Article

Protein Structural Variation in Computational Models and Crystallographic Data

Dmitry A. Kondrashov,[1,*] Adam W. Van Wynsberghe,[2] Ryan M. Bannen,[1] Qiang Cui,[2,3] and George N. Phillips, Jr.[1,4,5]
[1] Department of Biochemistry
[2] Graduate Program in Biophysics
[3] Department of Chemistry and Theoretical Chemistry Institute
[4] Department of Computer Science
University of Wisconsin, Madison, Madison, WI 53706, USA
[5] Lab address: http://phillips-lab.biochem.wisc.edu/
*Correspondence: dkon@biochem.wisc.edu
DOI 10.1016/j.str.2006.12.006

SUMMARY

Normal mode analysis offers an efficient way of modeling the conformational flexibility of protein structures. We use anisotropic displacement parameters from crystallography to test the quality of prediction of both the magnitude and directionality of conformational flexibility. Normal modes from four simple elastic network model potentials and from the CHARMM force field are calculated for a data set of 83 diverse, ultrahigh-resolution crystal structures. While all five potentials provide good predictions of the magnitude of flexibility, all-atom potentials have a clear edge at prediction of directionality, and the CHARMM potential has the highest prediction quality. The low-frequency modes from different potentials are similar, but those computed from the CHARMM potential show the greatest difference from the elastic network models. The comprehensive evaluation demonstrates the costs and benefits of using normal mode potentials of varying complexity.

INTRODUCTION

The native state of a protein is an ensemble of conformers, deviating to some extent from the average coordinates reported as the experimental structure. Knowledge of the static structure is not sufficient for understanding the functional mechanisms, which often depend on the flexibility of protein structures. Experimental observation of conformational motion of biomolecules is becoming possible, thanks to experimental innovation, but remains a formidable challenge. Crystals can be subjected to time-resolved experiments (Moffat, 2001), but the range of applications is limited to reactions that can be triggered by light or trapped by clever manipulations. Nuclear magnetic resonance spectroscopy can be used to determine both the structure and the dynamics of proteins (Lindorff-Larsen et al., 2005),

but it is limited both by the maximum size of protein structures and by the difficulty of discrimination of slowly or quickly exchanging dynamics (Palmer et al., 2001). Mass spectrometry coupled with hydrogen/deuterium exchange and proteolysis has been used to determine changes in the relative solvent accessibility of amide hydrogens (Lanman and Prevelige, 2004), and single-molecule experiments using optical trapping have resulted in spectacular observations of the motion of motor proteins (Abbondanzieri et al., 2005). In general, direct measurement of molecular motion remains laborious and limited.

Computer simulations of biological macromolecules enable detailed explorations of the conformational ensemble near the native state (Karplus and Kuriyan, 2005). However, the computational cost of molecular dynamics with all-atom force fields limits the accessible timescale of simulations, particularly of large molecular assemblies. Thus approximate methods, such as normal mode analysis (NMA), are often used to efficiently describe the allowed conformational ensemble of protein structures (Brooks and Karplus, 1983; Go et al., 1983; Levitt et al., 1985). The decomposition into modes with different frequencies reduces the dimensionality of the problem, as a few lowest-frequency modes describe the most dominant directions of motion (Teodoro et al., 2003). These global modes have been used to predict protein flexibility (Cui et al., 2004; Van Wynsberghe et al., 2004) and to study the mechanism of conformational transitions necessary for protein function (Ma and Karplus, 1997). Simple coarse-grained potentials, such as elastic network models (ENMs), provide an efficient description of a protein structure by connecting atoms or residues within a certain distance with identical harmonic potentials (Tirion, 1996). Despite the extreme simplification, these models capture the basic topology of a structure and generate predictions of the flexibility and preferred modes of motion of proteins that are in general agreement with experimental data (Bahar and Rader, 2005).

The study of protein conformational dynamics requires interplay between experiment and computation. A readily available measure of conformational mobility is the

Debye-Waller temperature factor, or B factor, which models the variance in atomic position from the scattering data. It has been used as a source of information on protein flexibility for decades (Frauenfelder et al., 1979), and as computational methodologies have matured, studies over large numbers of crystal structures have shown good agreement with computations, specifically with ENM results (Kundu et al., 2002). While the classic B factor has long been a routine parameter in protein structure refinement, until recently few crystal data sets contained sufficiently many observations (unique reflections) to allow determination of anisotropic displacement parameters (ADPs). These parameters model the probability distribution of atomic positions as a Gaussian function with ellipsoidal contours, and have been shown to significantly improve the refinement statistics for crystal structures of biological macromolecules (Dauter et al., 1997; Esposito et al., 2000; Longhi et al., 1997) at resolution better than 1.2 Å. ADPs have been used in a few studies as a qualitative indicator of the directionality of prevalent motion in a protein structure (Wilson and Brunger, 2000), but this source of experimental information has not been systematically exploited.

The proliferation of various simple ENM-like models for macromolecular fluctuation begs the question of their relative fidelity and reliability, but no systematic comparison of the methods has been undertaken, to the best of our knowledge. Recently, systematic assessments of individual ENM potentials were published: a validation of a model based on Cα coordinates (anisotropic network model; ANM) using B factors from a diverse set of crystal structures (Eyal et al., 2006), and a study showing that adding residue-specific parameters into the same model leads to large improvement in B factor prediction (Hamacher and McCammon, 2006). Our group has used 98 highest-resolution crystal structures in the Protein Data Bank (PDB) for systematic evaluation of prediction of the magnitude of motion in protein structures using an isotropic ENM model, and demonstrated that using atomic information and strengthening the model parameter for covalent interaction resulted in better prediction quality (Kondrashov et al., 2006). In the present work, we compare the quality of prediction of the magnitude and direction of structural variance for the most commonly used anisotropic ENM potentials, and we introduce a new one to better model different chemical interactions. Comparison between ADPs and computational variance matrices allows a quantitative evaluation of the merits and drawbacks of different potentials. We also investigate the effect of the choice of potential on global dynamic properties such as the correlation matrix.

