Structure
Article
Cell PRESS

Thorough Validation of Protein Normal Mode Analysis: A Comparative Study with Essential Dynamics

Manuel Rueda,[1,3] Pablo Chacón,[4] and Modesto Orozco[1,2,3,5,*]
[1] Molecular Modeling and Bioinformatics Unit, Institut de Recerca Biomèdica, Parc Cientific de Barcelona, 08028 Barcelona, Spain
[2] Structural Bioinformatics Unit, Instituto Nacional de Bioinformática, Parc Cientific de Barcelona, Josep Samitier 1-5, 08028 Barcelona, Spain
[3] Departament de Bioquímica i Biologia Molecular, Facultat de Biologia, Universitat de Barcelona, Avgda Diagonal 645, Barcelona 08028, Spain
[4] Centro de Investigaciones Biológicas, Ramiro de Maeztu 9, Madrid 28040, Spain
[5] Computational Biology Program, Barcelona Supercomputer Centre, Jordi Girona 31, Edifici Torre Girona, Barcelona 08028, Spain
*Correspondence: modesto@mmb.pcb.ub.es

DOI 10.1016/j.str.2007.03.013

SUMMARY

The deformation patterns of a large set of representative proteins determined by essential dynamics extracted from atomistic simulations and coarse-grained normal mode analysis are compared. Our analysis shows that the deformational space obtained with both approaches is quite similar when taking into account a representative number of modes. The results provide not only a comprehensive validation of the use of a low-frequency modal spectrum to describe protein flexibility, but also a complete picture of normal mode limitations.

INTRODUCTION

Flexibility is a key determinant of protein function (Daniel et al., 2003; Eisenmesser et al., 2002; Hinsen et al., 1999; Luo and Bruice, 2004; Ma and Karplus, 1998; Remy et al., 1999; Sacquin-Mora and Lavery, 2006; Waldron and Murphy, 2003; Yang and Bahar, 2005), but, due to different technical problems, its systematic study has been possible only in recent years. From these analyses we know that the essential deformation space of proteins is related to the conformational space sampled by evolution in protein families (Leo-Macias et al., 2005; Qian et al., 2004), that side chains at protein-binding sites are "entropically trapped" even in the holo form of proteins (Bartlett et al., 2002), and that interactions at binding sites alter the entire dynamics of the protein (Ming and Wall, 2006). Many studies have shown the close relationship between protein function, or even catalysis, and collective dynamics (Yang and Bahar, 2005). Similarly, other studies have demonstrated that deformability patterns are guiding some allosteric transitions responsible for cooperativity in proteins (Gerstein et al., 1994; Ma and Karplus, 1998) All of these evidences suggest that a hidden flexibility code has been printed by evolution in the structure of biological macromolecules in order to optimize their biological action (Qian et al., 2004). The knowledge of protein flexibility is then crucial for understanding protein function and evolution.

Different experimental approaches have been developed to examine protein flexibility, but they are still not of general applicability and, in most cases, provide only rough measures such as atomic B factors. This has fuelled the use of theoretical approaches to study the deformability of equilibrium structures of macromolecules. Two main algorithms are used to compute essential deformations: (i) essential dynamics (ED), and (ii) normal mode analysis (NMA). In ED (Amadei et al., 1993), the deformation modes are obtained by diagonalization of the (mass-weighted) covariance matrix obtained from molecular dynamics (MD) or Brownian dynamics (BD) simulations, while, in NMA (Cui and Bahar, 2006; Go et al., 1983; Levitt et al., 1985), the deformation modes are obtained by diagonalization of the (mass-weighted) Hessian matrix. In the first case, a real potential trying to reproduce the physics of macromolecular interactions is used, without an a priori decision about the minimum energy structure of the macromolecule. In the second case, the known structure of the macromolecule is defined as a minimum, and the detailed atomic potentials are often replaced by simple harmonic or quasi-harmonic potentials between interacting atoms or residue pairs (Cui and Bahar, 2006; Tirion, 1996). Thus, despite the similarity, there are intrinsic differences between the way in which NMA and ED define the essential deformation pattern.

Ten years ago, Tirion pioneered the use of simplified potentials to study the deformation modes (Tirion, 1996). This idea was further extended to use coarse-grained (C_α) protein representation by several research groups, including those of Bahar (Bahar et al., 1997), Haliloglu (Haliloglu et al., 1997), Hinsen (Hinsen, 1998), Sanejouand (Tama and Sanejouand, 2001), Jernigan (Song and Jernigan, 2006), and Brooks (Zheng et al., 2006). Web-based database systems such as MolmovDB (Alexandrov et al.,

Table 1. Number of Eigenvectors Needed to Explain 90% of the Variance and Comparative Measures of Deformational Patterns Obtained with NMA and ED of Selected Proteins Grouped by Size and CATH Categories

Size	Protein (Residues)	CATH	Number of Eigenvectors, 90% Variance[a]	γ (90% ED)	Z Score (90% ED)	γ (50 Eigenvectors)	Z Score (50 Eigenvectors)
Small	1OPC (99)	α	46/104/18	0.56/0.59	70/74	0.63/0.68	82/92
	1FAS (61)	β	53/54/12	0.52/0.52	31/32	0.63/0.70	25/36
	1FVQ (72)	$\alpha+\beta$	78/84/30	0.62/0.68	49/57	0.68/0.75	59/72
Medium	1OOI (124)	α	130/145/37	0.59/0.66	135/155	0.63/0.71	128/150
	1BSN (138)	β	70/108/9	0.50/0.52	41/43	0.59/0.63	70/79
	1CHN (126)	$\alpha+\beta$	28/129/15	0.48/0.52	124/136	0.61/0.66	134/148
Big	1GND (430)	α	520/448/30	0.53/0.56	283/300	0.56/0.61	328/361
	1CZT (160)	β	139/152/38	0.58/0.64	112/130	0.60/0.65	102/114
	1SUR (213)	$\alpha+\beta$	172/198/19	0.58/0.61	192/202	0.63/0.68	201/221
Multidomain	1BR5 (267)	-	353/284/85	0.62/0.68	223/253	0.58/0.64	187/211
	2PIA (321)	-	366/331/96	0.60/0.65	209/235	0.57/0.62	179/198
	1E9S (2545)	-	3598/3114/790	0.55/0.60	2434/2681	0.42/0.44	2204/2327

For the last four columns, two definitions of the "important space" were used: (i) eigenvectors needed to explain 90% variance, and (ii) the first 50 eigenvectors (values in those columns are listed as distance cutoff NMA·ED/inverse exponential NMA·ED).

[a] Values in this column are listed as distance cutoff NMA/inverse exponential NMA/ED.

