JCTC Journal of Chemical Theory and Computation

Approaching Elastic Network Models to Molecular Dynamics Flexibility

Laura Orellana,[†,‡] Manuel Rueda,[†,§] Carles Ferrer-Costa,[†] José Ramón Lopez-Blanco,[||] Pablo Chacón,[||] and Modesto Orozco*[,†,‡]

Joint Research Program in Computational Biology from the Institute for Research in Biomedicine Barcelona (IRBB) and Barcelona Supercomputing Center (BSC), Barcelona, Spain, Chemical and Physical Biology, Centro de Investigaciones Biológicas, Madrid, Spain, Departament de Bioquímica i Biologia Molecular, Universitat de Barcelona, Barcelona, Spain, and Skaggs School of Pharmacy, University of California—San Diego, La Jolla, California 92093

Received February 16, 2010

Abstract: Elastic network models (ENMs) are coarse-grained descriptions of proteins as networks of coupled harmonic oscillators. However, despite their widespread application to study collective movements, there is still no consensus parametrization for the ENMs. When compared to molecular dynamics (MD) flexibility in solution, the ENMs tend to disperse the important motions into multiple modes. We present here a new ENM, trained against a database of atomistic MD trajectories. The role of residue connectivity, the analytical form of the force constants, and the threshold for interactions were systematically explored. We found that contacts between the three nearest sequence neighbors are crucial determinants of the fundamental motions. We developed a new general potential function including both the sequential and spatial relationships between interacting residue pairs which is robust against size and fold variations. The proposed model provides a systematic improvement compared to standard ENMs: Not only do its results match the MD results—even for long time scales—but also the model is able to capture large X-ray conformational transitions as well as NMR ensemble diversity.

1. Introduction

Protein functions largely depend on the intrinsic flexibility of their structures; even processes such as ligand binding or catalysis, in which the overall shape or surface properties play a dominant role, are coupled to local movements of the polypeptide backbone.[1] The intrinsic deformability of different protein families seems to guide structural changes along evolution, and deformation patterns (i.e., the large-scale motions) are extremely conserved in proteins displaying a common function.[2] Unfortunately, despite promising advances,[3] the experimental study of large-scale dynamics is still difficult, and a large amount of information comes from theoretical calculations. Among the computational approaches to tackle the question of protein flexibility, molecular dynamics (MD)[4−6] is probably the most accurate, since it is based on a rigorous physical formalism and a thorough parametrization from quantum-mechanical and experimental measurements. Although the high computational cost still limits atomistic simulations to the nanosecond to microsecond time scale, the principal component analysis (PCA) of MD trajectories—also called the essential dynamics (ED)[7] approach—provides valuable information on large-scale functional motions, as we will discuss below. An alternative to MD to reach biologically relevant time and length scales is coarse-grained (CG) models,[8] which simplify both the protein representation and the potential functions. Among these methods, the elastic network models (ENMs) are

* Corresponding author phone: (+34)934037156; fax: (+34)934037157; e-mail: modesto.orozco@irbbarcelona.org.

[†] Joint Research Program in Computational Biology from the IRBB and BSC.

[‡] Universitat de Barcelona.

[§] University of California—San Diego.

[||] Centro de Investigaciones Biológicas.

10.1021/ct100208e
Published on Web 08/20/2010

probably the most widely used ones.[9] The ENM potential is defined by a network of springs connecting the C^{α} atoms in a topology matrix $\mathbf{\Gamma}$ (known as the Kirchhoff matrix) of inter-residue contacts, where the ijth element is equal to -1 if nodes (i.e., residues) i and j are within the cutoff distance r_c or 0 otherwise and the diagonal elements (iith) are equal to residue connectivity:

$$\Gamma_{ij} = \begin{cases} -1 & \text{if} \quad d_{ij} \le r_c \\ 0 & \text{if} \quad d_{ij} > r_c \end{cases} \qquad \Gamma_{ii} = -\sum_{k|k\neq i}^{N} \Gamma_{ik} \qquad (1)$$

The topology of the C^{α} network may be alternatively expressed in terms of a stiffness matrix, whose elements are the Hookean force constants, K_{ij}, acting between any pair of nodes i, j:

$$K_{ij} = 0.5\xi\Gamma_{ij} \qquad (2)$$

where ξ is a constant which may or may not have the same value for all pairs, depending on the model. Hence, the overall potential energy of the network is given by

$$E = \sum_{i\neq j} K_{ij}(R_{ij} - R_{ij}^0)^2 \qquad (3)$$

where R_{ij} and R_{ij}^0 are the instantaneous and reference (equilibrium) distances between each pair of α-carbons i and j. The functional in eq 3 can be implemented into Monte Carlo or dynamics algorithms[10] to obtain ensembles of accessible configurations or within the elastic network normal mode analysis approach (NMA)[11,12] to build the Hessian matrix of the potential. Within the anisotropic network model (ANM)[13] approach, diagonalization of the Hessian directly yields a set of eigenvectors and eigenvalues (in energy or frequency units) which together define the near-equilibrium harmonic deformability space. In spite of this extreme simplicity, the lowest frequency modes of the ENMs provide descriptions of large-scale flexibility in good agreement with empirical and theoretical data, being especially well-suited to trace cooperative domain and segment movements. However, it cannot be ignored that ENMs are based on a harmonic, near-equilibrium approach and a rigid topology and thus have problems in capturing large anharmonic motions that can (in principle) be traced by MD simulations. Furthermore, there is no consensus parametrization, and the diverse models are often fitted to each particular problem.

Many attempts have been made to improve the robustness and generality of ENMs, for example, developing methods where atoms are grouped into rigid blocks,[14] scaling differently covalent and noncovalent contacts,[15] or using Markovian approaches[16] to define the coarse-graining. The use of an isotropic constant and a cutoff is appealing for its simplicity, but can lead to different outcomes depending on the selected threshold for interactions[17] (see also the Results and Discussion). Therefore, to avoid the use of an arbitrary cutoff, the discrete Hamiltonian is sometimes replaced by continuous functions that scale the force constants with an inverse power of the inter-residue distance. For example, Hinsen et al.[18] derived a function for the spring strength by fitting to a local minimum from a single 1.5 ns MD simulation of one protein. This force constant definition proposed stronger couplings for backbone neighbors and a sixth power of distance for the rest of the interactions. The distinction of short- and long-range terms was, however, dependent on a short 4 Å cutoff, and the formulation also included a protein-fitted scaling factor for the global energetics. Kovacs et al.[19] proposed a simpler sixth-power exponential which did not require any cutoff or scaling factor. Other authors have also used several distance-dependent force constants,[20] including sometimes specific short-range terms (see ref 21, also based on MD) or, alternatively, bond cutoffs for chain neighbors.[22] Recently, Jernigan et al.[23] suggested an inverse-square function for the reproduction of B factors, but they found that the resulting stronger long-range cohesion prevented discrete domains from moving properly.

Attempts to refine and improve ENMs have a common drawback: the lack of reliable experimental data on protein flexibility in solution, mostly coming from nuclear magnetic resonance (NMR) spectroscopy relaxation measurements[24] and to a lesser extent neutron scattering data,[25] both available for very few proteins. Therefore, in most studies so far, ENMs have been validated by fitting the calculated atomic fluctuations to B factors found in the crystal, in some cases to the degree of reaching an almost perfect fit.[26] Nevertheless, the use of X-ray B factors as a reference for flexibility in solution has been highly controversial,[27−29] since they are subject to crystal packing effects, among other biases such as internal static disorder or refinement errors.[30] Other indirect sources of flexibility data for calibration and benchmarking have been the study of the environment-dependent conformational space of proteins[23,31] and, more recently, the analysis of NMR ensembles,[32,33] including comparisons with their RMSDs.[34] However, in the first case no guarantee exists that conformational changes induced, for example, by the presence of other molecules match the intrinsic deformation pattern of apo proteins. Furthermore, principal components predicted from PCA of selected NMR ensembles agree with the ED modes,[35] but caution must be taken since local diversity of NMR structures may also be a sign of experimental uncertainty due to missing data. In summary, there is a dramatic lack of direct experimental information on protein flexibility in solution, which hinders the validation of current models. As a consequence, concerns exist in their real generality and physical sense and in whether a small improvement compensates for an increase in model complexity and the need for adjusting more ad hoc parameters.

On the basis of the previous paragraph, it seems reasonable to use MD simulations as reference data for refinement of ENMs. Surprisingly, only a few authors have explored the use of MD data for ENM parametrization. MD simulations render detailed flexibility information on the correlated motions for the time scale sampled, as shown in comparisons with NMR fast motions.[36] Current MD simulations reproduce accurately high-quality direct NMR information on protein flexibility (RDCs and S^2 parameters) in the few proteins for which these measurements are available.[37] On the other hand, MD displays excellent correlation with B factors, even though MD B factors are systematically larger, especially for very flexible residues (which appear "frozen" in the crystal

lattice[37]). MD captures both short- and large-scale flexibilities, the latter being extracted from ED treatment of the collected trajectory,[38] allowing the characterization of collective anharmonic displacements often related to function.[39−41] As pointed out above, these so-called *essential* modes also correlate extremely well (both in directions and variance distribution) with the principal components from selected NMR ensembles[35] and thus can be expected to provide a quite realistic picture of large-scale flexibility in solution.

