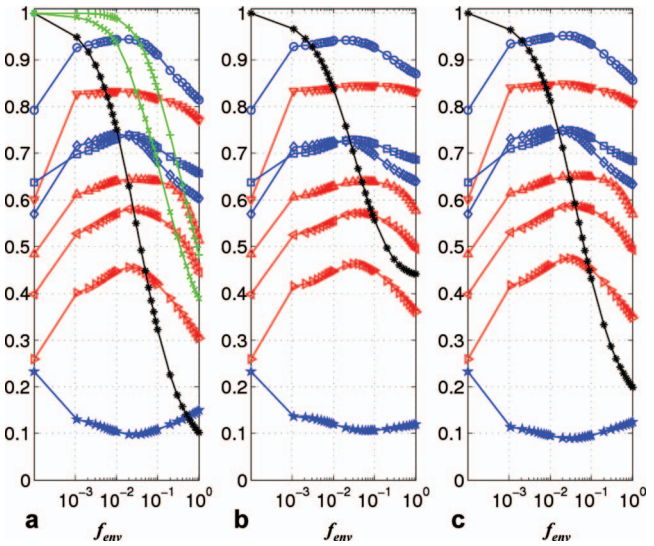

FIG. 3. The results of ADP modeling averaged over 83 PDB structures using ANM under three boundary conditions: (a) fixed environment, (b) free environment, and (c) buffered environment. Shown here are Pearson correlations of diagonal ($\lhd$), off-diagonal ($\rhd$), all elements ($\bigtriangledown$) of ADPs and B factors ($\bigtriangleup$), and directional metrics including f_{ncc} ($\bigcirc$), cc_{mod} ($\diamond$), KL distance ($\star$), dot product ($\square$), and the fractional contribution of rigid-body motions f_{RT} (*) as a function of f_{env}. Also shown in panel (a) are the average maximum overlap $\langle O_{max} \rangle$ ($\times$) and average cumulative overlap $\langle O_{cumu} \rangle$ (+).

To further assess the significance of our finding of optimal ADP modeling at small f_{env}, we investigated how the f_{env}-dependence of ADP modeling quality varies among the 83 PDB structures. Using ANM combined with fixed environment, we calculated all metrics (see Sec. II) and then averaged the three directional metrics (cc_{mod}, dot product, and KL distance) over the ADPs of each structure with anisotropy ≤ 0.5 (see Fig. 4). It is found that the improvement of ADP modeling from $f_{env}=0$ to 0.02 is achieved for 99%, 96%, 100%, 94%, 86%, 95%, 88%, and 99% of all structures as assessed by the Pearson correlations for all, diagonal, and off-diagonal ADP elements and B factors, f_{ncc}, cc_{mod}, dot product, and KL distance. Similarly, the improvement from $f_{env}=1$ to 0.02 is achieved for 85%, 93%, 88%, 81%, 80%, 96%, 89%, and 96% of all structures as assessed by the Pearson correlations for all, diagonal, and off-diagonal ADP elements and B factors, f_{ncc}, cc_{mod}, dot product, and KL distance. So the finding of optimal ADP modeling at $f_{env}=0.02$ (compared to $f_{env}=0$, 1) holds for $>80\%$ PDB structures of our data set. The p-value of this finding is estimated to be $\sim 2.1 \times 10^{-8}$, which indicates its statistical significance.

B. How much do the rigid-body motions contribute to ADP?

The importance of rigid-body motions to protein dynamics at crystalline states remains controversial. The earlier modeling of B factors and ADPs by ENM usually ignored the contribution from rigid-body motions.[30,33,35] One study compared a set of models for crambin at 0.83 Å resolution

