Effects of Surface Water on Protein Dynamics Studied by a Novel Coarse-Grained Normal Mode Approach

Lei Zhou* and Steven A. Siegelbaum*†

*Department of Neuroscience, and †Department of Pharmacology, Howard Hughes Medical Institute, Columbia University,
New York, New York

ABSTRACT Normal mode analysis (NMA) has received much attention as a direct approach to extract the collective motions of macromolecules. However, the stringent requirement of computational resources by classical all-atom NMA limits the size of the macromolecules to which the method is normally applied. We implemented a novel coarse-grained normal mode approach based on partitioning the all-atom Hessian matrix into relevant and nonrelevant parts. It is interesting to note that, using classical all-atom NMA results as a reference, we found that this method generates more accurate results than do other coarse-grained approaches, including elastic network model and block normal mode approaches. Moreover, this new method is effective in incorporating the energetic contributions from the nonrelevant atoms, including surface water molecules, into the coarse-grained protein motions. The importance of such improvements is demonstrated by the effect of surface water to shift vibrational modes to higher frequencies and by an increase in overlap of the coarse-grained eigenvector space (the motion directions) with that obtained from molecular dynamics simulations of solvated protein in a water box. These results not only confirm the quality of our method but also point out the importance of incorporating surface structural water in studying protein dynamics.

INTRODUCTION

One major goal of studies of protein structure-function relationships is to identify their macroscopic correlated motions, and how these motions change in response to various external perturbations, such as ligand-binding. A variety of experimental approaches, including x-ray crystallography, NMR spectroscopy, and single-molecule biophysical techniques, have provided insights into macroscopic protein motions by monitoring the structural alterations of the same protein under different conditions. On the other hand, theoretical studies, such as molecular dynamics (MD) simulations and normal mode analysis (NMA), can also provide valuable information about internal protein motions (1,2).

Standard MD simulations sample the conformational space of a protein using the definitions for atomic interactions from various force fields and usually include explicitly treated water to reproduce solvent effects (3,4). Correlated protein motions can then be extracted from the MD simulations through diagonalizing the covariance matrix obtained from a section of the MD trajectory. This is also referred to as essential dynamics (5), principal component analysis (PCA) (6), or quasiharmonic analysis (7,8), due to the complex and anharmonic nature of protein dynamics. However, the size of the system, especially with explicitly treated water molecules, has provided a great computational challenge, generally limiting the timescale of MD simulations for large macromolecules to the nanosecond range, significantly shorter than the biologically relevant timescale of conformational changes that may require milliseconds or longer. Therefore, inefficient sampling is still a significant obstacle to extracting meaningful correlated motions from MD simulations (9,10).

Classical all-atom normal mode analysis (AANM) offers the ability to overcome some of the computational cost of MD simulations. AANM makes the simplifying assumption that protein motions can be described by harmonic motions around a local minimum on the protein energy surface. Starting with an initial protein structure, standard AANM requires an extensive minimization of the system's potential energy followed by the calculation of the Hessian matrix, whose $3N \times 3N$ ($N =$ the number of atoms) elements represent the second derivative of the potential energy function along the Cartesian coordinates. Diagonalization of the mass-weighted Hessian matrix can then be used to generate the eigenvectors and eigenvalues of the matrix, which provide, respectively, information about the directions of the various correlated motions within the protein and their amplitudes at a given frequency (11–14).

However, the application of AANM to biological macromolecules has been limited by the requirements of physical memory to store the all-atom Hessian matrix and the significant CPU time to diagonalize the very large matrix. Therefore, in practice, AANM is normally applied to protein systems containing at most a few hundred residues, in most cases without explicitly treated water molecules. However, since solvent has important and complex interactions with the solute molecule, the explicit treatment of solvents is thought to be essential to faithfully reproduce protein dynamics. For MD simulations, it has been a standard practice to simulate a protein molecule in a box filled with explicitly treated water

Submitted June 22, 2007, and accepted for publication December 28, 2007.

Address reprint requests to Lei Zhou, Department of Neuroscience, Columbia University, 1051 Riverside Drive, New York, NY 10032. Tel.: 212-543-5584; Fax: 212-795-7997; E-mail: lz150@columbia.edu.

Editor: Angel E. Garcia.

0006-3495/08/05/3461/14 $2.00

doi: 10.1529/biophysj.107.115956

molecules and use periodic boundary conditions. However, performing AANM on such a system to extract protein motions is usually beyond the capabilities of currently available computational hardware and software.

To date, there have been only a few published AANM studies involving explicitly treated water molecules within a distance of several Angstroms of the protein surface (15–17). These studies showed that incorporating surface water is helpful to reproduce experimental observations, including the *B*-factors determined by x-ray crystallography. However, given the scarcity of these studies, novel techniques, with the ability to efficiently incorporate solvent effects and provide a complete survey of the vibrational spectrum, are still needed to improve the efficiency of AANM for large systems.

Technically, even though the storage of a Hessian matrix has become less of an obstacle due to the introduction of sparse matrix techniques, the diagonalization of the all-atom matrix is still a challenge and new algorithms are being continuously added to various linear algebra packages. These include the method of diagonalization in a mixed basis (18,19) implemented in CHARMM (20) and the iterative Lanczos/Arnoldi factorization method (21) implemented in GROMACS (22), two widely used simulation packages. Nevertheless, these iterative numerical methods are still very time consuming and can only yield a small fraction of the total eigenvectors, usually those corresponding to the lowest vibrational frequencies.

Fortunately, the low-frequency vibrational modes are closely related to large-amplitude correlated protein motions with minimum energy costs, which usually reflect the conformational changes relevant to protein function (1,23). Indeed, the collective motions represented by these eigenvectors are in good agreement with independent experimental measurements (24–26). However, pinpointing the most functionally relevant individual mode is not a trivial task. In addition, it has been suggested that a combination of modes is required for a reasonable mapping of the correlated motions (1,13,27). Moreover, recent studies showed that the modes of higher frequencies are also important, because energy input from external perturbations can shift the distribution of different modes to higher frequencies (13,28). Thus, a complete survey of the eigenvector space and corresponding eigenvalues is important for various theoretical applications, such as the calculation of thermodynamic configuration entropy and heat capacity.

As an alternative to classical AANM, coarse-grained approaches have been pursued to reduce the size of the system and improve computational efficiency (14,29–31). The block normal mode method (BNM) is an effective coarse-grained NMA approach that treats proteins as a system of rigid blocks (32–34). However, BNM still relies on a complex all-atom representation and starts from the same all-atom Hessian matrix as AANM. An important breakthrough came with the introduction of the elastic network model (ENM), which simplifies the complex atomic interactions to potential energy functions with only a single parameter (1 kcal/mol/Å^2) for C-α atoms, thus bypassing the time-consuming energy minimization steps (35,36). ENM (or isotropic Gaussian network model) reflects the intrinsic protein dynamics embedded in the overall molecular topology and effectively reproduces certain aspects of the atomic fluctuations determined by NMR and x-ray crystallography (37–39). The corresponding model used in NMA is referred to as the anisotropic network model (ANM) (40). Despite the dramatic simplifications, ENM is widely applied to large macromolecules and assemblies beyond the reach of traditional methods (41–45).

However, there is a trade-off between accuracy and speed in these coarse-grained methods. Much effort has gone into comparing results from these approximate methods with the results of classical AANM, the parent method, or the results of MD simulations, as a reference (32,33,42,46). Based on the degree of similarity between the low-frequency eigenvectors of AANM and the corresponding eigenvectors of coarse-grained methods, BNM has been found to produce more accurate results than ENM (32,33,45). This is not surprising because BNM starts from an extensively energy-minimized system described by an all-atom force field and then projects the all-atom Hessian matrix to the space of predefined blocks. In the limit, this method allocates only one residue in each block, providing the highest possible resolution in the implementation of the BNM method (however, at the greatest computational cost). Such an approach is implemented in the most recent version of CHARMM. Nonetheless, even BNM results show significant deviations from the AANM approach. Moreover, no coarse-grained method is able to incorporate the contributions from explicitly treated water molecules.

Here we implemented a novel coarse-grained normal mode method (CGNM) based on a partition scheme of the all-atom Hessian matrix to extract the correlated motions in the subspace of C-α atoms. We carried out our initial analysis on the 120-amino-acid cyclic nucleotide-binding domain (CNBD) and adjacent upstream 90-amino-acid cytoplasmic C-linker region from the HCN2 hyperpolarization-activated cyclic nucleotide-regulated cation channel (47). High resolution x-ray crystallography has shown that in the presence of cyclic nucleotides, this isolated soluble protein domain forms a fourfold symmetric tetrameric assembly with one cyclic nucleotide bound in the CNBD of each of the four subunits (48).

In this study, we report that CGNM provides a more accurate description of the motions of the HCN2 CNBD, as well as that of four other proteins, compared to two other coarse-grained methods, ENM and BNM, based on the degree of similarity of the results from the three coarse-grained approaches with the results of a full AANM analysis. It is important to note that we found that CGNM also allowed us to incorporate explicitly treated surface water molecules into protein motions projected in the subspace of the relevant atoms (C-α atoms in this study). Furthermore, a comparison of our CGNM results containing a layer of surface water with

MD results on the same protein in a water-filled box demonstrates the importance of incorporating such surface structural water in studying protein dynamics.

METHODS

Classical AANM and coarse-grained CGNM analyses

We used a representative snapshot from MD simulations based on the x-ray crystal structure of the HCN2 channel C-terminus protein (PDB ID 1Q5O) (48) as the starting structure for the NMA described here. Briefly, the MD simulations, which we described in detail elsewhere (49), were performed as follows. We used the GROMOS96 force field from the GROMACS package (22,50). The whole system contains four subunits and each subunit contains 8636 protein atoms, 4 cAMP molecules, 23,654 water molecules, and 12 chloride ions to balance the charge in the system. For the bound ligand, cAMP, we used the topology generated by the PRODRG server and the partial charges defined in the GROMOS96 force field (50). We used the flexible SP3 water model in the simulation (51). The distance between the protein and each side of the rectangle box was set at 10 Å. The particle mesh Ewald method (52), with a cutoff distance of 10 Å, was used for the electrostatic potential energy. Before MD simulations, we applied basic energy minimization steps (steepest descent (SD) and conjugate gradient (CG)) to optimize the starting system and remove any nonphysical contacts. During the first 500 ps of the MD simulation, the positions of the heavy atoms in the protein were fixed so that the system, especially the explicit water molecules, can be further optimized. After these steps, we carried out a normal MD simulation with a time step of 2 fs and collected the trajectory every 0.5 ps.