In a previous related work,[42] we performed a thorough comparison between the collective motions predicted by ED and different ENMs. We found that the space defined by the first, most relevant NMA eigenvectors captures the backbone flexibility as given by ED, with the inverse function proposed by Kovacs outperforming the original cutoff approach. However, despite these good correlations with the ED eigenspace, the main motions in NMA are often spread out into multiple modes of similar energy, instead of being concentrated in a few modes as detected in ED. In other words, ED displays higher flexibility, describing collective motions in fewer modes than NMA. Note that this discrepancy cannot be corrected by scaling uniformly the spring constants, since the variance distribution pattern along the energy spectra is fundamentally different (see the discussion below). In this paper we tried to find solutions to this problem by deriving a refined EN-NMA model based on comparison with atomistic MD simulations for a large number of proteins. The proposed ED-refined ENM method (in the following ed-ENM) provides results closest to those of MD, is able to reproduce flexibility in NMR ensembles, and can trace efficiently biologically relevant deformations observed in the Protein Data Bank. The ed-ENM is freely available through the Web site[58] http://mmb.pcb.ub.es/Flexserv. Improvement with respect to standard elastic network models is consistent in all the metrics considered.

2. Methods

2.1. Elastic Network Normal Mode Analysis. ENM can be considered a generalization of the bead-and-strings Rouse polymer model, but contrary to this simple scheme where only chained monomers are coupled, current ENMs connect all α-carbons within a given threshold. Thus, all interactions within the cutoff are harmonic and uniform (irrespective of their chemical nature), and all interactions outside are negligible. By relying on the Cartesian distance as the sole criterion, ENMs are not able to distinguish between close chain neighbors and remote contacts. To derive a more physically sound model, we explored alternative approaches, where the $C^{\alpha}-C^{\alpha}$ interaction strength depends on their topological relationship. After extensive testing of different potential functionals, we analyzed in detail three models that represent increasing levels of topological complexity and constant scaling: (i) a cutoff model with a uniform constant, the most widespread approach (standard defaults in ref 43); (ii) a noncutoff model using an exponential decay function, as developed by Kovacs and co-workers;[19] (iii) a hybrid cutoff model with sequential weighted springs for the first (M) neighbors, while the rest are represented by an inverse function of the Cartesian distance.

To obtain the weights of the spring constants for the first sequential neighbors in an unbiased way, we computed the residue−residue "apparent" stiffness constants obtained from MD assuming the harmonic oscillator model:

$$K_{ij}^{\mathrm{app}} = \frac{k_{\mathrm{B}}T}{\langle [R_{ij} - R_{ij}^{0}]^2 \rangle} \tag{4}$$

where k_{B} is the Boltzmann constant, T is the temperature, and $R_{ij} - R_{ij}^{0}$ is the oscillation in the interaction distance from average values. These constants were fitted to an inverse exponential function using a nonlinear regression routine for a small protein set (see the Results and Discussion):

$$K_{ij}^{\mathrm{ap}}(S_{ij}) = \frac{C^{\mathrm{seq}}}{S_{ij}^{n_{\mathrm{seq}}}} \tag{5}$$

where S_{ij} stands for the distance in sequence between residues i and j. The optimum exponent determining the shape of the variation is used in the rest of the study, while the constant C^{seq} is further refined to match "real" instead of "apparent" force constants (see the Results and Discussion). A similar strategy was used to derive the distance dependence for nonsequential interactions:

$$K_{ij}^{\mathrm{app}}(d_{ij}) = \frac{C^{\mathrm{cart}}}{d_{ij}^{n_{\mathrm{cart}}}} \tag{6}$$

where d_{ij} is the distance between residues i and j in a given conformation; in our implementation $d_{ij} = |R_{ij}^{0}|$. In the ed-ENM, the network topology is defined by a fully connected matrix for the first M neighbors, and contrary to standard pure continuum methods, we introduce a size-dependent cutoff to annihilate artifactual distant interactions (see the Results and Discussion). Thus, given a pair of residues i and j with sequential distance $S_{ij} > 0$ and Cartesian distance d_{ij}, the ijth element of the hybrid inter-residue contact matrix is

$$\Gamma_{ij}\begin{cases} S_{ij} \le M & \Gamma_{ij} = 1 \\ S_{ij} > M & \begin{cases} \Gamma_{ij} = 1 & \text{if} \quad d_{ij} \le r_{\mathrm{c}} \\ \Gamma_{ij} = 0 & \text{otherwise} \end{cases} \end{cases} \tag{7}$$

where Γ_{ij} always has $2M + 1$ nonzero diagonal entries defining neighbor chained contacts. Accordingly, the force constants K_{ij} are dependent not only on the Cartesian but also on the sequential distance:

$$K_{ij}\begin{cases} S_{ij} \le M & K_{ij} = C^{\mathrm{seq}}/S_{ij}^{n_{\mathrm{seq}}} \\ S_{ij} > M & \begin{cases} \text{if} \quad d_{ij} \le r_{\mathrm{c}} \quad \text{then} \quad K_{ij} = (C^{\mathrm{cart}}/d_{ij})^{n_{\mathrm{cart}}} \\ K_{ij} = 0 \quad \text{otherwise} \end{cases} \end{cases} \tag{8}$$

where values for all terms ($n^{\mathrm{seq}} = 2$ and $C^{\mathrm{seq}} = 60$ kcal/(mol·Å^2); $n^{\mathrm{cart}} = 6$ and $C^{\mathrm{cart}} = 6$ kcal/(mol·Å^2); in energy units) are obtained by fitting to apparent force constants and structural variance profiles. On the basis of MD simulations, a limit of $M = 3$ was used for sequential interactions, and the cutoff radius (r_{c}) was found to be dependent on the size (see the Results and Discussion).

2.2. Molecular and Essential Dynamics. MD simulations for several proteins (see above) were titrated, neutralized, hydrated, minimized, heated, and equilibrated for at least 0.5 ns. Trajectories were collected for at least 10 ns using three all-atom force fields (AMBER,[44] CHARMM,[45] and OPLS/AA[46]). The three trajectories obtained were combined to create a *metatrajectory* which is expected to collect much of the equilibrium dynamics of proteins (control simulations were also performed considering the individual trajectories). The noise arising from irrelevant short-range vibrations was filtered to obtain large-scale motions by ED:[7] the MD trajectory snapshots were aligned to the original X-ray reference structure (or the average of the NMR ensemble) to compute a common average structure and used to build a covariance matrix whose diagonalization (PCA) yields a set of eigenvectors and eigenvalues representing the essential, large-scale movements (further details in ref 37). To check the ability of the ENMs to capture deformations happening on longer time scales, we extend several calculations to long (0.1 μs) or very long (0.5−1 μs) time scales, using in this case only the AMBER force field as discussed below.

2.3. Training Proteins. Initial training of the model was performed by taking six highly representative proteins (PDB 1I6F, 1PHT, 1AGI, 1JLI, 1BSN, and 1SUR) of different sizes (60−200 residues) which were present in our μMODEL subset of the MODEL database (http://mmb.pcb.ub.es/MoDEL; see ref 37). Parameters were adjusted using as reference the dynamics metatrajectories described above.

2.4. Test Proteins. The model was first tested against the rest of the proteins (from 32 to 400 residues) contained in the μMODEL set, composed of 32 proteins representing the main metafolds. Larger and multidomain proteins (PDB 3ADK, 1BUD, 1SSX, 1PPO, 1DUA, 1QLJ, and 1PMI) were added, including some extremely large proteins (1SQC (619 residues), 1E5T (710), and 1J0M (747)) and a multimeric complex (1E9S (2545)). To test the performance of ed-ENM on long time scales and avoid any bias introduced by the length of the simulations, we analyzed extended MD trajectories (0.1 μs) for 2GB1, 1CEI, 1CQY, and 1OPC, up to the microsecond (0.5−1 μs) for the last two proteins (1CQY, 1OPC) plus 1UBQ and 1KTE. Long trajectories as well as standard trajectories for large proteins were obtained only with AMBER.

2.5. Comparison Metrics. The ENMs' ability to reproduce MD flexibility was tested considering a wide variety of metrics to cover different aspects.

2.5.1. Relative Deformational Amplitude. The size and complexity of the protein deformation space were characterized by (i) the *structural variance*, (ii) the *number of modes needed to explain 90%* of this variance, (iii) the *variance profile* with respect to the number of modes, (iv) the "*reduced variance*" defined as the variance explained by the first five modes, which for most average-sized proteins accounts for 70−80% of the total variance (see Figure S1 in the Supporting Information; similar findings in ref 31), and (v) the *strength of the softer deformation modes*. Note here that the ED eigenvalues obtained by diagonalization of the Cartesian covariance matrix (describing the mode amplitude) appear in distance units, but can be converted into energy units (kcal/(mol·Å^2), i.e., mode strength) for comparison with NMA modes by using

$$K_\nu = \frac{k_B T}{\lambda} \tag{9}$$

where ν stands for a given mode, k_B is Boltzmann's constant, T is the temperature, and λ stands for the associated eigenvalue (in square distance, Å^2).