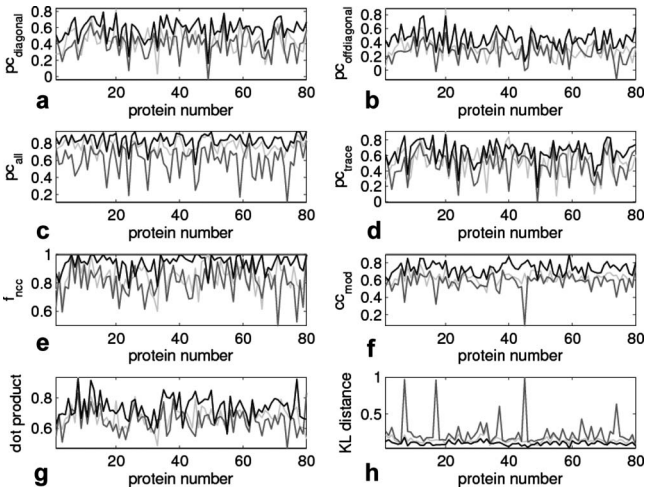

FIG. 4. The variation of ADP modeling results among 83 PDB structures calculated using ANM and fixed environment. Panels (a)–(h) show Pearson correlations of diagonal, off-diagonal, all elements of ADP and B factors, and directional metrics including f_{ncc}, $\mathrm{cc}_{\mathrm{mod}}$, dot product, and KL distance. The results of $f_{\mathrm{env}}=0$, 0.02, and 1 are colored gray, black, and light gray, respectively. Because three PDB structures do not have ADPs with anisotropy ≤ 0.5, only results for 80 PDB structures are shown.

using TLS model and concluded that rigid-body libration of entire crambin contribute 60% of the overall mobility.[47] In a recent study, vGNM was proposed to take into account the contribution of rigid-body motions and allow the amplitude of low-frequency modes to be variables,[36] and it was found that rigid-body motions account for nearly 60% of the total atomic fluctuations.[36] However, a recent comparison of molecular dynamics simulations with crystallographic B factors estimated that rigid-body motions contribute only 20%–30% of total positional variance in B factors.[48]

To address the above controversy, we calculated the fractional contribution of six translational and rotational modes to the total positional variance in a protein structure:

$$f_{RT}=\frac{\sum_{m=1}^{6} V_{RT,m}^{T}\langle u_P u_P^{T}\rangle V_{RT,m}}{\mathrm{Trace}\langle u_P u_P^{T}\rangle}, \tag{17}$$

where $V_{RT,m}$ is the eigenvector of the mth translational and rotational mode ($1\leq m\leq 6$) of a protein structure, $\langle u_P {u_P}^{T}\rangle$ is the covariance matrix given in Eqs. (7)–(9). f_{RT} is calculated as a function of f_{env} (see Fig. 3). Because the rigid-body motions of a protein structure are restrained by the crystal contact interactions characterized by f_{env}, f_{RT} decreases as f_{env} increases from 0 to 1. For $f_{\mathrm{env}}=0$, the protein structure is unrestrained by crystal contacts, so the translations and rotations dominate the thermal fluctuations because they can be excited without energy cost (this anomaly was usually removed by excluding the contributions of translational and rotational modes to $\langle u_P {u_P}^{T}\rangle$).[30,33,35] Depending on the ENM scheme and boundary condition, f_{RT} ranges from 50% to 70% when the optimal ADP modeling is attained (with f_{env} within $0.02\sim 0.06$). Our result is in agreement with several previous studies,[36,47] and it confirms that rigid-body motions contribute significantly to the atomic fluctuations of protein structures in crystalline states, so they must be considered for correct modeling of ADPs.

C. How robust are low-frequency modes to crystal packing?

Numerous studies demonstrated the importance of low-frequency modes in describing protein functional motions.[26,27] Therefore, it is essential to assess to what extent they are affected by crystal packing. To this end, we compared the following two sets of low-frequency modes: (A) the 10 lowest nonzero modes solved for an isolated protein structure ($f_{\mathrm{env}}=0$) (the six translational and rotational zero modes for isolated protein structure are excluded from comparison) and (B) the 16 lowest nonzero modes solved for the entire protein-environment system with $f_{\mathrm{env}}\in(0,1]$. We only consider ANM combined with fixed environment (see Sec. II), because the normal modes solved for free or buffered environment are not directly comparable with the modes for isolated protein structure due to their different dimension.

To assess the *individual* similarity between the above two sets of modes (A and B), we compute the average maximal overlap $O_{\max}=\Sigma_{m=1}^{10}\max_{1\leq n\leq 16}(O(m,n))/10$, where $O(m,n)$ is the absolute value of the dot product between the eigenvectors of mode m from A set and mode n from B set. To assess the *collective* similarity between the two sets of modes, we calculate the average cumulative overlap $O_{\mathrm{cumu}}=\Sigma_{m=1}^{10}\Sigma_{n=1}^{16}O(m,n)^2/10$. $O_{\max}$ and O_{cumu} are calculated and averaged over 83 PDB structures to get $\langle O_{\max}\rangle$ and $\langle O_{\mathrm{cumu}}\rangle$ as a function of f_{env} [see Fig. 3(a)].

