Insights into Equilibrium Dynamics of Proteins from Comparison of NMR and X-Ray Data with Computational Predictions

Lee-Wei Yang,[1,3] Eran Eyal,[1] Chakra Chennubhotla,[1] JunGoo Jee,[2] Angela M. Gronenborn,[2] and Ivet Bahar[1,*]

[1] Department of Computational Biology
[2] Department of Structural Biology
School of Medicine, University of Pittsburgh, Biomedical Science Tower 3, 3051 Fifth Avenue, Pittsburgh, PA 15213, USA
[3] Present address: Laboratory of Molecular Design, Institute of Molecular and Cellular Biosciences, University of Tokyo, 1-1-1, Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan.
*Correspondence: bahar@ccbb.pitt.edu
DOI 10.1016/j.str.2007.04.014

SUMMARY

For a representative set of 64 nonhomologous proteins, each containing a structure solved by NMR and X-ray crystallography, we analyzed the variations in atomic coordinates between NMR models, the temperature (B) factors measured by X-ray crystallography, and the fluctuation dynamics predicted by the Gaussian network model (GNM). The NMR and X-ray data exhibited a correlation of 0.49. The GNM results, on the other hand, yielded a correlation of 0.59 with X-ray data and a distinctively better correlation (0.75) with NMR data. The higher correlation between GNM and NMR data, compared to that between GNM and X-ray B factors, is shown to arise from the differences in the spectrum of modes accessible in solution and in the crystal environment. Mainly, large-amplitude motions sampled in solution are restricted, if not inaccessible, in the crystalline environment of X-rays. Combined GNM and NMR analysis emerges as a useful tool for assessing protein dynamics.

INTRODUCTION

X-ray crystallography and solution NMR are two major techniques broadly used for determining the atomic structures of biomolecules. The three-dimensional (3D) structures derived by the two techniques for a given protein usually exhibit the same backbone topology/fold, whereas they may differ in their local structural features such as surface loop conformations and side-chain rotational states, due to crystal packing or environmental effects (Billeter, 1992; Brünger, 1997; Engh et al., 1993; Davy et al., 1998; Powers et al., 1993). While a wealth of studies have been published to date on the comparison of X-ray and NMR "structures," no systematic study of the "dynamics" of X-ray and NMR structures, as implied by the resolved structures, has been conducted to date.

A measure of conformational flexibility of proteins under native state conditions is the ensemble of conformations sampled near the global energy minimum. In particular, the mean-square variations in the coordinates of amino acids about their mean (native) positions provide an experimentally detectable measure of equilibrium dynamics. The temperature factors (B factors) measured by X-ray crystallography and the NMR-derived order parameters (Yang and Kay, 1996) extracted from relaxation data (Wagner, 1993; Kay, 1998; Eisenmesser et al., 2002) contain this information, correlating with thermal vibrations. In this study, we set out to assess whether for experimentally determined structures, irrespective of methodology, similar or distinct equilibrium dynamics can be discerned. We aimed to understand the molecular origins of any differences by comparing them to computational predictions.

Recent computational studies based on normal mode analysis (NMA) showed that the 3D structure uniquely defines the collective motions accessible near native state conditions (Cui and Bahar, 2006). Beginning with the Gaussian network model (GNM) (Bahar et al., 1997; Haliloglu et al., 1997), several studies have demonstrated that residue fluctuations predicted by simple elastic network (EN) models agree with experimental B factors (Kundu et al., 2002; Yang et al., 2006; Kondrashov et al., 2006). Other studies showed that computational predictions based on EN models or variants thereof are consistent with the order parameters derived from NMR relaxation experiments (Haliloglu and Bahar, 1999; Temiz et al., 2004; Ming and Bruschweiler, 2006).

The GNM describes the *intrinsic dynamics* of proteins. The intrinsic dynamics refers to the motions defined by the structure, or by the topology of interresidue contacts, in the folded state. This type of topology-driven or structure-induced dynamics is expected to be perturbed in the presence of environmental effects, such as interactions with solvent or lipid molecules, or intermolecular contacts (e.g., in the crystal form). The fluctuation dynamics provided by the GNM will be accurate to the extent that

such perturbing environmental effects do not play a dominant role. In other words, we expect our predictions to agree better with experimental data if the molecules are experiencing *minimally restricting environments*.

Based on the large number of Protein Data Bank (PDB) (Berman et al., 2000) structures, determined both in solution (by NMR) and in the crystal (by X-ray), as well as recent advances in computational characterization of equilibrium dynamics using coarse-grained normal mode analyses (Cui and Bahar, 2006; Chennubhotla et al., 2005; Ma, 2005), we are now in a position to systematically explore similarities and differences in the equilibrium dynamics of proteins in the two different environments.

We therefore examined the temperature factors measured by X-ray crystallography and the root-mean-square deviations (rmsds) in residue positions exhibited by NMR models deposited for the same protein, and compared these data with the residue fluctuations predicted by the GNM. The NMR rmsds generally reflect the "uncertainties" in atomic coordinates resulting from the methodological approaches inherent to NMR structure determination. Calculations performed here for a representative set of PDB structures show that these rmsds closely correlate with the fluctuation dynamics predicted by the GNM. Interestingly, the quantitative agreement between theory (GNM) and NMR data is significantly better than that between GNM and X-ray data. The two sets of experimental data, on the other hand, exhibit moderate correlation. The differences between NMR and X-ray data are explained in light of the accessibility and inaccessibility of theoretically predicted modes of relaxation to molecules in solution or in the crystal, respectively.