To carry out the normal mode calculations based on the all-atom force fields, we first performed an extensive energy minimization to ensure that the starting structure represents a local minimum on the energy surface. To achieve this, we applied SD and CG followed by the limited-memory Broyden-Fletcher-Goldfarb-Shanno method (22) at double precision numerical accuracy to the representative snapshot structure from the MD simulations obtained above. During these energy minimization steps, the electrostatic energy was described by a switch function with the distance for normal treatment set at 15 Å and the cut-off distance set at 18 Å (53).

The key step in the analysis was then to calculate the Hessian matrix for the entire system, containing the second derivatives of the potential energy functions ($\partial^2 \mathbf{V}/\partial x_i \partial x_j$). The matrix was then partitioned into four sections to extract the C-α components according to the equation

$$H_{all} = \begin{vmatrix} H_{xx} & H_{xy} \\ H_{yx} & H_{yy} \end{vmatrix}; \tag{1}$$

$$H'_{xx} = H_{xx} - H_{xy} \times H_{yy}^{-1} \times H_{yx}. \tag{2}$$

Here, H_{xx}, H_{yy}, $H_{xy,}$ and H_{yx} submatrices contain the elements representing the interactions of, respectively, relevant to relevant atoms, nonrelevant to nonrelevant atoms, relevant to nonrelevant atoms, and nonrelevant to relevant atoms. In the CGNM method, the energetic contributions of all interactions with and between nonrelevant atoms (the non-C-α atoms here) are incorporated into a simplified Hessian matrix for the relevant-atom subspace, H'_{xx}, using Eq. 2. The theoretical basis for deriving the C-α atom motions based on the atomic fluctuations from classical AANM was published by Berendsen and colleagues (54). A similar equation was used to extract the effective force constant matrix for C-α atoms (55) and discussed in the GROMACS discussion board (www.gromacs.org) in 2005 for the purpose of comparing AANM and MD-based PCA in the C-α subspace. Moreover, a recent study by Eom et al. (56) used a very similar method to obtain a coarse-grained approximation to ENM. After basic matrix manipulations, we found that our partitioning approach (Eqs. 1 and 2) is identical to that of Eom et al. The major difference between our study and that of Eom et al. is that we have applied a coarse-grained approximation to classical NMA based on all-atom force fields, whereas Eom et al. aimed to improve the efficiency of ENM (which they referred to as the Gaussian network model).

To sort the Hessian into relevant and nonrelevant parts, we first converted the sparsely-stored mass-weighted Hessian matrix into a double precision ASCII file. We then generated an index file in which the indices for all C-α atoms (relevant atoms) were arranged at the beginning followed by non-C-α atoms. Each entry in the sparse Hessian matrix was read into the program and allocated to a new position, using the index file as a key for sorting. The C-α component (xx part) was stored in a dense matrix format. The symmetrical xy and yx parts were stored in a coordinate format for a sparse matrix. Non-C-α components (yy) were stored in a row-major format for a sparse matrix. We used a direct solving routine from the PARDISO package (57) and standard LAPACK and BLAS routines for matrix calculations.

Based on the eigenvectors and the corresponding eigenvalues, the following equation was used to calculate the mean-square fluctuation (MSF) (Å^2):

$$\langle \Delta X_k^2 \rangle = \frac{k_B T}{m_k} \cdot \sum_i \frac{Y_{ki}^2}{\varpi_i^2}, \tag{3}$$

where k is the atom index, i is the eigenvector index, m_k is the atom mass, and ω is vibrational frequency.

The following equation was used to calculate the configurational entropy based on the eigenvalues from ENM, CGNM, or PCA (58,59):

$$\begin{aligned} S_{vib} &= k_B \cdot \sum_{i=7}^{3N} \left(\frac{\alpha}{e^{\alpha} - 1} - \ln\left(1 - \frac{1}{e^{\alpha}}\right) \right) \\ \alpha &= \frac{h \cdot \varpi_i}{2\pi \cdot k_B T}, \end{aligned} \tag{4}$$

where h is Planck's constant, k_B is Boltzmann's constant, and ω is the vibrational frequency.

Normal mode analysis based on elastic network model (ENM)

C-α atom coordinates from the energy minimized structures were directly used in the NMA based on the potential energy function defined by the elastic network model (ENM or anisotropic network model (ANM)) (40). For ENM, we used the default settings of the force constant (1 kcal/mol/Å^2) and cutoff distance (13 Å).

Block normal mode analysis

The same all-atom Hessian matrix was projected onto a subspace of rigid blocks, each of which contained a single residue for the protein or a single cAMP molecule for the bound ligand, to pursue the highest resolution possible with this method. The degrees of freedom equal six times the number of blocks. The Fortran code of DIAGRTB (v2.52) was used in this research with a modification of the size of the array, LRWORK, from 32,000,000 to 200,000,000, so that larger systems could be accommodated (32,33).

PCA based on MD simulations

We used g_covar from GROMACS to perform PCA on a section of the MD trajectory. Overall rotational and translational motions were removed by fitting the protein structure of each time frame to a reference structure (starting frame). For the MD simulations at low temperatures, we reduced the system temperature with a simulated annealing protocol and then collected the MD trajectories after a 200-ps equilibration at the corresponding temperature. For each PCA, we used a 2-ns-long MD trajectory containing 4000 frames. The eigenvalue outputs from the PCA analysis represent the vibrational amplitude and were converted into the square of angular velocity by the equation

$$\varpi_{\mathrm{i}}^2 = \frac{k_{\mathrm{B}}T}{\langle \Delta X_{\mathrm{i}}^2 \rangle \cdot m_{\mathrm{k}}}. \tag{5}$$

The anharmonic factor for each eigenvector from PCA was calculated by the equation

$$\alpha_{\mathrm{i}}^2 = \frac{\langle \Delta X_{\mathrm{i}}^2 \rangle}{\langle \Delta X_{\mathrm{i}}^2 \rangle^{\mathrm{har}}} = \left(\frac{\varpi_{\mathrm{i}}^{\mathrm{har}}}{\varpi_{\mathrm{i}}}\right)^2 = \frac{\dfrac{1}{\varpi_{\mathrm{i}}^2}}{\sum\limits_{\mathrm{j}=1}^{3\mathrm{N}-6} \dfrac{M_{\mathrm{ij}} \cdot M_{\mathrm{ij}}}{\varpi_{\mathrm{j}}^2}}, \tag{6}$$

where i is the index for PCA eigenvectors, j is the index for NMA eigenvectors, $(\Delta X_{\mathrm{i}}^2)^{\mathrm{har}}$ is the harmonic mean-square fluctuation as described by the NMA eigenvectors, and M_{ij} is the dot product between the ith PCA eigenvector and jth NMA eigenvector (60).

Alignment of correlated coordinate systems

The reference structures from two eigenvector systems were first aligned to the mass center of the molecule. A rotation matrix was then calculated based on two aligned reference structures. The second set of eigenvectors was rotated by the equation

$$V'_{\mathrm{klm}} = \sum_{\mathrm{n}=1}^{3} R_{\mathrm{mn}} \cdot V_{\mathrm{kln}}, \tag{7}$$

where k is the eigenvector index, l is the atom index, and m and n are the indices of xyz dimension (12). The following parameters for the overlap analysis, including dot product (Eq. 8), spanning coefficient (Eq. 9) (33,46), and cumulative overlap factor (COF) (Eq. 10) (54), were calculated based on these aligned eigenvector sets.

$$|M_{\mathrm{ij}}| = \left| \sum_{\mathrm{m}=1}^{\mathrm{N}} \sum_{\mathrm{n}=1}^{3} V1_{\mathrm{imn}} \cdot V2_{\mathrm{jmn}} \right|; \tag{8}$$

$$\mathrm{SPAN}_{\mathrm{i} \leq 100} = \sum_{\mathrm{j}=1}^{100} (M_{\mathrm{ij}})^2; \tag{9}$$

$$\mathrm{COF}_{\mathrm{X}} = \frac{\sum\limits_{\mathrm{k}=1}^{\mathrm{X}} \sum\limits_{\mathrm{i}=1}^{\mathrm{k}} \sum\limits_{\mathrm{j}=1}^{\mathrm{k}} M_{\mathrm{ij}} \cdot M_{\mathrm{ij}}}{X}. \tag{10}$$

Unit conversion among different simulation packages

AANM, CGNM, and BNM use the same mass-weighted Hessian matrix; therefore, the corresponding orthogonal eigenvector output should still be mass-weighted. However, the default output of eigenvectors is not mass-weighted in the GROMACS program and not strictly orthogonal. We modified the source code of GROMACS to generate mass-weighted orthogonal eigenvectors for AANM analysis. We converted the GROMACS eigenvalues (ω_{G}^2, based on the mass-weighted Hessian matrix, in units of kJ/mol/nm^2/amu) into the square of angular velocity (in units of s^{-2}) by multiplying the Gromacs eigenvalues by the conversion factor of 10^{-24}, based on the relation

$$\begin{aligned} &\frac{\mathrm{kJ}}{\mathrm{mol} \cdot \mathrm{nm}^2 \cdot \mathrm{amu}} \\ &= \frac{10^3 \times \mathrm{J}}{6.022 \times 10^{23} \times 10^{-18} \times \mathrm{m}^2 \times 1.66 \times 10^{-27} \times \mathrm{kg}} \\ &= 10^{24} \times \mathrm{s}^{-2}. \end{aligned} \tag{11}$$

The following equation was used for calculating the MSF (Å^2):

$$\begin{aligned} \langle \Delta X_{\mathrm{k}}^2 \rangle &= k_{\mathrm{B}}T \cdot \sum_{\mathrm{i}} \frac{Y_{\mathrm{ki}}^2}{m_{\mathrm{k}} \cdot \varpi^2} = \frac{1.38 \times 10^{-23} \mathrm{J} \times k^{-1} \times 300k}{13 \times 1.66 \times 10^{-27} \mathrm{kg}} \cdot \sum_{\mathrm{i}} \frac{Y_{ki}^2}{\varpi^2} \\ &= \frac{19.18 \times 10^4 \times \mathrm{m}^2 \times \mathrm{s}^{-2}}{10^{24} \mathrm{s}^{-2}} \cdot \sum_{\mathrm{i}} \frac{Y_{\mathrm{ki}}^2}{\varpi_{\mathrm{G}}^2}, \\ &= 19.18 \times 10^{-20} \times \mathrm{m}^2 \cdot \sum_{\mathrm{i}} \frac{Y_{\mathrm{ki}}^2}{\varpi_{\mathrm{G}}^2} = 19.18 \times \mathring{\mathrm{A}}^2 \cdot \sum_{\mathrm{i}} \frac{Y_{\mathrm{ki}}^2}{\varpi_{\mathrm{G}}^2} \end{aligned} \tag{12}$$

where ω_{G}^2 is the eigenvalue of the Gromacs unit, k is the atom index, and i is the eigenvector index.