$\langle O_{\max}\rangle$ ($\langle O_{\mathrm{cumu}}\rangle$) is found to decrease from 1 to 0.39 (0.48) as f_{env} increases from 0 to 1 [see Fig. 3(a)]. At $f_{\mathrm{env}}=0.02$, $\langle O_{\max}\rangle\sim 0.88$ and $\langle O_{\mathrm{cumu}}\rangle\sim 0.97$, which indicate that the low-frequency modes at $f_{\mathrm{env}}=0.02$ are highly similar to the modes for isolated protein structure. In particular, the finding that $\langle O_{\mathrm{cumu}}\rangle\sim 1$ suggests that the essential subspace spanned by the low-frequency modes is nearly invariant despite weak crystal contact interactions. The robustness of low-frequency modes to crystal packing implies that proteins in crystalline states have intact functional motions and therefore can remain functionally active (see Ref. 45).

D. How does crystal packing affect the optimal parametrization of ANM?

In previous studies, it was found that the optimal fitting of B factors by ANM is attained at a high cutoff distance $R_c=15$ Å–24 Å,[30] which is beyond the range of C_α–C_α distances between contacting residues (4.4–12.8 Å, see Ref. 49). Such inconsistency casts doubt on the validity of ENM parametrization by fitting B factors without considering crystal packing effects.[40] To address this issue, we evaluated the quality of ADP modeling by ANM as a function of R_c for isolated protein structure ($f_{\mathrm{env}}=0$) and weak protein-environment interactions ($f_{\mathrm{env}}=0.02$). Interestingly, the R_c-dependence differs significantly between the two cases (see Fig. 5).

At $f_{\mathrm{env}}=0$, the quality of ADP modeling assessed by the four average Pearson correlations (for all, diagonal, off-diagonal ADP elements, and B factors) and average KL distance improves as R_c increases from 7 to 20 Å [see Fig. 5(a)]. The average $\mathrm{cc}_{\mathrm{mod}}$ also peaks at 20 Å. So the optimi-

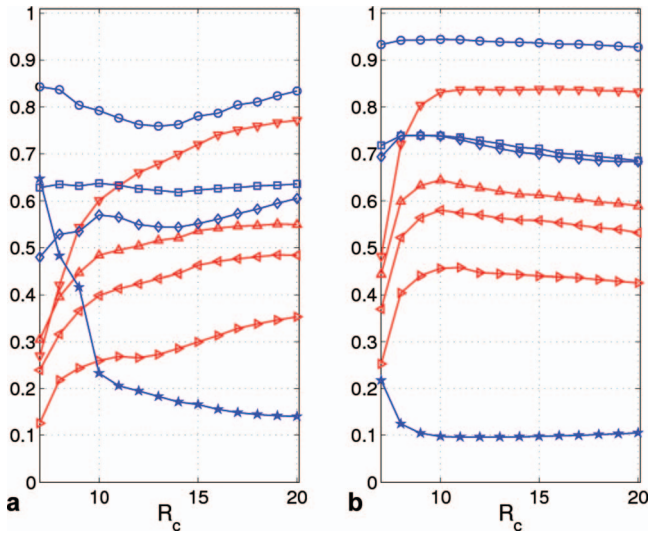

FIG. 5. R_c-dependence of ADP modeling results averaged over 83 PDB structures using (a) ANM for isolated structure ($f_{env}=0$) and (b) ANM with fixed environment and weak protein-environment interactions ($f_{env}=0.02$). Shown here are Pearson correlations of diagonal ($\triangleleft$), off-diagonal ($\triangleright$), all elements ($\triangledown$) of ADPs and B factors ($\triangle$), and directional metrics including f_{ncc} ($\bigcirc$), cc_{mod} ($\diamond$), KL distance ($\star$) and dot product ($\square$) as a function of R_c.

zation of ADP fitting by ANM without considering crystal packing leads to a high $R_c \geq 20$ Å in agreement with previous findings.[30,40]

At $f_{env}=0.02$, all metrics peak near $R_c=10$ Å (except the average KL distance and the average Pearson correlation for all ADP elements, which are flat for 10 Å $\leq R_c \leq 20$ Å) [see Fig. 5(b)]. Thus an optimal fitting of ADPs is attained near $R_c=10$ Å.

