



Published in final edited form as:

Proteins. 2012 May ; 80(5): 1425–1435. doi:10.1002/prot.24040.

Structural features that predict real-value fluctuations of globular proteins

Michal Jamroz^{1,2}, Andrzej Kolinski¹, and Daisuke Kihara^{2,3,4,*}

¹Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warszawa, Poland ²Department of Biological Sciences, College of Science, Purdue University, West Lafayette, Indiana 47907, USA ³Department of Computer Science, College of Science, Purdue University, West Lafayette, Indiana 47907, USA ⁴Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, Indiana 47907, USA

Abstract

It is crucial to consider dynamics for understanding the biological function of proteins. We used a large number of molecular dynamics trajectories of non-homologous proteins as references and examined static structural features of proteins that are most relevant to fluctuations. We examined correlation of individual structural features with fluctuations and further investigated effective combinations of features for predicting the real-value of residue fluctuations using the support vector regression. It was found that some structural features have higher correlation than crystallographic B-factors with fluctuations observed in molecular dynamics trajectories. Moreover, support vector regression that uses combinations of static structural features showed accurate prediction of fluctuations with an average Pearson's correlation coefficient of 0.669 and a root mean square error of 1.04 Å. This correlation coefficient is higher than the one observed for the prediction by the Gaussian network model. An advantage of the developed method over the Gaussian network models is that the former predicts the real-value of fluctuation. The results help improve our understanding of relationships between protein structure and fluctuation. Furthermore, the developed method provides a convenient practical way to predict fluctuations of proteins using easily computed static structural features of proteins.

Keywords

protein flexibility; protein dynamics; structure-dynamics relationship; support vector regression; molecular dynamics; fluctuation prediction

Introduction

Thanks to worldwide efforts in structural genomics^{1–3} we now know over 75 thousand protein tertiary structures⁴. This number is only a small fraction as compared with of the number of known protein sequences. Computational methods can predict structures for more than a half of newly sequenced proteins by means of template-based modeling with a sufficiently high accuracy^{5–8}. For some of the remaining proteins, it is possible to predict their structures in a *de novo* fashion if they are small and structurally simple^{9–14}. Thus, the problem of protein structure prediction is practically gradually being solved, and it may be

*Corresponding Author, dkihara@purdue.edu, Tel: (765) 496-2284, Fax: (765) 496-1189.

The authors declare no conflict of interest.

completely solved in the near future. Obviously, for the most difficult (and “atypical”) cases of monomeric structures and to a much larger extent for the plethora of possible protein-protein (and protein-nucleic acid, protein-carbohydrate, etc.) complexes structure prediction will remain a challenging task for decades^{9;15-17}. The knowledge of protein tertiary structures facilitates fast developments in various branches of molecular medicine and biotechnology^{18;19}. It, however, becomes more and more obvious that in order to understand the underlying molecular mechanisms of life we need to see biomolecules “in action”.

Protein dynamics, resulting from a specific flexibility of their structures, has drawn much attention recently in both theoretical and experimental molecular biology. Studies of dynamics of protein structures and their assemblies are important for understanding the mechanisms of protein function in various cellular processes^{20;21}, in particular, ligand binding, enzymatic reactions²², conformational diseases²³, and protein-protein interaction²⁴. The understanding of protein flexibility is also important for practical applications such as development of computer-aided methods of enzyme design^{25;26} and drug development²⁷.

In X-ray protein crystallography, which determines the Cartesian coordinates of atoms in proteins, uncertainties/fluctuations of atomic positions are provided in the form of B-factors²⁸. The B-factor measures the mobility of atoms, but it also reflects some inherent aspects of crystallographic techniques. Moreover, fluctuations estimated by B-factors are influenced by the molecular environment of the crystal structure. Protein mobility in solution could differ qualitatively from that in a crystal. Eastman et al. showed that B-factors are an accurate measure of fluctuations for stable parts of proteins, but significantly underestimate motion in flexible regions²⁹. Somewhat more straightforward measures of structure fluctuations could be derived from NMR experiments, although resulting estimates can be flawed by various limitations of actual measurements and by the computational schemes of their interpretation³⁰⁻³³. Therefore, these methods do not fully reflect actual fluctuations of proteins.

Molecular Dynamics (MD) is the most straightforward method for theoretical studies of dynamic aspects of molecular systems. Due to the progress in computing technology it is now practical to simulate protein systems in a timescale of tens of nanoseconds. Nevertheless, such simulations remain costly. With a significantly less computational requirement, the internal motion of a protein can be approximated by the normal mode analysis of a harmonic model of proteins³⁴. Another possibility is to employ simulations using coarse grained representations of protein structures. A simple approach is the Gaussian Network Model (GNM) and its derivatives³⁵⁻³⁸. Long-time simulation at an intermediate resolution can be achieved using simplified protein models, such as UNRES³⁹ and CABS⁴⁰. These models enable a low resolution study of dynamics (or stochastic dynamics) in timescales by a few orders of magnitude longer than possible by all-atom MD⁴¹⁻⁴⁴. A weak point of studying dynamics using coarse grained models is a lack of straightforward scaling between the models' time and the real time. Thus, all-atom MD simulations should be always used as a reference for coarse grained dynamics.