2.5.2. Deformational Space Overlap. Hess's metric[47−49] was used to estimate the similarity of NMA and ED deformation spaces:

$$\gamma_{XY} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^{ED}\cdot\nu_j^{NMA})^2 \tag{10}$$

where the indexes i and j stand for the orders of the eigenvectors (ν, ranked according to their variance contribution) and m stands for the number of eigenvectors in the "important space", defined as the minimum number needed to explain a certain variance threshold. We considered here two definitions of the important space to guarantee a representative number of eigenvectors in the calculations: (i) eigenvectors needed to explain the 90% variance ($\gamma_{90\%}$) and (ii) the first 50 eigenvectors (γ_{50}). Additionally, the similarity between the first 10 eigenvectors from the ED and normal mode subspace was computed (γ_{10}). However, the similarity index in eq 10 presents two shortcomings: (i) the similarity increases with the important space size and (ii) the index is not sensitive to the eigenvector permutation. To solve the first, we refer to Hess's indexes to background models using Z_{score}:

$$Z_{score} = \frac{(\gamma_{AB}(obsd)) - (\gamma_{AB}(random))}{std(\gamma_{AB}(random))} \tag{11}$$

where 500 physically meaningful random models were obtained by diagonalization of a covariance matrix derived from discrete molecular dynamics (DMD) simulations performed using a Hamiltonian containing covalent bonds plus a hard sphere potential.[10] To evaluate the impact of permutation, we computed dot products between eigenvector pairs, determining the difference in rank between the ones showing the largest overlap, and used Perez's similarity index, which weighs the similarity of each pair of eigenvectors by their associated Boltzmann factor (see ref 50):

$$\xi_{AB} = \frac{2\sum_{i=1}^{i=z}\sum_{j=1}^{j=z}\left[(\nu_i^A\cdot\nu_j^B)\dfrac{\exp\left\{-\dfrac{(\Delta x)^2}{\lambda_i^A}-\dfrac{(\Delta x)^2}{\lambda_j^B}\right\}}{\sum_{i=1}^{i=z}\exp\left\{-\dfrac{(\Delta x)^2}{\lambda_i^A}\right\}\sum_{j=1}^{j=z}\exp\left\{-\dfrac{(\Delta x)^2}{\lambda_j^B}\right\}}\right]^2}{\sum_{i=1}^{i=z}\left(\dfrac{\exp\left\{-2\dfrac{(\Delta x)^2}{\lambda_i^A}\right\}}{\left(\sum_{i=1}^{i=z}\exp\left\{-\dfrac{(\Delta x)^2}{\lambda_i^A}\right\}\right)^2}\right)^2+\sum_{j=1}^{j=z}\left(\dfrac{\exp\left\{-2\dfrac{(\Delta x)^2}{\lambda_j^B}\right\}}{\left(\sum_{j=1}^{j=z}\exp\left\{-\dfrac{(\Delta x)^2}{\lambda_j^B}\right\}\right)^2}\right)^2} \tag{12}$$

where the common displacement (Δx) is selected as the minimum value for which the impact outside the important space is negligible. An additional metric that helps in determining the similarity between MD and NMA-based eigenvectors is the "spread" index by Hinsen:[51]

$$s_i = (\sum_j j^2 {\eta_{ij}}^2 - (\sum_j j {\eta_{ij}}^2)^2)^{1/2} \quad (13)$$

where $\eta_{ij} = \nu_i^A \cdot \nu_j^B$. Note that for two identical sets of modes ${\eta_{ij}}^2 \neq 0$ only if the $i = j$ spread becomes equal to 0. Higher values indicate the distribution of the eigenvector i on a larger number of eigenvectors j in the B space.

2.5.3. Relative Distribution of the Deformational Pattern. The flexibility distribution along the residues can be analyzed from different metrics. A powerful one is Brüschweiler's "collectivity" index,[52] which evaluates the amount of residues involved in every motion k:

$$\kappa_k = \frac{1}{N} \exp\{-\sum_{i=1}^{N} {u_{k,i}}^2 \log {u_{k,i}}^2\} \quad (14)$$

where N is the total number of residues in the protein and ${u_{k,i}}^2$ is given by

$${u_{k,i}}^2 = \frac{{\nu_{k,X}}^2 + {\nu_{k,Y}}^2 + {\nu_{k,Z}}^2}{m_i} \quad (15)$$

where m_i is the mass of residue i. The large-scale motions tend to be the more collective ones. The B factors for each residue i, B_i, were evaluated from average thermal fluctuations, $\langle \Delta {r_i}^2 \rangle$, under mode k:

$$B_i = (8\pi^2/3)\langle (\Delta r_i)^2 \rangle$$

where

$$\langle (\Delta r_i)^2 \rangle = (3k_B T/\xi)[\Gamma^{-1}]_{ii} = (3k_B T/\xi) \sum_k [{\lambda_k}^{-1} \nu_k \nu_k^T]_{ii} \quad (16)$$

They were also processed to determine Lindemann's indexes,[53] a useful metric providing information on the macroscopic behavior (liquid or solid) of proteins:

$$\Delta_L = \frac{(\sum_i \langle \Delta {r_i}^2 \rangle / N)^{1/2}}{a'} \quad (17)$$

where a' is the most probable nonbonded near-neighbor distance (taken as 4.5 Å). To avoid noise introduced by high-frequency modes, B factors and Lindeman's indexes have been computed by summing the contributions of the first 50 modes (negligible differences are expected if more modes are considered).

2.5.4. Dot Product against X-ray Transition Vectors. Systems selected for analysis belong to a benchmark of conformational transitions (http://sbg.cib.csic.es/Software/NMAFIT), formed by 54 transition problems from the macromolecular motions database MolMovDB,[54] with displacements greater than 2 Å C^α rmsd (the average displacement was 6.3 Å with a standard deviation of 3.4 Å). We present results for 10 different motions between open/closed pairs; note that each open/closed pair presents two different transition problems. The ability of ed-ENM to predict these biologically relevant transitions was estimated by the accumulated normalized dot products between the 5 (γ_5) and 10 (γ_{10}) first eigenvectors of the corresponding closed/open form, which have been shown to describe the conformational change[31,54] with respect to the multidimensional vector driving the transition (see eq 10; here $m = 1$ for the first subspace; thus, here γ denotes a dot product between a single vector and a subspace, as opposed to the deformational space overlap in section 2.5.2). As an additional metric, we determined the rank distance between the transition vector and the best overlapped eigenvector (a value of 0 indicates that it is the first one).

2.5.5. Dot Product against Principal Components from NMR Ensembles. To have a qualitative approximation to the flexibility present in NMR multiple structures, we selected 26 ensembles from the Protein Data Bank having at least 10 conformers and spanning a wide size range. Each structure was coarse-grained to the C^α level and then aligned to its average. The closest structure to this initial one was used as a template for a second alignment and computation of the final average structure, which was the reference for subsequent ANM and PCA. The performance of the different ENMs to describe the diversity of the structural ensemble is measured by the accumulated normalized dot products (as given in eq 10) between the 5 (γ_5) and 10 (γ_{10}) first eigenvector pairs from each subspace (i.e., a deformational space overlap as in section 2.5.2) and also by the value of the dot product for the best overlapped pair (γ_{max}) (a vector to vector inner product).

3. Results and Discussion

3.1. Optimization of the Method. As described above, MD trajectories of a small set of proteins were used to formulate the model, which was later tested against a larger set. The key elements to explore in the training phase were (i) the function for the force constant distance dependence, (ii) the effects of disconnecting/connecting sequential and spatial relationships, (iii) the optimal threshold for distant interactions, and (iv) the pre-exponential factors (C^{seq} and C^{cart}; see eq 8 for an explanation) used to scale residue−residue stiffness. Our purpose was to define a limit for relevant contacts to infer general connectivity rules. Nevertheless, a multiparametric fitting of all these elements to MD may lead to an overtrained method, and thus, we decided to follow a conservative stepwise strategy to guarantee its generality and physical sense.

3.1.1. Definition of a Sequential Threshold for Nearest-Neighbor Interactions. We first analyzed the distance dependence of the apparent inter-residue force constant detected in MD. The results in Figure 1 (left) show that in the limit of uncoupled oscillators (see eq 4) the apparent force constant decays exponentially with the $C^\alpha - C^\alpha$ distance; similar findings were obtained by Hinsen et al.[18] However, there are evident deviations at distances corresponding to $i \rightarrow i + 1$ residue interactions (close to 3.8 Å) and to a lesser extent

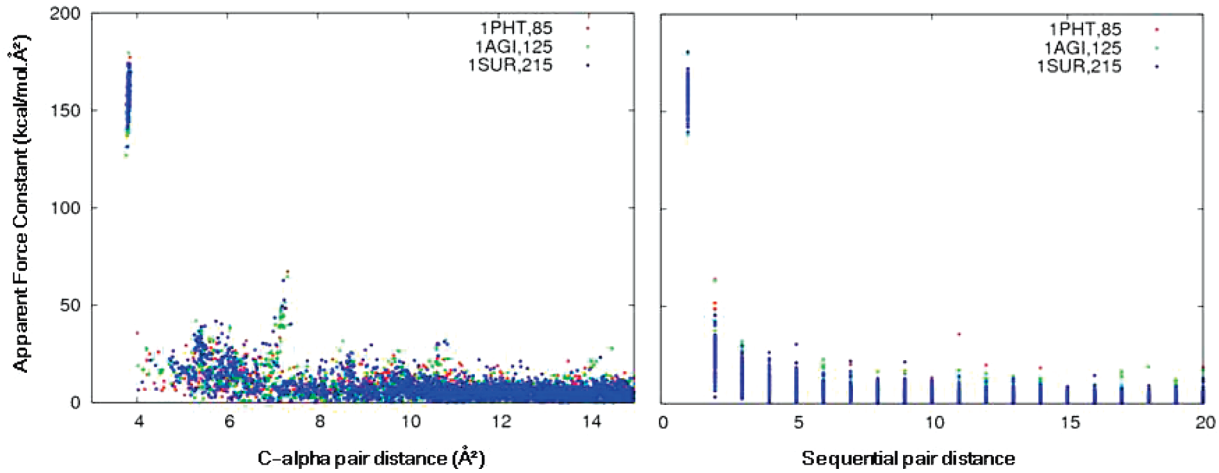

Figure 1. Dependence of the apparent residue−residue force constant (K_{ij}^{app}; see eq 4) on the residue−residue distance in Cartesian (left, Å) and sequence (right) space as determined for MD in three proteins of different sizes from the training set (see the Methods).