RESULTS

Analysis of Crystallographic Data

The present study evaluates predictions of five coarse-grained normal mode potentials using a set of anisotropic displacement parameters from ultrahigh-resolution crystal structures. The PDB (Berman et al., 2000) was
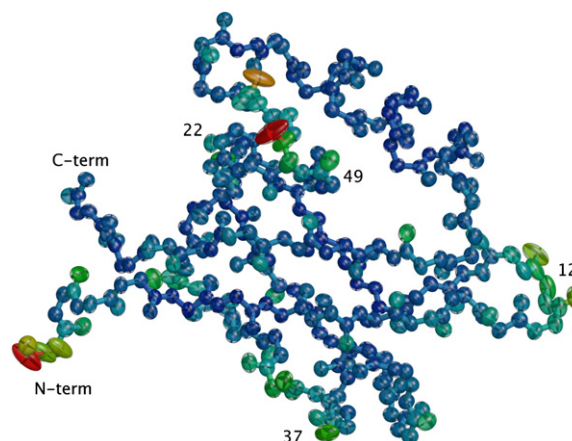

Figure 1. Example of a High-Resolution Protein Structure Showing Anisotropic Temperature Factors for Backbone Atoms

The ellipsoids represent 90% probability volume of atomic position, with color varying from immobile (blue) to more mobile (red). Most of the backbone atoms are not mobile and isotropic, with a few loop residues (residue numbers labeled) showing clear directional preference in positional distribution. PDB ID code 1R6J.

searched for all X-ray crystal structures of proteins with chain length of at least 50 residues, with resolution at or beyond 1 Å, with the restriction that the structures have less than 50% sequence identity. Eighty-three such structures were deposited with anisotropic displacement parameters, containing a total of 17,763 protein residues. Excluding those with disordered Cα atoms or those involved in intermolecular crystal contacts, both of which have an effect on the ADP, left 12,348 residues with usable ADPs. The anisotropic displacement parameters are commonly represented as ellipsoids in crystal structures, as shown in Figure 1, and contain information about both the magnitude and the preferred direction of atomic variation in the crystal. The anisotropy of the ellipsoid, defined as the ratio of the smallest to the largest eigenvalue of ellipsoid matrix (Trueblood et al., 1996), is a measure of deviation from spherical shape. We separated the structures by refinement software and found different distributions of anisotropy for the Cα ADPs. Sixty-eight structures were refined using SHELX (Sheldrick and Schneider, 1997), and the rest were determined using Refmac from the CCP4 suite (CCP4, 1994). The Cα ADPs in the Refmac set had a mean anisotropy of 0.64, compared with 0.51 for the SHELX set (Figure 2), suggesting that the crystallographic restraints used in the two programs have significant effects on resulting ADPs. As sphere-like ellipsoids contain little directional information, a subset of ADPs with anisotropy of less than 0.5 was chosen, leaving 4642 ADPs to compare with the computational predictions of directionality of variance.

Normal Mode Potentials

Elastic network models are dependent on two parameters: the cutoff distance (Tirion, 1996), which separates
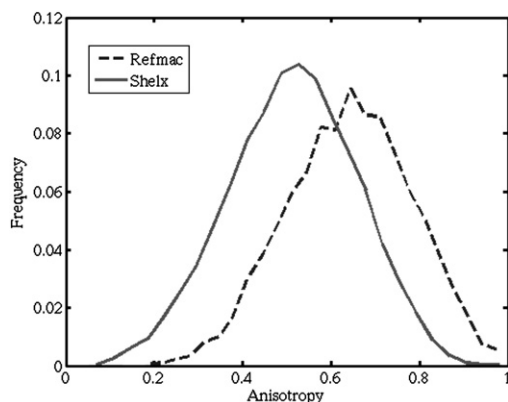
Figure 2. Distributions of Anisotropy Parameters in Structures Refined with SHELX and Refmac Software
The large difference is likely due to different default restraints on the anisotropic parameters in the two.

atom pairs deemed in contact from those which are not interacting, and a force constant for the interaction between contacting atoms We have shown recently that a stronger force constant between covalently bound residues resulted in greatly improved variance prediction quality for an isotropic ENM (Kondrashov et al., 2006), compared with the single force constant Gaussian network model (Bahar et al., 1997). In this work, we introduce a new ENM method, called the distance-based network model (DNM), with multiple force constants for atomic contacts, as described in Experimental Procedures. It is clear that atom pairs closer than 2.3 Å are covalently bound and thus have stronger interactions than those 5 Å apart. To mimic the chemistry, several discrete distance ranges were defined, and force constants were varied to optimize the agreement with ADPs. We found that similar results were obtained if the force constants for each category were set to the reciprocal of the total number of contacts in this range. Because the number of atomic contacts grows with distance, this ensures that interactions between atoms farther away are represented by weaker force constants than those in close proximity. The atomic interactions are added up with the appropriate force constants for each residue, producing a residue-level model based on atomic interactions, with no additional free parameters, as the force constants are defined based on the contact matrices. The only parameter not defined from the structure is the maximum cutoff distance considered, and we optimized it by comparing prediction quality in calculations with a range of cutoffs from 5 Å to 11 Å. The analysis for agreement with magnitudes and directions of ADP ellipsoids with the model predictions is shown in Table S1 (in the Supplemental Data available with this article online). With the exception of the 5 Å cutoff the results were very similar, and 9 Å was selected as the optimal cutoff distance.

We also tested four existing normal mode potentials, three of the ENM variety and one based on the CHARMM force field. The ElNemo method (ElN) (Suhre and Sanejouand, 2004) depends on the cutoff distance between atoms, and the atomic contact matrices are combined into rigid-motion blocks on a residue level. We varied the cutoff distance from 5 Å to 11 Å (Table S2), and found the best results at 5 Å, compared with the default value of 8 Å. The ANM (Atilgan et al., 2001) depends on a cutoff distance between Cα atoms, and we evaluated the results for a range from 10 Å to 16 Å (see Table S3). The variation is also relatively small, but there is an opposite trend between quality of prediction of direction and magnitude of motion. The best cutoff for directionality prediction was at 10 Å, while the best agreement in magnitude was with a 16 Å cutoff, in contrast to the previously used value of 13 Å (Atilgan et al., 2001). Normal modes were computed using an atomistic force field (CHARMM) with rigid-body blocks for residues, referred to as block normal modes (BNM) (Li and Cui, 2002). The inclusion of nonprotein ligands and cofactors in the CHARMM force field resulted in significant increases in quality of prediction, and thus all the nonprotein residues for which CHARMM libraries could be found were added to the models. The last method used is the harmonic Cα potential (HCA) with a distance-dependent force constant (Hinsen et al., 2000), as implemented in the molecular modeling toolkit (MMTK) (Hinsen, 2000).