Since we compared the results of different methods in the subspace of C-α atoms, mass-weighting will not affect the eigenvector results of ENM. However, a mass factor is needed for the calculation of vibrational frequencies and atomic fluctuation amplitudes. To our knowledge, there is no standard method for converting the units to compare the ENM results directly with other calculations (e.g., AANM, BNM, etc.) without scaling. Here, we tentatively added a mass factor corresponding to the mass of C-α atom so that the angular velocity in units of s^{-1} and MSF in units of Å^2 can be generated using a force constant of 1 kcal/mol/Å^2. The eigenvalues were converted into the square of the angular velocity by multiplying by the factor

$$\begin{aligned} &\frac{\mathrm{kcal}}{\mathrm{mol} \cdot \mathring{\mathrm{A}}^2 \times 13 \times \mathrm{amu}} \\ &= \frac{10^3 \times 4.184 \times \mathrm{J}}{6.022 \times 10^{23} \times 10^{-20} \times \mathrm{m}^2 \times 13 \times 1.66 \times 10^{-27} \mathrm{kg}} \\ &= \frac{0.6948 \times \mathrm{kg} \times \mathrm{s}^{-2}}{13 \times 1.66 \times 10^{-27} \mathrm{kg}} = 3.22 \times 10^{25} \times \mathrm{s}^{-2}. \end{aligned} \tag{13}$$

The eigenvalue output from PCA analysis (default GROMACS in units of nm^2; no mass weighting; C-α only) was converted into the square of angular velocity by the equation

$$\begin{aligned} \varpi^2 &= \frac{k_{\mathrm{B}}\mathrm{T}}{\langle \Delta X^2 \rangle \times \mathrm{mass}} = \frac{1.38 \times 10^{-23} \mathrm{J} \times \mathrm{k}^{-1} \times 300 \times \mathrm{k}}{\mathrm{nm}^2 \times 13 \times \mathrm{amu}} \\ &= \frac{4.14 \times 10^{-21} \times \mathrm{J}}{10^{-18} \times \mathrm{m}^2 \times 13 \times 1.66 \times 10^{-27} \times \mathrm{kg}} \\ &= 0.192 \times 10^{24} \times \mathrm{s}^{-2}. \end{aligned} \tag{14}$$

The experimental B-factor obtained through x-ray crystallography can be directly converted to atomic fluctuation (MSF, in Å^2) using the equation (61)

$$\mathrm{MSF} = \frac{3}{8 \times \pi^2} \mathrm{B}_{\mathrm{factor}}. \tag{15}$$

RESULTS

Comparison of AANM with coarse-grained ENM, BNM, and CGNM approaches

Our goal in this study was to develop a coarse-grained approximation to classical all-atom normal mode (AANM) analysis (11,12). We have implemented a novel coarse-grained normal mode analysis (CGNM) that decreases the computational cost associated with AANM by partitioning the all-atom Hessian matrix containing the second derivative of the potential energy function into relevant and nonrelevant components, here focusing on the C-α atoms (see Methods, Eq. 2). To assess the accuracy of our method, we first compared the results of AANM, the standard for these comparisons, with those of CGNM, as well as with results from two other coarse-grained approaches, ENM and BNM (32,40). As ENM and BNM treat proteins in a vacuum in the absence of water, we first compared the four methods under these dehydrated conditions. In the following section, we consider the effects on CGNM results of adding surface water.

As NMA is based on a harmonic approximation of the protein energy surface near an ideally global minimum, it first requires an extensive minimization of the potential energy of the starting protein structure. Here we used a representative structure of the HCN2 C-terminus protein obtained from a 20-ns-long MD trajectory based on the original crystal structure (48). This procedure allows the protein structure to be efficiently equilibrated in the same force field used by subsequent NMA (GROMOS96) (50), as reasonably long MD simulations should optimize the loop conformations and allow for small-scale rearrangement of secondary structures (62,63).

We first removed all water molecules from the representative MD snapshot structure. After an extensive minimization of the system, the final structure containing only the protein and cAMP atoms was used to generate the all-atom Hessian matrix, which was then iteratively diagonalized to produce the AANM result, providing the reference for comparison with the coarse-grained methods. Due to the limitation of computational resources, only a small fraction of the total eigenvectors and corresponding eigenvalues were calculated (2000, or 8% of 26,232). All technical details are given in Methods and Table 1. Briefly, ENM starts from the C-α atom coordinates and generates a complete set of orthogonal eigenvectors. BNM and CGNM methods started from the same all-atom Hessian matrix used by AANM. Whereas BNM simplifies the calculation through projecting the all-atom Hessian matrix into predefined rigid blocks, CGNM relies on a matrix-partitioning scheme to integrate the energetic contributions from non-C-α atoms into the motions of C-α atoms. Since the eigenvector outputs of BNM are in the all-atom space, they were projected to the C-α atom subspace for comparison purposes. This was followed by a normalization step that makes each eigenvector unitary ($V_i \cdot V_i^T = 1$) but not strictly orthogonal ($V_i \cdot V_j = 0$, i $\neq$ j). The eigenvector outputs of CGNM are naturally orthogonal in the C-α subspace and thus were directly used in the overlap analysis.

The results of ENM, BNM, and CGNM were compared to the results of AANM in terms of the overlap of the resulting eigenvectors, representing the direction of correlated motion, and eigenvalues, representing the amplitude or the frequency of each motion. Three different methods were used to check the overlap between the eigenvectors from AANM versus a given coarse-grained method. First, a direct view of overlap was obtained from a plot of the inner product between each pair of eigenvectors (Eq. 8). Such plots confirm previous studies that BNM generates results closer to those of AANM than does ENM; this is shown by the tighter clustering of points near the ideal diagonal relationship for the BNM versus AANM plot (33,46) (Fig. 1, *A* and *B*). It is important to note that CGNM provides an even better match (tighter diagonal clustering) with the AANM results than does BNM (Fig. 1 *C*). Second, we quantified the overlap between two sets of eigenvectors using the spanning coefficient (Eq. 9), representing the overlap between each AANM eigenvector with a group of eigenvectors from each coarse-grained analysis (33,45,54). The nearly straight line of the spanning coefficient curve of CGNM up to a frequency of 10 cm^{-1} indicated that the 70 or so lowest-frequency AANM eigenvectors can be almost completely mapped by the first 100 eigenvectors of CGNM (Fig. 2 *A*). However, this close mapping only extends as far as the first ~10 or ~15 AANM eigenvectors for ENM or BNM, respectively (Fig. 2 *A*).

A potential bias of using spanning coefficients is that an arbitrary number (100 here) of eigenvectors needs to be predefined, because the spanning coefficient involving all eigenvectors is theoretically equal to 1. This makes the spanning coefficient less meaningful when comparing systems of different dimensions of freedom. To circumvent this difficulty, we calculated COF, a factor for the overlap between two pools of eigenvectors as a function of pool size (54) (Eq. 10, Fig. 2 *B*). Consistent with the other methods of

TABLE 1 Comparison of parameters used in different NMA approaches

	AANM	ENM	BNM	CGNM
Residues	804	804	804	804
Atoms	8636	804	8636	8636
Starting Hessian matrix size	$25,908^2$	2412^2	$25,908^2$	$25,908^2$
Working Hessian matrix size	$25,908^2$	2412^2	4824^2	2412^2
Practical/theoretical eigenvector set	2000/25,908	2412/2412	4824/4824	2412/2412
Eigenvector dimension	25,908	2412	25,908	2412
C-α only component extraction	Yes	No	Yes	No
Orthogonality of C-α component	No, but normalized	Yes	No, but normalized	Yes
CPU time (3.4 Ghz Xeon, sequential implementation)	~67 h	~1 h	~5 h	~7 h
Peak physical memory (Mbyte)	~1577	44	~1400	~1900

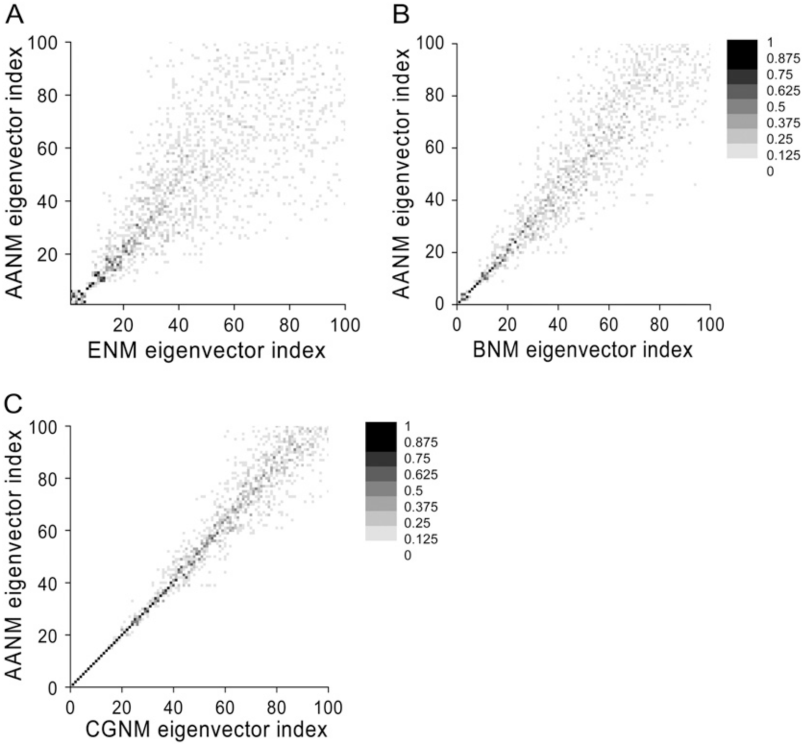

FIGURE 1 Grayscale plot showing the absolute value of the dot product between each pair of eigenvectors. (*A*) Gray scale plot showing the absolute value of the dot product between the eigenvectors from ENM analyses (*x* axis) versus AANM analyses (*y* axis). The quality of correspondence between eigenvectors is indicated by the darkness of the symbol and its proximity to the diagonal. (*B*) Results for BNM (*x*) versus AANM (*y*). (*C*) Results for CGNM (*x*) versus AANM (*y*).

comparison, the COF results show that CGNM significantly outperforms the other two methods: the space represented by the first 100 eigenvectors from AANM overlaps 95% of that of CGNM versus 85% of BNM and only 65% of ENM.