to $i \rightarrow i + 2$ and $i \rightarrow i + 3$ sequence interactions (around 6−10 Å). The singular nature of these nearest-neighbor interactions becomes evident in a plot of the apparent force constant dependence on the sequential distance (Figure 1, right). Fitting of force constants to this sequence distance for close chain neighbors reveals an order 2, inverse square exponential relationship ($n_{seq} = 2$ in eq 8) which has been incorporated into the algorithm for $i \rightarrow i + 1$ to $i \rightarrow i + 3$ contacts. This formalism defines the relative strength of the interactions between the first three neighbors as approximately 10^2:10^1:10^0 (in order of magnitude), a ratio that we found important to capture mode directionality. The pre-exponential factor (C^{seq}) appearing in eq 8 cannot be taken directly from MD apparent force constant profiles in Figure 1 and must be refitted to avoid over-restriction of the movement (see the discussion below). To further explore the effects of connectivity, other definitions of the chained residues were analyzed in simpler networks, where an increasing range of sequential contacts was weighted over a background of binary 1/0 contacts for cutoffs from 7 to 25 Å, confirming the i, $i + 3$ limit for main chain interactions (see Figure S3 in the Supporting Information). These simple networks also show that the overlaps follow a peak distribution around maximal values (from 8 to 15 Å) which becomes wider and shifts to higher cutoffs as the chain length increases. It is worth note that the one-neighbor sequence list (topologically equivalent to the constants scaling as 100:1:1 proposed by Jeong et al.[22]) gives suboptimal results, suggesting that $i \rightarrow i + 2$ and $i \rightarrow i + 3$ backbone contacts must be clearly weighted over the background defined by a cutoff of ≥8 Å. The extension of the sequential singularity to $i \rightarrow i + 5$ interactions did not yield any improvement, as could be anticipated from Figure 1. Interestingly enough, the deletion of distant sequential interactions in a fully connected, continuous network had negligible effects (see Figure S4, top, in the Supporting Information), and conversely, a subminimal NMA model, where only sequential-based $i \rightarrow i + 1$ to $i + 3$ level interactions were included, provided a quite striking agreement with ED modes (FigureS4, bottom). These findings suggest that interactions between sequence neighbors (related to torsional angles defining the secondary structure) are very important to define the preferred directions of large-scale motions, and therefore, proteins behave as robust networks of reduced connectivity regarding their near-equilibrium dynamics.

3.1.2. Scaling of the Force Constant Energies and the Distance Threshold for Spatial Interactions. When sequential interactions are removed from Figure 1 (right), the apparent force constants are found to decay with the distance following an order 6 exponential ($n_{cart} = 6$ in eq 8). This sixth-order inverse power law mirrors the distance dependence of the weak, long-range electrostatic interactions determining the 3D fold. Such a dependence, previously proposed by other authors,[18,19,23] was incorporated into the method, whereas the pre-exponential factor (C^{cart} in eq 8) was further refined against structural variance plots to scale the energy; a size-dependent distance cutoff was introduced to avoid over-restraint of the motions (see the discussion below). In summary, the ed-ENM model treats the strong covalent interaction between nearest-neighbor residues with an order 2 sequentially decaying power law, whereas long-range contacts follow the well-known sixth power law. Once this optimal function was determined, we fitted the force constants by comparison with ED estimates of the (i) total variance, (ii) variance profile, and (iii) reduced variance in order to scale the amplitude distribution of the modes. As mentioned above, we explored values for the sequential, C^{seq} (in the range of 40−200 kcal/(mol·Å^2)), and Cartesian, C^{cart} (in the range of 2−12 kcal/(mol·Å^2)), constants, finding optimal agreement in the training set for $C^{seq} = 60$ kcal/(mol·Å^2) and $C^{cart} = 6$ kcal/(mol·Å^2). The results are robust to changes of ±10 kcal/(mol·Å^2) in C^{seq} and ±1 kcal/(mol·Å^2) in C^{cart}, particularly regarding the mode directions (see Figure S5 in the Supporting Information).

Analysis of Figure 1 and inspection of the ED of training trajectories reveal that there is a threshold distance from which the apparent restriction in the movement of two pairs of residues is very small and can be explained only by indirect interactions (see Figure S2 in the Supporting Information). This recommends the use of a cutoff to eliminate restrictions to protein movement due to distant negligible interactions. We systematically compared the

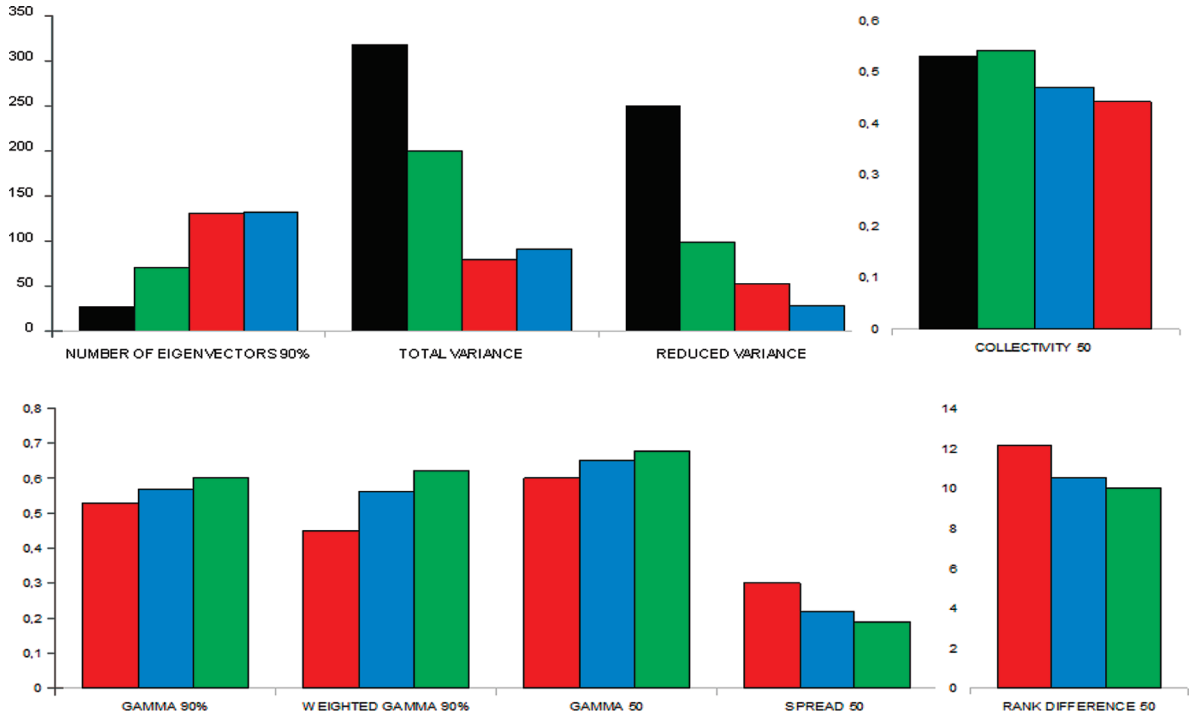

Figure 2. Different metrics for the comparison between MD and ENM-NMA: black, reference MD simulations; green, present ed-ENM model; red, standard cutoff model; blue, Kovac's formalism.

cutoff, continuous, and mixed cutoff/continuous approaches for a wide range of threshold distances (2−25 Å). Optimum cutoff values were determined by analyzing the dot products between ENMs and ED modes (eq 10) and the relative variance profiles, the magnitudes greatly independent of the value of the force constants. The best results were obtained with mixed continuous/cutoff approaches, and the optimum distance threshold was found to be roughly dependent on the protein size. The size dependence of the cutoff is also clear in the simpler networks tested in Figure S3 in the Supporting Information. We found that the optimal cutoff can be formulated as an approximate logarithmic function of the chain length, starting with a minimal value of 8 Å for the smallest proteins (see the function in Figure S6 in the Supporting Information). This soft size-dependent cutoff (resulting in a practical range from 10 to 16 Å for average to big proteins) removes irrelevant contacts, without affecting important structural details. Subsequently, we will always use this automatic procedure for the cutoff definition in our ed-ENM, avoiding then an arbitrary selection for each protein.