A number of computational methods for predicting protein fluctuations have been published; however, almost all of them evaluated their prediction results mainly in comparison with the crystallographic B-factor of proteins. As discussed above, the B-factor does not fully capture the mobility of proteins in solution. As we show in this work the fluctuations observed in MD and the B-factor correlate rather poorly, as was also concluded in a previous work²⁹.

There are a series of works that use GNM or its variants for predicting B-factors of proteins^{35;38;45;46}. Micheletti et al. extended GNM by adding C β atoms (β GM). The fluctuations of residues predicted by β GM were compared with the fluctuations from the MD simulation of HIV-1 protease⁴⁷. The self-consistent pair contact probability method, which is similar in its spirit to GNM, was used to predict fluctuations and compared with B-factors⁴⁸. The Zhou group developed an all-atom mean-field model to predict fluctuations⁴⁹.

Structural features of proteins were also investigated that can indicate fluctuations represented by B-factors. These features include solvent accessibility of residues⁵⁰, distance from a residue to the center of mass of the protein⁵¹, eigenvectors of the square distance matrix⁵², and predicted local fragment structures⁵³. An alternative direction pursued was to predict B-factors from protein sequences. Machine learning methods, such as Support Vector Machine^{54;55}, the random forest algorithm⁵⁶, or an artificial neural network⁵⁷, were used to predict fluctuations using sequence information and structural features that can be predicted from sequences, such as the secondary structure and the accessible surface area of residues.

In this work we used Support Vector Regression (SVR) to investigate the relationship between protein structure and dynamics. We employed various structural characteristics as well as structure fluctuation profiles predicted by GNM as input for SVR. The target reference is the dynamics observed in long MD simulations for a representative set of 592 globular proteins. To the best of our knowledge, this is the first time that protein fluctuations have been investigated on such a large dataset of MD simulations. In this context we also analyzed differences of protein dynamics deduced from the B-factors and the in-solvent dynamics computed by MD simulations. A more practical purpose of this work is to provide a fast (essentially instantaneous in comparison to MD) and reliable method that can be used for predicting fluctuations of protein structures. Unlike existing works mentioned above, we predict the real-value of residue fluctuations rather than simply showing correlation between predicted and actual fluctuations values. Remarkably, our method predicts fluctuation highly accurately with an average error of less than 1.1 Å. The correlation coefficient of our prediction with the actual fluctuations observed in MD simulations is higher than that of GNM. We also found that some of the static structural features, such as residue contact number, have higher correlation with the residue fluctuation in MD simulation than B-factors do. The developed software for predicting fluctuation, named flexPred, has been made freely available for the academic community.

Materials and Methods

Dataset of molecular dynamics trajectories

The MD trajectories of proteins were selected from MoDEL (Molecular Dynamics Extended Library)⁵⁸. Out of 1897 entries in the database, the following entries were discarded: trajectories for protein structures solved by nucleic magnetic resonance (NMR), those which include more than one protein chain in the simulation, and trajectories for proteins whose length differ from the corresponding entries in the Protein Data Bank (PDB)⁴. These MD trajectories were computed using AMBER⁵⁹, GROMACS⁶⁰, or NAMD⁶¹ force fields. If more than one simulation is available for a protein, we used the first one with an earlier entry date in the database. The MoDEL trajectory files were uncompressed with the PCASuite software⁶². 837 trajectories remained after this filtering process. From this subset, we removed redundant proteins using the PISCES server⁶³ with a sequence identity cutoff of 35%. The final number of trajectories is 592. This dataset contains proteins from all main classes in the CATH database⁶⁴: 111 proteins in the α class (18.75%), 149 proteins in the β class (25.17%), 256 in the $\alpha\beta$ class (43.24%), and 76 in the few secondary structure

class (12.84%). The length of the proteins ranges from 21 to 994 residues (Figure 1). The simulation time was 10 ns for most of the proteins (96.11%), while the rest of the proteins had shorter trajectories: 5 nanoseconds (ns) (0.33%), 2 ns (2.36%), 1 ns (0.5%), and one protein each with 6.5 ns, 6.0 ns, 5.5 ns, and 4.5 ns.