RESULTS

Calculations were performed for a set of n = 64 pairs of protein structures sharing at least 95% sequence identity, one member of the pair being determined by X-ray crystallography and the other by NMR (for details, see section A in Supplemental Data available with this article online). The GNM (see Experimental Procedures) was used to calculate (1) the rms fluctuations $<(\Delta R_i)^2>^{1/2}_{GNM-X}$ of residues (represented by their C^α atoms) around their equilibrium positions for each X-ray structure and (2) their counterpart, $<(\Delta R_i)^2>^{1/2}_{GNM-N}$, for the NMR structures. These two sets of results are termed *theoretical* results. We also compiled two sets of *experimental* data, namely, $<(\Delta R_i)^2>^{1/2}_{X-ray}$ based on the B factors $B_i = (8\pi^2/3) <(\Delta R_i)^2>_{X-ray}$ reported in the PDB for each X-ray structure, and $<(\Delta R_i)^2>^{1/2}_{NMR}$ based on the rmsds in the C^α coordinates of NMR models with respect to the coordinate-average model for each protein determined by NMR. Calculations repeated for different NMR models as the reference demonstrated that the results were not sensitive to the choice of reference. We denote the correlation coefficient between NMR and X-ray rms data for a given pair of protein structures as σ_{NX}, that between NMR data and GNM predictions as σ_{NG}, and, finally, that between X-ray data and GNM predictions as σ_{XG}.
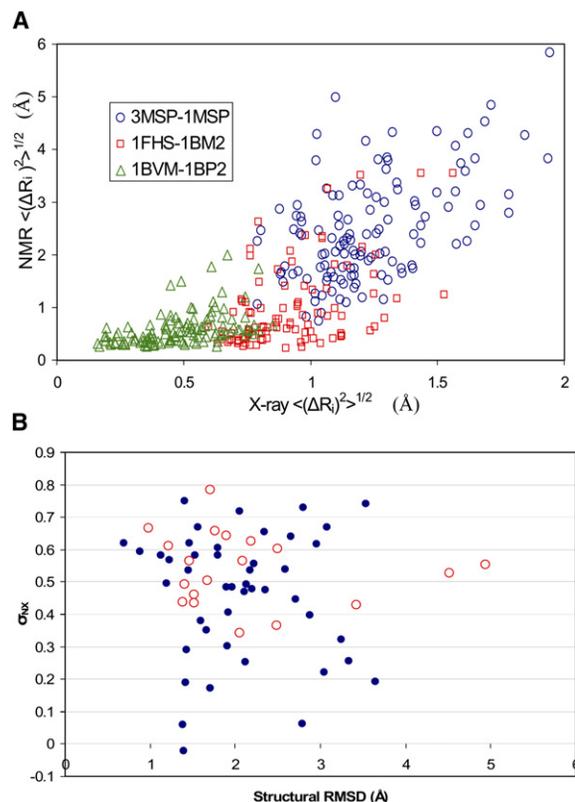
Figure 1. Comparison of Rmsd from Mean Positions Observed in Solution NMR and in X-Ray Crystallographic Experiments

(A) $<(\Delta R_i)^2>^{1/2}_{NMR}$ values corresponding to the NMR structures for motile sperm protein (PDB ID code 3MSP), SRC homology domain (PDB ID code 1FHS), and bovine pancreatic phospholipase (PDB ID code 1BVM) are compared with those, $<(\Delta R_i)^2>^{1/2}_{X-ray}$, reported for their X-ray counterparts PDB ID codes 1MSP, 1BM2, and 1BP2, respectively. Each point represents the rms variations in the position of a given C^α atom inferred from NMR and X-ray data. The results for all the aligned residues of the 64 protein pairs (not shown for clarity) yield the linear regression equation $<(\Delta R_i)^2>^{1/2}_{NMR} = 2.22 <(\Delta R_i)^2>^{1/2}_{X-ray} - 0.49$; that is, the NMR models exhibit, on average, rms fluctuations twice as large as those observed in X-ray structures.

(B) The correlation coefficients σ_{NX} for residue fluctuations in the two experimental data sets for each protein plotted against the corresponding structural rmsds (rmsd$_{N-X}$) for the NMR and X-ray structures. The mean correlation coefficient averaged over all proteins and its standard deviation is $<\sigma_{NX}> = 0.485 \pm 0.022$, with a standard deviation of 0.178. The correlation coefficients exhibit no detectable dependence on rmsd$_{N-X}$.

How Do the Rmsds between NMR Models Compare with X-Ray Crystallographic B Factors?

Figure 1A compares the rmsds deduced from NMR models, $<(\Delta R_i)^2>^{1/2}_{NMR}$ (ordinate), and from X-ray B factors, $<(\Delta R_i)^2>^{1/2}_{X-ray}$ (abscissa) for three example proteins. We note that $<(\Delta R_i)^2>^{1/2}_{NMR}$ is roughly twice as large as $<(\Delta R_i)^2>^{1/2}_{X-ray}$. The larger size deviations in residue coordinates inherent to NMR models are consistent with previous observations on α-amylase (Billeter, 1992; Powers et al., 1993). We found this trend to hold in general
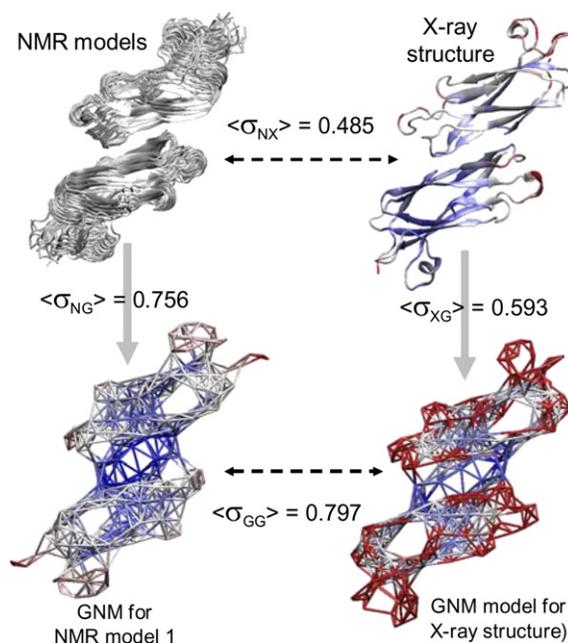
for the complete set of 64 pairs (see Figure 1A legend for more details).

The correlation coefficients σ_{NX} between the two sets of experimental data for the three proteins analyzed in Figure 1A are 0.535, 0.492, and 0.437. By repeating the same type of comparative analysis for the complete set of 64 pairs of structures, we obtained an average correlation of $<\sigma_{NX}> = 0.485$, with a standard deviation of $\delta_{NX} = \pm 0.178$. The standard error ε_{NX} in the mean value is $\delta_{NX}/n^{1/2} = 0.022$.

Next, we investigated whether the observed differences in the fluctuation dynamics could be attributed to differences in the mean coordinates as determined by X-ray and NMR. If this were the case, a higher correlation σ_{NX} should be observed for a given pair when the corresponding $rmsd_{N-X}$ between the NMR and X-ray structures is small. The results displayed in Figure 1B demonstrate, however, that this is not the case, and no discernible dependence ($R^2 = 0.004$) of σ_{NX} on $rmsd_{N-X}$ is noted. Calculations for protein pairs exhibiting equal sequence length (represented by the open circles) were performed to rule out any size bias. Again, no dependence of σ_{NX} on $rmsd_{N-X}$ was observed ($R^2 = 0.018$). Therefore, we believe that the detected differences for the two experimental data sets may reflect the types of motions (dynamics) sampled by the protein in the two different environments (crystal and solution). This notion is tested by GNM calculations described below.