Comparison of Crystallographic and Computational Variance

Anisotropic covariance tensors were computed from 100 lowest-frequency normal modes (excluding the trivial rotation and translation modes; see Experimental Procedures) from 83 structures, and fidelity of both magnitude and direction prediction was assessed. Magnitude prediction quality was measured by linear correlation between isotropic ADPs (B factors) and the predicted isotropic variances over each structure. Two different measures were used for directional agreement, the absolute value of the dot product between the largest axes of the anisotropic ADPs (ellipsoids) and the volume overlap fraction, as defined in Experimental Procedures. These two measures were employed to compare pairs of corresponding residues, and the reported numbers are the statistics over all sufficiently anisotropic ellipsoids from all 83 structures. Table 1 shows that prediction quality was markedly different for the magnitude and direction of motion. All the models had average isotropic correlations of 0.66–0.68, with the exception of 0.61 for HCA. On the other hand, there was considerable variation in the directional agreement of ADP ellipsoids. The two measures of directional agreement, the dot product and the overlap fraction, largely showed the same trend, with HCA and ANM displaying relatively weak agreement, while ElN, DNM, and BNM show considerably higher prediction quality, with CHARMM-based BNM having an edge over the ENM methods.

The mean absolute value of the dot product is easy to interpret as a measure of the angle between the preferred direction and in the experimental and computed ellipsoids. The average value of 0.65 for BNM corresponds

Table 1. Prediction Quality from 100 Lowest-Frequency Modes Using Different NMA Potentials

	Dot[a]	Overlap[a]	Isotropic[b]
Random[c]	0.5	0.3	0
HCA	0.599 (0.199)	0.520 (0.167)	0.617 (0.131)
ANM (16 Å)[d]	0.556 (0.190)	0.525 (0.172)	0.676 (0.111)
EIN (5 Å)[d]	0.641 (0.208)	0.583 (0.184)	0.680 (0.128)
DNM (9 Å)[d]	0.650 (0.209)	0.575 (0.181)	0.655 (0.136)
CHARMM BNM	0.655 (0.211)	0.608 (0.188)	0.658 (0.128)

[a] Mean and standard deviation of comparisons with ADPs from individual Cα atoms.
[b] Mean and standard deviation for isotropic correlation over 83 entire structures.
[c] Computed for randomly oriented ADPs with anisotropy 0.5.
[d] Choice of optimal cutoff parameter is in parentheses.

Table 2. Prediction Quality of CHARMM BNM from Different Numbers of Lowest-Frequency Modes

	Variance Fraction[a]	Dot Product[b]	Overlap Fraction[b]	Isotropic Correlation[c]
3 modes	0.272	0.632	0.153	0.531
5 modes	0.348	0.634	0.305	0.557
10 modes	0.473	0.640	0.451	0.600
20 modes	0.623	0.643	0.536	0.622
30 modes	0.720	0.645	0.563	0.632
40 modes	0.791	0.646	0.578	0.640
50 modes	0.847	0.648	0.589	0.645
60 modes	0.894	0.650	0.594	0.650

[a] Variation is represented by the indicated subset of modes as a fraction of the variance from 100 lowest-frequency modes, excluding rigid-body modes.
[b] Mean comparisons with ADPs from individual Cα atoms.
[c] Mean isotropic correlation over 83 entire structures.

to an angle of 48° while the average of 0.56 for ANM corresponds to an angle of 56°, but this does not tell the whole story because it only compares one principal axis of the ellipsoids. The overlap coefficient is the volume fraction shared by two ellipsoids of unit volume, and this quantity varies appreciably from 0.52 for HCA to 0.61 for BNM. We tested the hypothesis that predictions agree no better than expected from a random uniform distribution of ellipsoid direction, for which the mean dot product is 0.5 and the mean overlap fraction is 0.3 (when anisotropy is fixed at 0.5). Almost all of the structures with a reasonable sample of usable ADPs (with anisotropy < 0.5) showed better than random agreement in overlap fraction (p < 0.01) (see Supplemental Data). For the method with lowest agreement, HCA, 11 structures did not meet this criterion, and only 4 had more than 10 sufficiently anisotropic ADPs, with the highest at 24. The results for the best-performing BNM method had only 4 structures where the null hypothesis could not be rejected, all of which had only 5 or fewer usable ADPs.

To illustrate the importance of including a large subset of normal modes for accurate variance prediction, we performed the computations with different numbers of modes from CHARMM BNM, shown in Table 2. In NMA, the reciprocal of the eigenvalue (frequency squared) represents the contribution of the mode to the total variance, and the first column shows the cumulative fraction of variance of the first 100 modes represented by the subset. The first 10 modes account for nearly half of the variance, but the prediction quality for all three measures is considerably lower than for the full 100 modes, and shows monotonic improvement with inclusion of additional modes. The effect is dramatic for the overlap fraction, largely due to the contribution of higher-frequency modes to "rounding" of the computed ellipsoids, leading to higher overlap volume with the relatively isotropic ADPs. However, the dot product and the isotropic correlation, which are independent of anisotropy, show consistent improvement with inclusion of additional higher-frequency modes, showing that cal-

culations using only a handful of lowest-frequency modes are likely imprecise.