Definition of fluctuation

The fluctuation of amino acid residue i is defined in two ways. It can be defined as a root mean square deviation (RMSD) of the mean position of an atom in an MD trajectory:

$$\sqrt{\langle(\Delta R_i)^2\rangle}^{MD} = \sqrt{\frac{1}{T} \sum_{t_j=1}^T (x_i(t_j) - \langle x_i \rangle)^2} \quad (\text{Eq. 1})$$

$x_i(t_j)$ is the Cartesian coordinates of the $\text{C}\alpha$ atom of residue i at time t_j in the trajectory, T is the number of time frames in the trajectory, and $\langle x_i \rangle$ is the average position of the $\text{C}\alpha$ atom of residue i in the trajectory. We also used the coordinates in the PDB file as the reference:

$$\sqrt{\langle(\Delta R_i)^2\rangle}^{ref} = \sqrt{\frac{1}{T} \sum_{t_j=1}^T (x_i(t_j) - x_i^{ref})^2}, \quad (\text{Eq. 2})$$

where x_i^{ref} is the coordinates of the $\text{C}\alpha$ atom of residue i in the PDB file. The distance of residue positions is computed after superimposing the PDB structure on each frame. If alternative positions of the atom are recorded in the PDB files, the first position of the atom was used. As shown in Figure 2, these two definitions give similar fluctuations of residues, but not identical. The correlation coefficient of the two fluctuation values is 0.86. The fluctuation value is smaller when the mean of a trajectory is used as the reference (Eq. 1) in almost all the cases (99.9 %). Unless noted, we use the second definition of fluctuation (Eq. 2) in the results that will be shown below, because we compare the fluctuations from MD with B-factors and GNM, both of which are attributed to PDB structures.

Structural features of proteins

We considered the following static protein structural features.

1. B-factor (temperature factor)²⁸. The B-factor reflects dynamic motion, the static disorder of the atom in the crystal structure, and also errors in model building. The B-factor values are taken from the PDB file.
2. Square of the distance between a residue and the protein center of mass, which is defined as follows:

$$r_i^2 = \left(x_i - \frac{1}{N} \sum_{j=1}^N x_j \right)^2, \quad (\text{Eq. 3})$$

where x_i is the position of the $\text{C}\alpha$ atom of residue i . A previous paper showed that this parameter has good correlation with the B-factor^{51;52}.

3. Residue contact number, which is defined as the number of surrounding residues whose $\text{C}\alpha$ atom is closer than a cutoff distance. The contact number was also shown to correlate well with the B-factor^{65;66}.
4. Number of hydrophobic/hydrophilic residue contacts, where the number of residue contacts is separately counted for hydrophobic and hydrophilic residues.

Hydrophobic/hydrophilic residues are those which have a positive/negative value on the Kyte-Doolittle hydrophobicity scale ⁶⁷.

5. Solvent accessibility surface area (\AA^2). This parameter is defined as water exposed surface of a residue. We used the DSSP program ⁶⁸ to compute the accessibility surface area of amino acids, which are then normalized with the value in the tripeptide with glycines on both sides of the target amino acid residue ⁶⁹.
6. Residue depth, which is defined as the distance of the $C\alpha$ atom or the average distance of all the atoms in a residue to the closest water molecule ⁷⁰. Protein surface was computed with the MSMS program ⁷¹. The *hsexpo* program was used to compute residue depth ⁷².
7. Lower/upper half-sphere exposure of a residue ⁷², which is defined as the number of contacts within a half-sphere of a radius of 13 \AA centering at either the $C\alpha$ or the $C\beta$ atom of the residue. The sphere is divided into half by a plane perpendicular to the $C\alpha$ - $C\beta$ vector.
8. Secondary structure. Each residue is classified into eight classes, i.e. seven secondary structure types defined by DSSP ⁶⁸ or other.
9. Fluctuations predicted by the Gaussian Network Model (GNM) ^{35;36}. GNM is a coarse-grained model where $C\alpha$ atoms are connected by springs. GNM has been used for investigating protein dynamics including the prediction of B-factor values of proteins ³⁸. We downloaded GNM codes from the Jernigan laboratory (<http://ribosome.bb.iastate.edu/>). Fluctuations were computed with a residue contact distance cutoff of 16 \AA ⁷³ and without using cutoff ³⁸. Residue contacts in a protein are represented as the Kirchhoff matrix in GNM:

$$\Gamma_{ij} = \begin{cases} -1 & \text{if } i \neq j \text{ and } r_{ij} \leq r_c \\ 0 & \text{if } i \neq j \text{ and } r_{ij} > r_c \\ -\sum_{i,i \neq j}^N \Gamma_{ij} & \text{if } i=j \end{cases}, \quad (\text{Eq. 4})$$

where r_{ij} is the distance between two atoms, i and j , and $r_c (= 16 \text{\AA})$ is the cutoff value. GNM without cutoff uses the following modified Kirchhoff matrix:

$$\Gamma_{ij} = \begin{cases} r_{ij}^{-2} & \text{if } i \neq j \\ -\sum_{i,i \neq j}^N \Gamma_{ij} & \text{if } i=j \end{cases}, \quad (\text{Eq. 5})$$

In both methods, the average fluctuation of residue i over time is defined by:

$$\langle (\Delta R_i)^2 \rangle = C (\Gamma_{ii}^{-1}), \quad (\text{Eq. 6})$$

where C is constant.