Based on the results shown above, it is clear that the CGNM method generates a more accurate set of eigenvectors than does BNM or ENM. Next, we checked the accuracy of different coarse-grained methods through calculating the atomic fluctuations based on the eigenvalues and eigenvectors of the Hessian matrices, still using the results from AANM as a reference. MSF or root mean-square fluctuation (RMSF) was used to provide a direct measure of the atomic vibrational amplitude. Both MSF and RMSF values can be used to compare computational results with experimental measures of motion, such as *B*-factors (see Methods, Eq.15).

To gain insight into atomic fluctuations we first plotted the eigenvalues from individual coarse-grained methods against the corresponding values from AANM (Fig. 2 *C*). Over a large range of eigenvalues, there is a nearly linear relationship between the results of AANM and those of CGNM or BNM, suggesting a close relationship. In contrast, such ENM results did not show a close agreement with those of AANM. Next, we used the eigenvalues and the eigenvectors of the vibrational modes to calculate the atomic fluctuations for each atom (Eq. 3) (Fig. 2 *D*). As predicted from the eigenvalues, the ENM fluctuation results (*blue circles*) deviate significantly from the other results. Both the CGNM (*red circles*) and BNM results (*green circles*) are in good agreement with the results from AANM on a residue-by-residue basis. Based on the correlation coefficient *R* values, the CGNM results (0.996) show a slightly better agreement with AANM compared to BNM (0.969). In contrast, both CGNM and BNM correlations are significantly better than that of ENM (0.838). To exclude possible errors introduced by the extraction of C-α components and the normalization step associated with AANM and BNM, we performed independent calculations using the original all-atom eigenvectors, which yielded identical results (Supplementary Material, Fig. S1).

NMA results incorporating a layer of explicit water on the protein surface

Next, we expanded CGNM to incorporate the effect of explicitly treated surface water molecules on protein dynamics, an area that to date has not been addressed by other coarse-grained normal mode methods and is computationally expensive for classical AANM. Previous experimental studies showed that the thickness of the surface structural water layer ranged from 3 Å for lysozyme, determined by x-ray and neutron scattering (64,65), to 5 Å for lactose, determined by terahertz spectroscopy (66). Here, we treated the case of a 4-Å-thick layer of explicit water on the protein surface, a compromise that enables the calculations to stay within the limits of currently available computational resources (Fig. 3, *A* and *B*). The all-atom Hessian matrix used in the CGNM calculations incorporated the interactions among all protein atoms, cAMP ligands, and explicitly treated water molecules. The corresponding hydration level is 0.56 (water mass/protein mass) and the system is of significant size (8636 protein atoms, 108 cAMP atoms, and 8817 water atoms). As a result,

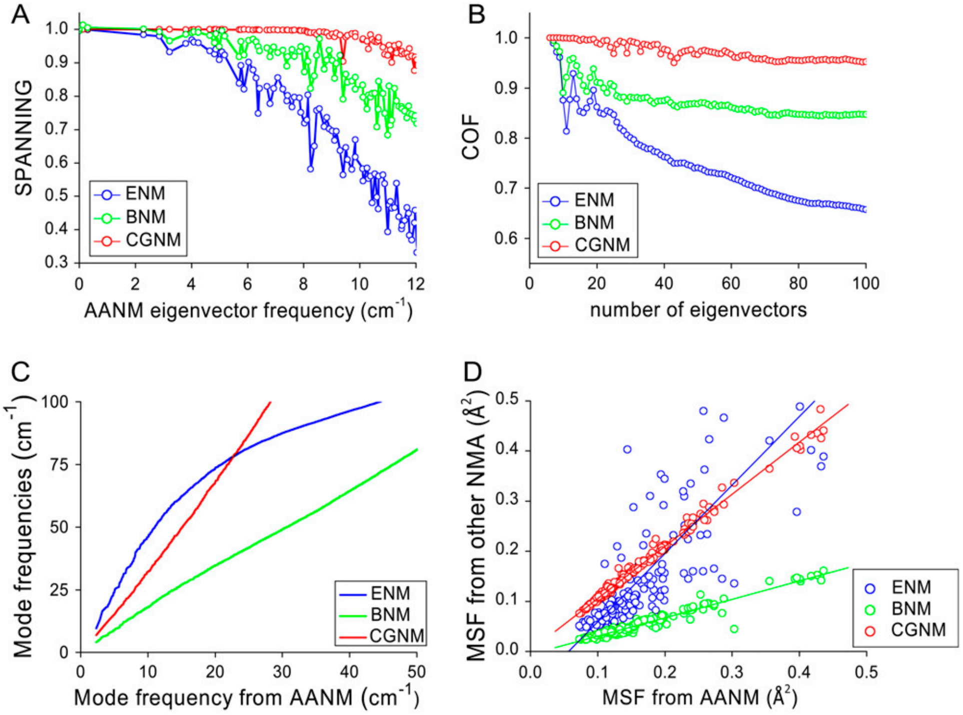

FIGURE 2 Quantification of the similarity of results between AANM and three coarse-grained NMA methods. (*A*) Spanning coefficients for each eigenvector of AANM analysis were calculated based on the first 100 eigenvectors of ENM (*blue*), BNM (*green*), and CGNM (*red*). The index for AANM eigenvector was converted to vibrational frequency. (*B*) COF of eigenvectors from AANM versus coarse-grained methods. (*C*) Cross plots of eigenvalues from coarse-grained versus AANM results. (*D*) Cross plot of MSF values from coarse-grained versus AANM results for each C-α atom. The linear equation $Y = A + B \times X$ was used to fit comparisons between different methods. CGNM: $B = 1.04$, $R = 0.996$; BNM: $B = 0.37$, $R = 0.969$; and ENM: $B = 1.36$, $R = 0.838$, where R is the correlation coefficient.

we could solve for <0.5% (250) of a total of over 52,683 AANM eigenvectors and eigenvalues using local computational resources. The BNM method could not be tested under these conditions because it does not currently include a method for allocating surface water molecules to specific blocks.

Inclusion of surface water led to a significant difference in AANM results compared to those obtained using AANM in the absence of water (Fig. 3, *C* and *D*). It was not surprising that AANM results with water also diverged from those obtained with CGNM or ENM in the absence of water. We were impressed that CGNM results in the presence of surface water showed a good agreement with those obtained using AANM in the presence of surface water, with an increased overlap of eigenvectors indicated by a twofold increase in spanning coefficients (~80%) compared to values obtained with the other methods in the absence of water (~40% (Fig. 3 *C*)). Improvement was also observed in the larger COF values with CGNM (~93%) versus those with the other methods (~75% using the first 100 eigenvectors (Fig. 3 *D*)).

Using AANM with surface water results as a standard, we next checked the accuracy of the RMSF values for each C-α atom using ENM, BNM, or CGNM (Fig. 4). CGNM with water (*red curve*) faithfully reproduced the pattern of the corresponding AANM results (*black curve*, Fig. 4 *A*). The shift in absolute amplitude is due to the different number of eigenvectors used (2406, or 99.7%, for CGNM; 244 modes, or 0.5%, for AANM). The striking similarity is reflected in the high *R*-factor of the CGNM data versus the AANM data (0.925 (Fig. 4 *D*, *left*)), which is much greater than that for ENM (0.780) and slightly greater than BNM (0.915, limited to calculations without water).

The effect of solvent molecules on protein dynamics is an important issue that has been addressed by experimental and computational approaches. Previous studies using AANM revealed that inclusion of surface water dampened the amplitude of atomic fluctuations (16,67). We found a similar effect of surface water using CGNM, in which the average fluctuations of C-α atoms with surface water (0.11 Å^2) is significantly smaller than that of protein alone (0.16 Å^2), providing further support for the ability of CGNM to incorporate surface water in protein dynamics.

Are the CGNM results with a 4-Å layer of surface water molecules comparable to results based on MD simulations, in which the protein is fully embedded in a $102 \times 102 \times 81$ Å^3 box filled with both surface and bulk water (Fig. 3 *A*)? MD simulations of the protein at 300 K did a reasonably good job of reproducing the absolute amplitude and overall pattern of RMSF values from x-ray crystallographic *B*-factors (Fig. 4 *B*). However, the RMSF values from MD simulations at 300 K are approximately three times larger than the CGNM results (Fig. 4 *A*). Moreover, the *R* factor between MD results and CGNM results with surface water is only 0.70 (Fig. 4 *E*). This deviation is likely caused by the contribution of random, diffusive motions that are included in the MD simulations but are ignored by the harmonic treatment of motions in all NMA approaches.