2005), ProMode (Wako et al., 2004), or iGNM (Yang et al., 2005) give access to numerous examples of the good correlation between low-frequency normal modes and the collective, large-amplitude observed motions in proteins. These tools are complemented by other web servers such as Elnémo (Suhre and Sanejouand, 2004), Webnm@ (Hollup et al., 2005), AD-ENM (Zheng and Doniach, 2003), Movies (Cao et al., 2004), UMMS (Jang et al., 2006), NOMAD-ref (Lindahl et al., 2006), oGNM (Yang et al., 2006), or Dfprot (Garzón et al., 2007), which also provide online normal-mode calculation with a variety of extra functionalities. Using these and other tools, NMA has been used to simulate protein deformations at extended-length scales (Bahar and Rader, 2005; Ma, 2005), or even to represent the flexibility of very low-resolution structures in which C_α cannot be located (Chacon et al., 2003; Kong et al., 2003). However, the extended use of the technique cannot hide the fact that coarse-grained NMA calculations still need validation by comparison with more detailed atomistic simulations based on physical potentials. Previous comparisons between NMA and MD simulations were limited to a few proteins and to short MD trajectories, often obtained in nonphysiological environments (Hayward et al., 1994, 1997). Thus, 10 years after the first coarse-grained NMA formulation, no evidence exists that the essential deformation space obtained by NMA properly represents that obtained in atomistic MD simulations in explicit water. Furthermore, both the individual predictive power of deformation modes obtained by NMA analysis and the similarity of the macroscopic flexibility properties of proteins derived by NMA and MD calculations are unknown. The lack of this benchmarking generates uncertainty about the quality and reliability of NMA-derived results to describe the flexibility of proteins in solution.

In this paper, we present a very wide comparison of coarse-grained NMA and atomistic MD-derived ED simulations on a data set containing all protein metafolds (Rueda et al., 2007). The study presented here, the largest done to our knowledge, represents a massive use of supercomputer resources (~100 CPU years) and provides a complete picture of the quality and limitations of coarse-grained NMA approaches.

RESULTS AND DISCUSSION

Stiffness Analysis

It is not possible to directly compare the size of the fluctuations given by the two methods. On one hand, the total MD variance depends on the length of the simulation (see Table S2 in the Supplemental Data available with this article online). On the other hand, the total variance explained by the NMA directly depends on the choice of the spring constant (see Figure S1), which was chosen to reproduce experimental B factors, mimicking then the reduced flexibility allowed by the crystal lattice (Rueda et al., 2007). Note that the rigidification of the system expected in NMA calculations performed with standard parameters can be corrected by reducing the Cα-Cα force constant (see Figure S2) or, alternatively, by scaling down the force constant associated with the normal modes (when necessary, the later approach was used in experiments described in this paper). In any case, we should emphasize that much caution is needed before translating NMA-detected fluctuations into sampling variances,

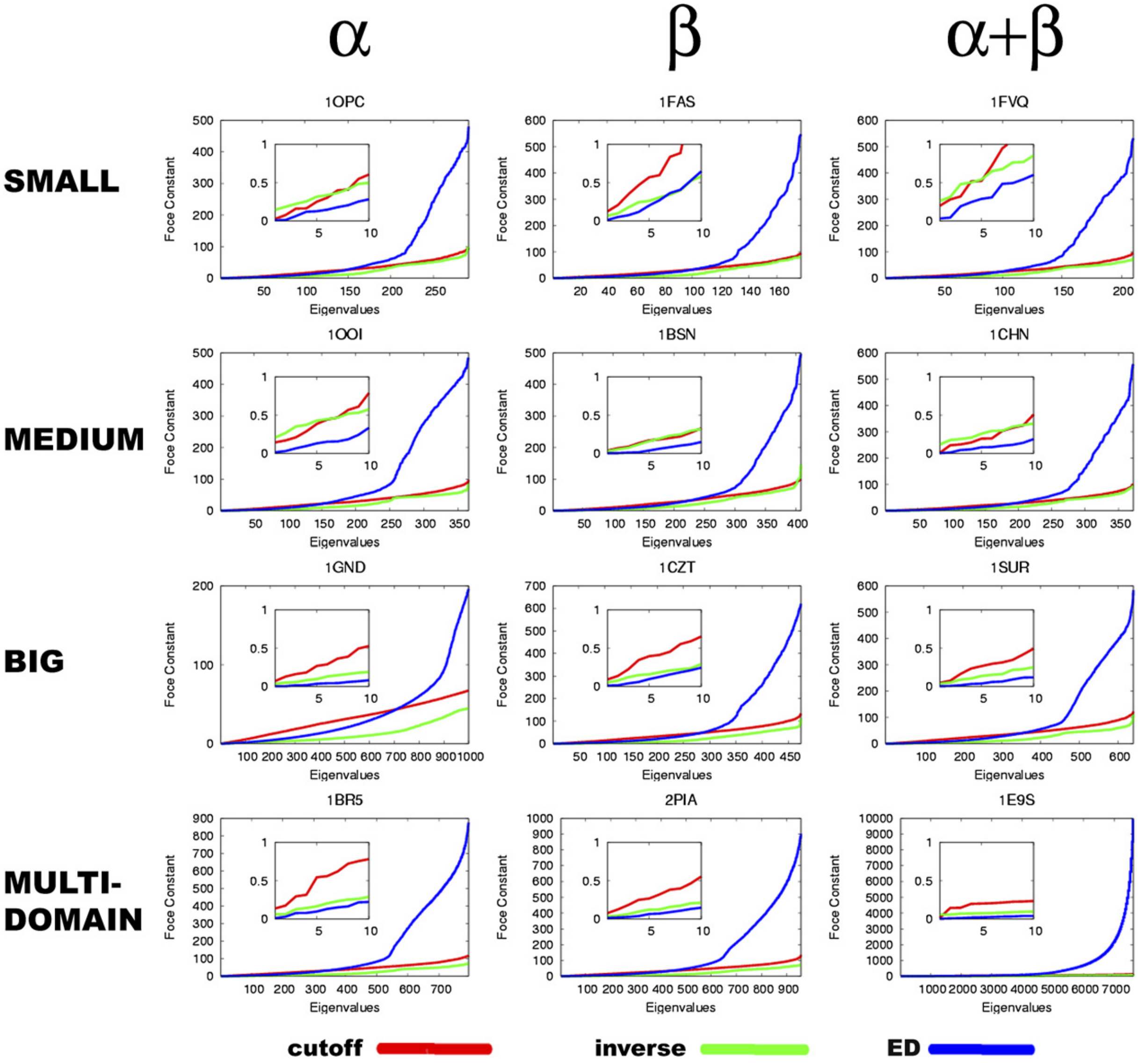

Figure 1. Force Constant Associated with Harmonic Deformations of Selected Proteins as Determined by ED and the Two NMA Methods

The force constant is measured as kcal/mol Å^2, harmonic deformations are ordered by rank, and selected proteins are classified according to size and CATH annotation. The insert corresponds to the first ten eigenvectors.

because the technique is not designed to properly reproduce total variance.