Support Vector Regression

We combined the structural features listed above to predict fluctuations using Support Vector Regression (SVR). The LIBSVM package ⁷⁴ with Gaussian kernels was used. Since it was not feasible to test all the possible combinations of features, features were added or changed one at a time starting from the one which has the largest correlation coefficient with residue fluctuation. We performed five-fold cross validation using the dataset of trajectories.

The default set of parameters in *libsvm*, $C = 64.0$, $\gamma = 1$, and $\epsilon = 0.5$, was used, which was shown to perform best among others tested in the first few feature combinations in the five-fold cross validation (data not shown).

Evaluation of fluctuations prediction

Pearson's correlation coefficient was used to examine how well individual features or predicted fluctuations correlated with actual fluctuations in the MD trajectories. Average correlation coefficients were computed using all the trajectories in the dataset.

In addition, the error of predicted fluctuations was quantified as the RMSD to the reference trajectory fluctuation:

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\Delta R_i^{pred} - \sqrt{\langle (\Delta R_i)^2 \rangle^{ref}} \right)^2}, \quad (\text{Eq. 7})$$

where N is the length of the protein, ΔR_i^{pred} is predicted, and $\sqrt{\langle (\Delta R_i)^2 \rangle^{ref}}$ is actual fluctuation (Eq. 2) of residue i .

Availability of the developed program

The program for predicting the fluctuation of residues in a protein structure is made freely available for the academic community at <http://kiharalab.org/flexPred/>. Both the web server and the source code written in Python are available. It takes a PDB file of a query protein for input data and outputs a predicted fluctuation value for each residue. The computational time for a protein is typically within a couple of seconds to twenty seconds depending on the length of the protein.

Results and Discussion

The relationships between structural features and residue fluctuations are examined in several aspects. First, we compare the correlation coefficient of individual static structural features with actual fluctuations. Then, we explore different combinations of features to make accurate prediction of fluctuations using SVR. Then, the accuracy of the fluctuation prediction by SVR and by GNM is further examined. Finally, we also consider the structural variation of models by NMR in comparison with prediction as well as the fluctuations observed in MD trajectories.

Correlation of static structural features of proteins with fluctuations

In Table 1, we compared the correlation coefficient of individual structural features with the fluctuation of residues observed in the MD trajectories. Eight different distance cutoff values, 6, 8, to 16 Å, were used for the residue contact number. The top of the table shows the correlation of the B-factor (0.484). Interestingly, several static structural features, namely, the distance to the center of mass, the contact number computed with the cutoff of 12–22 Å, have more significant correlation with the fluctuations than the B-factor. Among the static features, the largest correlation coefficients were observed for the residue contact number (15 and 16 Å). These results indicate that the motion of chains in the MD trajectories is better captured by the coarse-grained topological structures of proteins rather than the B-factor.

As a reference, we also show the correlation of the fluctuations predicted by GNM (bottom rows of Table 1). GNM showed higher correlation than the other structural features. Note

that GNM actually simulates dynamic motion of protein structures; thus it has a different nature from the other static features compared in the table. Consistently with the previous work by Yang et al.³⁸, GNM without using a distance cutoff showed higher correlation than GNM with a distance cutoff.

Since the residue contact number (with a 16 Å cutoff) and the square of distance to the center of mass showed two largest correlation coefficients among the static structure features examined, we used these two features as the basis for combinations of input features for training SVR in the next section.

SVR models for predicting residue fluctuation using static structure features

Next, we employed SVR to predict the fluctuation of residue positions in the MD trajectories using various combinations of static structural features. Fluctuation predictions by GNM (at the bottom of Table 1) were not included as features. Five-fold cross validation was performed, in which SVR parameters were trained on four-fifths of the dataset, while prediction was made for the rest of the one-fifth of the dataset. This procedure was repeated five times to make prediction for all data in the dataset. Starting from the combination of the residue contact number (with 16 Å cutoff) and the square of distance to the center of mass, which are the two features that showed the highest correlation with fluctuations (Table 1), seventeen different feature combinations were tested by adding one feature at a time (Table 2).