Thermal Fluctuations Predicted by the GNM Correlate Well with B Factors, and Even Better with the Rmsds between NMR Models

Figures 2 and 3 describe the calculation scheme adopted for each pair of NMR and X-ray structures for a sample protein, the motile major sperm protein (MSP) from *Ascaris suum*. The upper two structures in Figure 2 illustrate the NMR (left) and X-ray (right) models. The NMR data comprise a best-fit superposition of conformers, and the X-ray structure is color coded according to the B factors reported in the PDB (blue, low B factors; red, high B factors). The lower two structures depict residue fluctuations calculated by the GNM for the first NMR model (left) and the X-ray structure (right), color coded by size. Figure 3 displays the residue fluctuation profiles for MSP, with the top panels depicting deviations in C^α coordinates plotted as a function of amino acid position i along the sequence, $<(\Delta R_i)^2>_{NMR}$ for the NMR ensemble (Figure 3A) and $<(\Delta R_i)^2>^{1/2}_{X-ray}$ for the X-ray structures (Figure 3B). The theoretical counterparts predicted by the GNM are shown in Figures 3C and 3D, respectively. The middle panels in Figure 3 compare the experimental and theoretical results. The correlation coefficients σ_{NG} (between NMR and GNM) and σ_{XG} (between X-ray and GNM) are found to be 0.909 and 0.596, respectively. Repeating this protocol for the set of 64 pairs of proteins yielded an average correlation of $<\sigma_{NG}> = 0.746 \pm 0.138$ between $<(\Delta R_i)^2>^{1/2}_{GNM-N}$ and $<(\Delta R_i)^2>^{1/2}_{NMR}$, and $<\sigma_{XG}> = 0.593 \pm 0.151$ between $<(\Delta R_i)^2>^{1/2}_{GNM-X}$ and $<(\Delta R_i)^2>^{1/2}_{X-ray}$ (Table 1; Figure 2). σ_{NG} values examined

Figure 2. Overview of the Calculation Scheme Conducted for All Proteins, Illustrated for Motile Major Sperm Protein from *Ascaris suum*

MSP is a dimeric β protein solved by NMR (Haaf et al., 1998) and X-ray (Bullock et al., 1996). The upper two structures depict the NMR models (left) and the X-ray structure (right) (PDB ID codes 3MSP and 1MSP, respectively). The X-ray structure is color coded according to the B factors reported in the PDB. The lower two diagrams are the GNM representations of the respective structures, color coded according to mobilities, from blue to red with increasing sizes of motions. The average correlation coefficients between the residue fluctuations derived from experimental data (rmsds between NMR models or B factors) or computed by the GNM are indicated by the $<\sigma>$ values (see also Table 1).

as a function of the rmsds between the NMR models, repeated for all proteins, showed that higher NMR rmsds do not necessarily imply a decrease in the correlation with GNM predictions. On the contrary, a more diverse set of NMR models seems to exhibit a stronger correlation with GNM fluctuations, as shown in Figure S1.

This analysis firmly establishes that the variations in amino acid positions derived from NMR models correlate with GNM predictions. Interestingly, the correlation between GNM predictions and NMR data is higher than that observed between GNM and X-ray data. The origins of this difference will be explored next.

X-Ray Structures Contain No Significant Contributions from Large-Scale Motions, whereas NMR Models Reflect Such Motional Characteristics

Prior to analyzing the origins of the differences between X-ray and NMR data sets, we examined the correlation $<\sigma_{GG}>$ between the two sets of *theoretical* results, $<(\Delta R_i)^2>^{1/2}_{GNM-N}$ and $<(\Delta R_i)^2>^{1/2}_{GNM-X}$, for each pair of structures. An average correlation of 0.797 was found for all pairs. This number provides a direct measure of
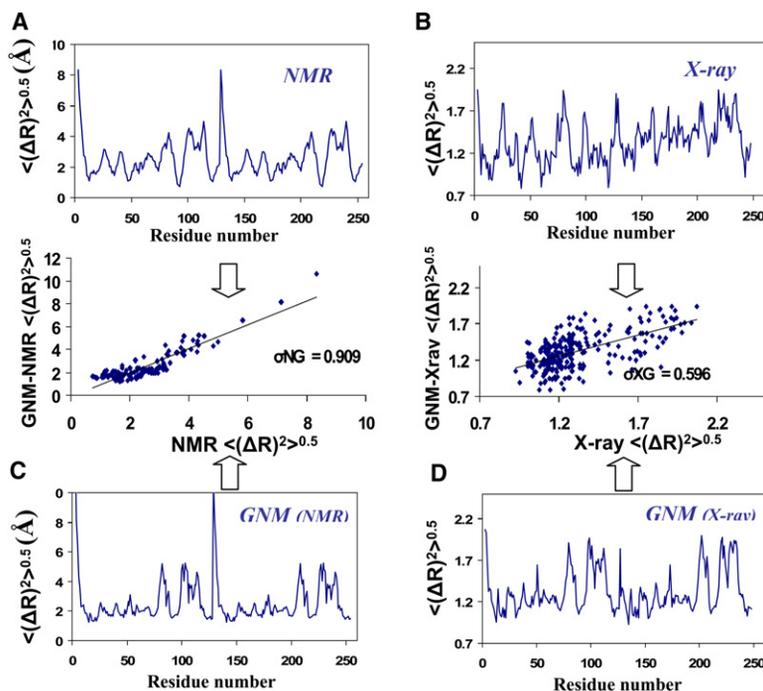
Figure 3. Schematic Description of the Calculation Scheme Adopted in the Present Study

(A) $<(\Delta R_i)^2>^{0.5}_{NMR}$, the rmsd between the 20 NMR models (in the left diagram) deposited for MSP shown as a function of residue index $1 \leq i \leq 252$.

(B) Rms fluctuations, $<(\Delta R_i)^2>^{1/2}_{X-ray}$, revealed by the B factors in the X-ray structure of MSP (i.e., as a function of residue index i).