Our analysis shows that fewer modes are needed in ED to capture the same threshold of variance (see Table 1). Thus, a number of NMA eigenvectors approximately equal to the number of protein residues is needed to explain 90% variance, while a much smaller number is needed when using ED eigenvectors (see Table 1). Quite interestingly, the point that divides the low- and high (ED)-deformation modes is clearly found to be (approximately) 1.8 times the number of residues (see Figure 1), and such a sharp division among modes does exist in NMA calculations. In summary, the essential deformation space reported by ED is wider, but simpler (i.e., defined by a smaller number of essential deformations), than that suggested by the simpler NMA treatments. There is then a systematic and fundamental difference between NMA and ED that cannot be ignored (note that this limitation holds if other C_α-C_α force constants are used; see Figure S3). However, if we limit our interest to the low modes (see inserts in Figure 1), NMA is found to describe ED deformability quite well, especially after suitable scaling (see Figure S2).

Detailed analysis of our results (see Figure 1) shows that the nature of the protein, as quoted by the CATH family, does not have any obvious influence on the relative ED/NMA stiffness, and that the effect of the protein size (or the presence of several domains) is to expand the size of the "important space" without adding any apparent

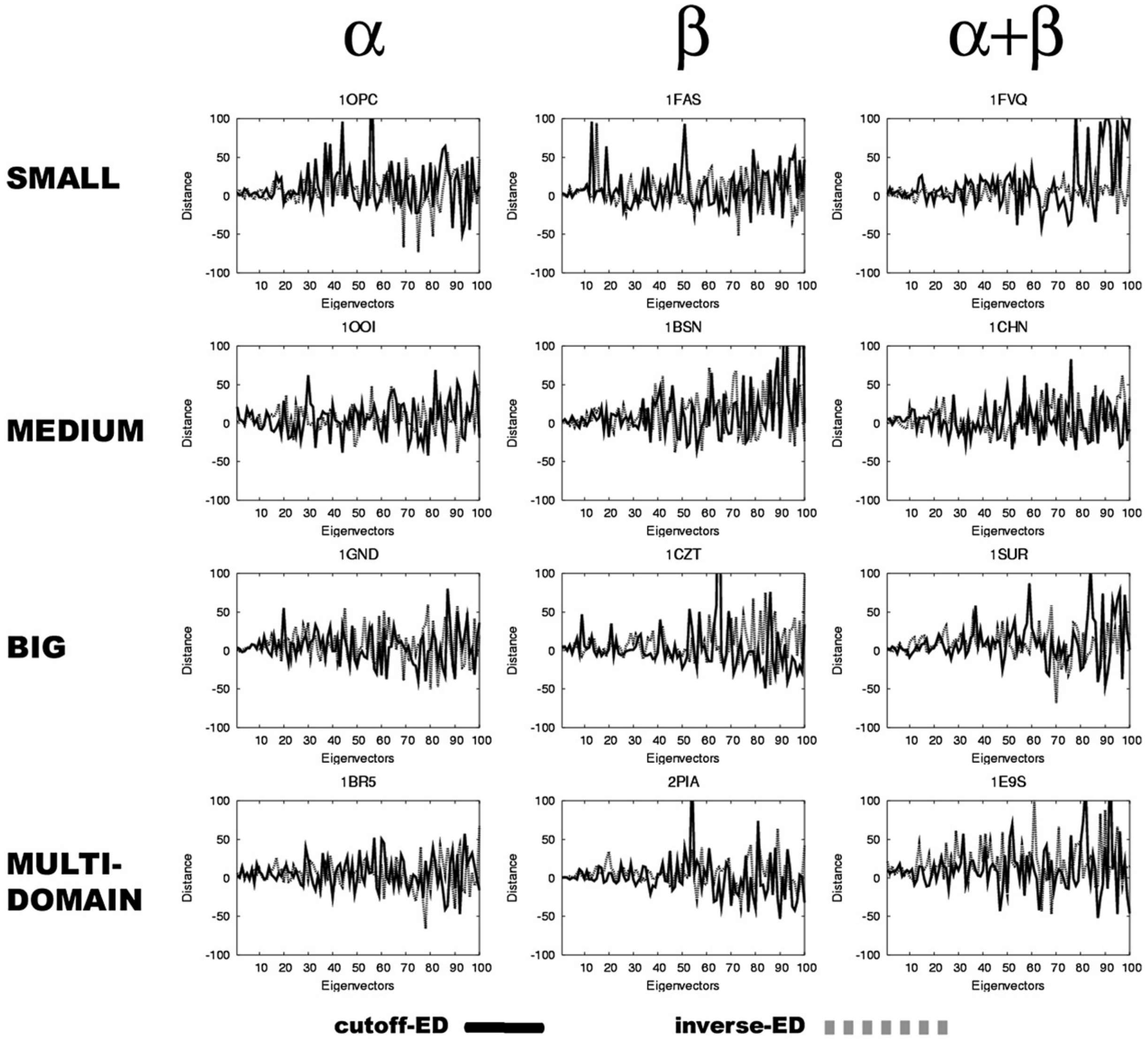

Figure 2. Difference in Rank between a Given NMA Eigenvector and the One from ED Displaying the Best Overlap with It

difficulty for NMA to describe the protein flexibility. Finally, the presence/absence of specific interactions, such as saline bridges or disulfide bridges, does not introduce advantages/problems in the ability of NMA to describe deformation modes.

Analysis of the Deformation Pattern

There is a poor pair correspondence between essential modes determined from NMA and ED, as noted in eigenvector-eigenvector dot products that are generally small along the diagonal (see selected examples in Figure S4). The correspondence becomes worse as the size of the important space increases, as noted in the rank difference between optimally overlapped vectors (see Figure 2). Similar findings are reached from Gerstein's indices along the diagonal ($i = j$ in Equation 13; see Figure S5), which are not far from those of a random model (O = 0.5), indicating that the atomic displacements in diagonal NMA eigenvectors do not correlate with the corresponding ED atomic displacements. It is worth noting that the difference between NMA and ED eigenvectors not only stems from a permutation in the rank of identical eigenvectors (see Figure 2), but also to the spread of each NMA eigenvector in many ED modes. This is illustrated in Figure 3 and Figure S4, which also show that, in general, better results are obtained when the inverse exponential NMA is used.

Though the preceding results warn against the use of individual NMA eigenvectors to describe major flexibility patterns, they do not necessarily imply that the information contained in the ED essential deformation space is not contained in the NMA space. This can be investigated by computing the γ index for a given "important" space and the associated Z score. After inspection of the cumulative variance versus eigenvector rank and a similarity