3.2. Validation of the Method. *3.2.1. Validation against MD Flexibility Data for a Representative Benchmark.* As described above, we analyzed the behavior of the new model by comparing our ed-ENM with MD metatrajectories in an extended set of proteins. The reference standard methods used for comparison were the original cutoff approach and the sixth-power exponential function developed by Kovacs et al. (see the Methods). In all cases NMA calculations were performed by taking the MD-averaged structure as a reference to allow direct comparison between the normal modes and MD. All ENMs considered reproduce the ED flexibility pattern reasonably well. Average results for the full μMODEL set displayed in Figure 2 and more detailed results for representative proteins in Table 1 illustrate that any of the ENMs are able to capture the overall features of the MD samplings as expected. Similarity indexes with respect to ED (for 90% variance; see eq 10) are in the range of 0.5−0.6 (0.6−0.7 if the index is computed considering 50 eigenvectors), with highly significant associated Z_{score} values (see Table 1 and Figures S7 and S8 in the Supporting Information). These similarity indexes are in fact not far from those obtained by comparing MD trajectories from different force fields among them in the range of 0.7−0.8 (see Table S1 in the Supporting Information; see also ref 54). However, there are considerable differences in the performance of the different methods, and the ed-ENM leads to a moderate but significant increase of around 3−5% in the average similarity index, from 0.54/0.57 to 0.60 in $\gamma_{90\%}$ and from 0.61/0.65 to 0.68 for γ_{50}. Most noticeably, the greatest improvement when using ed-ENM is centered on the prevalent eigenvectors, as shown by the variance-weighted Perez similarity index (see eq 12) for the 90% threshold, which increases from 0.45/0.56 to 0.62 (see $\gamma_{90\%}$ in Table 1), and similar increases in the raw overlap between the eigenspaces defined by the first 10 low-frequency modes (see γ_{10} in Table 1). This close correspondence between MD and ed-ENM lowest frequency motions also becomes clear in the corresponding spread values, lower than those obtained with standard methods (see the average in Figure 2, bottom, and profiles for a few proteins, Figure 4, left).

Analysis of total variances and variance profiles reveals some of the most serious shortcomings of the standard ENMs. First, they underestimate the MD total variance (by a factor of 3−4-fold; see Table 1 and Figure 2, top right), which means that in ENM samplings the structure is too rigid, and this cannot be detected when using crystal flexibility as a reference. Note that the ENM-MD deviation

Table 1. Comparative Measurements of Flexibility Patterns Obtained with NMA and ED of Selected Proteins

PDB code (CATH)	total variance[a]	no. of eigenvectors[a] (90% variance)	similarity (γ_{10})[c]	similarity ($\gamma_{90\%}$)[b,c]	Z_{score}[c] (90% variance)	similarity (γ_{50})[c]	Z_{score}[c] (50 eigenvectors)	Pearson coefficient[c,d]
1OPC 99 (α)	201/67/56/140	19/46/96/44	0.46/0.49/0.48	0.56/0.59/0.61 0.49/0.60/0.60	26/29/31	0.63/0.68/0.70	93/104/109	0.50/0.59/0.65 0.33/0.25/0.39
1CSP 67 (β)	86/45/46/73	20/30/61/38	0.51/0.54/0.61	0.62/0.64/0.68 0.61/0.68/0.72	37/39/44	0.64/0.70/0.72	64/75/79	0.46/0.55/0.71 0.49/0.54/0.62
1SDF 67 (α + β)	460/76/92/556	7/15/38/9	0.48/0.53/0.53	0.43/0.43/0.49 0.16/0.22/0.52	23/23/28	0.66/0.63/0.67	48/43/50	0.76/0.77/0.79 –
1OOI 124 (α)	131/38/53/103	37/131/133/74	0.28/0.36/0.40	0.59/0.66/0.68 0.47/0.21/0.69	20/34/38	0.63/0.71/0.72	127/149/151	0.40/0.61/0.60 0.23/0.46/0.65
1BFG 126 (β)	85/27/52/75	54/166/143/94	0.44/0.49/0.51	0.62/0.67/0.71 0.66/0.73/0.74	37/50/61	0.62/0.66/0.70	145/158/170	0.39/0.58/0.59 0.30/0.30/0.50
1CHN 126 (α + β)	359/138/71/160	15/29/118/62	0.46/0.47/0.49	0.48/0.52/0.53 0.38/0.52/0.55	19/23/24	0.61/0.66/0.68	131/146/151	0.54/0.68/0.74 0.35/0.62/0.53
1IL6 166 (α)	840/43/105/252	9/164/139/77	0.50/0.50/0.49	0.49/0.50/0.50 0.09/0.28/0.43	27/28/28	0.60/0.66/0.66	95/109/109	0.68/0.81/0.83 –
1CZT 158 (β)	197/42/112/146	38/140/140/97	0.42/0.49/0.49	0.58/0.65/0.69 0.54/0.70/0.72	42/54/61	0.60/0.65/0.69	111/124/134	0.51/0.56/0.72 0.66/0.67/0.77
1GND 430 (α + β)	1022/83/248/484	30/521/409/214	0.45/0.51/0.51	0.53/0.56/0.58 0.27/0.32/0.65	23/25/27	0.56/0.61/0.62	330/363/370	0.75/0.77/0.72 0.48/0.57/0.53
1BR5 267 (α)	185/47/150/274	85/353/261/146	0.40/0.44/0.45	0.62/0.68/0.68 0.56/0.73/0.72	41/59/59	0.58/0.64/0.64	200/225/225	0.65/0.71/0.73 –
2PIA 321 (β)	255/69/210/364	96/366/305/162	0.54/0.59/0.60	0.60/0.65/0.66 0.56/0.63/0.71	33/43/46	0.57/0.62/0.62	170/189/189	0.55/0.60/0.62 0.49/0.52/0.50
2HVM 273 (α + β)	376/32/112/183	44/449/307/184	0.41/0.45/0.45	0.55/0.61/0.60 0.27/0.55/0.61	33/43/42	0.56/0.62/0.61	177/200/196	0.68/0.84/0.81 –

[a] Values in the cells always correspond to the MD/cutoff NMA/Kovac/ed-ENM method. [b] Values in the first line of the cells correspond to the standard Hess metrics (eq 10) and values in the second line to the Perez index (eq 12). In every line the results displayed correspond to the cutoff NMA/Kovac/ed-ENM method. [c] Values in the cells correspond to the cutoff NMA/Kovac/ed-ENM method. [d] Values in the first line of the cells correspond to correlations against ED atomic fluctuations and values in the second line to correlations against experimental *B* factors. In every line the results displayed correspond to the cutoff NMA/Kovac/ed-ENM method.

in variance cannot be fully explained by the fact that we are using an MD metatrajectory as a reference, since it is also evident in single trajectories (see Table S1 in the Supporting Information). Interestingly, the deviation in variance with respect to MD simulations is not uniform for the entire deformation space (which would allow the correction by scaling force constants), but it is larger for the first essential movements, as shown by the reduced variance (see Figure 2, top left). In other words, the MD deformation space is larger (in terms of variance) but less complex (i.e., fewer eigenvectors are required to explain a given variance threshold) than the space described by standard ENMs (see Figure 3, left). The reason for this behavior is clear from the analysis of the variance profiles and the force constants (K_ν in eq 9) associated with essential deformations. The standard ENMs and MD simulations distribute variance along the different modes in a different way: while MD defines a small number of soft, highly collective movements which concentrate most of the variance, in ENMs the deformability is distributed along a larger number of eigenvectors. In summary, not only is the total variance different, but MD and standard ENMs also differ in how this variance is partitioned between modes as discussed before, and this is something that cannot be corrected by scaling a uniform spring constant, since it is more related to the topological properties of the network.

All metrics indicate that ed-ENM yields a remarkable improvement in the total variance and, more important, in the balance of deformation movements as noted in the reduced variance, force constant (K_ν in eq 9; Figure 3, right) profiles, and complexity (i.e., number of eigenvectors to capture a certain variance threshold) of the deformation space (see Table 1; μMoDEL averages in Figure 2, top). It is worth noting that the improvement obtained by using the ed-ENM model is mainly focused on the softest, low-frequency modes and is constant for all the size ranges of proteins considered and for all structural families, as shown by selected examples in Table 1. These soft modes of deformation are highly cooperative, involving a great part of the molecule, as shown by their collectivity degree.[52] The amount of residues involved in essential movements is similar in ENM and MD according to the Brüschweiler index (0.4−0.6 average for the first 50 modes), but there is a uniform tendency of standard NMA to less collective movements (Figure 2, top right), a situation that is corrected in ed-ENM, possibly due to the strongest nearest-neighbor coupling. Projection of the

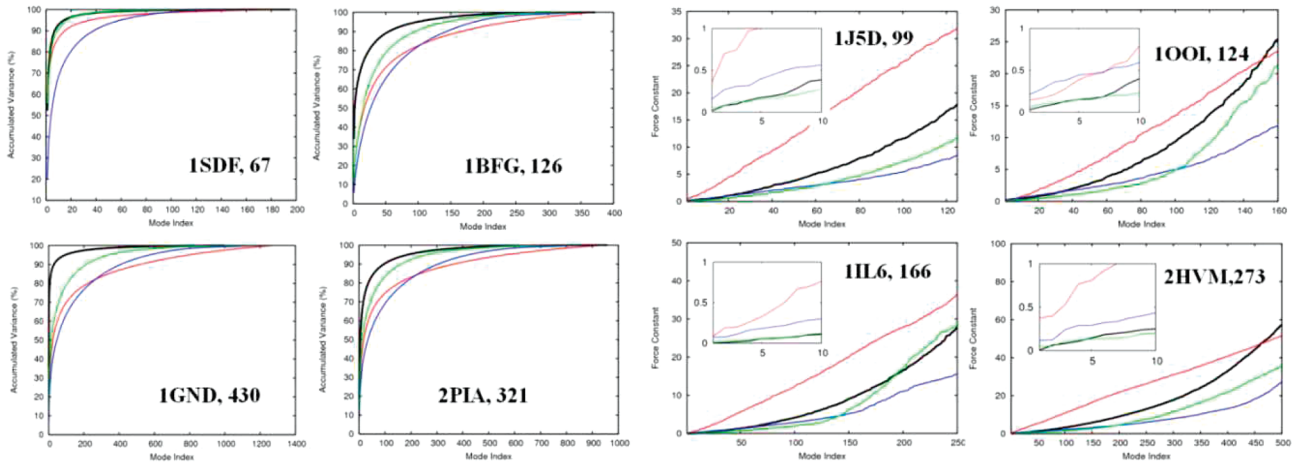

Figure 3. Cumulative variance with respect to the number of eigenvectors (left) and strengths of the essential deformation modes (right, K_ν in eq 9) computed by the different methods for some typical proteins (the inset corresponds to a zoom of the first eigenvalues). Illustrative proteins of different sizes and secondary structure compositions are displayed (the name and number of the residues are shown in each graph). The color code is as in Figure 2.