Effect of Potential on Global Dynamic Ensembles

Comparison of the normal modes produced by different methods revealed a clear distinction between the harmonic ENM models and the CHARMM force field BNM. We used the modes computed from all 83 structures to investigate how the dynamic ensemble predictions depend on the use of the potential. The overlap measure described in Experimental Procedures was used to compare the 17 lowest-frequency modes from all five models. Figure 3 shows the agreement between individual modes for all ten pairs of potentials, averaged over all 83 structures. The highest agreement was observed for the lowest-frequency modes, but the overlap measure dropped below 0.5, depending on the pair of methods, at some point in the first 15 modes. This demonstrated that the details of potential play a secondary role at lowest-frequency modes, which are dominated by contact topology and shape of the molecular structure. A second observation is the distinctiveness of modes derived from CHARMM-based BNM, which showed much lower overlap with ENM-based methods (dotted lines) than overlap among modes from ENM-type potentials (solid lines), with the single exception of the overlap between ANM and EIN. Because the latter is an all-atom potential, it is reasonable that it should be closer to chemistry-based BNM than to Cα-based ANM. We tested the possibility that minimization of structures prior to BNM is responsible for the difference in BNM modes by calculating DNM modes from the minimized structures. The resulting average overlap with BNM was 0.76 as opposed to 0.75 for BNM with DNM from unminimized structures, still much lower than DNM agreement with other methods. This suggests that the chemical information present in the all-atom CHARMM potential plays a role in determining the lowest-frequency modes, in addition to the topology of the structure.
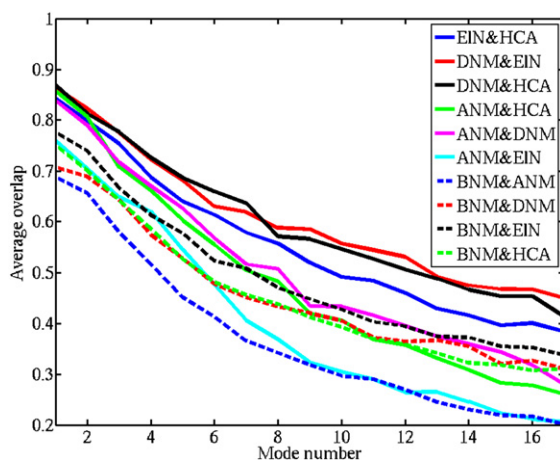
Figure 3. Overlap Scores for Individual Normal Modes from Different Potentials Averaged over All 83 Structures

Each curve is a comparison between a pair of potentials for the 17 lowest-frequency modes. The solid curves compare different ENM-like potentials, while the dotted curves compare CHARMM-based BNM results with those from ENM potentials.

To illustrate the differences between the chemical force field and ENM, we chose a small, well-studied structure from the data set, a PDZ2 domain from syntenin (PDB ID code 1R6J), and computed the correlation matrices (see Experimental Procedures) from the 90 low-frequency modes of ANM and BNM. Figure 4 shows correlation matrices computed from ANM and BNM modes. In general they look quite similar, with major features determined by the secondary-structure elements: antiparallel β sheets appear as positive bands perpendicular to the diagonal, and the two helices result in a thickening of the diagonal band. While the pattern of secondary structures is clear in both potentials, there are evident differences. First,

the magnitude of correlation is at least two times weaker in ANM (see the color bar), and the secondary-structure features are not as clear, due to the inclusion of residues as far as 16 Å away. Second, due to identical force constants for distant and proximal interactions, the diagonal band is considerably weaker in the ANM plot than in BNM, which has a more realistic representation of covalent bonds and other main-chain interactions. Both potentials capture the effect of gross topology, but the effects of specific chemistry are hidden in the fine details of the BNM correlation matrix.

DISCUSSION

We analyzed five different coarse-grained potentials used to model the conformational flexibility of protein structures. These were evaluated both by validation against experimental data and by comparison among the different potentials. To our knowledge, this is the first systematic attempt to use anisotropic displacement parameters to validate computational predictions, and it behooves us to note the challenges arising from using this data source. The reliability of ADPs has been tested before (Merritt, 1999), with good agreement in ellipsoid shape observed between independently determined structures of the same protein; we found the same to be true for structures of myoglobin in four different crystal forms (D.A.K., W. Zhang, R. Aranda IV, B. Stec, and G.N.P., unpublished data). This shows that ADPs are robust experimental parameters, and to minimize the noise contributions we used the highest-resolution crystal structures available. However, quantitative comparison between ADPs and computational predictions is not straightforward, due to contributions of experimental noise, model error (Kuriyan et al., 1986), rigid-body motion of the entire molecule (Kuriyan and Weis, 1991), and specifics of crystal
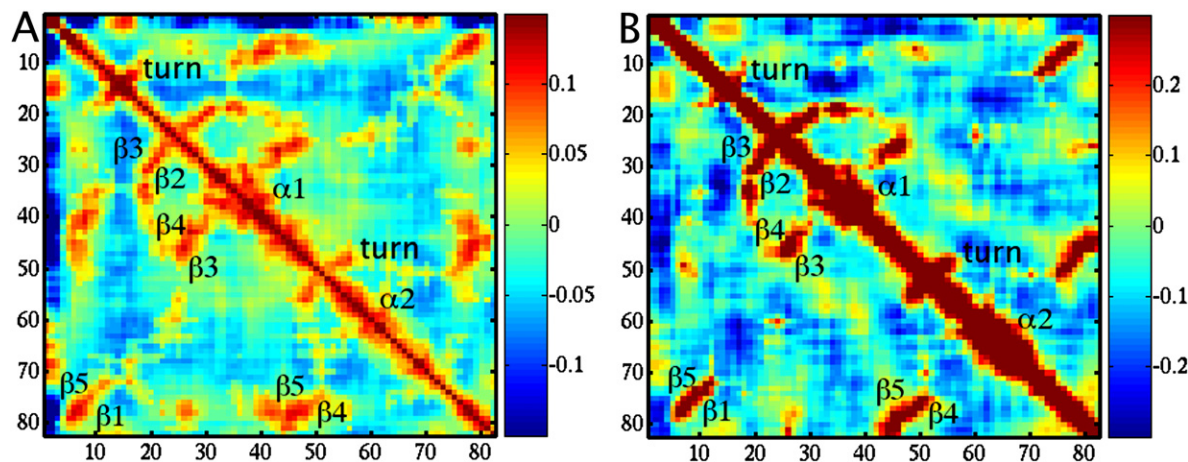

Figure 4. Correlation Matrices Generated from Normal Mode Analyses of a PDZ Domain

The plots show correlation between residues with indices shown on x and y axes, blue color indicating negative correlation and red signifying positive, with the range shown in the color bars. Secondary-structure elements are labeled in sequence order. PDB ID code 1R6J.