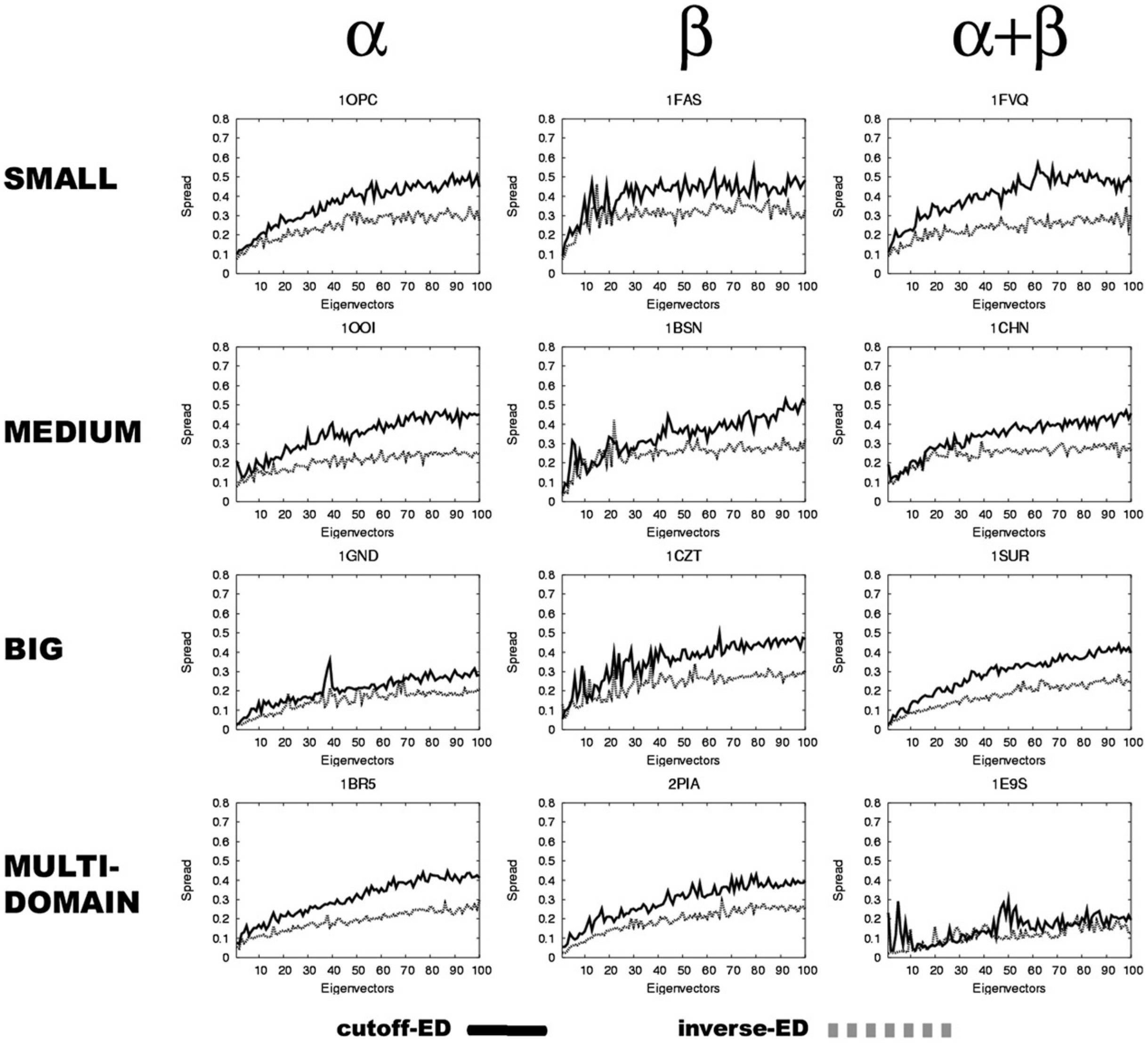

Figure 3. Spread of NMA Eigenvectors in the ED Space for Selected Proteins
See Equation 14.

index for selected examples (see Figure S6), we have considered two definitions for "important" space: (i) the first 50 eigenvectors (a value that always represents a large percentage of variance), and (ii) the number of ED eigenvectors needed to explain 90% of the MD variance. Similarity indices obtained for the first 50 eigenvectors are in the range of 0.4–0.7 (see Table 1), with larger values obtained for smaller proteins. The similarity indices become less dependent on the protein size (0.5–0.6; see Table 1) when the "important" space is defined by using the eigenvectors needed to explain 90% of the variance; however, in any case, the differences obtained by using the two definitions of the "important" space are small. It is worth noting that the correspondence in the movements remains unaffected when using longer simulation times (see Table S2). As noted in the very large Z score values, all similarity measures are far from random noise, indicating that, despite the poor pair correlation between eigenvectors, the essential deformation space of proteins measured by NMA and ED is reasonably similar. Detailed analysis of deformation pattern again shows the slightly superior performance of the inverse exponential algorithm with respect to the standard distance cutoff procedures. According to our results, the use of larger intraresidue cutoff values for the distance-based approach does not improve the similarity values (see Figure S7). Finally, we notice that there is a lack of any differential trend in the performance of NMA models regardless of the CATH family, protein size, or the presence/absence of disulfide or saline bridges in the structures.

Sampling of the Cartesian space modeled by activating the first 50 ED modes in a Metropolis Monte Carlo program reproduces the original sampling from the MD simulation (α of ~1.9 and Ω of ~0.99), without significant

Table 2. Cross-Rmsd Distance and Similarity Index between Cartesian Samplings Obtained from Monte Carlo in Important Spaces Defined by the First 50 Eigenvectors Obtained by ED and the 2 Versions of NMA Considered in This Paper

Size	Protein	α	Ω
Small	distance cutoff	1.90	0.97
	inverse exponential	1.91	0.98
Medium	distance cutoff	2.04	0.97
	inverse exponential	2.09	0.98
Big and multidomain	distance cutoff	1.70	0.97
	inverse exponential	1.70	0.98

Calculations are performed considering only C_α, and values shown here correspond to averages obtained for all of the proteins in each category. For cross-rmsd distance (α in Å), see Equation 16; for similarity index (Ω), see Equation 17.

changes if alternative definitions of the important space are used, indicating that the Monte Carlo procedure can properly capture the space obtained by MD simulation in spite of the drastic reduction in space dimensionality (from $3N - 6$ in MD to 50 in Monte Carlo simulations). Application of this technique, but by now using the sampling obtained by ED eigenvectors as a reference, allowed us to recognize the excellent similarity of important deformation spaces obtained from ED and (scaled) NMA calculations. Thus, ED/NMA cross-rmsd differences (see Equation 16) are ~1.8–2.0 Å (Table 2), which matches the normal self-cross rmsd generated in a MD trajectory by thermal noise and yields similarity indices close to 1.

To finish our study, we focus our analysis on the residue level, by comparing the C_α B factors derived from NMA and ED models, which also show a very good correspondence (Spearman's correlation coefficients ~0.7–0.8 for the set of proteins considered here). If no scaling of NMA forces is done, NMA-derived B factors are always smaller than those predicted by MD; however, after the scaling procedure, NMA and ED B factor profiles are not only qualitatively close, but are also quantitatively close (see a few examples in Figure 4). It is interesting to note that the profile of B factors is well preserved if larger intraresidue cutoff values are used within the distance cutoff NMA procedure (see Figure S8). Finally, it is worth noting (see Table 3) that, after this scaling, NMA-computed atomic displacements are able to capture the macroscopic character of proteins that emerge as a solid core surrounded by a near-liquid environment (Rueda et al., 2007; Zhou et al., 1999).

In summary, even though each individual eigenvector obtained in NMA has a small value, their combination generates an extremely correct representation of the C_α conformational space of proteins, as defined by "state of the art" atomistic MD simulations. As found systematically throughout this paper, this finding is independent of the protein family or size, suggesting that this is a general behavior in proteins, and that, bearing in mind its intrinsic limitations, NMA can be safely used to trace the flexibility of proteins.