Among the seventeen feature combinations examined, all except for two (the feature set 1 and set 17) showed higher correlation with actual fluctuations than GNM (Table 1). The largest correlation coefficient, 0.669, was achieved for the feature set 15, which uses the residue contact numbers with different distance cutoffs. In terms of average RMS, all the feature combinations predicted residue fluctuations within an RMS of 1.1 Å, ranging from 1.042 Å to 1.092 Å. The smallest RMS was achieved for feature sets 6, 7, 12, 13, and 14, which combine the residue contact numbers, the square distance from the center of mass, and the B-factor. Sets 6 and 7 additionally used information about the secondary structure. The RMS and the average correlation coefficients (Table 2) correlate moderately with a correlation coefficient of 0.627 (Fig. 3). Figure 4 shows the distribution of the average correlation coefficients between predicted and actual fluctuations (Fig. 4A) and the average RMS (Fig. 4B) for each protein, which were predicted using feature set 12. Remarkably, the majority (70%) of proteins fluctuations were predicted within an RMS of 1.0 Å. The strong advantage of the developed SVR models is that they predict the real-value of fluctuation, unlike GNM which predicts only the relative magnitude of residue fluctuations that need to be rescaled to obtain actual fluctuation values.

Incorporating dynamic (GNM) features to SVR models

We further investigated whether adding GNM as an input feature can improve fluctuations prediction with SVR. We used $\langle(\Delta R_i)^2\rangle$ for the fluctuations from GNM (Eq. 6) without a distance cutoff because it has higher correlation with the actual fluctuations than $\sqrt{\langle(\Delta R_i)^2\rangle}$ does. To each of the feature sets examined in Table 2 we added $\langle(\Delta R_i)^2\rangle$ predicted by GNM and performed five-fold cross validation. The resulting fluctuation prediction with and without GNM was compared in terms of the correlation coefficient (Fig. 5A) and the RMS (Fig. 5B) with the actual fluctuations.

Adding GNM in the feature set made slight improvement in the RMS of the predicted fluctuations (Fig. 5B) except for one case (feature set 12), lowering RMS on average by 0.010. However, small consistent deterioration of the correlation coefficient was observed

(Fig. 5A) when GNM was added. The average decrease in the correlation coefficient is 0.013. Thus, GNM did not make significant contribution to improving fluctuation prediction.

Comparison of SVR model prediction results with B-factor fluctuation values

In Figure 6, we show four examples of actual and predicted fluctuations as well as fluctuations derived from the B-factors. For residue i with a B-factor of B_i , the fluctuation is defined as

$$\sqrt{\langle(\Delta R_i)^2\rangle}^{Bfactor} = \sqrt{\frac{3B_i}{8\pi^2}}. \quad (\text{Eq. 8})$$

The fluctuations from the B-factor was also rescaled to achieve a smaller RMS with the actual fluctuations (i.e. fluctuations from MD trajectories) as follows

$$\sqrt{\langle(\Delta R_i)^2\rangle}_{rescaled} = \sqrt{\langle(\Delta R)^2\rangle}_{min} + \alpha \left(\sqrt{\langle(\Delta R)^2\rangle}_{max} - \sqrt{\langle(\Delta R)^2\rangle}_{min} \right) \frac{\sqrt{\langle(\Delta R_i)^2\rangle}^{Bfactor} - \sqrt{\langle(\Delta R)^2\rangle}_{min}}{\sqrt{\langle(\Delta R)^2\rangle}_{max} - \sqrt{\langle(\Delta R)^2\rangle}_{min}},$$

where $\sqrt{\langle(\Delta R)^2\rangle}_{max}$ and $\sqrt{\langle(\Delta R)^2\rangle}_{min}$ are the maximum and the minimum values of actual fluctuations, and $\sqrt{\langle(\Delta R)^2\rangle}_{max}^{Bfactor}$ and $\sqrt{\langle(\Delta R)^2\rangle}_{min}^{Bfactor}$ are the maximum and the minimum fluctuation values computed from B-factor values (Eq. 8) in the protein. α is a weighting factor explored from 0.1 to 1.0 with an interval of 0.1 to seek smaller RMS for the actual fluctuations (Table 3). In Figure 6, α is set to 1.0 for the plots of “Fluctuation from B-factor, rescaled”. Note that this rescaling obviously changes the RMS but does not change the correlation coefficient to the actual fluctuation. The actual fluctuations in the MD trajectories are defined by Equation 2 and predictions were made using feature set 15 in Table 2. The right panel of each protein visualizes the magnitude of actual fluctuations in a color scale from blue to red with blue showing small while red for large fluctuation.