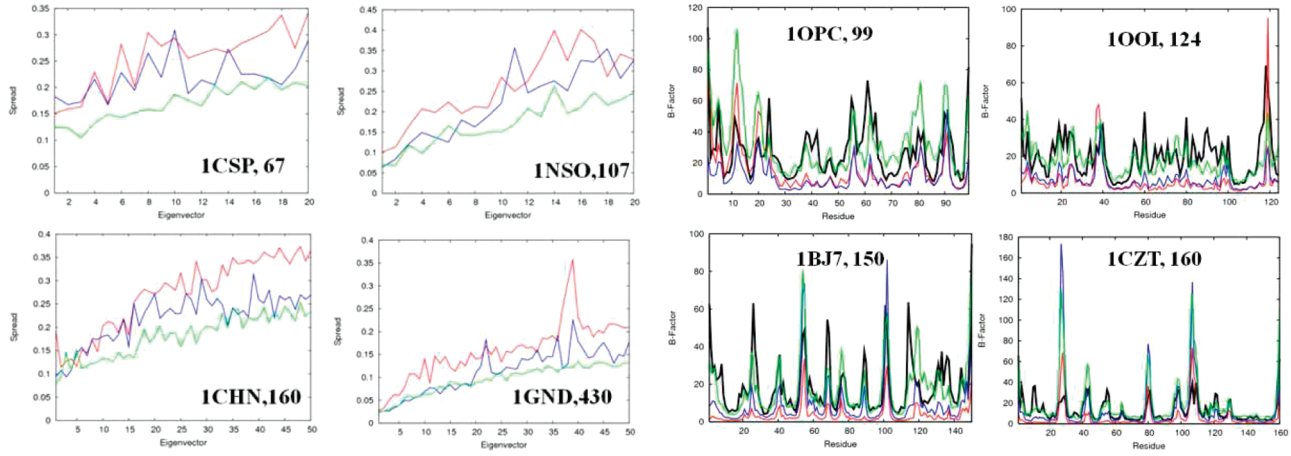

Figure 4. Spread of the eigenvectors in the ED eigenspace for randomly selected proteins (left). *B* factor profiles (Å^2) computed by the different methods for a selected number of proteins (right). Values obtained considering in all cases movements along the first 50 eigenvectors. The color code is as in Figure 2.

collective modes on individual residues allowed us to estimate residue fluctuations in solution (see the Methods, eq 16). As previously reported,[10,42] all ENMs reproduce (see Table 1) the MD atomic fluctuations reasonably well, with Pearson's correlation factors in the range of 0.5−0.6 (Spearman's coefficients typically 0.7−0.8). However, when individual fluctuation distributions are compared (see Figure 4, right), the shortcomings of standard ENMs become evident in a flattening of the profiles, resulting from the problems of ENM in capturing local but large nonharmonic deformations. It is also worth noting that even our interest was not in the description of flexibility in the crystal but that in solution, where the ed-ENM approach also yields a slight improvement (see Figure 4 and Table 1) in the X-ray *B* factor profiles. We also found that it is possible to raise the correlations for *B* factors by increasing the distance threshold (unpublished data), but as a result the structure becomes stiffened and the accuracy decreases in other global flexibility measurements, such as the similarity index, variance profiles, or overlap with transition vectors (see below). A simple postprocessing of positional fluctuations allows the derivation of Lindemann's index (see the Methods, eq 17), a key descriptor to analyze the macroscopic nature of proteins. The results in Table S2 in the Supporting Information illustrate the superiority of the ed-ENM with respect to the standard methods to estimate the absolute MD-derived Lindemann index. The ed-ENM nicely reproduces the core/surface (solid/liquid) asymmetry of proteins and the different macroscopic behavior of the main classes of secondary structures.

3.2.2. Robustness for Large Proteins and Extended Simulations. All results reported to this point suggest that the ed-ENM provides a better approximation of protein flexibility in solution when compared to standard ENM models. There are, however, two reasons for concern that have not yet been addressed regarding the behavior of the model on the biologically relevant time and length scales: (i) What happens when large proteins are considered (larger than those analyzed during the calibration)? (ii) What happens when the new ed-ENM is compared with the flexibility description obtained from long trajectories, where the protein is expected to display larger nonharmonic deformations? To answer the first question, we extended our study to several large and multimeric proteins (from 600 to 2500 residues), finding that ed-ENM captures well their fundamental dynamics (see

Table 2. Rmsd (Å) between X-ray Conformations, Overlaps (%) between Essential Deformation Spaces[a] and the Transition Vector, and Rank of Maximum Overlap[b] for the Cutoff, the Inverse Exponential Model, and the ed-ENM[c]

length (no. of residues) (CATH)	PDB code	rmsd	γ_5[d]	γ_{10}[d]	rank difference[d,e]
101	1L5E (open)	8.8	0.76/0.43/0.81	0.81/0.76/0.85	0 (0.70)/0 (0.66)/0 (0.65)
	1L5B (closed)		0.83/0.80/0.81	0.86/0.85/0.87	1 (0.27)/1 (0.28)/2 (0.55)
148	1CFD (open)	10.2	0.88/0.93/0.94	0.93/0.94/0.95	1 (0.38)/0 (0.62)/1 (0.55)
	1CFC (closed)		0.83/0.89/0.89	0.93/0.92/0.94	1 (0.55)/0 (0.45)/1 (0.55)
214	4AKE (open)	8.3	0.90/0.90/0.92	0.93/0.92/0.93	0 (0.67)/0 (0.38)/0 (0.67)
	1AKE (closed)		0.55/0.57/0.64	0.61/0.68/0.71	0 (0.32)/0 (0.36)/0 (0.40)
219	1NBV (H) (open)	2.2	0.69/0.69/0.70	0.73/0.72/0.73	2 (0.68)/0 (0.29)/0 (0.32)
	1CBV (H) (closed)		0.68/0.69/0.71	0.72/0.71/0.72	2 (0.38)/2 (0.40)/0 (0.37)
271	1URP (open)	7.7	0.96/0.93/0.95	0.96/0.95/0.97	1 (0.94)/1 (0.72)/1 (0.80)
	2DRI (closed)		0.83/0.82/0.88	0.86/0.88/0.92	0 (0.62)/0 (0.56)/1 (0.71)
317	1CKM (A) (open)	4.3	0.93/0.91/0.93	0.94/0.93/0.95	0 (0.86)/0 (0.44)/0 (0.88)
	1CKM (B) (closed)		0.21/0.49/0.57	0.65/0.73/0.78	6 (0.29)/2 (0.14)/4 (0.23)
320	3DAP (open)	5.8	0.89/0.90/0.93	0.94/0.92/0.95	0 (0.75)/1 (0.58)/0 (0.68)
	1DAP (closed)		0.20/0.18/0.27	0.44/0.62/0.78	9 (0.19)/7 (0.33)/4 (0.22)
401	9AAT (open)	2.2	0.15/0.07/0.55	0.68/0.64/0.71	5 (0.26)/5 (0.45)/4 (0.44)
	1AMA (closed)		0.07/0.08/0.60	0.68/0.67/0.76	6 (0.30)/5 (0.39)/6 (0.30)
452	1BNC (open)	5.4	0.84/0.85/0.87	0.87/0.90/0.90	0 (0.83)/0 (0.71)/0 (0.81)
	1DV2 (closed)		0.70/0.69/0.76	0.77/0.80/0.85	4 (0.22)/0 (0.40)/0 (0.48)
517	1RKM (open)	5.8	0.93/0.92/0.93	0.94/0.94/0.95	0 (0.91)/0 (0.84)/0 (0.92)
	2RKM (closed)		0.62/0.64/0.67	0.68/0.75/0.73	1 (0.32)/0 (0.52)/0 (0.42)
avg	open		0.66/0.67/0.70	0.75/0.75/0.76	0.9 (0.69)/0.6 (0.57)/0.6 (0.67)
	closed		0.51/0.53/0.58	0.65/0.67/0.69	3.0 (0.35)/1.7 (0.38)/1.7 (0.42)

[a] Considering the first 5 and 10 eigenvectors. [b] See the Methods for a description of the different metrics. [c] Examples from the NMAfit benchmark. [d] Values in these columns are for the cutoff/inverse/ed-ENM. [e] In parentheses, the dot product of the maximal overlap vector is given.