Conclusions

A very wide systematic comparison of essential deformation modes performed thanks to a very large database of atomistic MD trajectories of representative proteins has allowed us to get a proper picture of the quality of limitations of simple NMA techniques compared to MD simulations. Results obtained here are very stable, irrespective of the protein family and size, and we are quite sure that they can be safely translated to the entire proteome. We found that individual NMA eigenvectors have small value, but that the "important" space defined by the first, most-relevant NMA eigenvectors provides an extremely correct picture of the trace flexibility of proteins in aqueous solution.

EXPERIMENTAL PROCEDURES

The Benchmark

A total of 30 proteins representative of all protein metafolds were selected as described elsewhere (see Table S1 and Day et al. [2003] and Rueda et al. [2007]); additional PDB entries were added to account for large or multidomain proteins. Thus, the comparative study includes mono- and multidomain proteins of very different sizes—from very small (31 residues) to extremely large (2545 residues)—with different folds, amino acid compositions, secondary structures, topology, and stability. We can expect then that consequences derived from this massive analysis can be safely translated to the entire proteome.

Normal Mode Analysis

Even though other approaches are available (Doruker and Jernigan, 2003; Jeong et al., 2006; Tama and Sanejouand, 2001; Zheng et al., 2006), we have used here the standard elastic network NMA approximation based on the use of C_α-C_α distances as descriptors of molecular deformations. The molecular Hamiltonian defining the energy necessary to distort a protein from its equilibrium geometry (the crystallographic or the MD-averaged conformation) is given by the following pairwise Hookean spring potential (Tirion, 1996) between C_α:

$$E = \sum K_{ij}(d_{ij} - d_{ij}[eq])^2, \tag{1}$$

where K_{ij} is a distance-dependent force constant (Equation 2) that restrains the C_α-C_α interaction ij at the equilibrium distance, $d_{ij}(eq)$ (Doruker et al., 2000), taking from the MD-averaged structure.

$$K_{ij} = \delta(d_{ij})C, \tag{2}$$

where C is a constant equal for all interactions (10 kcal/mol Å^2 [Suhre and Sanejouand, 2004]) and $\delta(d_{ij})$ is 1 when d_{ij} is smaller than a threshold distance (values of ~8–9 Å are used) and 0 otherwise.

This approximation, referred to as distance cutoff, provides, despite its simplicity, reasonable descriptions of large-scale molecular motions (Bahar and Rader, 2005; Ma, 2005), but it presents a source of arbitrariness regarding the need to use a cutoff to remove springs for distant interactions. Thus, other approaches have been developed to define continuum functions for the spring constant. Among others (Hinsen et al., 1999), Kovacs et al. (2004) have developed a simple function that assumes an inverse exponential relationship between the distance and the force constant (see Equation 3). The approach, which provides good results in several examples (Kovacs et al., 2004), maintains the simplicity of the original method, allowing us to avoid the problems intrinsic to the use of an empirical cutoff.

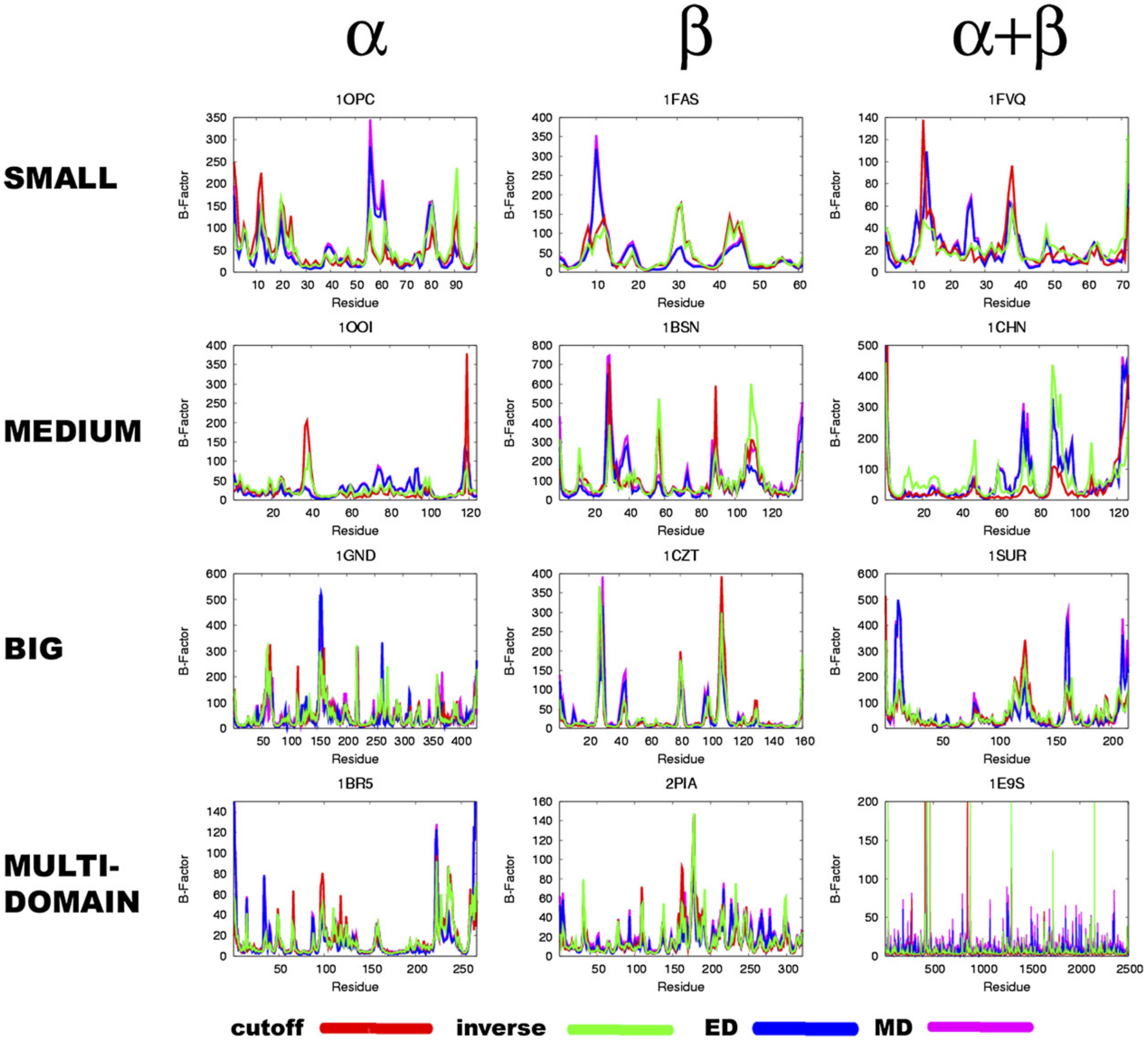

Figure 4. Examples of B Factor, C_α, Profiles Predicted by Activating the First 50 Modes in NMA and ED Simulations

As a reference, the MD values (typically almost superposed by the ED values) are shown. NMA values shown here were obtained after the scaling of force constants (see text).

$$K_{ij} = C\left(\frac{d_{ij}^{0}}{d_{ij}}\right)^{6} + as_{ij}, \qquad (3)$$

where C is a stiffness constant (taken as 40 kcal/mol Å^2), and d_{ij}^0 is a fitted constant, taken as the mean C_α-C_α distance between consecutive residues.