(A) Correlation from anisotropic network model.

(B) Correlation from CHARMM-based block normal modes. Note that the range in (A) is half that of (B).

environment, such as crystal contacts between copies of the protein packed in the lattice (Phillips, 1990) and collective lattice modes (Clarage et al., 1992). Further, the ADP represents the best fit of a Gaussian distribution to the electron density of an atom, but anharmonic and multimodal positional distributions are expected for protein atoms, especially in mobile regions, such as the surface. Only atoms with pronounced anisotropy are used for directional comparison, which tend to lie in mobile regions with poorer electron density (see Figure 1) and which are not adequately modeled with a single conformer (DePristo et al., 2004). Thus, it is likely that many atoms in our directional data set are not adequately modeled by the Gaussian ADP model. Despite these caveats, our results show good agreement between the predicted and computed ADPs: for virtually all structures, the overlap fraction between BNM predictions and ADPs is significantly higher than the expectation from a uniform random variable. This suggests that the influence of the factors listed above is not sufficient to overwhelm the important contribution of intramolecular conformational flexibility. This is consistent with a recent comparison of molecular dynamics simulations with crystallographic B factors which estimated that rigid-body motions contribute only 20%–30% of total positional variance in B factors (Meinhold and Smith, 2005). The agreement between computation and experiment serves to validate both the interpretation of the experimental data and the reliability of computational predictions.

In our analysis, we combined multiple low-frequency normal modes to generate the anisotropic variance for each residue from a large number of modes, weighted by the calculated frequencies, and compared the result with the crystallographic variation. This method has been used in previous work applying normal modes to crystallographic refinement (Kidera and Go, 1990), but is not in common use for validating normal modes with experimental displacements. Instead, the procedure is often used to project low-frequency modes individually onto a conformational change, and to obtain a cumulative projection coefficient. This, however, is impossible to do without prior knowledge of the conformational change in the structure, and gives only an agreement between the subspace spanned by several modes and the conformational change. Our approach does not presume any knowledge beyond the initial structure, and measures agreement with the entire normal mode ensemble, rather than with individual modes.

This is also, as far as we know, the first large-scale comparative study of coarse-grained normal mode methods. Comparison of the modes from different potentials reveals a distinct split between the ENM methods and BNM, as seen in Figure 3. This suggests that the chemical information absent in the ENM potential is observable in the BNM results, although there is significant similarity at low-frequency modes due to the shape of the structure reflected in both potential types. The observation opens up a possibility of separating the effect of gross protein structure from that of detailed residue chemistry as reflected by the CHARMM force field. A careful comparison of ENM predictions with those from normal modes with chemical force field could potentially be used to determine residues whose chemistry plays a key role in the dynamic coupling in the structure, and which would therefore be especially sensitive to mutation. The visual comparison of the correlation patterns from ANM and BNM demonstrates that the chemical effects are subtle in comparison to the topological features captured by both BNM and ANM, and all the other methods.

The choice of computational strategy to address a given problem involves balancing computational efficiency against model detail. Fast calculations are meaningless if they give unreliable results, and extremely accurate calculations are of no use if they cannot be completed in a reasonable time frame. Normal mode analysis is based on a choice to limit the model to the neighborhood of the potential minimum. Further simplification of using an elastic network model potential instead of a physical, all-atom potential is another concession toward efficient calculation and away from physical reality. We found that prediction quality of the magnitude of flexibility is similar for CHARMM BNM and all ENM models, with the exception of HCA. This is consistent with a recent comparison of different levels of ENM potentials which found that addition of all-atom coordinates resulted in only small improvements in B factor agreement (Sen et al., 2006). The results once again demonstrate the robustness of the elastic network models, and suggest that the main factor in determining macromolecular flexibility is the number of local contacts, determined by the shape of the molecule (Halle, 2002). In prediction of directionality of motion, there is a clear difference between methods that are based only on $C\alpha$ coordinates (HCA and ANM) and those that consider all atoms. CHARMM-based BNM has the best directional agreement as measured by the overlap fraction, while our new method, DNM, and EIN come close to matching this standard. This suggests that an all-atom ENM potential can give an accurate representation of the conformational ensemble of a protein near the native state, but the inclusion of chemical forces improves the model.

We must also consider the cost, both computational and human, required by the different methods. One of the main differences between ENM techniques and BNM is that the latter requires an initial minimization step (see Experimental Procedures). If minimization is not complete, subsequent diagonalization will lead to spurious modes with large, negative frequencies; one must be careful to only pick productive modes when using results from BNM, whereas elastic network models are at a local minimum by construction. Further, the initial setup with an all-atom potential requires attention to the individual oddities of each structure: disulfide bonds, nonstandard residues, bound ligands, or cofactors. Each of these issues must be dealt with individually, thus making automation of the calculations more difficult. Compared with ENM models, in which most of these details are ignored, CHARMM-based normal modes require a great deal of human effort.

The present results indicate that anisotropic temperature factors from high-resolution crystal structures contain a measure of internal molecular flexibility, and can be used as a source of dynamic information and as a test for computational methodologies. Comparison of different methods indicates that elastic network models can describe the conformational ensemble of protein structures with accuracy approaching that of CHARMM, but that there is a substantial spread in prediction quality of different ENM potentials. Using an exclusively Cα-based potential results in a large sacrifice in prediction quality of directionality, but the lowest-frequency modes are robust across the methods. The information may help those studying interactions within biological molecules choose the appropriate level of complexity for the system of interest and for the level of detail required of the prediction.