Table S3 and Figures S9 and S10 in the Supporting Information). This confirms that the method can be transferred to analyze large systems, difficult to tackle by MD simulations. The second challenge was to compare the ed-ENM modes to those derived from long MD trajectories (from 0.1 to 0.5−1 μs), where nonharmonic movements are likely to have more impact on the dynamics. Once again, all the metrics demonstrate the robustness and generality of the ed-ENM, in particular, in correcting the splitting of the soft modes observed in standard approaches (see Table S4 and Figures S11 and 12 in the Supporting Information). However, though the variance descriptors remain at the same order of magnitude, there is a uniform tendency for all ENMs to lower the similarity indexes when the time span of the MD is exteded (see similarity index values falling from 0.6−0.7 to 0.4−0.5 for 1CQY and 1OPC in Table S4). This is not surprising since, in a longer trajectory, the structures are able to explore a wider conformational subspace and thus undergo anharmonic departures from equilibrium that cannot be fully captured by any NMA-based approach as discussed above.

3.2.3. Validation against Empirical Flexibility Data from X-ray and NMR. Finally, we tested the method against experimental data on flexibility from both X-ray conformer transitions and PCA of selected NMR ensembles (see the Methods). First, we analyzed the ability of ed-ENM to predict functional important closed/open transitions between X-ray conformers. These large-scale rearrangements involve cooperative motions of domains or subunits, behaving as rigid clusters but preserving the overall fold; in this case the local cohesion prevails over interdomain, long-range interactions. Hence, a great shortcoming in continuum ENM approaches is the over-restriction of displacements between domains, as noticed before.[23] On the other hand, ENM cutoff approaches display a difficult balance between violation of dihedral constraints for lower distance thresholds and over-restriction of motions if increased. We expected that the combination of a sixth power law with a soft size-dependent cutoff, together with the strongest, inverse-square cohesion limited to neighbors, would allow more natural internal movements. To verify our hypothesis, we studied a benchmark of selected conformational transitions from the macromolecular motion database MolMovDB[54] (http://www.molmovdb.org). Average results for the full benchmark (54 structures) and detailed data for 10 selected cases are displayed in Table 2: 4 structures undergoing large transitions (rmsd > 7 Å) and 6 more with local, less dramatic changes (rmsd = 2−6 Å). The results show that all ENMs encode the functional transitions in their intrinsic flexibility, but the ed-ENM provides the best agreement between the transition vector and the harmonic deformation space. In the open forms, considering only the first 5 modes, the overlaps range from around 0.60 to 0.95 (average 0.7) and from 0.70 to 0.97 (average 0.76) if the harmonic space is extended to 10 modes (see γ_5 and γ_{10} in Table 2); note that random deformations would yield overlaps around 0.08 (5 eigenvectors) and 0.16 (10 eigenvectors). There is a systematic trend to better performance of the ed-ENM (2−5%) regardless of the extent of the transition, particularly remarkable when considering only the first five dominant modes. The greatest improvement using the ed-ENM is achieved for the closed forms, more difficult to treat since they can be easily overconstrained by long-range springs: in this case γ_5 increases by nearly 10% (from 0.50 in standard approaches to almost 0.60). The agreement is particularly surprising in the most challenging cases, where other ENMs fail dramatically (see, for example, the *closed* → *open* transition for 1CKM (B), 1AMA, and 1DAP). These notable differences

Table 3. Cumulative Overlaps between the First 10 (γ_{10}) Normal Modes and PCs from NMR Ensembles and Largest Overlap (γ_{max}) between an ENM Mode and the Best Overlapped PC for Each Set (See Eq 10) for 26 Proteins

PDB code	*N*	*M*	$\gamma_{(5)}$[a]			$\gamma_{(10)}$[a]			γ_{max}		
1RO4	58	35	0.56	0.52	0.58	0.57	0.53	0.62	0.59	0.33	0.72
1E9T	59	59	0.51	0.48	0.63	0.53	0.54	0.58	0.48	0.57	0.64
1BW5	66	50	0.69	0.59	0.74	0.56	0.56	0.62	0.53	0.63	0.78
2EOT	74	32	0.52	0.42	0.61	0.56	0.41	0.55	0.57	0.76	0.54
1A6X	87	49	0.55	0.44	0.56	0.41	0.37	0.48	0.54	0.37	0.95
1BVE	99	28	0.47	0.52	0.49	0.33	0.36	0.37	0.81	0.7	0.88
1Q06	101	55	0.57	0.58	0.57	0.51	0.60	0.57	0.51	0.58	0.77
2CZN	103	38	0.62	0.55	0.66	0.50	0.53	0.51	0.87	0.65	0.70
1A90	108	31	0.36	0.44	0.49	0.38	0.40	0.42	0.65	0.49	0.83
2BO5	120	44	0.54	0.37	0.58	0.55	0.45	0.56	0.73	0.57	0.68
1E5G	120	50	0.71	0.70	0.69	0.60	0.64	0.63	0.93	0.90	0.96
1CMO	127	43	0.70	0.61	0.70	0.56	0.52	0.59	0.56	0.60	0.52
1ITI	133	31	0.53	0.65	0.59	0.46	0.51	0.44	0.78	0.78	0.89
1C89	134	40	0.70	0.76	0.80	0.55	0.60	0.63	0.53	0.69	0.63
1XSB	153	39	0.46	0.43	0.49	0.44	0.38	0.43	0.89	0.41	0.92
1BF8	205	20	0.47	0.55	0.54	0.43	0.47	0.48	0.86	0.78	0.88
1BY1	209	20	0.55	0.55	0.56	0.42	0.46	0.47	0.50	0.65	0.53
1N6U	212	22	0.63	0.61	0.59	0.58	0.58	0.58	0.64	0.37	0.60
2JZ4	299	20	0.53	0.51	0.61	0.41	0.40	0.45	0.49	0.80	0.74
2D21	370	20	0.60	0.56	0.65	0.50	0.47	0.50	0.67	0.62	0.62
avg			0.56	0.54	0.61	0.49	0.49	0.53	0.65	0.61	0.74

[a] Values in these columns are for the cutoff/inverse/ed-ENM.

are related to the concentration of the conformational change in the first dominant eigenvectors. Accordingly, the rank differences are often also smaller and the best overlapped eigenvectors closest to the transition direction. In summary, the ed-ENM displays a higher cooperativity and less dispersion of the motions—as in the above comparison with ED—and thus traces the functional changes with fewer modes.

Finally, we analyzed the ability of ed-ENM to approach the structural diversity of NMR ensembles, which in a first approach can be related (not in a fully rigorous manner) to the experimental flexibility pattern. The analysis of 20 selected NMR multiple structures shows striking correlations with the three ENMs (see Table 3), which confirms previous results[31] and supports the validity of ENMs to sample the near-equilibrium conformational space in solution. It is also clear that the ed-ENM method outperforms the other two ENM approaches, especially when considering only the first 5 eigenvectors whose overlap γ_5 increases from 0.56/0.54 to 0.61 and the best overlapped pair (γ_{max}), which increases from a 0.65/0.61 average to 0.74, reaching values near 0.90 (see 1BVE, 1ITI, and 1BF8) or even above (1A6X, 1E5G, and 1XSB). In more than half of the proteins (11 cases), the best overlapped vector is found in the ed-ENM method, followed by the inverse (5 cases) and cutoff (4 cases) approaches, following the trend observed in the rest of the tests. In conclusion, the ed-ENM seems to provide a significant and systematic improvement in the description of protein dynamics (as deduced from structural diversity in NMR ensembles) with respect to the two most used ENM implementations.

4. Conclusions

The ability of the elastic network NMA models to predict qualitatively the intrinsic motions of proteins has been widely demonstrated in the past few years. In comparison with MD, ENMs tend to yield a sparser pattern of flexibility, related to their harmonic character, and then, the information required for a realistic description of a functional motion is dispersed into a higher number of modes.[42] Another problem of ENMs has been the lack of consensus in the refinement, mainly due to the scarcity of direct measurements of protein flexibility. In previous studies we demonstrated that MD gives an accurate picture of flexibility in solution.[37] In this work we have used atomistic simulations as an alternative source for ENM refinement to extract connectivity rules and obtain a realistic scaling of the force constants. These constraints led to the formulation of a new ED-refined ENM (ed-ENM), based on a simple hybrid potential considering chain topology, which has been validated against a database of MD trajectories. The method proposes a simple and robust scaling of the local backbone and long-range contacts, avoiding any arbitrary, free parameters. A soft size-dependent cutoff is applied to eliminate noise from irrelevant contacts and increase computational efficiency when dealing with large systems. Our goal was not to reproduce any particular flexibility measurement (such as *B* factor profiles), but rather to develop a general method able to trace protein flexibility better than or at least as well as the best performing standard approach for the widest range of descriptors and the largest variety of protein sizes and folds. As discussed above, higher scores for individual flexibility measurements can be achieved by problem-specific adjustment of the ENMs, but only compromising accuracy in other aspects. For example, large cutoffs boost correlations with *B* factors, but the resulting more rigid structures cannot display large conformational transitions. Clearly, when considering all the flexibility measurements presented here, the ed-ENM outperforms standard approaches in the representation of both local and global flexibility for a wide range of proteins and without any ad hoc adjustments. Comparisons to submicrosecond MD suggest that ed-ENM is flexible enough to partially capture

nonharmonic deformations. The method is robust, general, and transferable and can describe large conformational transitions required for biological activity. Finally, we have demonstrated that the method introduced here captures the flexibility of NMR structural ensembles with remarkable precision. Therefore, the bulk of results presented demonstrate that the ED-refined ENM can be a useful alternative to well-established coarse-grained NMA methods. The ability of a minimalist model based on close-chain neighbor interactions both to match molecular dynamics and to trace these complex transitions strongly supports the hypothesis that local covalent topology encodes an important part of the intrinsic flexibility pattern of proteins and thus guides biologically relevant conformational changes. Though this idea may appear somewhat counterintuitive, given the importance of long-range interactions for the 3-D fold, it is just outlining the fact that conformational transitions usually involve motions of rigid residue clusters that maintain the local fold and that this local fold is dependent on the nearest-neighbor contacts, sterochemically restrained. Thus, not only the global contact topology[11,54−57] but also the inner topology defined from nearest-neighbor contacts plays a great role in the determination of these lowest energy, intrinsically favored modes.