The finding of optimal parametrization of ANM at $f_{env}=0.02$ is physically more meaningful, because $R_c=10$ Å falls well within the range of C_α–C_α distances between contacting residues. Our finding supports the importance of considering crystal packing effects for the modeling of ADPs.[40]

E. Further evaluation of interprotein and intraprotein residue-residue interactions

Based on the optimization of ADP modeling, we found that the interprotein interactions are much weaker than the intraprotein interactions in protein crystals. This finding is somewhat unexpected because both interprotein and intraprotein interactions involve the same chemical forces at the atomic level. To further assess the above finding, we used a statistical potential developed by Miyazawa and Jernigan[50] to evaluate the residue-residue contact energy between neighboring proteins and within a protein for the list of 83 protein structures. Following Ref. 50, for a protein structure and its neighboring molecules, we calculated the contact energy between pairs of residues whose side-chain centers are within 6.5 Å. The average contact energy between residues of neighboring proteins (termed $e_{interprotein}$) and residues within a protein (termed $e_{intraprotein}$) are calculated and reported in Table I. To assess the favorability of calculated energy values, we also calculated a reference energy $e_{ref}=\Sigma_{n=1}^{20}\Sigma_{m=1}^{20}f_n f_m e_{nm}$, where n and m are indices for amino acid

TABLE I. Comparison of average residue-residue contact energy between neighboring proteins ($e_{interprotein}$), within a protein ($e_{intraprotein}$), and the reference energy (e_{ref}). The energy is calculated using the Miyazawa–Jernigan statistical potential (Ref. 50). For details, see Sec. III of main text.

	Average residue-residue contact energy		
PDB code	$e_{intraprotein}$	$e_{interprotein}$	e_{ref}
1a6m	−3.94	−2.15	−3.04
1byi	−3.72	−2.71	−3.10
1c75	−3.43	−1.83	−2.76
1c7k	−3.09	−1.83	−2.66
1ea7	−3.24	−2.37	−2.82
1eb6	−3.20	−1.98	−2.65
1exr	−3.58	−2.03	−2.90
1f94	−3.42	−3.00	−3.09
1f9y	−3.97	−2.29	−3.23
1g4i	−3.45	−2.59	−2.84
1g66	−3.30	−2.19	−2.83
1g6x	−3.46	−2.98	−2.99
1ga6	−3.22	−2.34	−2.83
1gkm	−3.68	−2.95	−3.15
1gqv	−3.57	−2.24	−2.93
1gvk	−3.51	−2.54	−3.08
1gwe	−3.29	−2.84	−2.82
1hj9	−3.44	−2.13	−2.93
1i1w	−3.57	−1.93	−2.97
1iqz	−3.41	−3.23	−2.87
1iua	−3.35	−2.36	−2.75
1ix9	−3.64	−1.98	−2.98
1ixh	−3.45	−1.76	−2.84
1j0p	−2.73	−1.58	−2.38
1jfb	−3.82	−2.29	−3.03
1k5c	−3.40	−1.91	−2.84
1kth	−3.06	−2.90	−2.74
1kwf	−3.41	−1.98	−2.91
1l9l	−3.85	−1.92	−2.82
1lkk	−3.73	−2.38	−2.98
1lni	−3.54	−2.49	−2.93
1lug	−3.64	−1.94	−2.92
1m1q	−3.02	−2.11	−2.57
1m40	−3.63	−2.28	−3.02
1mc2	−3.34	−2.36	−2.81
1mj5	−3.75	−2.28	−3.11
1muw	−3.60	−3.03	−3.02
1mwq	−3.48	−2.62	−3.01
1n4w	−3.48	−2.33	−3.02
1n55	−3.70	−3.00	−3.05
1nki	−3.98	−2.14	−3.24
1nls	−3.53	−2.47	−2.97
1nwz	−3.77	−2.18	−2.95
1o7j	−3.59	−2.07	−3.06
1oai	−3.83	−2.14	−2.79
1od3	−3.64	−2.44	−2.99
1ok0	−3.12	−3.06	−2.91
1pq7	−3.47	−2.06	−2.91
1r2m	−4.00	−3.25	−3.29
1r6j	−4.08	−2.81	−3.07
1rb9	−3.29	−2.02	−2.64
1rtq	−3.35	−2.12	−2.86
1sfd	−3.77	−2.58	−3.01
1ssx	−3.31	−2.21	−2.92
1tg0	−3.73	−2.51	−2.91
1tqg	−4.29	−2.49	−3.22