Since diffusive motions are greater at higher temperatures, we examined the ability of NMA to more accurately correspond to MD simulation results at lower temperatures. We

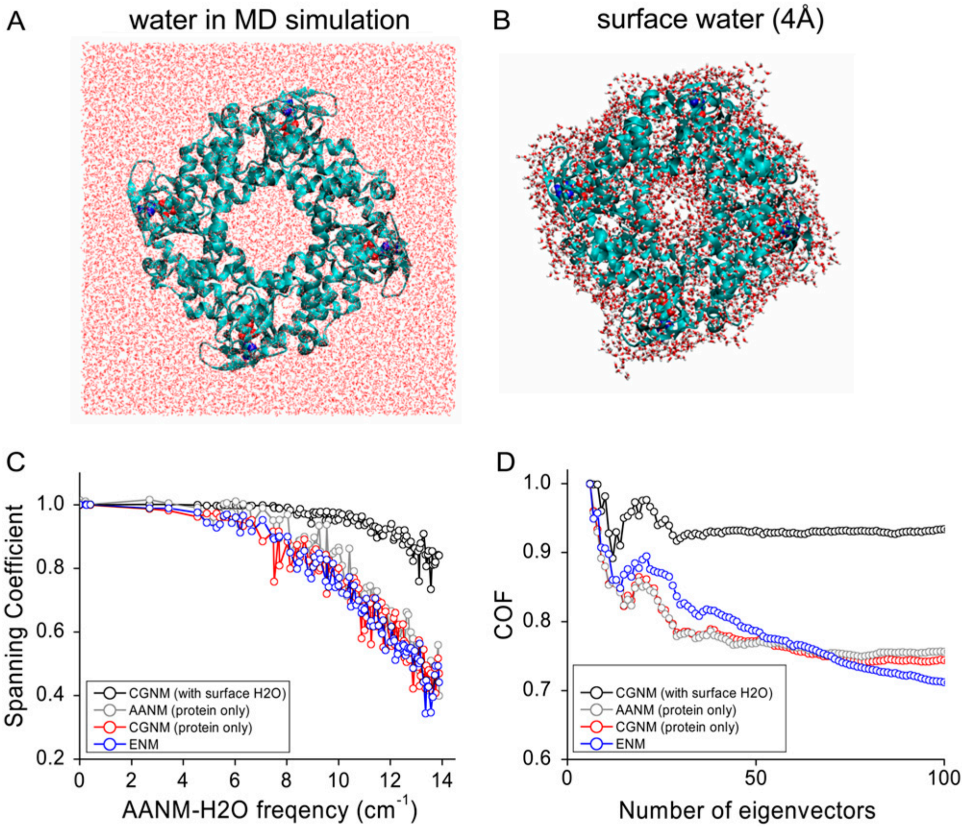

FIGURE 3 Spanning coefficient and cumulative overlap factor curves for different NMA methods with and without surface water molecules. (*A*) In the MD simulation system, the tetrameric HCN2 CNBD/C-linker domain was put in the center of a box containing explicitly treated water ($n = 23{,}658$; *red lines*). (*B*) Energy-minimized structure of protein with a surface layer of explicitly treated water molecules ($n = 2939$). (*C*) Spanning coefficient for each AANM (with water) eigenvector was calculated using the first 100 eigenvectors of CGNM with water (*black*), AANM without water (*gray*), CGNM (protein only, *red*), and ENM (protein only, *blue*). (*D*) COF curves showing the overlap between two pools of eigenvectors as a function of the number of eigenvectors involved.

first collected MD trajectories at 120 K and 180 K, respectively. The RMSF based on the MD simulation at 120 K was much smaller than that at 300 K; however, the convergence with CGNM results was not improved (0.69) (Fig. 4 *B*). Interestingly, the agreement between CGNM and MD simulation results at 180 K was much improved, in terms of both the absolute value of fluctuations and overall pattern, with the *R*-factor increased to 0.85 with a slope factor of 0.82 (Fig. 4 *C*). These results indicate that at certain low temperatures (180 K), the atomic fluctuations can be largely accounted for by harmonic motions involving the protein plus surface water but do not necessarily involve the bulk solvent that is also present in the MD simulations.

Comparing MD simulations with CGNM, with and without water, at different temperatures

To further examine the effect of surface water on protein dynamics, the distribution of vibrational-mode frequencies was plotted for both CGNM and MD simulations at the three different temperatures. Previous studies have shown that explicit water has a complex influence on protein dynamics, including temperature-dependent frictional dampening and temperature-independent shifting of the vibrational modes to higher frequencies (68–70). However, it is not clear whether these effects are due to an interaction of the protein with surface structural water versus an interaction that also requires the presence of bulk solvent (71). Here, using CGNM, we find that surface water alone is sufficient to shift fluctuations to higher frequencies (Fig. 5 *A*), an effect that is observed at all three temperatures, thus confirming its temperature-independent character. It is most likely that this effect represents a static interaction of a cagelike structure formed by the interaction of surface water with exposed residues on the protein surface (65).

To isolate the potential influence of the anharmonic, diffusive protein motions captured by MD simulations but not by CGNM, we compared the frequency distributions of vibrational modes between CGNM with 4 Å surface water and the MD simulations with bulk solvent (~10 Å from protein surface plus periodic boundary condition). The distribution of vibrational modes from MD simulations at low temperatures (180 K and 120 K) was quite similar to those from CGNM with water (Fig. 5 *B*). However, MD simulations, but not CGNM, revealed a significant shift to lower frequencies upon raising the temperature to 300 K. This is in good agreement with previous studies suggesting that the shift to low frequencies is related to the anharmonic nature of protein dynamics (Fig. 5 *C*) and that the contributions from bulk solvent are more prominent at high temperatures (300 K) and for low-frequency modes (69,70).

As a final test of the various NMA approaches, we compared the orthogonal sets of eigenvectors in the subspace of C-α atoms derived from the three coarse-grained methods

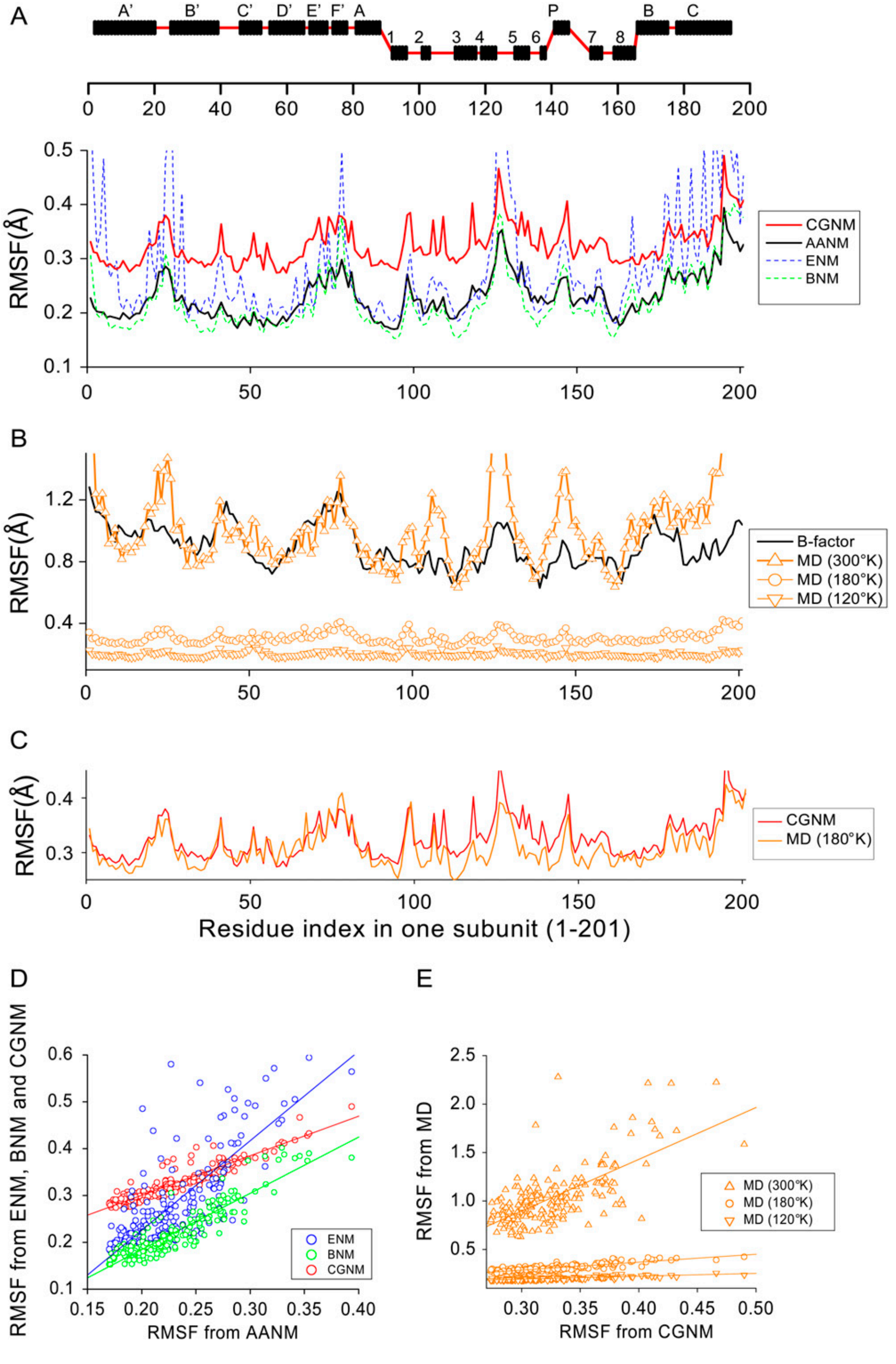

FIGURE 4 Comparison of RMSF values from various NMA approaches, MD simulations, and values converted from crystallographic *B*-factors. (*A*, *top*) Secondary structure along the primary sequence of a single subunit. (*A*, *bottom*) Comparison of RMSF values based on the first 244 of 52,683 eigenvectors from AANM with surface water (*black line*) with RMSF values based on the complete set of eigenvectors (2406) from CGNM with surface water (*red line*). The RMSF values for ENM (*blue*) and BNM (*green*) are also shown for comparison purposes. (*B*) Comparison of RMSF values converted from crystallographic *B*-factors (*black line*) with values determined from MD trajectories (*orange*) collected at 300 K (*triangles*), 180 K (*circles*), and 120 K (*inverted triangles*). (*C*) RMSF values from CGNM with water (*red*) are in good agreement with those from MD simulations at 180 K (*orange*). (*D*) Cross plots of RMSF results based on AANM with surface water versus RMSF results from ENM (*blue*), BNM (*green*), and CGNM (*red*). Results were fit with the linear equation $Y = A + B \times X$, where R is the correlation coefficient: ENM: $B = 1.91$, $R = 0.780$; BNM: $B = 1.20$, $R = 0.915$; CGNM: $B = 0.84$, $R = 0.925$. (*E*) Cross plots of RMSF results based on CGNM with surface water and RMSF results based on MD simulations at different temperatures. Linear least-squares fit results: 300 K: $B = 5.33$, $R = 0.70$; 180 K: $B = 0.82$, $R = 0.85$; 120 K: $B = 0.30$, $R = 0.69$).