Once the Hamiltonian is defined, the diagonalization of the mass-weighted Hessian (H_m, see Equation 4) yields the eigenvectors (the essential deformation modes) and the associated eigenvalues (λ) or vibrational frequencies. If the mass matrix is taken as the unit, the eigenvalues appear in energy units.

$$H_m = M^{-1/2} H M^{-1/2}, \qquad (4)$$

where H is the Hessian matrix, and M is the diagonal mass matrix.

Essential Dynamics, ED

This approach starts from the anharmonic representation of the macromolecular system provided by the force field (Equations 5 and 6). After equilibration, MD will provide a Boltzmann's ensemble of the macromolecular conformational space (i.e., the trajectory).

$$E = E_{bonded} + E_{nonbonded}, \qquad (5)$$

$$E_{bonded} = \sum_{bonds} K_s (l - l_0)^2 + \sum_{angles} K_s (\theta - \theta_0)^2 + \sum_{torsions} \sum_{i=1}^{3} \frac{V_i}{2}(1 + \cos(i\phi - \xi)), \qquad (6)$$

$$E_{non-bonded} = \sum_{a,b} \frac{Q_a Q_b}{r_{ab}} + \sum_{a,b} \left(\frac{C_{ab}}{r_{ab}}\right)^{12} - \left(\frac{D_{ab}}{r_{ab}}\right)^{6}, \qquad (7)$$

Table 3. Average Values for Lindemann's Indices for Buried and Exposed Residues

Location	Distance Cutoff	Inverse Exponential	ED	MD
Buried	0.20	0.21	0.20	0.22
Exposed	0.35	0.35	0.35	0.37

The averages are given for all proteins. The values were computed from the Cartesian samplings in the important space defined by the first 50 eigenvectors of ED and the 2 versions of NMA used in this study. The real MD values are shown as reference.

where *l* and θ stand for bond lengths and angles, respectively; the subscript 0 represents equilibrium values; K_s and K_b are the associated force constants; Φ is a torsion angle; V_i is the potential associated with the Fourier terms used to represent torsions; ξ is the phase angle; *Q* is an atomic charge; *C* and *D* denote the van der Waals parameters; and r_{ab} stands for interatomic distance.

Diagonalization of the (Cartesian or mass-weighted) covariance matrix yields a set of eigenvalues and the corresponding eigenvectors, which represent the essential deformation of the molecule. Note that the eigenvalues obtained by diagonalization of the Cartesian covariance matrix appear in distance units, but can be easily transformed into energy units by using:

$$k_l = \frac{k_b T}{\lambda_l}, \tag{8}$$

where λ_l stands for the l^{th} eigenvalue, k_b is the Boltzmann's factor, and *T* is the absolute temperature (note that k_l is associated with a mode that affects the entire protein, thus differing from the force constants that modulate C_α-C_α interactions; Equations 1–3).

It is worth noting that even though the meaning of the essential deformation modes obtained by NMA and ED is similar, the way in which they are obtained is different. In NMA, we assume that (i) the reference structure corresponds to a free energy minimum, (ii) no other minima are significantly populated, and (iii) all of the thermal macromolecular motions around the reference structure are Gaussian in nature (i.e., harmonic in energy). None of these assumptions exist in explicit-solvent ED simulations based on physical potentials.

Statistical Comparison between NMA and ED

The deformability of proteins predicted by NMA and ED can be examined by analyzing the respective sets of eigenvalues and eigenvectors. Several complementary aspects have been addressed to quantify the degree of similarity between the deformation patterns.

The Relative Amplitude of the Deformation Space

The size and complexity of the accessible deformation space were characterized by different measures, such as (i) the variance, (ii) the strength of the softer deformation modes (harmonic force constants; see Equation 8), and (iii) the number of modes needed to explain 90% of the structural variance.

Deformational Space Overlap

The simplest way to analyze overlap between two essential deformation spaces is to compare their corresponding eigenvectors (υ) by using Hess's metrics, as shown in Equation 9 (Hess, 2000; Noy et al., 2006; Orozco et al., 2003):

$$\gamma_{XY} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\nu_i^X \bullet \nu_j^Y\right)^2, \tag{9}$$

where *X* and *Y* stand here for two methods (NMA and ED), the indices *i* and *j* stand for the orders of the eigenvectors (ranked according to their contribution to the structural variance), and *m* stands for the number of eigenvectors in the "important space," which is defined as the minimum number of eigenvectors needed to explain a certain variance threshold.

Note that the similarity index depends on the size of the important space (for m = 3N − 6 [N = number of C_α], the similarity index will be always be equal to 1), which means that similarity indices need to be referred to a background model to derive Z score indices like that shown in Equation 10:

$$Z_{score} = \frac{(\gamma_{XY}(observed)) - (\gamma_{XY}(random))}{std(\gamma_{XY}(random))}, \tag{10}$$

where the random models were obtained by diagonalization of a pseudo covariance matrix obtained by random permutation of the C_αs for each snapshot. The standard deviation in Equation 10 was obtained by considering 500 different random models.

It is worth noting that a good similarity index (Equation 9) might be due to three different situations: (i) the ideal case of a perfect one-to-one correspondence between eigenvectors of the two important spaces, (ii) a good correspondence between permuted eigenvectors (example: the first eigenvector of space X fits perfectly with the tenth eigenvector of space Y), or (iii) a perfect spread of a given eigenvector from X into many others from the space defined by Y. To study these possibilities, we computed the dot products between eigenvector X/Y pairs, determining the difference in rank between the eigenvectors showing the largest overlap and also the eigenvector "spread function" (see [Hinsen, 1998] and Equation 11):

$$s_i = \sqrt{\sum_{j=1}^{m} j^2\eta_{ij}^2 - \left(\sum_{j=1}^{m} j\eta_{ij}^2\right)^2}, \tag{11}$$

where $\eta_{ij} = \upsilon_i^X \bullet \upsilon_j^Y$ and *m* = 3N − 6 (N = number of C_α; if not all the modes Y are available, the overlaps must be scaled to ensure that $\sum_i \eta_{ij}^2 = 1$). Note that for two identical sets of modes, η_{ij}^2 is a value other than zero for only *i* = *j*, and the spread becomes 0.