The first example, retrovirus coat protein (PDB ID: 1mof) (Figs. 6A, B), exhibits a large fluctuation at two termini and at the end of the long helix. Prediction by SVR captured fluctuating residues and the magnitude fairly well with a correlation coefficient of 0.80 and an RMS of 1.55 Å. The fluctuations derived from B-factor have lower correlation with the actual fluctuations (correlation coefficient of 0.69) with a larger RMS of 1.91 Å even after rescaling. In the second example (Figs. 6C, D) of homing endonuclease PI-PfuI (PDB ID: 1dq3), overall fluctuation is not large but shows high peaks of fluctuation at loop regions. The predicted fluctuations have a correlation coefficient of 0.81 while the fluctuations from B-factor have a moderate correlation of 0.50. The third example, DNA-binding protein gp32 (PDB ID: 1gpc) (Figs. 6E, F), has the largest fluctuation at the loop of residues 150–160 and over 3 Å fluctuation at the other loop regions, which are captured well by the prediction. Predicted fluctuations have a correlation coefficient of 0.78 and a small RMS of 1.04 Å. In contrast, the correlation of fluctuations from B-factor is 0.55 with a larger RMS of 1.93 Å. The last example, MTCP-1 (PDB ID: 1a1x) (Figs. 6G, H) is a β -barrel protein with a long loop at residues 50 to 60. Relatively large fluctuation was observed at the N-terminus and at the loop regions that connect β -strands (e.g. residues 35–40), which are well predicted. The overall RMS of the prediction is 0.79 Å and the correlation coefficient with the actual fluctuations is 0.82, better than the fluctuations derived from B-factors.

Consistent with Table 1, the fluctuations from B-factors correlate only moderately with the actual fluctuations. Fluctuations computed from B-factors using Eq. 8 have always a larger

RMS than the SVR prediction. The agreement of the fluctuations from B-factors can be improved if it is rescaled individually for each protein as shown in the second column from the right in Table 3; however, the value of the optimal scaling factor α differs from protein to protein and thus cannot be known beforehand. In contrast, our prediction by SVR has a significantly higher correlation with the actual prediction and it predicts the real-value of the fluctuations satisfactorily without any rescaling.

MD fluctuations and fluctuations from NMR models

The MoDEL database also contains simulations of protein structures determined by NMR. We selected 140 non-redundant protein structures determined by NMR that contain more than ten models in their PDB files. Redundant proteins were removed by considering sequence identity according to the PISCES database⁶³. Using the 140 proteins, we compared fluctuations observed in the NMR models, MD trajectories, and the predicted fluctuations. The results are summarized in Table 4. The fluctuation prediction was carried out using feature set 16, which does not contain the B-factor term (NMR structures do not have B-factors).

It is shown that the prediction has a significant correlation (0.739) with the structural variation of the models derived from NMR. Interestingly, the correlation coefficient between the prediction and NMR is highest among the other two pairs, prediction vs. MD and NMR vs. MD.

Conclusion

We used a large number of MD trajectories of non-homologous proteins as references and examined static structural features of the proteins that are most relevant to fluctuations. We examined the correlation of individual structural features with fluctuations and then investigated effective combinations of features for SVR to predict the real-value of fluctuation of residues. The main findings of this work are summarized as follows. First of all, two types of structural features, the distance to the center of mass of the protein and the residue contact number, showed a higher correlation coefficient with fluctuations than B-factor does. Combinations of static features used as input for SVR achieved accurate prediction of fluctuations with a correlation coefficient of 0.67 and RMS of 1.042 Å. This correlation coefficient is higher than GNM has to the actual fluctuation. Our method predicts the structural variation of NMR models also well. The current study demonstrates that flexibility of proteins is inherently coded in coarse-grained static protein structural features, even more than in the crystallographic B-factors. Thus, protein motion is determined by its static structure that is coded by its sequence, which could be considered as an extension of the Anfinsen's dogma⁷⁵. Indeed series of studies on GNM has also demonstrated that motion of a protein is determined by its structure. However, the current work further shows that static structural features can predict the real-value of fluctuations, which GNM has not been shown to be able to do. As the importance of protein dynamics has been more recognized for biological function, the prediction method we developed has also a practical value in the wide areas of biology and biotechnology.

Acknowledgments

We thank Dr. Jordi Camps (Centre Nacional d'Anàlisi Genòmica, Spain) and Dr. Tim Meyer (Institute for Research in Biomedicine, Spain) for help with the PCAsuite software and the MoDEL database. MJ acknowledges the support from a Project operated within the Foundation for Polish Science MPD Programme co-financed by the EU European Regional Development Fund. This work was supported by a grant from the National Science Foundation (IIS0915801). DK also acknowledges grants from the National Institutes of Health (R01GM075004 and R01GM097528) and the National Science Foundation (DMS0800568, EF0850009).