(C and D) Rms fluctuations computed by the GNM for the NMR structure, $<(\Delta R_i)^2>_{GNM-N}$ (C), and rms fluctuations computed by the GNM for the crystal structure, $<(\Delta R_i)^2>^{1/2}_{GNM-X}$ (D). The two middle plots show the comparison of the experimental and theoretical results for the NMR (left) and X-ray (right) models. The correlation coefficient, σ_{NG}, between $<(\Delta R_i)^2>^{1/2}_{NMR}$ and $<(\Delta R_i)^2>^{1/2}_{GNM-N}$ is 0.909 (left), and that, σ_{XG}, between $<(\Delta R_i)^2>^{1/2}_{X-ray}$ and $<(\Delta R_i)^2>^{1/2}_{GNM-X}$ is 0.596 (right). PDB ID codes 3MSP and 1MSP share 100% sequence identity and rmsd of 1.45 Å for the C^α atoms.

the sensitivity of GNM results depending on whether the X-ray or NMR structure coordinates are used for the calculations. Figure S2 presents more details on the sensitivity of this correlation to the similarity between the two structures used in the calculations. The correlation between the two sets of predicted fluctuations tends to decrease with increasing dissimilarity between the X-ray and NMR structures, as can be expected. The high average correlation of 0.797 is indeed consistent with the similarities in structure (only 15 out of 64 pairs exhibited structural rmsd$_{N-X}$ values larger than 2.6 Å). Both the insensitivity of the σ_{NX} values to rmsd$_{N-X}$ (Figure 1B) and the high correlation $<\sigma_{GG}>$ (0.797) suggest that structure-induced perturbations are barely responsible for the weak correlation of 0.485 between NMR and X-ray data.

We next examined the two data sets with respect to the mode spectra provided by the GNM. Essentially, we removed the contribution from the slowest modes of motion by computing the fluctuations in the absence of the contributions from these modes.

Calculations for all 64 pairs of proteins resulted in the curves displayed in Figure 4. The dependence of the average correlations between GNM results and X-ray (top) and NMR (bottom) data on the successive exclusion of slow modes from GNM calculations is shown. The abscissa indicates the number of modes included in the predictions, N′ referring to all modes and N′ − k all but the lowest-frequency k modes. The thermal fluctuations $<(\Delta R_i)^2>^{1/2}_{GNM-X}$ evaluated by the GNM without including the contribution from the global (i.e., lowest-frequency) mode yielded an average correlation $<\sigma_{XG}>$ of 0.589 with X-ray crystallographic fluctuations. Interestingly, this value is very close to the one (0.593) computed with all modes (including the slowest), indicating that X-ray structures do not adequately sample the slowest-mode motions in the crystal. Note that this is the *average* correlation computed for all structures. Examination of the individual cases showed that σ_{XG} increased in some cases and decreased in others. Upon further removal of additional modes, for example the slowest two, four, six,

Table 1. Average Correlation Coefficients between NMR-, X-Ray-, and GNM-Derived Rmsds

	$<(\Delta R_i)^2>^{1/2}_{NMR}$	$<(\Delta R_i)^2>^{1/2}_{X-ray}$	$<(\Delta R_i)^2>^{1/2}_{GNM-N}$	$<(\Delta R_i)^2>^{1/2}_{GNM-X}$
$<(\Delta R_i)^2>^{1/2}_{NMR}$	1	0.485 ± 0.178	**0.746** ± 0.138	0.581 ± 0.189
$<(\Delta R_i)^2>^{1/2}_{X-ray}$	0.485 ± 0.178	1	0.543 ± 0.162	**0.593** ± 0.151
$<(\Delta R_i)^2>^{1/2}_{GNM-N}$	**0.746** ± 0.138	0.543 ± 0.162	1	0.797 ± 0.143
$<(\Delta R_i)^2>^{1/2}_{GNM-X}$	0.581 ± 0.189	**0.593** ± 0.151	0.797 ± 0.143	1

The standard deviations ± δ are listed next to the mean values. The errors ε in the mean values are equal to ±δ/$n^{1/2}$ with n = 64, such that ε < 0.04 in all cases. $<(\Delta R_i)^2>^{1/2}_{NMR}$ and $<(\Delta R_i)^2>^{1/2}_{X-ray}$ are the rms deviations of residue i in the NMR and X-ray structures, respectively; $<(\Delta R_i)^2>^{1/2}_{GNM-X}$ and $<(\Delta R_i)^2>^{1/2}_{GNM-N}$ are the rms fluctuations predicted by the GNM, based on X-ray and the first model of NMR structures, respectively. The numbers that are not in bold refer to crosscorrelations between NMR and X-ray experiments/calculations.
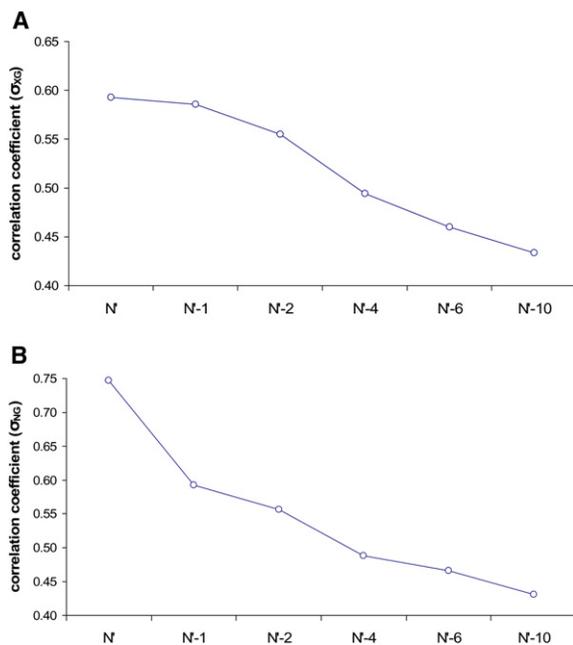
Figure 4. Variation in $<\sigma_{XG}>$ and $<\sigma_{NG}>$ as a Function of the Number of Excluded GNM Slow Modes, Averaged over All 64 Protein Pairs

The abscissa indicates the number of modes taken into consideration. N' refers to the complete set of nonzero modes; N' − k refers to all modes except the slowest k modes. Note that the differences in average correlation coefficients as a function of the number of included modes included are statistically significant, with the exception of that between the N and N − 1 values for σ_{XG}, verified by paired Student's t test.

(A) Excluding the contribution of the slowest GNM mode does not decrease the correlation $<\sigma_{XG}>$, whereas additional removal of slow modes from the computations reduces the correlation with X-ray results.