Relative Distribution of Deformational Pattern

Additional measures were performed to capture similarities in the atomic distribution of the deformation map that are not evident in eigenvector metrics based on the dot product. A first index designed to capture these similarities was developed by Gerstein and coworkers (O_{XY}^{i}; see Equation 12 and Krebs et al. [2002]) and is based on the direct comparison of the components of eigenvectors *i* and *j* on a given residue (the k^{th} residue of a total of *N*):

$$O_{XY}^{i-j} = \frac{1}{N}\sum_{k=1}^{N}\left|v_k^{i,X} \bullet v_k^{j,Y}\right|. \tag{12}$$

A complementary estimate of the similarity at the atomic level of the deformation space can be obtained by generating Cartesian pseudo trajectories by activating normal mode deformations by using a Metropolis Monte Carlo algorithm with a Hamiltonian defined as shown in Equation 13. The displacements obtained can then be projected to the Cartesian space to generate pseudo –trajectories.

$$E_X = \sum_{i=1}^{m'} k_i^X \Delta D_i^X, \tag{13}$$

where the sum can be extended from 1 (*m′* = 1; useful when we want to compare pairs of eigenvectors) to the entire important space (*m′* = *m*), and ΔD_i^X stands for a displacement along a given mode (*i*) in space *X*.

The (pseudo)trajectories obtained by the Metropolis procedure can then be compared with simple metrics, such as the direct or normalized cross-rmsd (Equations 14 and 1 5), which determines the degree of similarity between the structures that are reasonably sampled in two different (pseudo)trajectories (A and B). Furthermore, they can be used to obtain atomic measures of mobility (at a given temperature), such as B factors, or estimates of the macroscopic flexibility properties of proteins, like Lindemann's index (Equation 16) (Rueda et al., 2007; Zhou et al., 1999).

$$\alpha_{AB} = \frac{1}{M_A M_B} \sum_{k=1}^{M_A} \sum_{k=1}^{M_B} \left(\frac{1}{N} \sum_{l=1}^{3N} (x_{Al} - x_{Bl})^2 \right)^{1/2}, \qquad (14)$$

where N is the number of atoms, and M is the number of frames.

$$\Omega_{AB} = \frac{\alpha_{AA} + \alpha_{BB}}{2\alpha_{AB}}, \qquad (15)$$

$$\Delta_L = \frac{\left(\sum_i \langle \Delta r_i^2 \rangle / N \right)^{1/2}}{a'}, \qquad (16)$$

where a' is the most probable nonbonded near-neighbor distance, N is the number of atoms, and $\langle \Delta r^2 \rangle$ stands for the mean square displacements of the atoms from their equilibrium position.

Technical Details

In all cases included in the benchmark, at least 10 ns trajectories were obtained by using the isothermal-isobaric periodic boundary simulations in explicit water and ions and the Particle Mesh Ewald (Darden et al., 1993) technique to account for long-rang electrostatic interactions. The quality of MD simulations is dependent on the quality of the force field used. Thus, for each protein (see Table S1), simulations were repeated by using three different force fields (AMBER parm-99 [Cornell et al., 1995; Wang and Cieplak, 2000], CHARMM22 [MacKerell et al., 1995, 1998], and OPLS/AA [Damm et al., 1997; Jorgensen et al., 1996; Kaminski et al., 1994, 2001]). Due to the strong similarity among the trajectories obtained with these force fields (Rueda et al., 2007), the three trajectories for each protein were combined to obtain a "meta-trajectory" of 30 ns, which was then used for ED calculations. In all cases, individual MD trajectories correlate very well with the meta-trajectory (data available upon request). Previous studies (Rueda et al., 2007) show that, for the selected proteins, reasonable sampling of equilibrium conformation is obtained within the 10 ns simulation time, but selected cases were studied by using longer trajectories (see below). We find that, for our purposes, the same results are obtained if 10 or 100 ns samplings are considered (see Table S2). For computational reasons multidomain proteins were studied only with the parm-99 force field (these large systems were studied for 100 ns).

MD calculations were performed by using both AMBER8.0 (Case et al., 2004) and NAMD2.6 (Kale et al., 1999). All calculations were carried out on the *MareNostrum* supercomputer at the Barcelona Supercomputer Center within the MODEL project (http://mmb.pcb.ub.es/MODEL), as well as in workstations in our laboratory.

NMA calculations were performed by using Elnémo (Suhre and Sanejouand, 2004) and DFprot (Garzón et al., 2007). Normal mode analysis was performed by using MD-averaged conformations as reference structures in combination with the two formalisms noted above with the following default parameters: (i) force constant (C in Equation 2) equal to 10 kcal/mol Å^2 with a distance cutoff of 8 (small and medium) or 9 Å (large and multidomain), and (ii) inverse exponential formalism (Equation 3) with $a = 0$, $C = 40$ kcal/mol Å^2, and $d_{ij}^0 = 3.8$ Å. As noted, we also explored the behavior of NMA when either the force constant or the cutoff was changed.

Supplemental Data

Supplemental Data include figures and tables and are available at http://www.structure.org/cgi/content/full/15/5/565/DC1/.

ACKNOWLEDGMENTS

The authors are very grateful to Prof. F. Javier Luque for reading the manuscript carefully. We also thank Dr. Angel R. Ortiz and Dr. Konrad Hinsen for helpful suggestions and Elizabeth Villa and the rest of the Theoretical and Computational Biophysics group at the University of Illinois at Urbana-Champaign for their help with the NAMD simulation of protein TrwB (1E9S). This work has been supported by the Spanish Ministry of Education and Science (BIO2006-01602, BFU2004-01282, and GEN2003-20642-CO9-07) and the National Institute of Bioinformatics (Structural Bioinformatics Node). Calculations were performed on the *MareNostrum* supercomputer at the Barcelona Supercomputer Center.

Received: January 29, 2007
Revised: March 27, 2007
Accepted: March 29, 2007
Published: May 15, 2007

REFERENCES

Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D., and Gerstein, M. (2005). Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool. Protein Sci. *14*, 633–643.

Amadei, A., Linssen, A.B., and Berendsen, H.J. (1993). Essential dynamics of proteins. Proteins *17*, 412–425.

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

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

Bartlett, G.J., Porter, C.T., Borkakoti, N., and Thornton, J.M. (2002). Analysis of catalytic residues in enzyme active sites. J. Mol. Biol. *324*, 105–121.

Cao, Z.W., Xue, Y., Han, L.Y., Xie, B., Zhou, H., Zheng, C.J., Lin, H.H., and Chen, Y.Z. (2004). MoViES: molecular vibrations evaluation server for analysis of fluctuational dynamics of proteins and nucleic acids. Nucleic Acids Res. *32*, W679–W685.

Case, D.A., Pearlman, D.A., Caldwell, J.W., Cheatham, T.E., III, Ross, W.S., Simmerling, C.L., Darden, T.L., Marz, K.M., Stanton, R.V., Cheng, A.L., et al. (2004). AMBER8 (San Francisco: University of California).

Chacon, P., Tama, F., and Wriggers, W. (2003). Mega-Dalton biomolecular motion captured from electron microscopy reconstructions. J. Mol. Biol. *326*, 485–492.

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

Cui, Q., and Bahar, I. (2006). Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems (Boca Raton, FL: CRC Press).

Damm, W., Frontera, A., Tirado-Rives, J., and Jorgensen, W.L. (1997). OPLS all-atom force field for carbohydrates. J. Comput. Chem. *18*, 1955–1970.