Reference List

1. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science*. 2006; 311:347–351. [PubMed: 16424331]
2. Todd AE, Marsden RL, Thornton JM, Orengo CA. Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol*. 2005; 348:1235–1260. [PubMed: 15854658]
3. Westbrook J, Feng Z, Chen L, Yang H, Berman HM. The Protein Data Bank and structural genomics. *Nucleic Acids Res*. 2003; 31:489–491. [PubMed: 12520059]
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000; 28:235–242. [PubMed: 10592235]
5. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*. 2006; 34:D291–D295. [PubMed: 16381869]
6. Kihara D, Skolnick J. Microbial Genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. *Proteins*. 2004; 55:464–473. [PubMed: 15048836]
7. Zhang Y. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol*. 2008; 18:342–348. [PubMed: 18436442]
8. Chen H, Kihara D. Effect of using suboptimal alignments in template-based protein structure prediction. *Proteins*. 2011; 79:315–334. [PubMed: 21058297]
9. Das R, Baker D. Macromolecular modeling with rosetta. *Annu Rev Biochem*. 2008; 77:363–382. [PubMed: 18410248]
10. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science*. 2005; 309:1868–1871. [PubMed: 16166519]
11. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci U S A*. 2001; 98:10125–10130. [PubMed: 11504922]
12. Kihara D, Zhang Y, Lu H, Kolinski A, Skolnick J. Ab initio protein structure prediction on a genomic scale: application to the *Mycoplasma genitalium* genome. *Proc Natl Acad Sci U S A*. 2002; 99:5993–5998. [PubMed: 11959918]
13. Borreguero JM, Skolnick J. Benchmarking of TASSER in the ab initio limit. *Proteins*. 2007; 68:48–56. [PubMed: 17444524]
14. Trojanowski S, Rutkowska A, Kolinski A. TRACER. A new approach to comparative modeling that combines threading with free-space conformational sampling. *Acta Biochim Pol*. 2010; 57:125–133. [PubMed: 20309433]
15. Venkatraman V, Sael L, Kihara D. Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. *Cell Biochem Biophys*. 2009; 54:23–32. [PubMed: 19521674]
16. Puton T, Kozlowski L, Tuszyńska I, Rother K, Bujnicki JM. Computational methods for prediction of protein-RNA interactions. *J Struct Biol*. 2011
17. Ritchie DW. Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci*. 2008; 9:1–15. [PubMed: 18336319]
18. Hillisch A, Pineda LF, Hilgenfeld R. Utility of homology models in the drug discovery process. *Drug Discov Today*. 2004; 9:659–669. [PubMed: 15279849]
19. Takeda-Shitaka M, Takaya D, Chiba C, Tanaka H, Umeyama H. Protein structure prediction in structure based drug design. *Curr Med Chem*. 2004; 11:551–558. [PubMed: 15032603]
20. Teilum K, Olsen JG, Kragelund BB. Functional aspects of protein flexibility. *Cell Mol Life Sci*. 2009; 66:2231–2247. [PubMed: 19308324]
21. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. *J Mol Graph Model*. 2001; 19:26–59. [PubMed: 11381529]
22. Hammes GG, Benkovic SJ, Hammes-Schiffer S. Flexibility, Diversity, and Cooperativity: Pillars of Enzyme Catalysis. *Biochemistry*. 2011