(B) Excluding the slowest GNM modes significantly decreases the correlation $<\sigma_{NG}>$ with the NMR rmsd data.

and ten modes, on the other hand, $<\sigma_{XG}>$ values decreased to 0.555, 0.496, 0.459, and 0.434, respectively (Figure 4A), suggesting that these modes do contribute to the fluctuations observed by X-ray crystallography.

The equivalent test performed for the NMR data set revealed significantly different behavior. In particular, the correlation $<\sigma_{NG}>$ averaged over all protein pairs decreased significantly, from 0.746 to 0.598, upon removal of the first mode contributions. Interestingly, this degree of correlation is comparable to the one observed for the X-ray sets (regardless of inclusion/exclusion of the slowest mode), lending further support to the notion that X-ray data barely report the slowest motional modes (or largest amplitude) that are accessible in solution. In other words, the higher correlation between experimental and computational results in the case of NMR structures appears to be associated with the effective contribution of the slowest motional mode to NMR data. Further removal of slow modes resulted in gradual decreases of $<\sigma_{NG}>$ to 0.561, 0.490, 0.468, and 0.431, respectively (Figure 4B).

DISCUSSION

In the present study, we present a comparative analysis of residue fluctuations (or rmsd) data near equilibrium coordinates derived from three different sources: NMR models, X-ray crystallographic B factors, and theoretical (GNM) predictions, and explore whether/how these data sets correlate. The results show that the NMR rmsds and GNM fluctuation profiles are correlated, whereas a poor-to-moderate correlation is observed between B factors and both NMR and GNM data. These direct observations resulted from the statistical comparison of the data deposited in the PDB and from automated application of the GNM. While it is intuitively compelling to think that motional data would be more relevant in solution, given the approximation in both the construction of NMR models and in the GNM, it is important to carefully assess and discuss the possible causes and implications of the observed high correlation (0.75) between NMR rmsds and GNM data, as well as the relatively poor correlation (0.49) between NMR and X-ray data for the same proteins. These two points are considered in the following sections.

The Discrepancy in Fluctuation Dynamics Revealed by X-Ray and NMR

Our present analysis shows that (1) NMR rmsds (C^α coordinates) and crystallographic B factors exhibit only a moderate correlation of 0.49, and (2) the difference between the two experimental data sets can be explained on the basis of the motional modes that are sampled in the two environments (solution and crystal). Overall, the fluctuation behavior inferred from X-ray data appears to contain little, if any, contribution from the global mode, as omission of the global mode's contribution from the theoretical predictions did not lead to a marked decrease in the level of agreement between theoretical and experimental data sets. This is in contrast to NMR data sets, for which removal of the global mode had a drastic effect, lowering the level of agreement between theory and experiment to a range comparable to that found with the X-ray data.

Impediments to sampling the slowest (or largest-amplitude) modes in the crystals may be caused by intermolecular contacts, low temperature, or immobilized water molecules. Indeed, our recent systematic analysis also indicated that B factors are significantly lower than theoretically predicted for regions involved in crystal contacts (Eyal et al., 2005), consistent with previous results reported by Phillips and coworkers (Kundu et al., 2002) and even Billeter's early findings for α-amylase (Billeter, 1992).

In this context, it may be worth pointing out that our earlier study on the dynamics of 1250 nonhomologous X-ray structures revealed better agreement between theory and X-ray crystallographic B factors for experimental data collected at higher temperatures: the correlation between the theoretical and experimental B factors, B_{GNM} and B_{exp}, indeed increased from 0.57 at <200K to 0.62 at 297K (Yang

et al., 2006). In fact, we did observe an increase in $<\sigma_{XG}>$ from N′ to N′ − 1 for those proteins resolved at the low temperature (<190K) and a decrease in that for the high-temperature (>277K) ones, which is consistent with our intuition that the slowest mode is involved and exercised more under a relaxed (high-temperature-softened) environment. The increase and decrease here are, however, statistically insignificant. Readers should note that the accessibility of slow modes is protein geometry and crystal dependent. Instead of increasing, the unchanged $<\sigma_{XG}>$ for the 64 proteins from N′ to N′ − 1 reflects the fact that slow modes can be exercised in crystals to some extent.

A more realistic comparison between X-ray and NMR data could be to consider an ensemble of X-ray structures deposited for the same protein, for example, the multiconformer refinements of X-ray data. As recently pointed out by Blundell, Terwilliger, and coworkers, an ensemble of models may provide a more suitable representation of a crystal structure, and this may become particularly important for medium- and low-resolution structures where a single parameter (B factor) per residue cannot adequately account for, or distinguish between, structural uncertainties, spatial heterogeneities, and equilibrium dynamics (Furnham et al., 2006). The rmsds in residue positions in such X-ray models may be larger than those calculated from the Debye-Waller temperature factors, and could exhibit anisotropic/anharmonic variations (Eyal et al., 2007) much like the NMR ensembles.

Why Do Theoretical Results Correlate Well with NMR Data?

Overall, NMR data appear to provide a better measure of equilibrium dynamics as calculated by the GNM, compared to X-ray crystallographic B factors. However, it is disputable whether or not the NMR conformers truly reflect conformational motions. Does the agreement between theory and experiment originate from similar assumptions adopted in both structure determination/refinement and the GNM? Do rmsds from NMR convey information on residue fluctuations near native state conditions?

To answer these questions, let us first examine how the ensemble of models is derived in NMR structure determination. The common approach is to use a set of measurable constraints usually consisting of interproton distances extracted from NOESY experiments and peptide torsion angles from measurements of three-bond J couplings. The NMR models deposited in the PDB are solved from a joint knowledge of experimentally determined distance constraints and empirical force field. The model quality is of combined effects such as restraint optimization and conformational averaging/relaxation (Brünger, 1997). GNM dynamics, on the other hand, are analytically solved, exclusively based on a single representative structure topology. It is fully controlled by the N × N Kirchhoff matrix Γ of C^{α} contact topology (see Experimental Procedures). Γ differs from the 3N × 3N Hessian *H* used in normal-mode analysis and energy minimization, and the associated potential in the GNM differs from that in the

elastic network NMA. Notably, the former takes account of both distance and orientation changes between residues, whereas the latter exclusively depends on distances (Chennubhotla et al., 2005).

We note that all the NMR models in an ensemble agree more or less equally well with the experimental restraints and exhibit comparable energies. The question then becomes how well these models sample the conformational space near native state. How robust are the experimental data? The validity of our results indeed depends on the robustness of the experimental data, and the results should be interpreted in this context. This issue is discussed in the following section.