Daniel, R.M., Dumm, R.V., Finney, J.L., and Smith, J.C. (2003). The role of dynamics in enzyme activity. Annu. Rev. Biophys. Biomol. Struct. *32*, 69–92.

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

Day, R., Beck, D.A., Armen, R.S., and Daggett, V. (2003). A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. Protein Sci. *12*, 2150–2160.

Doruker, P., and Jernigan, R.L. (2003). Functional motions can be extracted from on-lattice construction of protein structures. Proteins *53*, 174–181.

Doruker, P., Atilgan, A.R., and Bahar, I. (2000). Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to α-amylase inhibitor. Proteins Struct. Funct. Genet. *40*, 512–524.

Eisenmesser, E.Z., Bosco, D.A., Akke, M., and Kern, D. (2002). Enzyme dynamics during catalysis. Science *295*, 1520–1523.

Garzón, J.I., Kovacs, J.A., Abagyan, R., and Chacon, P. (2007). Dfprot: a webtool for predicting local chain deformability. Bioinformatics, in press. Published online February 3, 2007. 10.1093/bioinformatics/btm014.

Gerstein, M., Lesk, A.M., and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. Biochemistry *33*, 6739–6749.

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

Haliloglu, T., Bahar, I., and Erman, B. (1997). Gaussian dynamics of folded proteins. Phys. Rev. Lett. *79*, 3090–3093.

Hayward, S., Kitao, A., and Go, N. (1994). Harmonic and anharmonic aspects in the dynamics of Bpti: a normal-mode analysis and principal component analysis. Protein Sci. *3*, 936–943.

Hayward, S., Kitao, A., and Berendsen, H.J.C. (1997). Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. Proteins Struct. Funct. Genet. *27*, 425–437.

Hess, B. (2000). Similarities between principal components of protein dynamics and random diffusion. Phys. Rev. E Stat. Phys. Plasmids Fluids Relat. Interdiscip. Topics *62*, 8438–8448.

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

Hinsen, K., Thomas, A., and Field, M.J. (1999). Analysis of domain motions in large proteins. Proteins *34*, 369–382.

Hollup, S.M., Salensminde, G., and Reuter, N. (2005). WEBnm@: a web application for normal mode analyses of proteins. BMC Bioinformatics *6*, 52.

Jang, Y., Jeong, J.I., and Kim, M.K. (2006). UMMS: constrained harmonic and anharmonic analyses of macromolecules based on elastic network models. Nucleic Acids Res. *34*, W57–W62.

Jeong, J.I., Jang, Y., and Kim, M.K. (2006). A connection rule for α-carbon coarse-grained elastic network models using chemical bond information. J. Mol. Graph. Model. *24*, 296–306.

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

Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. (1999). NAMD2: greater scalability for parallel molecular dynamics. J. Comput. Phys. *151*, 283–312.

Kaminski, G., Duffy, E.M., Matsui, T., and Jorgensen, W.L. (1994). Free energies of hydration and pure liquid properties of hydrocarbons from the OPLS all-atom model. J. Phys. Chem. *98*, 13077–13082.

Kaminski, G.A., Friesner, R.A., Tirado-Rives, J., and Jorgensen, W.L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J. Phys. Chem. B *105*, 6474–6487.

Kong, Y., Ming, D., Wu, Y., Stoops, J.K., Zhou, Z.H., and Ma, J. (2003). Conformational flexibility of pyruvate dehydrogenase complexes: a computational analysis by quantized elastic deformational model. J. Mol. Biol. *330*, 129–135.

Kovacs, J.A., Chacon, P., and Abagyan, R. (2004). Predictions of protein flexibility: first-order measures. Proteins *56*, 661–668.

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

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., and Ortiz, A.R. (2005). An analysis of core deformations in protein superfamilies. Biophys. J. *88*, 1291–1299.

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

Lindahl, E., Azuara, C., Koehl, P., and Delarue, M. (2006). NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. Nucleic Acids Res. *34*, W52–W56.

Luo, J., and Bruice, T.C. (2004). Anticorrelated motions as a driving force in enzyme catalysis: the dehydrogenase reaction. Proc. Natl. Acad. Sci. USA *101*, 13152–13156.

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

Ma, J., and Karplus, M. (1998). The allosteric mechanism of the chaperonin GroEL: a dynamic analysis. Proc. Natl. Acad. Sci. USA *95*, 8502–8507.

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

MacKerell, A.D., Jr., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B *102*, 3586–3616.

Ming, D., and Wall, M.E. (2006). Interactions in native binding sites cause a large change in protein dynamics. J. Mol. Biol. *358*, 213–223.

Noy, A., Meyer, T., Rueda, M., Ferrer, C., Valencia, A., Perez, A., Orozco, M., de La Cruz, X., and Luque, F.J. (2006). Data mining of molecular dynamics trajectories of nucleic acids. J. Biomol. Struct. Dyn. *23*, 447–456.

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

Qian, B., Ortiz, A.R., and Baker, D. (2004). Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proc. Natl. Acad. Sci. USA *101*, 15346–15351.

Remy, I., Wilson, I.A., and Michnick, S.W. (1999). Erythropoietin receptor activation by a ligand-induced conformation change. Science *283*, 990–993.

Rueda, M., Ferrer-Costa, C., Meyer, T., Perez, A., Camps, J., Hospital, A., Gelpi, J.L., and Orozco, M. (2007). A consensus view of protein dynamics. Proc. Natl. Acad. Sci. USA *104*, 796–801.

Sacquin-Mora, S., and Lavery, R. (2006). Investigating the local flexibility of functional residues in hemoproteins. Biophys. J. *90*, 2706–2717.

Song, G., and Jernigan, R.L. (2006). An enhanced elastic network model to represent the motions of domain-swapped proteins. Proteins *63*, 197–209.

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

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

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

Wako, H., Kato, M., and Endo, S. (2004). ProMode: a database of normal mode analyses on protein molecules with a full-atom model. Bioinformatics *20*, 2035–2043.

Waldron, T.T., and Murphy, K.P. (2003). Stabilization of proteins by ligand binding: application to drug screening and determination of unfolding energetics. Biochemistry *42*, 5058–5064.

Wang, J., and Cieplak, P. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J. Comput. Chem. *21*, 1049–1074.

Yang, L.-W., and Bahar, I. (2005). Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure *13*, 893–904.

Yang, L.W., Liu, X., Jursa, C.J., Holliman, M., Rader, A.J., Karimi, H.A., and Bahar, I. (2005). iGNM: a database of protein functional motions based on Gaussian Network Model. Bioinformatics *21*, 2978–2987.

Yang, L.W., Rader, A.J., Liu, X., Jursa, C.J., Chen, S.C., Karimi, H.A., and Bahar, I. (2006). oGNM: online computation of structural dynamics using the Gaussian Network Model. Nucleic Acids Res. *34*, W24–W31.

Zheng, W., and Doniach, S. (2003). A comparative study of motor-protein motions by using a simple elastic-network model. Proc. Natl. Acad. Sci. USA *100*, 13253–13258.

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

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