(ENM, BNM, and CGNM) versus the eigenvectors based on PCA of MD simulation trajectories (Fig. 5 *D*). At 300 K, the COF curves show a poor overlap between the eigenvectors at frequencies <25 cm^{-1} from the MD simulations versus those obtained from all three NMA methods. At increasing frequencies, there is a gradual increase in the overlap of the NMA results with the PCA modes, consistent with the idea that frequencies $>$40–50 cm^{-1} correspond to more ‘‘harmonic’’ protein vibrations (60).

Does the use of CGNM with a layer of water molecules on the protein surface make any significant difference in the overlap of eigenvectors with MD simulations? Careful examination of the COF curve suggests that indeed the results from CGNM with water are slightly but consistently better

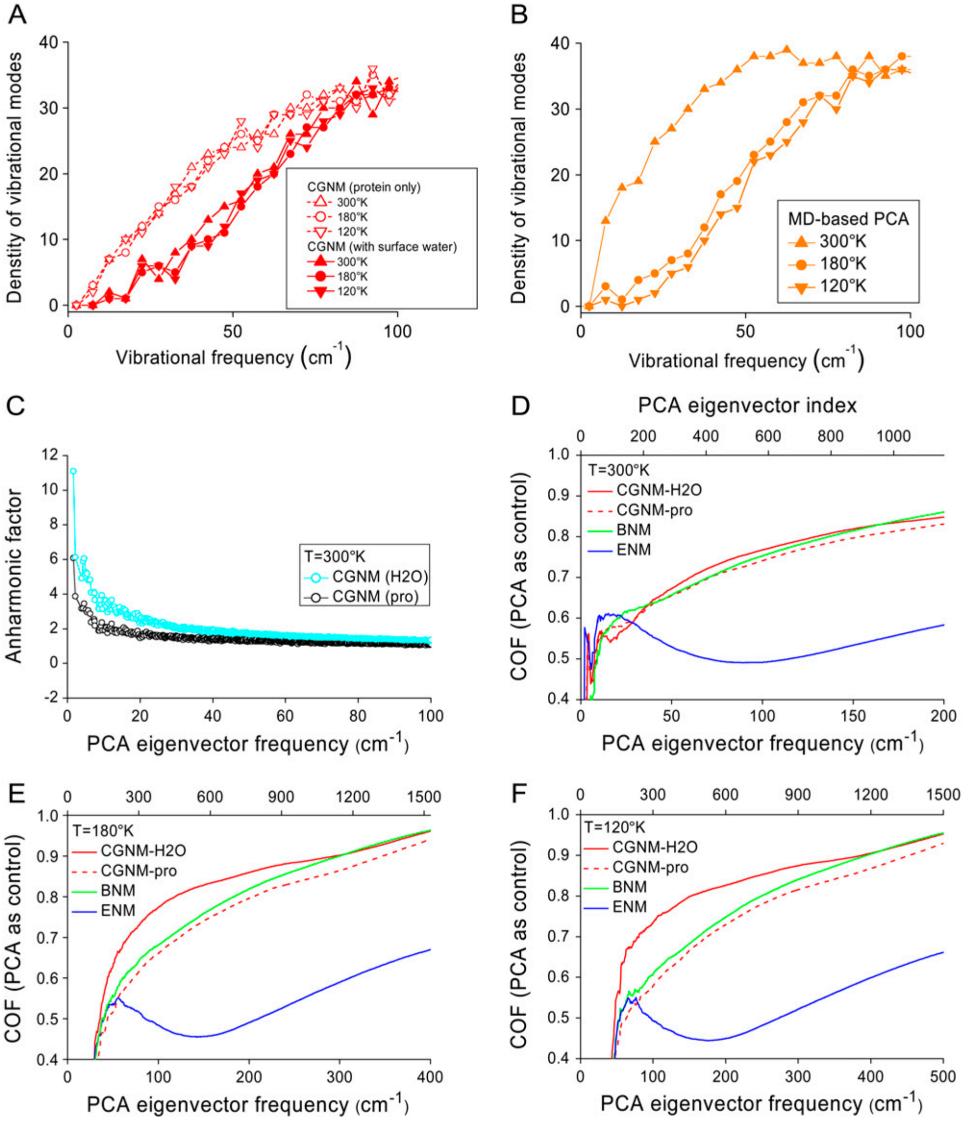

FIGURE 5 Effects of solvent on protein dynamics determined by CGNM and MD approaches. (*A*) Distributions of the vibrational mode densities for CGNM without water (*open symbols*) and CGNM with water (*solid symbols*) at three temperatures, 300 K (*triangles*), 180 K (*open circles*), and 120 K (*inverted triangles*). (*B*) Distributions of the vibrational mode densities for MD simulations at 300 K (*triangles*), 180 K (*circles*), and 120 K (*inverted triangles*). All plots are shown in brown to be consistent with Fig. 4. (*C*) Anharmonic factor for the PCA modes from MD simulations at 300 K based on harmonic CGNM modes without (*black*) or with (*blue*) surface water. (*D*) COF curves showing the overlap between the eigenvectors from MD simulations at 300 K and CGNM with water (*red solid line*), CGNM without water (*red dashed li*ne), BNM (*green*), or ENM (*blue*). (*E*) COF results showing the overlap between MD results and CGNM, BNM, and ENM results at 180 K. (*F*) COF results showing the overlap between MD results and CGNM, BNM, and ENM results at 120 K.

than the CGNM results without water at 300°K. An even greater improvement in overlap with the MD simulations was observed when using CGNM with water at lower temperatures (180 K (Fig. 5 *E*) and 120 K (Fig. 5 *F*)). The increased overlap between CGNM with surface water and the MD simulations at the two lower temperatures (but especially at 120 K) suggests not only that CGNM is able to incorporate, at least partially, the contributions from explicit surface water, but also that surface water makes a significant contribution to protein vibrational modes with frequencies >50 cm^{-1}.

DISCUSSION

In this article, we implemented a matrix-partitioning scheme to extract the C-α components from the all-atom Hessian matrix, thus providing a novel coarse-grained NMA approach, which we termed CGNM. This method generated more accurate results than did other coarse-grained NMA methods, including ENM and BNM, based on a comparison with results obtained using classical AANM. However, CGNM retained the benefits of a great reduction in computational cost with the two other coarse-grained approaches. The flexibility in partitioning the all-atom Hessian matrix into relevant versus nonrelevant groups makes it straightforward to scale the scope of analysis, for example, from C-α atoms only to inclusion of all backbone atoms, depending on the size of the system and the available computational resources. In this manner, we were able to model the contributions from explicitly treated surface water to protein motion, which is beyond the reach of other coarse-grained NMA methods. Thus, the CGNM method represents a novel coarse-grained NMA approach that can be used to obtain more accurate results for systems of significant size.

Multiple lines of evidence indicate that CGNM produces more accurate results than ENM or BNM, using AANM results as a reference. Overlap plots and spanning coefficients clearly show that CGNM outperformed the other coarse-grained methods for the first 100 or so individual eigenvectors, which are of great functional significance because they represent the directions of protein conformational changes with highest amplitude, slowest frequency, and least energetic cost. COF, which has the advantage of representing the overlap of two groups of eigenvectors, confirmed that CGNM more closely reproduces classical AANM results than do ENM or BNM methods. We also confirmed that CGNM outperforms ENM or BNM on dozens of other proteins, with a range in size from 200 to 1300 amino acids (results of four other sample proteins are shown in Figs. S2 and S3).

A comparison of eigenvalues and related atomic fluctuations among different NMA methods also revealed differences among the three coarse-grained methods. ENM generated a surprisingly good match to the AANM results given the dramatic simplifications in its potential energy functions. However, the results from BNM and CGNM were in much better agreement with the AANM results compared to ENM, indicating that the detailed chemical information embedded in the all-atom Hessian matrix used for AANM, BNM, and CGNM makes an important contribution. These results are in good agreement with a recent comprehensive comparison among NMA approaches of different complexity (41). Moreover, CGNM performed slightly better than BNM, as shown by the correlation coefficient (R) between their MSF values and the values obtained by AANM.

Why does CGNM yield more accurate results than BNM, even though both methods are derived from the same all-atom Hessian matrix? BNM is rooted in the rotation-translation block model, which projects the all-atom Hessian matrix into a subspace of rigid blocks. Even though BNM fully takes into account the coupled motions between different blocks, the method ignores the small high-frequency vibrations related to the intrinsic flexibility within each block (72). Moreover, during analysis, the intermediate BNM results in the subspace of rigid blocks must be projected first back onto the space for all atoms and then onto the subspace of C-α atoms. However, the center of mass for each block is different from the position of the C-α atoms and varies among different amino acid residues. In contrast, CGNM is based on partitioning the all-atom Hessian matrix through a simple but theoretically rigorous scheme, which is then used to derive the motions for the C-α atoms (55). The fact that CGNM implicitly incorporates energetic contributions from non-C-α atoms into C-α atoms may contribute to the greater accuracy of this method.