Theoretical Studies Demonstrate that, in General, the Topology of the Determined Structure Is Insensitive to Using the Full NOE Restraint Set

Brünger, for example, showed that an ensemble calculated with only 50% of the available experimental restraints deviates by only 0.75 Å for the heavy atoms from one based on the full set (Brünger, 1997). A low number of restraints, however, results in a larger rmsd for the ensemble. Therefore, a correct structure (fold) is independent of the size of the rmsd; small values solely indicate a higher precision. Indeed, a statistical analysis carried out for RECOORD, a database of 500+ rerefined structures using standardized protocols and algorithms, showed that the correlation between rmsd and "structural uncertainty" is 0.69, whereas other quality indicators such as nuclear Overhauser enhancement (NOE) completeness, number of restraints per residue, and Ramachandran map had little impact (Nederveen et al., 2005). The term "structural uncertainty" in this context is defined as the degree of insufficiency in positional information quantified by the QUEEN algorithm, using information theory (Nabuurs et al., 2003). Furthermore, it was shown that the rmsds may have been underestimated in many NMR structures (Nabuurs et al., 2003; Spronk et al., 2003), as indicated by an average increase in the backbone rmsd for the rerefined RECOORD structures (Nederveen et al., 2005). Because our results do not depend on the absolute values of the rmsds but rather their residue profile/distribution, any increase in rmsd values would not influence our results.

To check how sensitive the residue variation profiles are to a specific set of restraints, we conducted the following test for the IgG binding domain (PDB ID code 3GB1; Kuszewski et al., 1999). Vuister and coworkers showed that long-range NOE restraints account for 86.3% of the information important for structural certainty in the IgG binding domain and that the restraints between Leu5 and Phe52 in this structure contain the highest information content (Nabuurs et al., 2003). In order to reexamine the effect of these restraints on the distribution of the structural models derived from NMR data, we excluded from the set of restraints all those associated with the pair Leu5-Phe52 and rerefined the structure (PDB ID code 3GB1) using only the NOE distance and torsion angle restraints deposited in the PDB. Indeed, the omission of the Leu5-Phe52
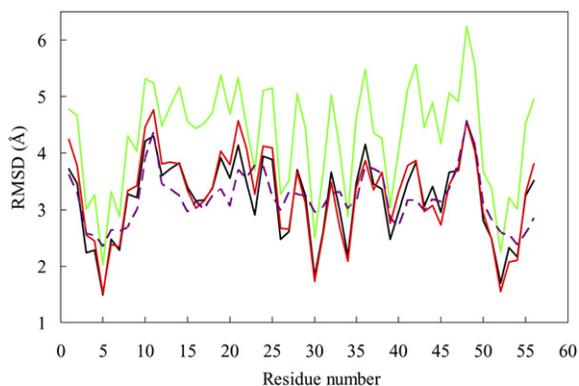
Figure 5. Backbone C^α Rmsds for Three Different Structural Ensembles of the IgG Binding Domain, PDB ID Code 3GB1, Structures

Structure I (black, solid) was determined with the complete set of NOE restraints; structure II (red, solid), by excluding the Leu5-Thr16 restraints; structure III (green, solid), by excluding the Leu5-Phe52 restraints. The GNM prediction (purple, dashed) is based on the first model of structure I. The average C^α rmsd in structures I, II, and III are 3.2, 3.3 (2% increase), and 4.3 Å (34% increase), respectively. The correlations between the rmsd values for structures I, II, and III and the corresponding GNM results are 0.78, 0.74, and 0.70, respectively. In addition, the correlation between the GNM-predicted fluctuation profiles of structures I and II is 0.96, of structures I and III is 0.98, and of structures II and III is 0.91. The restraints included or excluded in the refinement are listed in Supplemental Data.

restraints caused a 34% increase in rmsd per residue (from 3.2 to 4.3 Å) (Figure 5). Yet, despite the larger uncertainty for each residue, the resulting variation profile, $\langle(\Delta R_i)^2\rangle^{1/2}_{NMR}$, still exhibited a correlation of 0.70 with $\langle(\Delta R_i)^2\rangle^{1/2}_{GNM-N}$ compared to the value of 0.78 in the original refinement. We note that the structural rmsd between the model 1 of the rerefined ensemble and that of the original is 0.39 Å. Finally, the GNM calculations reported here used the first NMR model listed in the PDB for each protein. However, we verified that the average correlation between theory and NMR data, $\langle\sigma_{NG}\rangle$, remained essentially unchanged (from 0.746 ± 0.138 to 0.756 ± 0.144), if any of the first eight models deposited in the PDB was selected (we restricted this calculation to eight models, as the 1RCH ensemble contains only eight models). Using a model with the smallest rmsd to the others for computing $\langle(\Delta R_i)^2\rangle^{1/2}_{GNM-N}$, a value of 0.740 ± 0.149 is obtained for $\langle\sigma_{NG}\rangle$, again essentially identical to the above.

Thus, the rmsd *profile* is not sensitive to the exact procedure/restraints used in the NMR structure determination/refinement, although the *absolute* rmsd per residue may vary due to different refinement protocols. Hence, the high correlation found here between the NMR rmsd profile $\langle(\Delta R_i)^2\rangle^{1/2}_{NMR}$ and the GNM predictions $\langle(\Delta R_i)^2\rangle^{1/2}_{GNM-N}$ emerges as a robust feature.

In terms of GNM methodology, the number of long-range residue contacts determines the differences between the coordination numbers (i.e., diagonal elements of the Kirchhoff matrix) of residues, which in turn dominate