TABLE I. (*Continued.*)

PDB code	Average residue-residue contact energy		
	$e_{\text{intraprotein}}$	$e_{\text{interprotein}}$	e_{ref}
1tt8	−3.99	−2.54	−3.27
1u2h	−3.40	−2.59	−2.81
1ufy	−3.98	−3.07	−3.16
1ug6	−3.70	−2.72	−3.16
1unq	−3.44	−2.79	−2.86
1us0	−3.80	−1.99	−3.12
1v0l	−3.48	−1.95	−2.87
1v6p	−2.87	−2.34	−2.41
1vbw	−3.46	−2.35	−2.77
1vyr	−3.48	−2.00	−2.91
1vyy	−3.33	−3.27	−3.00
1w0n	−3.57	−2.51	−2.98
1x6z	−3.25	−2.38	−2.64
1x8q	−3.40	−1.92	−2.83
1xmk	−3.66	−2.63	−2.94
1y55	−3.55	−3.55	−3.08
1ylj	−3.44	−2.19	−2.88
1zk4	−3.45	−3.17	−2.93
1zzk	−3.62	−2.53	−2.89
2bt9	−3.18	−2.17	−2.76
2bw4	−3.51	−2.90	−3.02
2cws	−3.45	−2.14	−2.87
2f01	−3.09	−2.87	−2.78
2fdn	−3.60	−2.98	−2.99
2pvb	−3.79	−2.16	−2.96
3lzt	−3.57	−2.31	−2.90
7a3h	−3.52	−1.97	−2.93
Average	−3.52	−2.41	−2.92

types, f_n is the percentage of amino acid n in a given protein, and e_{nm} represents the Miyazawa–Jernigan contact energy[50] between amino acid n and m. A contact energy value is considered favorable (or unfavorable) if it is lower (or higher) than e_{ref}.

Among the 83 protein structures, it is found that $e_{\text{interprotein}} > e_{\text{intraprotein}}$ and $e_{\text{intraprotein}} < e_{\text{ref}}$ in all cases, while $e_{\text{interprotein}} < e_{\text{ref}}$ only in nine cases (see Table I). Therefore, the intraprotein residue-residue interactions are significantly stronger than the interprotein ones. The former are highly favorable in stabilizing protein native structures, while the latter are much less favorable. This finding strongly supports our modeling of ADPs based on weak protein-environment interactions (i.e., small f_{env}).

To further establish the proposed dependence of ADP modeling quality on the relative strength of interprotein and intraprotein interactions, we selected the following two subsets from the list of 83 protein structures. Subset I consists of ten structures with the highest values of $e_{\text{intraprotein}} - e_{\text{interprotein}}$, which have relatively strong interprotein interactions (i.e., comparable to intraprotein interactions). Subset II consists of ten structures with the lowest values of $e_{\text{intraprotein}} - e_{\text{interprotein}}$, which have very weak interprotein interactions (i.e., much higher than intraprotein interactions). Then we modeled the ADPs from each subset using a "homogeneous" two-component ENM (i.e., $f_{\text{env}}=1$) (for results, see Table II). Indeed, the modeling performance, as assessed by various metrics, is much better for subset I than subset II (see Table II). Compared with the results of ADP modeling at $f_{\text{env}}=0$ (see Table II), we found significant improvement from $f_{\text{env}}=0$ to 1 for subset I. In contrast, most metrics indicate either weak or no improvement from $f_{\text{env}}=0$ to 1 for subset II. The above results strongly support our proposal because the use of $f_{\text{env}}=1$ assumes comparable strength for interprotein and intraprotein interactions, which is true for subset I but not subset II. In sum, the above results not only support our finding that the interprotein interactions are much weaker than the intraprotein interactions in protein crystals, but also point to possible use of energy calculations based on statistical potentials to guide the tuning of f_{env} to optimize ADP modeling.