Acknowledgment. This work was supported by the Spanish Ministry of Education and Science (Consolider E-Science and Grant BIO2009-10964), the Spanish Ministry of Health (COMBIOMED project), the Fundación Marcelino Botín, and the National Institute of Bioinformatics. Calculations were performed on the MareNostrum supercomputer at the BSC. L.O. is funded by a predoctoral fellowship from the Spanish Health Ministry.

Supporting Information Available: Additional tables containing comparative metrics obtained with different force fields and the ed-ENM (Table S1), a comparison of Lindemann's coefficients (Table S2), and comparative metrics for large proteins and submicrosecond MD simulations to evaluate robustness in extended length (Table S3) and time (Table S4) scales and additional figures giving the percentage of variance captured by the first five modes in the benchmark (Figure S1), an illustration of the difference between direct and indirect interactions (Figure S2), similarity index profiles in sequential-based networks (Figure S3), similarity index profiles switching on/off the sequential constants (Figure S4), the robustness to changes in Cartesian/sequential constants (Figure S5), the size-dependent cutoff function (Figure S6), comparative similarity index and Z_{score} values for the benchmark (Figures S7 and S8), variance profiles and force constants for large proteins (Figures S9 and S10), and extended submicrosecond MD simulations (Figures S11 and S12). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Henzler-Wildman, K. A.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **2007**, *450*, 913–916.

(2) Velázquez-Muriel, J. A.; Rueda, M.; Cuesta, I.; Pascual-Montano, A.; Orozco, M.; Carazo, J. M. Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Struct. Biol.* **2009**, *9*, 6.

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

(4) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590.

(5) Dynamical simulation methods. In *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*; Brooks, C. L., III, Karplus, M., Pettitt, B. M., Eds.; Advances in Chemical Physics, Vol. LXXI; John Wiley & Sons Ltd.: New York, 1988; pp 33−58.

(6) Karplus, M.; Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679–6685.

(7) Amadei, A.; Linssen, A. B.; Berendsen, H. J. Essential dynamics of proteins. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 412–425.

(8) Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

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

(10) Emperador, A.; Carrillo, O.; Rueda, M.; Orozco, M. Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophys. J.* **2008**, *95*, 2127–2138.

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

(12) Case, D. A. Normal mode analysis of biomolecular dynamics. In *Computer Simulation of Biomolecular Systems*; Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1997; Vol. 3, pp 284−301.

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

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

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

(16) Chennubhotla, C.; Bahar, I. Markov methods for hierarchical coarse-graining of large protein dynamics. *J. Comput. Biol.* **2007**, *14*, 765–76.

(17) Sen, T. Z.; Jernigan, R. L. Optimizing the parameters of the Gaussian network model for ATP-binding proteins. In *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*; Cui, Q., Bahar, I., Eds.; CRC Press: Boca Raton, FL, 2006.

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

(19) Kovacs, J. A.; Chacon, P.; Abagyan, R. Predictions of protein flexibility: First-order measurements. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 661–668.

(20) Riccardi, D.; Cui, Q.; Phillips, G. N. Application of elastic network models to proteins in the crystalline state. *Biophys. J.* **2009**, *96*, 464–475.

(21) Moritsugu, K.; Smith, J. C. Coarse-grained biomolecular simulation with REACH, realistic extension algorithm via covariance Hessian. *Biophys. J.* **2007**, *93*, 3460–3469.

(22) Jeong, J. I.; Jang, Y.; Kim, M. K. A connection rule for a-carbon coarse-grained elastic network models using chemical bond information. *J. Mol. Graphics Modell.* **2006**, *24*, 296–306.

(23) Yang, L.; Song, G.; Jernigan, R. L. Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci.U.S.A.* **2009**, *106*, 12347–52.

(24) Wagner, G. NMR relaxation and protein mobility. *Curr. Opin. Struct. Biol.* **1993**, *3*, 748–754.

(25) Gabel, F.; Bicout, D.; Lehnert, U.; Tehei, M.; Weik, M.; Zaccai, G. Protein dynamics studied by neutron scattering. *Q. Rev. Biophys.* **2002**, *35*, 327–367.

(26) Erman, B. The Gaussian network model: Precise prediction of residue fluctuations and application to binding problems. *Biophys. J.* **2006**, *91*, 3589–3599.

(27) Hinsen, K. Structural flexibility in proteins: Impact of the crystal environment. *Bioinformatics* **2008**, *24*, 521–528.

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

(29) Soheilifard, R.; Makarov, D. E.; Rodin, G. J. Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. *Phys. Biol.* **2008**, *5*, 026008−026021.

(30) Carugo, O.; Argos, P. Reliability of atomic displacement parameters in protein crystal structures. *Acta Crystallogr., D: Biol. Crystallogr.* **1999**, *55*, 473–478.

(31) Tama, F.; Sanejouand, Y. H. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **2001**, *14*, 1–6.

(32) Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R. L. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* **2008**, *16*, 321–330.

(33) Yang, L.; Eyal, E.; Bahar, I.; Kitao, A. Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): Insights into functional dynamics. *Bioinformatics* **2009**, *25*, 606–614.

(34) Yang, L.; Eyal, E.; Chennubhotla, C.; Jee, J. G.; Gronenborn, A.; Bahar, I. Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure* **2007**, *15*, 741–749.

(35) Abseher, R.; Horstink, L.; Hilbers, C. W.; Nilges, M. Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. *Proteins: Struct., Funct., Genet.* **1999**, *31*, 370–382.

(36) Case, D. A. Molecular dynamics and NMR spin relaxation in proteins. *Acc. Chem. Res.* **2002**, *35*, 325–331.

(37) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Pérez, A.; Camps, J.; Hospital, A.; Gelpí, J. L.; Orozco, M. A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 796–801.

(38) Berendsen, H. J. C. Bio-molecular dynamics comes of age. *Science* **1996**, *271*, 954–955.

(39) Ichiye, T.; Karplus, M. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 205–217.

(40) Kitao, A.; Go, N. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164–9.

(41) Hayward, S.; Kitao, A.; Go, N. Harmonicity and anharmonicity in protein dynamics: A normal modes and principal component analysis. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 177–186.

(42) Rueda, M.; Chacón, P.; Orozco, M. Thorough validation of protein normal mode analysis: A comparative study with essential dynamics. *Structure* **2007**, *15*, 565–575.

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

(44) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(45) MacKerell, A. D.; Wiorkiewicz-Kuczera, J.; Karplus, M. An all-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.

(46) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(47) Hess, B. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E* **2000**, *62*, 8438–8448.

(48) Noy, A.; Meyer, T.; Ferrer, C.; Valencia, A.; Pérez, A.; de la Cruz, X.; López-Bes, J. M.; Pouplana, R.; Fernandez-Recio, J.; Luque, F. J.; Orozco, M. Data mining of molecular dynamics trajectories of nucleic acids. *J. Biomol. Struct. Dyn.* **2006**, *23*, 447–456.

(49) Orozco, M.; Perez, A.; Noy, A.; Luque, F. J. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* **2003**, *32*, 350–364.

(50) Pérez, A.; Blas, J. R.; Rueda, M.; López-Bes, J. M.; de la Cruz, X.; Orozco, M. Exploring the essential dynamics of B-DNA. *J. Chem. Theory Comput.* **2005**, *1*, 790–800.

(51) Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 417–429.

(52) Brüschweiler, R. Collective protein dynamics and nuclear spin relaxation. *J. Chem. Phys.* **1995**, *102*, 3396–3403.

(53) Zhou, Y.; Vitkup, D.; Karplus, M. Native proteins are surface-molten solids: Application of the Lindemann criterion for the solid versus liquid state. *J. Mol. Biol.* **1999**, *285*, 1371–1375.

(54) Krebs, W. G.; Alexandrov, V.; Wilson, C. A.; Echols, L.; Yu, H.; Gerstein, M. Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 682–695.

(55) Emperador, A.; Meyer, T.; Orozco, M. United-atom discrete molecular dynamics of proteins using physics-based potentials. *J. Chem. Theory Comput.* **2008**, *4*, 2001–2010.

(56) Nicolay, S.; Sanejouand, Y. H. Functional modes of proteins are among the most robust ones. *Phys. Rev. Lett.* **2006**, *96*, 078104.

(57) Zheng, W.; Brooks, B. R.; Thirumalai, D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7664–7669.

(58) Camps, J.; Emperador, A.; Carrillo, O.; Orellana, L.; Hospital, A.; Rueda, M.; Cicin-Sain, D.; D'Abramo, M.; Gelpí, J. L.; Orozco, M. FlexServ: An integrated tool for the analysis of protein flexibility. *Bioinformatics* **2009**, *25*, 1709–10.

CT100208E