the mean-square fluctuations of residues (Yang et al., 2005). The higher average correlation, $\langle\sigma_{NG}\rangle$ (0.746), with NMR data obtained with NMR structure-based GNM predictions, as opposed to that (0.581) found with X-ray structure-based GNM calculations, also reflects the dependence of the GNM results on structural coordinates. In fact, an average correlation $\langle\sigma_{GG}\rangle$ of 0.797 is found between the two sets of theoretical results obtained for NMR and X-ray structures (Table 1), and the correlation decreases with increasing rmsd between structural coordinates (Figure S2). If $\langle\sigma_{GG}\rangle$ was closer to unity (which would be the case if the coordinate variations between the X-ray and NMR structures of the same protein were very small), we would expect the value 0.581 to approach $\langle\sigma_{NG}\rangle$. We note that local structural changes usually affect the high-frequency modes, whereas the low-frequency modes are robust and insensitive to detailed coordinates but depend on the overall shape, or fold (Tirion, 1996; Haliloglu et al., 1997; Hinsen, 1998; Tama and Sanejouand, 2001; Doruker et al., 2002; Ma, 2005; Lu and Ma, 2005; Bahar and Rader, 2005; Cui and Bahar, 2006; Sanejouand, 2006; Nicolay and Sanejouand, 2006; Karplus, 2006; Zheng et al., 2006). Calculations performed for pairs (19 out of 64) whose members had equal sequence length (see Supplemental Data, section A) confirmed this behavior: an average correlation of 0.937 was found between the profiles of the lowest-frequency modes for these pairs, whereas the corresponding $\langle\sigma_{GG}\rangle$ was 0.834 for the same subset. The decrease in correlation between the GNM results and experimental data (from 0.593 to 0.543, or from 0.746 versus 0.581 when the structure used in GNM calculations differs from the one whose fluctuation behavior is experimentally observed) is thus attributed to the dependence of GNM modes on structural coordinates, an effect that is more pronounced toward the higher-frequency portion of the spectrum.

Thus, both NMR models and GNM results are heavily influenced by tertiary contacts. One therefore might argue that the agreement between variability data from NMR models and GNM fluctuations may be due to the fact that both methodologies incorporate contact topologies and calculate behavior compatible with the distribution of interresidue contacts. However, there are three major differences: first, the GNM takes into account the complete distribution of interresidue contacts; NMR data are based only on those restraints that can be measured, not necessarily representing the complete set. Second, the GNM yields an analytical, unique solution for a given architecture, based on fundamental statistical mechanical theory and methods; that is, the results are *physically* meaningful. NMR models, or the rmsd fluctuations inferred from the comparison of these models, are derived from solving a mathematical optimization problem. Third, the GNM imposes a harmonic potential on the coarse-grained representation, whereas refined NMR models are populated in local minima after simulated annealing and energy minimization over the anharmonic (considering the involvements of van der Waals and electrostatics)

potential that comprises empirical force field and NMR restraint-derived penalty functions. The fact that the two sets of data yield a satisfactory level of agreement supports the views that (1) NMR models should not be viewed solely as alternative solutions for the 3D structure of the examined protein but as an ensemble of conformations accessible under the experimental conditions of the structure determination, and (2) their rmsd values, although reflecting uncertainties in the coordinates, may contain physically meaningful contributions of equilibrium fluctuations that can be extracted using GNM calculations.

EXPERIMENTAL PROCEDURES

Representative Protein Sets

A set of 64 pairs of protein structures was extracted from the PDB as described in Supplemental Data. Each pair contains two proteins sharing at least 95% sequence identity, one determined by X-ray crystallography and the other by solution NMR. Where applicable, we align the NMR and X-ray sequences (Myers and Miller, 1988) and then perform comparisons only at sequence positions with reported 3D coordinates in both X-ray and NMR structure files.

The Gaussian Network Model

The structure is modeled as a network of N nodes, the positions of which are identified by the α carbons. Drawing on the statistical mechanical theory of polymer networks (Flory, 1976), node fluctuations are assumed to be isotropic and Gaussian. The topology of the network is described by an $N \times N$ Kirchhoff matrix, $\mathbf{\Gamma}$. The off-diagonal element Γ_{ij} of $\mathbf{\Gamma}$ is -1 if nodes i and j are within a cutoff distance, r_c, and zero otherwise. The diagonal elements represent the coordination number of each residue (or the degree of each node). Assigning a uniform spring constant, γ, to all contacts, the crosscorrelations between the fluctuations $\Delta \mathbf{R}_i$ and $\Delta \mathbf{R}_j$ of residues i and j are evaluated using Bahar et al. (1997) and Chennubhotla et al. (2005):

$$<\Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j> = (3 k_B T / \gamma) \left[\mathbf{\Gamma}^{-1} \right]_{ij}, \tag{1}$$

where k_B is the Boltzmann constant, T is the absolute temperature, and $[\mathbf{\Gamma}^{-1}]_{ij}$ is the ijth element of the inverse of $\mathbf{\Gamma}$ (Cui and Bahar, 2006). Setting $j = i$ in Equation 1, we obtain the rms fluctuation of residue i, $<(\Delta \mathbf{R}_i)^2>$, which is directly compared to the corresponding X-ray crystallographic B factor,

$$B_i \equiv (8\pi^2 / 3) <(\Delta \mathbf{R}_i)^2>, \tag{2}$$

reported in the PDB, thus providing a quantitative measure of correlation between computations and experimental data. Extensive application to PDB structures has shown that γ is of the order of 1 kcal/(mol Å^2) (Kundu et al., 2002; Yang et al., 2006).

The equilibrium dynamics of the structure results from the superposition of N − 1 nonzero modes found by the eigenvalue decomposition of $\mathbf{\Gamma}$ such that (Bahar et al., 1997)

$$<(\Delta R_i)^2> = \sum_{k=2}^{N} \frac{3 k_B T}{\gamma} \left(\frac{1}{\lambda_k} \mathbf{u}_k \mathbf{u}_k^T \right)_{ii}. \tag{3}$$

The elements of the kth eigenvector, $\mathbf{u}_k$, describe the displacements of the residues along the kth mode coordinate, and the kth eigenvalue, λ_k, scales with the square frequency of the kth mode. Note that $\lambda_1 = 0$, as $\mathbf{\Gamma}$ has a reduced rank N − 1. We can quantify the fluctuations driven by subsets of modes by including selected modes in the above summation.

Intermodel Rmsd Calculation in NMR Structures

The rmsd for each residue i in a given set of NMR models is calculated using

$$\text{rmsd}_i = \sqrt{\frac{\sum_{k=1}^{m} \left| \mathbf{r}_{i,k} - \bar{\mathbf{r}}_i \right|^2}{m}}, \tag{4}$$

where $\bar{\mathbf{r}}_i = \frac{\sum_{k=1}^{m} \mathbf{r}_{i,k}}{m}$ is the average position over m models, and $\mathbf{r}_{i,k}$ is the position vector of residue i in the kth model. We have calculated the average correlation coefficients $<\sigma_{NG}>$ between the above rmsds and the GNM-predicted fluctuations.