IV. CONCLUSION

We performed ENM-based modeling of atomic fluctuations in a protein structure that interacts with its crystalline environment. The modeling results are compared with the ADP data from a data set of 83 high-resolution crystal structures.[33,40] The main contributions of this study are as follows:

(1) We explored crystal packing effects using ENM with a new parameter f_{env} to tune the relative strength of interprotein and intraprotein interactions. We found that the optimal modeling of ADPs require significantly

TABLE II. ADP modeling results averaged over two subsets and the entire list of 83 PDB structures. Subset I consists of ten structures with relatively strong crystal contact interactions. Subset II consists of ten structures with very weak crystal contact interactions. Results of $f_{\text{env}}=0$ and 1 are shown. See Sec. II for definitions of the metrics for ADP modeling assessment.

Test cases	f_{env}	$\text{pc}_{\text{diagonal}}$	$\text{pc}_{\text{offdiagonal}}$	pc_{all}	pc_{trace}	f_{ncc}	cc_{mod}	Dot product	KL distance
Subset I: 1y55, 1ok0,1vyy,1kth, 1iqz,2f01,1zk4, 1f94,1gwe,1g6x	1.000	0.56	0.41	0.78	0.64	0.90	0.66	0.71	0.13
	0.000	0.41	0.23	0.56	0.49	0.81	0.51	0.64	0.46
Subset II: 1l9l, 1nki,1us0,1tqg, 1a6m,1lug,1oai, 1ixh,1f9y,1ix9	1.000	0.40	0.24	0.76	0.46	0.78	0.58	0.62	0.17
	0.000	0.42	0.26	0.63	0.53	0.77	0.56	0.59	0.25
All 83 structures	1.000	0.44	0.30	0.77	0.51	0.81	0.60	0.66	0.15
	0.000	0.40	0.26	0.60	0.48	0.79	0.57	0.64	0.23

weaker interprotein interactions than intraprotein interactions. This robust result is obtained for three different boundary conditions (fixed, free, and buffered environments) and three ENM schemes (ANM, DNM, and HCA). The fixed and free environments represent two opposite limits of treating the mobility of crystalline environment, while the buffered environment is in between these two limits. So our finding is unlikely to be an artifact of a particular choice of boundary condition or model parameter. Further evaluation of residue-residue contact energy using the Miyazawa–Jernigan statistical potential[50] supports the above finding.

(2) Despite the simplicity of our model, we achieved comparable ADP-modeling quality than previous studies that used more fitting parameters[51] or computationally expensive formulation of boundary conditions.[40] Our ADP modeling based on ANM and fixed environment boundary condition ($f_{env}=0.02$) yielded average pc_{trace} ~ 0.64, average $cc_{mod} \sim 0.74$, and average dot product ~ 0.74. In comparison, these metrics were found to be 0.55, 0.70, and 0.73 in a recent ADP-modeling study based on ANM and Born–von Ka´rma´n boundary condition.[40] It will be interesting to see if the combination of weak protein-environment interactions and Born–von Ka´rma´n boundary condition would lead to further improvement in ADP modeling.

(3) Our modeling protocol is highly efficient thanks to the use of a sparse linear-equation solver instead of more expensive eigensolver. Furthermore, the use of matrix inversion [see Eqs. (7)–(9)] accounts for contributions of all modes to ADPs, which helps to improve the modeling accuracy.[33]

Following Ref. 40, we also explored ANM combined with the periodic boundary conditions for the asymmetric unit, which result in worse ADP modeling results than the three boundary conditions considered here (data not shown). Therefore, it is not sufficient to only consider those normal modes that observe the crystallographic symmetry. Other contributions, either from lattice vibrations at $q \neq 0$ (Ref. 40) or symmetry-breaking modes of protein-environment system, should be counted for accurate modeling of ADPs.

The finding of optimal ADP modeling for weak protein-environment interactions has the following important implications:

(1) The rigid-body motions of a protein relative to its crystalline environment are only weakly restrained, so they can contribute significantly to thermal fluctuations as found in previous studies[36,47] (however, recent studies showed that internal motions may contribute more substantially in large flexible protein complexes, see Refs. 14 and 15). Therefore, the modeling of ADPs with internal motions alone is not justified and may lead to unphysical parametrization of ENM, while the proper incorporation of rigid-body motions can result in physically meaningful parametrization of ENM.[32,40] The finding of weak protein-environment interactions for optimal fitting of ADPs is consistent with our previous finding that the addition of a small fraction of GNM potential to ANM potential improves the fitting of B factors and crystallographically observed conformational changes.[52] In both formulations, large contributions of rigid-body motions are taken into account.