A key advantage of CGNM is its ability to incorporate the detailed chemical information imbedded in the protein structure, including explicitly treated structural water molecules on the protein surface. For most MD applications, it has been relatively standard to treat solvent molecules explicitly, which is required to reproduce the electrical and dynamic properties of solvents (70,73–76). Indeed, experimental and theoretical studies have found that the surface water molecules within a radial distance of 3–5 Å from the protein surface have very different physical-chemical properties from those in bulk solvent and play important roles in modulating protein motions. For example, the experimental observation that the density of the first hydration shell is ~5% higher than that of bulk water has been successfully reproduced by MD simulations (65,77,78). Ideally, classical AANM should be performed on the same protein-water system used in MD simulations. However, the size of the system limits the AANM method so that bulk solvent and surface water must often be omitted for proteins of significant size. Here, we applied CGNM to systems containing a layer of explicitly treated water molecules and found that it not only reproduced results based on classical AANM with explicit surface water, but also helped delineate some features of complex solvent effects.

The choice of a surface water layer of 4 Å in this study represents a balance that places a modest demand on computational resources but is consistent with experimental observations on protein surface water thickness, ranging from 3 Å for lysozyme (64,65) to 5 Å for lactose (66). We found that CGNM, which is based on a harmonic approximation to the energy surface, in the presence of surface water is able to reproduce MD results for atomic fluctuations of a fully solvated protein at 180 K. Interestingly, this temperature (180 K) is near the glass-transition point where diffusion starts to contribute significantly more to protein dynamics than does harmonic vibration (17,71,79–81). Moreover, spectroscopic experiments on bovine serum albumin showed that there is a significant dynamic change (glass transition) at around 170 K to 180 K, which might be due to formation of a rigid structure formed by water molecules covering the protein surface (79,82). These results corroborate this study, in which only the surface water is treated explicitly.

However, it is noticeable that even though CGNM results with water show an improved match with the atomic fluctuations from MD simulations compared to CGNM results on the dehydrated protein, the results from CGNM differ in important respects from those obtained using MD simulations or from experimentally determined crystallographic B-factors. This might be due to the complex nature of the protein energy surface and complex interactions between protein and solvent. The good agreement between the B-factors and the MD results at 300 K confirms the advantages of MD, a method that does not involve a harmonic approximation of the protein energy surface and explicitly treats all water molecules (Fig. 4 B).

The CGNM results are successful in reproducing previous observations that solvation increases protein vibrational frequencies and point to the role of surface water in this phenomenon (24,69,83). Moreover, these effects are likely to reflect temperature-independent interactions in which surface water molecules serve to fill in protein surface irregularities and stabilize polar side chains, forming a cagelike structure around the protein surface (83). In contrast, a comparison of

CGNM with surface water to MD simulations including bulk water indicate that bulk water molecules behave more like free water, acting to decrease the vibrational frequency of protein dynamics in a temperature-dependent manner (70). A recent experimental study of the influence of hydration on protein dynamics gives direct support to our results. Quasielastic neutron and light-scattering experiments show that adding an initial hydration layer ($h \approx 0.2$) increases the fast vibrational modes. Interestingly, further increasing the hydration level ($h > 0.2$) significantly activates slower processes (78). Therefore, these experimental observations are in good agreement with our simulation results showing the different contributions of solvent molecules to protein dynamics (70,84,85).

The poor overlap in the eigenvectors from various NMA approaches (no water or surface water only) with the MD simulation results (surface water and bulk water), especially for the low-frequency modes (25 cm^{-1}), is not surprising. A previous study using a jump-among-minima model, which divides protein motions into intra-substate motions and inter-substate jumps based on a multiple local minima model of the energy surface, generated a much better overlap with MD results than does NMA (86). Moreover, a mixture of harmonic NMA plus diffusive Brownian dynamics has been proven to be effective in reproducing the results of MD simulations and experimental observations (55,87). These studies suggest that the harmonic approximation of the protein energy surface and the neglect of solvent limits the ability of NMA approaches, including AANM, BNM, and CGNM, to reproduce the directionality of intrinsic anharmonic protein dynamics in the native state (54,84). However, for modes beyond 25 cm^{-1}, there is a gradual increase in the fidelity of CGNM and BNM, especially for CGNM with a layer of surface water. It is interesting that this frequency region is the same spectrum covered by terahertz absorption spectroscopy (1 THz = 33 cm^{-1}), where experimental results showed that solvation tends to enhance protein dynamics (88–90). Therefore, CGNM provides a convenient tool for modeling the contributions of surface water into protein dynamics at these higher frequencies. In principle, CGNM could be expanded to incorporate the effect of bulk solvent molecules in conjunction with other methods, such as the Langevin Model (71).

Thus far, the results presented for the HCN2 CNBD are for the cyclic-nucleotide bound state of the protein. However, we obtained similar results for the unliganded protein, using a representative snapshot from a 20-ns-long MD simulation with cAMP removed as the starting structure (Fig. S4–S6). One theoretical application of a complete set of eigenvectors and eigenvalues from NMA is the estimation of the configurational entropy (58,59). Taking advantage of the CGNM results for the unliganded protein versus the cAMP-bound protein in the subspace of C-α atoms, we estimated the entropy change of C-α atoms upon cAMP binding to be −127.8 J/mol without water or −174.3 J/mol with surface water (Table 2). Both values should be smaller than the estimate involving all atoms. However, the direction of the changes from the two independent calculations is consistent with previous MD results and the concept that ligand binding for hydrophilic or charged ligands (cAMP carries a negative charge) usually involves a reduction in the configurational entropy of the protein (91–93). Further improvements of the computational routine will focus on reducing the memory cost and use of more efficient routines for sparse matrix manipulation. With advances in computational algorithms, more memory-efficient and high-performance (sequential or parallel) routines could further improve this method and thus widen its application to more complex systems.

TABLE 2 Configurational entropy of C-α atoms based on NMA and PCA (300 K)

	Unliganded protein (8.31 J/mol/K)	cAMP-bound protein (8.31 J/mol/K)	Difference (8.31 J/mol/K)
ENM-protein	4103.92	4126.26	22.34
ENM-water	4154.64	4156.63	1.99
CGNM-protein	2555.63	2540.23	−15.40
CGNM-water	2390.10	2369.13	−20.97
PCA (10 ns)	3051.79	2957.70	−97.09
PCA (4 ns)	2756.74	2680.41	−76.33

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

We are grateful for the computational time provided by Pittsburgh Supercomputing Center through the National Resource Allocation Committee (grant MCB060032N) to L.Z. and S.A.S. This work was partially supported by grant NS36658 from the National Institutes of Health (to S.A.S.).

REFERENCES

1. Berendsen, H. J., and S. Hayward. 2000. Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.* 10:165–169.
2. Kitao, A., and N. Go. 1999. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* 9:164–169.
3. Karplus, M., and J. A. McCammon. 2002. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9:646–652.
4. McCammon, J. A., B. R. Gelin, and M. Karplus. 1977. Dynamics of folded proteins. *Nature.* 267:585–590.
5. Amadei, A., A. B. Linssen, and H. J. Berendsen. 1993. Essential dynamics of proteins. *Proteins.* 17:412–425.
6. Garcia, A. E. 1992. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* 68:2696–2699.
7. Ichiye, T., and M. Karplus. 1991. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins.* 11:205–217.
8. Janezic, D., R. M. Venable, and B. R. Brooks. 1995. Harmonic analysis of large systems. III. Comparison with molecular dynamics. *J. Comput. Chem.* 16:1554–1566.
9. Hess, B. 2000. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics.* 62:8438–8448.

10. Balsera, M. A., W. Wriggers, Y. Oono, and K. Schulten. 1996. Principal component analysis and long time protein dynamics. *J. Phys. Chem.* 100:2567–2572.

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

12. Brooks, B. R., D. Janezic, and M. Karplus. 1995. Harmonic analysis of large systems. I. Methodology. *J. Comput. Chem.* 16:1522–1542.

13. Ma, J. 2005. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure.* 13:373–380.

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

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

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

17. Yu, X., J. Park, and D. M. Leitner. 2003. Thermodynamics of protein hydration computed by molecular dynamics and normal modes. *J. Phys. Chem. B.* 107:12820–12828.

18. Durand, P., G. Trinquier, and Y. H. Sanejouand. 1994. A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers.* 34:759–771.

19. Perahia, D., and L. Mouawad. 1995. Computation of low-frequency normal modes in macromolecules: improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Comput. Chem.* 19:241–246.

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

21. Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. 1999. LAPACK Users' Guide, 3rd edition. Society for Industrial and Applied Mathematics, Philadelphia.

22. Lindahl, E., B. Hess, and D. van der Spoel. 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* 7:306–317 (online computer file).

23. McCammon, J. A., B. R. Gelin, M. Karplus, and P. G. Wolynes. 1976. The hinge-bending mode in lysozyme. *Nature.* 262:325–326.

24. Balog, E., J. C. Smith, and D. Perahia. 2006. Conformational heterogeneity and low-frequency vibrational modes of proteins. *Phys. Chem. Chem. Phys.* 8:5543–5548.

25. Karplus, M., Y. Q. Gao, J. Ma, A. van der Vaart, and W. Yang. 2005. Protein structural transitions and their functional role. *Phil. Trans. R. Soc. Lond.* 363:331–355; discussion, 355–356.

26. van der Spoel, D., B. L. de Groot, S. Hayward, H. J. Berendsen, and H. J. Vogel. 1996. Bending of the calmodulin central helix: a theoretical study. *Protein Sci.* 5:2044–2053.

27. Van Wynsberghe, A. W., and Q. Cui. 2006. Interpreting correlated motions using normal mode analysis. *Structure.* 14:1647–1653.

28. Petrone, P., and V. S. Pande. 2006. Can conformational change be described by only a few normal modes? *Biophys. J.* 90:1583–1593.

29. Brooks, B., and M. Karplus. 1985. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. USA.* 82:4995–4999.

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

31. Krebs, W. G., V. Alexandrov, C. A. Wilson, N. Echols, H. Yu, and M. Gerstein. 2002. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins.* 48:682–695.

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

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

34. Mouawad, L., and D. Perahia. 1993. Diagonalization in a mixed basis: a method to compute low-frequency normal modes for large macromolecules. *Biopolymers.* 33:599–611.

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

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

37. Yang, L. W., E. Eyal, C. Chennubhotla, J. Jee, A. M. Gronenborn, and I. Bahar. 2007. Insights into equilibrium dynamics of proteins from comparison of NMR and x-ray data with computational predictions. *Structure.* 15:741–749.