In order to examine how the NMR rmsd$_i$ profiles of residues are sensitive to the choice of the reference NMR model, we compared profiles using different NMR models as reference. These calculations showed that the reference model has minimal effect on the evaluated profiles; the correlation coefficients between the rmsd$_i$ profiles based on different reference models indeed remained higher than 0.95, confirming that the reference model has little effect on the resulting rmsd$_i$ profiles. We also repeated the comparison using as the reference NMR model one with the smallest rmsd to all others in the ensemble for a given protein. All calculations for the complete set of NMR structures confirmed that $<\sigma_{NG}>$ values are relatively insensitive to the choice of reference model.

Sequence and Structural Alignment

Protein sequence alignments were performed using global dynamic programming (Myers and Miller, 1988). Because atomic positions are not present for every residue in all PDB files due to experimental limitations in NMR and X-ray methodologies, we only compared the NMR rmsd and X-ray B$_{exp}$ for those positions for which coordinates were available. The structural alignment between paired NMR and X-ray structures was carried out using combinatorial extension (CE) software (Shindyalov and Bourne, 1998). Structural rmsds were computed for structurally aligned C$^\alpha$ traces.

Supplemental Data

Supplemental Data include two figures and Supplemental Experimental Procedures and can be found with this article online at http://www.structure.org/cgi/content/full/15/6/741/DC1/.

REFERENCES

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. Structure *2*, 173–181.

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

Billeter, M. (1992). Comparison of protein structures determined by NMR in solution and by X-ray diffraction in single crystals. Q. Rev. Biophys. *25*, 325–377.

Brünger, A.T. (1997). X-ray crystallography and NMR reveal complementary views of structure and dynamics. Nat. Struct. Biol. *4*, 862–865.

Bullock, T.L., Roberts, T.M., and Stewart, M. (1996). 2.5 Å resolution crystal structure of the motile major sperm protein (MSP) of *Ascaris suum*. J. Mol. Biol. *263*, 284–296.

Chennubhotla, C., Rader, A.J., Yang, L.-W., and Bahar, I. (2005). Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. Phys. Biol. *2*, S173–S180.

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

Davy, S.L., Osborne, M.J., and Moore, G.R. (1998). Determination of the structure of oxidized *Desulfovibrio africanus* ferredoxin I by ^{1}H NMR spectroscopy and comparison of its solution structure with its crystal structure. J. Mol. Biol. *277*, 683–706.

Doruker, P., Jernigan, R.L., and Bahar, I. (2002). Dynamics of large proteins through hierarchical levels of coarse-grained structures. J. Comput. Chem. *23*, 119–127.

Eisenmesser, E.Z., Bosco, D.A., Akke, M., and Kern, D. (2002). Enzyme dynamics during catalysis. Science *295*, 1480–1481.

Engh, R.A., Dieckmann, T., Bode, W., Auerswald, E.A., Turk, V., Huber, R., and Oschkinat, H. (1993). Conformational variability of chicken cystatin. Comparison of structures determined by X-ray diffraction and NMR spectroscopy. J. Mol. Biol. *234*, 1060–1069.

Eyal, E., Gerzon, T., Potpov, V., Edelman, M., and Sobolev, V. (2005). The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J. Mol. Biol. *351*, 431–442.

Eyal, E., Chennubhotla, C., Yang, L.-W., and Bahar, I. (2007). Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. Bioinformatics, in press. 10.1093/bioinformatics/btm186.

Flory, P.J. (1976). Statistical thermodynamics of random networks. Proc. R. Soc. Lond. A *351*, 351–380.

Furnham, N., Blundell, T.L., DePristo, M.A., and Terwilliger, T.C. (2006). Is one solution good enough? Nat. Struct. Mol. Biol. *13*, 184–185.

Haaf, A., LeClaire, L., III, Roberts, G., Kent, H.M., Roberts, T.M., Stewart, M., and Neuhaus, D. (1998). Solution structure of the motile major sperm protein (MSP) of *Ascaris suum*—evidence for two manganese binding sites and the possible role of divalent cations in filament formation. J. Mol. Biol. *284*, 1611–1624.

Haliloglu, T., and Bahar, I. (1999). Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. Proteins *37*, 654–667.

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

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

Karplus, M. (2006). Foreword. In Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems, *Volume 9*, Q. Cui and I. Bahar, eds. (London: CRC Press), pp. v–ix.

Kay, L.E. (1998). Protein dynamics from NMR. Nat. Struct. Biol. *5*, 513–517.

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

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

Kuszewski, K., Gronenborn, A.M., and Clore, G.M. (1999). Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. J. Am. Chem. Soc. *121*, 2337–2338.

Lu, M., and Ma, J. (2005). The role of shape in determining molecular motions. Biophys. J. *89*, 2395–2401.

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

Ming, D., and Bruschweiler, R. (2006). Reorientational contact-weighted elastic network model for the prediction of protein dynamics: comparison with NMR relaxation. Biophys. J. *90*, 3382–3388.

Myers, E.W., and Miller, W. (1988). Optimal alignments in linear space. Comput. Appl. Biosci. *4*, 11–17.

Nabuurs, S.B., Spronk, C.A., Krieger, E., Maassen, H., Vriend, G., and Vuister, G.W. (2003). Quantitative evaluation of experimental NMR restraints. J. Am. Chem. Soc. *125*, 12026–12034.

Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A., Nabuurs, S.B., Guntert, P., Livny, M., Markley, J.L., Nilges, M., et al. (2005). RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins *59*, 662–672.

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

Powers, R., Clore, G.M., Garrett, D.S., and Gronenborn, A.M. (1993). Relationships between the precision of high resolution protein NMR structures, solution order parameters and crystallographic B factors. J. Magn. Reson. B *101*, 325–327.

Sanejouand, Y.-H. (2006). Functional information from slow mode shapes. In Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems, Volume 9, Q. Cui and I. Bahar, eds. (London: CRC Press), pp. 91–109.

Shindyalov, I.N., and Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. *11*, 739–747.

Spronk, C.A., Nabuurs, S.B., Bonvin, A.M., Krieger, E., Vuister, G.W., and Vriend, G. (2003). The precision of NMR structure ensembles revisited. J. Biomol. NMR *25*, 225–234.

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

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

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

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

Yang, D., and Kay, L.E. (1996). Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: application to protein folding. J. Mol. Biol. *263*, 369–382.

Yang, L.-W., Liu, X., Jursa, C.J., Holliman, M., Rader, A.J., Karimi, H., and Bahar, I. (2005). *i*GNM: 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., and Bahar, I. (2006). *o*GNM: online computation of structural dynamics using the Gaussian network model. Nucleic Acids Res. *34*, W24–W31.

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