EXPERIMENTAL PROCEDURES

We use normal mode analysis (Brooks and Karplus, 1983; Go et al., 1983; Levitt et al., 1985) to predict the positional ensemble of protein structures. The different models use distinct potentials, all of which require the knowledge of protein structure. The Hessian matrix of the potential is diagonalized to find the normal modes, or eigenvectors u_i and the corresponding frequencies ω_i: $Hu_i = \omega_i^2 u_i$. The decomposition allows us to compute the covariance matrix, which is proportional to the pseudoinverse of the Hessian. Let δ_i be the deviation from the mean for component i; then the covariance between two deviations is

$$\langle \delta_i \delta_j \rangle = \frac{1}{2k_B T} \sum_k \frac{1}{\omega_k^2} u_{ik} u_{jk},$$

where brackets denote mean value, and u_i is the ith component of the kth normal mode with frequency ω_k. Note that the modes with the lowest frequencies make the greatest contribution to residue mobility, so a small fraction of all the modes is sufficient to obtain a good approximation of the sum. This allows us to compute anisotropic variances as 3 × 3 blocks around the diagonal of the covariance matrix (Kidera and Go, 1990).

We may also compute the correlation coefficient between the deviations of any two atoms, to generate the global correlation matrix

$$R(\delta_i, \delta_j) = \frac{\langle \delta_i \delta_j \rangle}{\sqrt{\langle \delta_i^2 \rangle \langle \delta_j^2 \rangle}}.$$

Elastic Network Models

Anisotropic network model (ANM) (Atilgan et al., 2001) is a version of elastic network model (ENM), based on connecting residues with Cα atoms within a cutoff distance R_c with spring-like interactions. The Hessian matrix is a 3N × 3N matrix, where N is the number of residues, consisting of 3 × 3 submatrices H_{ij} which depend on the direction of the vector between Cα atoms i and j, and are 0 if the Cα atoms are more than R_c apart. The diagonal submatrices are defined as follows: $H_{ii} = -\sum_j H_{ij}$. This defines a coarse-grained elastic network model of a protein structure with directional information. We implemented this algorithm using perl code to read PDB files and construct the Hessian, with MATLAB (The Mathworks, Natick, MA) scripts used for diagonalization.

We introduce two modifications to ANM, analogous to those we had previously proposed for isotropic models (Kondrashov et al., 2006), and term the new model distance network model (DNM). First, the connectivity of the elastic potential is based on distances between nonhydrogen atoms of residue pairs, instead of only the Cα atoms. The contacts from all atoms are added for each residue to yield an

interaction potential at the residue level. Second, we introduce different classes of residue interactions based on interatomic distances, with distinct Hookean spring constants. We use distance bins to define the interaction classes, specifically, covalent interactions are found by distance less than 2.3 Å, the next shell is up to 3.3 Å, followed by 5, 7, 9, and 11 Å. The Hessian matrix for each bin is defined exactly as for ANM above, with the difference that the equilibrium distance between two atoms has to be in the distance bin, while the coordinates (x_i, y_i, z_i) for residue i remain the Cα coordinates. If H_a is the contact matrix for class a, the total Hessian matrix for DNM is a linear combination of the matrices, with k_a as the interaction constant for each class:

$$H_{total} = \sum_a k_a H_a = \sum_a \frac{1}{tr(H_a)} H_a.$$

The constants k_a define the strength of interactions, and we chose to use the total number of contacts in each class as a normalization constant, $k_a = 1/[tr(H_a)]$. Thus, although DNM introduces several different interaction constants, these are defined from the contact matrices, and thus are not free parameters to be optimized. The only free parameter, as in other ENMs, is the cutoff distance for atomic contacts, which we vary from 5 Å to 11 Å, as described in Results. The implementation again used a combination of perl and MATLAB scripts.

The details of the normal mode analysis implementation of the molecular modeling toolkit (MMTK) have been described elsewhere (Hinsen, 2000). For this study, we used the harmonic Cα force field (HCA) (Hinsen et al., 2000), which defines different interaction constants for covalently bonded and noncovalently bonded Cα atoms. The model uses the reciprocal of distance to weight the harmonic interaction constants, and no parameters are varied from the default values. The MMTK calculations for all 83 structures in the data set are carried out in 2 hr on a single 2 GHz AMD Athlon processor with 2 GB of RAM.

ElNemo (EIN) (Suhre and Sanejouand, 2004) is an all-atom ENM, which constructs a contact matrix for all atoms within a certain radius and then treats blocks of one or more residues as rigid bodies using the rotation-translation blocking algorithm (Tama et al., 2000). The two main programs that constitute ElNemo, pdbmat and diagrtb, were kindly provided by the authors and installed on the local cluster. All blocking was done on a residue-by-residue basis and the interaction cutoff distance was varied from 4 Å to 11 Å. Running EIN on all 83 structures in the data set using eight different cutoff distances took roughly 1 day to complete on a 100 node cluster of 2.2 GHz Apple G5 processors with 4 GB of RAM.

Block Normal Modes with CHARMM

Block normal mode analysis (BNM), originally suggested by Tama et al. (2000) and subsequently improved by Li and Cui (2002), computes an all-atom Hessian which is then projected onto a blocked space spanned by the rotational and vibrational degrees of freedom of predefined blocks; in this work each residue is treated as a rotation-translation block, as in the ElNemo method above. For this level of coarse graining, the procedure reduces the Hessian storage space by approximately a factor of 25 and the diagonalization time by a factor of 125. The resulting blocked eigenvectors were then projected back to the all-atom space to give all-atom eigenvectors. This procedure perturbs the magnitudes of the eigenvalues, but in a linear fashion for the low-frequency modes (Li and Cui, 2002; Tama et al., 2000). The appropriate scale factor of 1.7 has been used in this work. Local minimization is performed to ensure that the linear term in the Taylor expansion of the potential is zero (Hayward, 2001). This minimization is completed using cycles of the adapted-basis Newton-Raphson method with gradually decreasing harmonic constraints to remove local steric clashes without perturbing the structure significantly. A final minimization with no harmonic constraints is performed until the root-mean-squared energy gradient reaches 0.01 kcal/mol/Å. The average minimization time for this set is approximately 8 min, but minimization times vary widely because of protein size: 54 s for the 52 residue PDB ID code 1RB9, and

73 min for the 325 residue PDB ID code 1O7J, all computed on a 1.8 GHz Athlon single-processor station with 1 GB of memory, running Red Hat Linux 7.2. In some of the systems studied, this level of minimization resulted in modes with large negative frequencies in addition to the normal six rotational-translational modes. In these cases, these modes were ignored for all subsequent calculations. The average diagonalization time for BNM was approximately 6 min, varying widely again: 48 s for PDB ID code 1RB9, and 51 min for PDB ID code 1O7J. All calculations are completed using the CHARMM suite of programs (Brooks et al., 1983; Neria et al., 1996). The extended atom CHARMM19 force field (Neria et al., 1996), modified for use with the EEF1 solvation model (Lazaridis and Karplus, 1999), is used for both minimizations and the BNM.