(2) Our finding lends support to the general notion that crystal packing only causes weak perturbations to the internal protein dynamics. So the conformational fluctuations relevant to protein functions can be studied by x-ray crystallography despite the effects of crystal packing.

We have seen small but noticeable improvement in ADP modeling from ANM to DNM and HCA (see Fig. 3 and Ref. 46), which supports the efforts to refine the force constants of ENM to better represent chemical forces in protein structures.[24,33] Given the previous finding that all-atom potential gives better prediction of directions of ADPs than C_α-based potentials,[33] it will be worthwhile to incorporate crystal contact interactions into all-atom potential to further improve the modeling of ADPs (Zheng, work in progress). Work in this direction will be greatly aided by the recent development of a highly efficient normal mode analysis protocol based on subsystem-environment partition.[53–55]

Continuing progress in the ENM-based modeling of ADPs will not only lead to deeper understanding of the dynamic basis of ADPs, but also help to improve the accuracy of ENM parameters. The latter is essential to the development and refinement of ENM-based techniques that probe protein dynamics of functional importance, which have been pursued in our recent studies.[56–59]

We caution that ENM is limited to the modeling of thermal fluctuations at harmonic limit. Future work is clearly needed to account for nonharmonic fluctuations and other contributions to ADPs beyond thermal fluctuations (such as static disorders, lattice defects, etc.). As indicated by a recent study, the thermal fluctuations may only make a small contribution to the crystallographic B factors.[41] If that is the case, there remains a long way to go before a satisfactory understanding of ADPs can be achieved.

ACKNOWLEDGMENTS

We acknowledge funding support from American Heart Association (Grant No. 0835292N).

[1] K. Henzler-Wildman and D. Kern, Nature (London) **450**, 964 (2007).
[2] R. Brüschweiler, Curr. Opin. Struct. Biol. **13**, 175 (2003).
[3] D. W. Li and R. Brüschweiler, J. Am. Chem. Soc. **131**, 7226 (2009).
[4] B. T. M. Willis, *Thermal Vibrations in Crystallography* (Cambridge University Press, London, 1975).
[5] C. Scheringer, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **33**, 879 (1977).
[6] E. A. Merritt, Acta Crystallogr., Sect. D: Biol. Crystallogr. **55**, 1997 (1999).
[7] M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**, 646 (2002).
[8] V. Tozzini, Curr. Opin. Struct. Biol. **15**, 144 (2005).
[9] V. Schomaker and K. Trueblood, Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem. **24**, 63 (1968).
[10] M. D. Winn, M. N. Isupov, and G. N. Murshudov, Acta Crystallogr., Sect. D: Biol. Crystallogr. **57**, 122 (2001).
[11] J. Painter and E. A. Merritt, Acta Crystallogr., Sect. D: Biol. Crystallogr. **62**, 439 (2006).
[12] R. Diamond, Acta Crystallogr., Sect. A: Found. Crystallogr. **46**, 425 (1990).