38. Temiz, N. A., E. Meirovitch, and I. Bahar. 2004. *Escherichia coli* adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling (15)N-NMR relaxation data. *Proteins.* 57:468–480.

39. Tama, F., and C. L. Brooks. 2006. Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu. Rev. Biophys. Biomol. Struct.* 35:115–133.

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

41. Kondrashov, D. A., A. W. Van Wynsberghe, R. M. Bannen, Q. Cui, and G. N. Phillips, Jr. 2007. Protein structural variation in computational models and crystallographic data. *Structure.* 15:169–177.

42. Rueda, M., P. Chacon, and M. Orozco. 2007. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure.* 15:565–575.

43. Tama, F., and C. L. Brooks 3rd. 2005. Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis. *J. Mol. Biol.* 345:299–314.

44. Tama, F., and Y. H. Sanejouand. 2001. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* 14:1–6.

45. Van Wynsberghe, A. W., and Q. Cui. 2005. Comparison of mode analyses at different resolutions applied to nucleic acid systems. *Biophys. J.* 89:2939–2949.

46. Li, G., and Q. Cui. 2004. Analysis of functional motions in Brownian molecular machines with an efficient block normal mode approach: myosin-II and Ca^{2+}-ATPase. *Biophys. J.* 86:743–763.

47. Santoro, B., D. T. Liu, H. Yao, D. Bartsch, E. R. Kandel, S. A. Siegelbaum, and G. R. Tibbs. 1998. Identification of a gene encoding a hyperpolarization-activated pacemaker channel of brain. *Cell.* 93:717–729.

48. Zagotta, W. N., N. B. Olivier, K. D. Black, E. C. Young, R. Olson, and E. Gouaux. 2003. Structural basis for modulation and agonist specificity of HCN pacemaker channels. *Nature.* 425:200–205.

49. Zhou, L., and S. A. Siegelbaum. 2007. Gating of HCN channels by cyclic nucleotides: residue contacts that underlie ligand binding, selectivity, and efficacy. *Structure.* 15:655–670.

50. van Gunsteren, W., S. R. Billeter, A. A. Eising, P. H. Huenenberger, P. Kruger, A. E. Mark, W. R. P. Scott, and I. G. Tironi. 1996. Biomolecular Simulation: The GROMOS96 Manual and User Guide. Vdf Hochschulverlag, Zurich.

51. Berendsen, H., J. Postma, W. van Gunsterenand, and J. Hermans. 1981. Interaction models for water in relation to protein hydration. D. Reidel, Boston.

52. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

53. van der Spoel, D., and P. J. van Maaren. 2006. The origin of layer structure artifacts in simulations of liquid water. *J. Chem. Theory Comput.* 2:1–11.

54. Amadei, A., B. L. de Groot, M. A. Ceruso, M. Paci, A. Di Nola, and H. J. Berendsen. 1999. A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells. *Proteins.* 35:283–292.

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

56. Eom, K., S. C. Baek, J. H. Ahn, and S. Na. 2007. Coarse-graining of protein structures for the normal mode studies. *J. Comput. Chem.* 28:1400–1410.

57. Schenk, O., and K. Gartner. 2006. On fast factorization pivoting methods for symmetric indefinite systems. *Electron. Trans. Numer. Anal.* 23:158–179.

58. Karplus, M., T. Ichiye, and B. M. Pettitt. 1987. Configurational entropy of native proteins. *Biophys. J.* 52:1083–1085.

59. Andricioaei, I., and M. Karplus. 2001. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* 115:6289–6292.

60. Hayward, S., A. Kitao, and N. Go. 1995. Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins.* 23:177–186.

61. Trueblood, K. N., H. B. Buergi, H. Burzlaff, J. D. Dunitz, C. M. Gramaccioli, H. H. Schultz, U. Shmueli, and S. C. Abrahams. 1996. Atomic displacement parameter nomenclature report of a subcommittee on atomic displacement parameter nomenclature. *Acta Crystallogr. A.* A52:770–781.

62. Fan, H., and A. E. Mark. 2003. Relative stability of protein structures determined by x-ray crystallography or NMR spectroscopy: a molecular dynamics simulation study. *Proteins.* 53:111–120.

63. Bois, P., B. Renaudon, M. Baruscotti, J. Lenfant, and D. DiFrancesco. 1997. Activation of f-channels by cAMP analogues in macropatches from rabbit sino-atrial node myocytes. *J. Physiol.* 501:565–571.

64. Svergun, D. I., S. Richard, M. H. Koch, Z. Sayers, S. Kuprin, and G. Zaccai. 1998. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA.* 95: 2267–2272.

65. Smith, J. C., F. Merzel, A. N. Bondar, A. Tournier, and S. Fischer. 2004. Structure, dynamics and reactions of protein hydration water. *Phil. Trans. R. Soc. Lond.* 359:1181–1189; discussion, 1189–1190.

66. Heugen, U., G. Schwaab, E. Brundermann, M. Heyden, X. Yu, D. M. Leitner, and M. Havenith. 2006. Solute-induced retardation of water dynamics probed directly by terahertz spectroscopy. *Proc. Natl. Acad. Sci. USA.* 103:12301–12306.

67. Teeter, M. M., and D. A. Case. 1990. Harmonic and quasiharmonic descriptions of crambin.:*J. Phys. Chem.* 94:8091–8097.

68. Hayward, S., A. Kitao, F. Hirata, and N. Go. 1993. Effect of solvent on collective motions in globular protein. *J. Mol. Biol.* 234:1207–1217.

69. Moritsugu, K., and J. C. Smith. 2005. Langevin model of the temperature and hydration dependence of protein vibrational dynamics. *J. Phys. Chem.* 109:12182–12194.

70. Hamelberg, D., T. Shen, and J. A. McCammon. 2006. Insight into the role of hydration on protein dynamics. *J. Chem. Phys.* 125:094905.

71. Moritsugu, K., and J. C. Smith. 2006. Temperature-dependent protein dynamics: a simulation-based probabilistic diffusion-vibration Langevin description. *J. Phys. Chem.* 110:5807–5816.

72. Essiz, S., and R. D. Coalson. 2006. A rigid-body Newtonian propagation scheme based on instantaneous decomposition into rotation and translation blocks. *J. Chem. Phys.* 124:144116.

73. Fox, T., and P. A. Kollman. 1996. The application of different solvation and electrostatic models in molecular dynamics simulations of ubiquitin: How well is the x-ray structure ''maintained''? *Proteins Struct. Funct. Genet.* 25:315–334.

74. Koehl, P. 2006. Electrostatics calculations: latest methodological advances. *Curr. Opin. Struct. Biol.* 16:142–151.

75. Paliwal, A., D. Asthagiri, D. Abras, A. M. Lenhoff, and M. E. Paulaitis. 2005. Light-scattering studies of protein solutions: role of hydration in weak protein-protein interactions. *Biophys. J.* 89:1564–1573.

76. Wagoner, J., and A. Baker Nathan. 2004. Solvation forces on biomolecular structures: a comparison of explicit solvent and Poisson-Boltzmann models. *J. Comput. Chem.* 25:1623–1629.

77. Merzel, F., and J. C. Smith. 2002. Is the first hydration shell of lysozyme of higher density than bulk water? *Proc. Natl. Acad. Sci. USA.* 99:5378–5383.

78. Roh, J. H., J. E. Curtis, S. Azzam, V. N. Novikov, I. Peral, Z. Chowdhuri, R. B. Gregory, and A. P. Sokolov. 2006. Influence of hydration on the dynamics of lysozyme. *Biophys. J.* 91:2573–2588.

79. Ostermann, A., R. Waschipky, F. G. Parak, and G. U. Nienhaus. 2000. Ligand binding and conformational motions in myoglobin. *Nature.* 404:205–208.

80. Vitkup, D., D. Ringe, G. A. Petsko, and M. Karplus. 2000. Solvent mobility and the protein ''glass'' transition. *Nat. Struct. Biol.* 7:34–38.

81. Steinbach, P. J., and B. R. Brooks. 1993. Protein hydration elucidated by molecular dynamics simulation. *Proc. Natl. Acad. Sci. USA.* 90:9135–9139.

82. Goddard, Y. A., J. P. Korb, and R. G. Bryant. 2006. Structural and dynamical examination of the low-temperature glass transition in serum albumin. *Biophys. J.* 91:3841–3847.

83. Enright, M. B., X. Yu, and D. M. Leitner. 2006. Hydration dependence of the mass fractal dimension and anomalous diffusion of vibrational energy in proteins. *Phys. Rev.* 73:051905.

84. Hayward, J. A., and J. C. Smith. 2002. Temperature dependence of protein dynamics: computer simulation analysis of neutron scattering properties. *Biophys. J.* 82:1216–1225.

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

86. Kitao, A., S. Hayward, and N. Go. 1998. Energy landscape of a native protein: jumping-among-minima model. *Proteins.* 33:496–517.

87. Meinhold, L., and J. C. Smith. 2007. Protein dynamics from X-ray crystallography: anisotropic, global motion in diffuse scattering patterns. *Proteins.* 66:941–953.

88. Whitmire, S. E., D. Wolpert, A. G. Markelz, J. R. Hillebrecht, J. Galan, and R. R. Birge. 2003. Protein flexibility and conformational state: a comparison of collective vibrational modes of wild-type and D96N bacteriorhodopsin. *Biophys. J.* 85:1269–1277.

89. Xu, J., K. W. Plaxco, and S. J. Allen. 2006. Collective dynamics of lysozyme in water: terahertz absorption spectroscopy and comparison with theory. *J. Phys. Chem.* 110:24255–24259.

90. Xu, J., K. W. Plaxco, and S. J. Allen. 2006. Probing the collective vibrational dynamics of a protein in liquid water by terahertz absorption spectroscopy. *Protein Sci.* 15:1175–1181.

91. Williams, D. H., E. Stephens, and M. Zhou. 2003. Ligand binding energy and catalytic efficiency from improved packing within receptors and enzymes. *J. Mol. Biol.* 329:389–399.

92. Homans, S. W. 2005. Probing the binding entropy of ligand-protein interactions by NMR. *ChemBioChem.* 6:1585–1591.

93. Gilson, M. K., and H. X. Zhou. 2007. Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct.* 36:21–42.