Measures of Agreement with Crystallographic Data

The data set was obtained by searching the PDB (Berman et al., 2000) for protein structures determined by X-ray crystallography to at least 1.0 Å resolution and containing at least 50 residues in a single chain. Structures with more than 50% identity were discarded, leaving 98 nonredundant proteins, of which 87 contained ANISOU cards (anisotropic displacement parameters); 4 more structures were discarded because they contained modified protein residues for which CHARMM libraries are not available. The resultant set is structurally diverse, with all major SCOP superfamilies (Murzin et al., 1995) represented, as shown in Table S1. All protein chains in the PDB files were kept in the model in order to best represent the crystal environment. Copies of the protein molecule surrounding the structure in the crystal are generated using the symexp command in PyMOL (DeLano, 2002). Residues with at least one atom less than 4 Å from an atom in a crystal copy were considered to be involved in crystal contacts, and were excluded from the comparison set. Further, ADPs from $C\alpha$ atoms with multiple conformations determined by an occupancy parameter with a value other than 1 were also excluded.

The anisotropic parameters are 3 × 3 matrices that define the variance of a three-dimensional Gaussian probability distribution for the position of each atom:

$$\rho(x) = \sqrt{\frac{\det(U^{-1})}{8\pi^3}} \exp\left[x^t \begin{pmatrix} U_{xx} & U_{xy} & U_{xz} \\ U_{xy} & U_{yy} & U_{yz} \\ U_{xz} & U_{yz} & U_{zz} \end{pmatrix} x\right].$$

The six components of ADPs, U_{xx} and so on, are reported in the PDB files in ANISOU cards (Berman et al., 2000). We compare the computationally predicted anisotropic parameters V with those from the crystal structures U. Prediction quality of the magnitude of variation is measured by linear correlation of the traces of the matrices U and V over the whole structure, which we call isotropic correlation. To compare directions of ellipsoids, we first divide all the matrices by their trace, to set all magnitudes to unity. Ellipsoids are described by their principal axes (eigenvectors) and the associated lengths (inverse eigenvalues); the ratio of the smallest to the largest eigenvalue is called its anisotropy (Trueblood et al., 1996). Directionality comparison was restricted to ellipsoids with anisotropy of less than 0.5, as directional comparison of near-spherical ellipsoids is meaningless. The simplest comparison of directionality is the absolute value of the dot product between the major directions. It is a rough estimate of agreement for two ellipsoids whose major axes are dominant, but has the virtue of simplicity. A more systematic measure of ellipsoid similarity was proposed by Merritt (1999), based on computation of the overlap integral between two probability densities. This measure, known to crystallographers as the real-space correlation coefficient, is defined for two three-dimensional Gaussian distributions with covariance matrices U and V as follows:

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

We also compare modes produced by the different normal mode potentials. To compare mode i (as ordered by frequency) from two methods, we take the average between the best agreement for mode i from method a with modes from method b, and the best agreement for mode i from method b with the modes from method a. We compare the modes similar in frequency ordering, specifically, only the modes no more than three indices higher or lower. The formula for overlap for mode i between methods a and b is

$$O_{a,b}(i) = \frac{1}{2} \max_{i-3 \leq j \leq i+3} u_i^a \bullet u_j^b + \frac{1}{2} \max_{i-3 \leq j \leq i+3} u_i^b \bullet u_j^a.$$

If the best agreement is between modes of the same index, then the two maxima are the same. Figures 2–4 were prepared using MATLAB (The Mathworks) and Figure 1 with Rastep and Raster3D (Merritt and Bacon, 1997).

Supplemental Data

Supplemental Data include eight tables and can be found with this article online at http://www.structure.org/cgi/content/full/15/2/169/DC1/.

REFERENCES

Abbondanzieri, E.A., Greenleaf, W.J., Shaevitz, J.W., Landick, R., and Block, S.M. (2005). Direct observation of base-pair stepping by RNA polymerase. Nature *438*, 460–465.

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys. J. *80*, 505–515.

Bahar, I., and Rader, A.J. (2005). Coarse-grained normal mode analysis in structural biology. Curr. Opin. Struct. Biol. *15*, 586–592.

Bahar, I., Atilgan, A.R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold. Des. *2*, 173–181.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Brooks, B., and Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. USA *80*, 6571–6575.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983). CHARMM—a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. *4*, 187–217.

CCP4 (Collaborative Computational Project, Number 4) (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D Biol. Crystallogr. *50*, 760–763.

Clarage, J.B., Clarage, M.S., Phillips, W.C., Sweet, R.M., and Caspar, D.L. (1992). Correlations of atomic movements in lysozyme crystals. Proteins *12*, 145–157.

Cui, Q., Li, G.H., Ma, J.P., and Karplus, M. (2004). A normal mode analysis of structural plasticity in the biomolecular motor F-1-ATPase. J. Mol. Biol. *340*, 345–372.

Dauter, Z., Wilson, K.S., Sieker, L.C., Meyer, J., and Moulis, J.M. (1997). Atomic resolution (0.94 Å) structure of *Clostridium acidurici* ferredoxin. Detailed geometry of [4Fe-4S] clusters in a protein. Biochemistry *36*, 16065–16073.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System (http://www.pymol.org).

DePristo, M.A., de Bakker, P.I., and Blundell, T.L. (2004). Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure *12*, 831–838.

Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G., Zagari, A., and Mazzarella, L. (2000). The ultrahigh resolution crystal structure of ribonuclease A containing an isoaspartyl residue: hydration and stereochemical analysis. J. Mol. Biol. *297*, 713–732.

Eyal, E., Yang, L.W., and Bahar, I. (2006). Anisotropic network model: systematic evaluation and a new web interface. Bioinformatics *22*, 2619–2627.