[13] A. Kidera and N. Go, Proc. Natl. Acad. Sci. U.S.A. **87**, 3718 (1990).
[14] B. K. Poon, X. Chen, M. Lu, N. K. Vyas, F. A. Quiocho, Q. Wang, and J. Ma, Proc. Natl. Acad. Sci. U.S.A. **104**, 7869 (2007).
[15] X. Chen, B. K. Poon, A. Dousis, Q. Wang, and J. Ma, Structure **15**, 955 (2007).
[16] M. Levitt, C. Sander, and P. S. Stern, J. Mol. Biol. **181**, 423 (1985).
[17] N. Go, T. Noguti, and T. Nishikawa, Proc. Natl. Acad. Sci. U.S.A. **80**, 3696 (1983).
[18] B. Brooks and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **80**, 6571 (1983).
[19] M. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).
[20] F. Tama and Y. H. Sanejouand, Protein Eng. **14**, 1 (2001).
[21] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).
[22] I. Bahar, A. R. Atilgan, and B. Erman, Fold. Des. **2**, 173 (1997).
[23] T. Haliloglu, I. Bahar, and B. Erman, Phys. Rev. Lett. **79**, 3090 (1997).
[24] K. Hinsen, A. Petrescu, S. Dellerue, M. Bellissent-Funel, and G. R. Kneller, J. Chem. Phys. **261**, 25 (2000).
[25] L. Yang, G. Song, and R. L. Jernigan, Proc. Natl. Acad. Sci. U.S.A. **106**, 12347 (2009).
[26] I. Bahar and A. J. Rader, Curr. Opin. Struct. Biol. **15**, 586 (2005).
[27] F. Tama and C. L. Brooks, Annu. Rev. Biophys. Biomol. Struct. **35**, 115 (2006).
[28] T. Z. Sen, Y. Feng, J. V. Garcia, A. Kloczkowski, and R. L. Jernigan, J. Chem. Theory Comput. **2**, 696 (2006).
[29] D. A. Kondrashov, Q. Cui, and G. N. Phillips, Jr., Biophys. J. **91**, 2760 (2006).
[30] E. Eyal, L.-W. Yang, and I. Bahar, Bioinformatics **22**, 2619 (2006).
[31] S. Kundu, J. S. Melton, D. C. Sorensen, and G. N. Phillips, Jr., Biophys. J. **83**, 723 (2002).
[32] R. Soheilifard, D. E. Makarov, and G. J. Rodin, Phys. Biol. **5**, 26008 (2008).
[33] D. A. Kondrashov, A. W. Van Wynsberghe, R. M. Bannen, Q. Cui, and G. N. Phillips, Jr., Structure **15**, 169 (2007).
[34] E. Eyal, C. Chennubhotla, L. W. Yang, and I. Bahar, Bioinformatics **23**, i175 (2007).
[35] L. Yang, G. Song, and R. L. Jernigan, Proteins **76**, 164 (2009).
[36] G. Song and R. L. Jernigan, J. Mol. Biol. **369**, 880 (2007).
[37] G. N. Phillips, Jr., Biophys. J. **57**, 381 (1990).
[38] L. W. Yang, E. Eyal, C. Chennubhotla, J. Jee, A. M. Gronenborn, and I. Bahar, Structure **15**, 741 (2007).
[39] L. Yang, G. Song, A. Carriquiry, and R. L. Jernigan, Structure **16**, 321 (2008).
[40] D. Riccardi, Q. Cui, and G. N. Phillips, Jr., Biophys. J. **96**, 464 (2009).
[41] K. Hinsen, Bioinformatics **24**, 521 (2008).
[42] W. D. Cornell, P. Cieplak, I. R. Bayly, I. R. Gould, K. M. Merz, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).
[43] Y. Chen, T. A. Davis, W. W. Hagner, and S. Rajamanickam, ACM Trans. Math. Softw. **35**, 1 (2008).
[44] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[45] S. J. Johnson, J. S. Taylor, and L. S. Beese, Proc. Natl. Acad. Sci. U.S.A. **100**, 3895 (2003).
[46] See supplementary material at http://dx.doi.org/10.1063/1.3288503 for Figs. S1 and S2.
[47] B. Stec, R. Zhou, and M. M. Teeter, Acta Crystallogr., Sect. D: Biol. Crystallogr. **51**, 663 (1995).
[48] L. Meinhold and J. C. Smith, Biophys. J. **88**, 2554 (2005).
[49] M. Cieplak and T. X. Hoang, Biophys. J. **84**, 475 (2003).
[50] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[51] M. Lu and J. Ma, Proc. Natl. Acad. Sci. U.S.A. **105**, 15358 (2008).
[52] W. Zheng, Biophys. J. **94**, 3853 (2008).
[53] W. Zheng and B. R. Brooks, Biophys. J. **89**, 167 (2005).
[54] H. L. Woodcock, W. Zheng, A. Ghysels, Y. Shao, J. Kong, and B. R. Brooks, J. Chem. Phys. **129**, 214109 (2008).
[55] J. Hafner and W. Zheng, J. Chem. Phys. **130**, 194111 (2009).
[56] W. Zheng, B. R. Brooks, and G. Hummer, Proteins **69**, 43 (2007).
[57] W. Zheng and D. Thirumalai, Biophys. J. **96**, 2128 (2009).
[58] W. Zheng and M. Tekpinar, BMC Structural Biology **9**, 45 (2009).
[59] W. Zheng, Proteins **76**, 747 (2009).