Frauenfelder, H., Petsko, G.A., and Tsernoglou, D. (1979). Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. Nature *280*, 558–563.

Go, N., Noguti, T., and Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc. Natl. Acad. Sci. USA *80*, 3696–3700.

Halle, B. (2002). Flexibility and packing in proteins. Proc. Natl. Acad. Sci. USA *99*, 1274–1279.

Hamacher, K., and McCammon, J.A. (2006). Computing the amino acid specificity of fluctuations in biomolecular systems. J. Chem. Theory Comput. *2*, 873–878.

Hayward, S. (2001). Normal mode analysis of biological molecules. In Computational Biochemistry and Biophysics, O.M. Becker, A.D. MacKerell, Jr., B. Roux, and M. Watanabe, eds. (New York: Marcel Dekker), pp. 153–167.

Hinsen, K. (2000). The molecular modeling toolkit: a new approach to molecular simulations. J. Comput. Chem. *21*, 79–85.

Hinsen, K., Petrescu, A.-J., Dellerue, S., Bellissent-Funel, M.-C., and Kneller, G.R. (2000). Harmonicity in slow protein dynamics. Chem. Phys. *261*, 25–37.

Karplus, M., and Kuriyan, J. (2005). Molecular dynamics and protein function. Proc. Natl. Acad. Sci. USA *102*, 6679–6685.

Kidera, A., and Go, N. (1990). Refinement of protein dynamic structure: normal mode refinement. Proc. Natl. Acad. Sci. USA *87*, 3718–3722.

Kondrashov, D.A., Cui, Q., and Phillips, G.N., Jr. (2006). Optimization and evaluation of a coarse-grained model of protein motion using X-ray crystal data. Biophys. J. *91*, 2760–2767.

Kundu, S., Melton, J.S., Sorensen, D.C., and Phillips, G.N., Jr. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys. J. *83*, 723–732.

Kuriyan, J., and Weis, W.I. (1991). Rigid protein motion as a model for crystallographic temperature factors. Proc. Natl. Acad. Sci. USA *88*, 2773–2777.

Kuriyan, J., Petsko, G.A., Levy, R.M., and Karplus, M. (1986). Effect of anisotropy and anharmonicity on protein crystallographic refinement—an evaluation by molecular-dynamics. J. Mol. Biol. *190*, 227–254.

Lanman, J., and Prevelige, P.E., Jr. (2004). High-sensitivity mass spectrometry for imaging subunit interactions: hydrogen/deuterium exchange. Curr. Opin. Struct. Biol. *14*, 181–188.

Lazaridis, T., and Karplus, M. (1999). Effective energy function for proteins in solution. Proteins *35*, 133–152.

Levitt, M., Sander, C., and Stern, P.S. (1985). Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. J. Mol. Biol. *181*, 423–447.

Li, G., and Cui, Q. (2002). A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca^{2+}-ATPase. Biophys. J. *83*, 2457–2474.

Lindorff-Larsen, K., Best, R.B., DePristo, M.A., Dobson, C.M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. Nature *433*, 128–132.

Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A., and Cambillau, C. (1997). Atomic resolution (1.0 angstrom) crystal structure of *Fusarium solani* cutinase: stereochemical analysis. J. Mol. Biol. *268*, 779–799.

Ma, J.P., and Karplus, M. (1997). Ligand-induced conformational changes in ras p21: a normal mode and energy minimization analysis. J. Mol. Biol. *274*, 114–131.

Meinhold, L., and Smith, J.C. (2005). Fluctuations and correlations in crystalline protein dynamics: a simulation analysis of staphylococcal nuclease. Biophys. J. *88*, 2554–2563.

Merritt, E.A. (1999). Comparing anisotropic displacement parameters in protein structures. Acta Crystallogr. D Biol. Crystallogr. *55*, 1997–2004.

Merritt, E.A., and Bacon, D.J. (1997). Raster3D photorealistic molecular graphics. Methods Enzymol. *277*, 505–524.

Moffat, K. (2001). Time-resolved biochemical crystallography: a mechanistic perspective. Chem. Rev. *101*, 1569–1581.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Neria, E., Fischer, S., and Karplus, M. (1996). Simulation of activation free energies in molecular systems. J. Chem. Phys. *105*, 1902–1921.

Palmer, A.G., Kroenke, C.D., and Loria, J.P. (2001). Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. In Nuclear Magnetic Resonance of Biological Macromolecules, Part B (San Diego: Academic Press), pp. 204–238.

Phillips, G.N., Jr. (1990). Comparison of the dynamics of myoglobin in different crystal forms. Biophys. J. *57*, 381–383.

Sen, T.Z., Feng, Y.P., Garcia, J.V., Kloczkowski, A., and Jernigan, R.L. (2006). The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models. J. Chem. Theory Comput. *2*, 696–704.

Sheldrick, G.M., and Schneider, T.R. (1997). SHELXL: high-resolution refinement. Methods Enzymol. *277*, 319–343.

Suhre, K., and Sanejouand, Y.H. (2004). ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res. *32*, W610–W614.

Tama, F., Gadea, F.X., Marques, O., and Sanejouand, Y.H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. Proteins *41*, 1–7.

Teodoro, M.L., Phillips, G.N., Jr., and Kavraki, L.E. (2003). Understanding protein flexibility through dimensionality reduction. J. Comput. Biol. *10*, 617–634.

Tirion, M.M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys. Rev. Lett. *77*, 1905–1908.

Trueblood, K.N., Burgi, H.B., Burzlaff, H., Dunitz, J.D., Gramaccioli, C.M., Schulz, H.H., Shmueli, U., and Abrahams, S.C. (1996). Atomic displacement parameter nomenclature—report of a subcommittee on atomic displacement parameter nomenclature. Acta Crystallogr. A *52*, 770–781.

Van Wynsberghe, A., Li, G.H., and Cui, Q. (2004). Normal-mode analysis suggests protein flexibility modulation throughout RNA polymerase's functional cycle. Biochemistry *43*, 13083–13096.

Wilson, M.A., and Brunger, A.T. (2000). The 1.0 Å crystal structure of Ca^{2+}-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. J. Mol. Biol. *301*, 1237–1256.