23. Chiti F, Dobson CM. Amyloid formation by globular proteins under native conditions. *Nat Chem Biol.* 2009; 5:15–22. [PubMed: 19088715]
24. Zacharias M. Accounting for conformational changes during protein-protein docking. *Curr Opin Struct Biol.* 2010; 20:180–186. [PubMed: 20194014]
25. Mandell DJ, Kortemme T. Backbone flexibility in computational protein design. *Curr Opin Biotechnol.* 2009; 20:420–428. [PubMed: 19709874]
26. Lassila JK. Conformational diversity and computational enzyme design. *Curr Opin Chem Biol.* 2010; 14:676–682. [PubMed: 20829099]
27. Lill MA. Efficient incorporation of protein flexibility and dynamics into molecular docking simulations. *Biochemistry.* 2011; 50:6157–6169. [PubMed: 21678954]
28. Debye P. Interferenz von Röntgenstrahlen und Wärmebewegung. *Annalen der Physik.* 1913; 348:49–92.
29. Eastman P, Pellegrini M, Doniach S. Protein flexibility in solution and in crystals. *Journal of Chemical Physics.* 1999; 110:10141–10152.
30. Ishima R, Torchia DA. Protein dynamics from NMR. *Nat Struct Biol.* 2000; 7:740–743. [PubMed: 10966641]
31. Baldwin AJ, Kay LE. NMR spectroscopy brings invisible protein states into focus. *Nat Chem Biol.* 2009; 5:808–814. [PubMed: 19841630]
32. Nilges M, Habeck M, O'Donoghue SI, Rieping W. Error distribution derived NOE distance restraints. *Proteins.* 2006; 64:652–664. [PubMed: 16729263]
33. Chalaoux FR, O'Donoghue SI, Nilges M. Molecular dynamics and accuracy of NMR structures: effects of error bounds and data removal. *Proteins.* 1999; 34:453–463. [PubMed: 10081958]
34. Brooks B, Karplus M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A.* 1983; 80:6571–6575. [PubMed: 6579545]
35. Haliloglu T, Bahar I, Erman B. Gaussian dynamics of folded proteins. *Physical Review Letters.* 1997; 79:3090–3093.
36. Tirion MM. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys Rev Lett.* 1996; 77:1905–1908. [PubMed: 10063201]
37. Bahar I, Erman B, Haliloglu T, Jernigan RL. Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. *Biochemistry.* 1997; 36:13512–13523. [PubMed: 9354619]
38. Yang L, Song G, Jernigan RL. Protein elastic network models and the ranges of cooperativity. *Proc Natl Acad Sci U S A.* 2009; 106:12347–12352. [PubMed: 19617554]
39. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comp Chem.* 1997; 18:849–873.
40. Kolinski A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol.* 2004; 51:349–371. [PubMed: 15218533]
41. He Y, Liwo A, Weinstein H, Scheraga HA. PDZ binding to the BAR domain of PICK1 is elucidated by coarse-grained molecular dynamics. *J Mol Biol.* 2011; 405:298–314. [PubMed: 21050858]
42. Kmiecik S, Kolinski A. Characterization of protein-folding pathways by reduced-space modeling. *Proc Natl Acad Sci U S A.* 2007; 104:12330–12335. [PubMed: 17636132]
43. Kmiecik S, Kolinski A. Folding pathway of the b1 domain of protein G explored by multiscale modeling. *Biophys J.* 2008; 94:726–736. [PubMed: 17890394]
44. Kmiecik S, Kolinski A. Simulation of chaperonin effect on protein folding: a shift from nucleation-condensation to framework mechanism. *J Am Chem Soc.* 2011; 133:10283–10289. [PubMed: 21618995]
45. Kondrashov DA, Cui Q, Phillips GN Jr. Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophys J.* 2006; 91:2760–2767. [PubMed: 16891367]
46. Lin TL, Song G. Generalized spring tensor models for protein fluctuation dynamics and conformation changes. *BMC Struct Biol.* 2010; 10 Suppl 1:S3. [PubMed: 20487510]

47. Micheletti C, Carloni P, Maritan A. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. *Proteins*. 2004; 55:635–645. [PubMed: 15103627]
48. Canino LS, Shen T, McCammon JA. Changes in flexibility upon binding: Application of the self-consistent pair contact probability method to protein-protein interactions. *J Chem Phys*. 2002; 117:9927–9933.
49. Pandey BP, Zhang C, Yuan X, Zi J, Zhou Y. Protein flexibility prediction by an all-atom mean-field statistical theory. *Protein Sci*. 2005; 14:1772–1777. [PubMed: 15987905]
50. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins*. 2009; 76:617–636. [PubMed: 19274736]
51. Shih CH, Huang SW, Yen SC, Lai YL, Yu SH, Hwang JK. A simple way to compute protein dynamics without a mechanical model. *Proteins*. 2007; 68:34–38. [PubMed: 17436324]
52. Kloczkowski A, Jernigan RL, Wu Z, Song G, Yang L, Kolinski A, Pokarowski P. Distance matrix-based approach to protein structure prediction. *J Struct Funct Genomics*. 2009; 10:67–81. [PubMed: 19224393]
53. Bornot A, Etchebest C, De Brevern AG. Predicting protein flexibility through the prediction of local structures. *Proteins*. 2011; 79:839–852. [PubMed: 21287616]
54. Gu J, Gribkov M, Bourne PE. Wiggle-predicting functionally flexible regions from primary sequence. *PLoS Comput Biol*. 2006; 2:e90. [PubMed: 16839194]
55. Chen P, Wang B, Wong HS, Huang DS. Prediction of protein B-factors using multi-class bounded SVM. *Protein Pept Lett*. 2007; 14:185–190. [PubMed: 17305606]
56. Hirose S, Yokota K, Kuroda Y, Wako H, Endo S, Kanai S, Noguchi T. Prediction of protein motions from amino acid sequence and its application to protein-protein interaction. *BMC Struct Biol*. 2010; 10:20. [PubMed: 20626880]
57. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. *Proteins*. 2005; 61:115–126. [PubMed: 16080156]
58. Meyer T, D'Abramo M, Hospital A, Rueda M, Ferrer-Costa C, Perez A, Carrillo O, Camps J, Fenollosa C, Repchevsky D, Gelpi JL, Orozco M. MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure*. 2010; 18:1399–1409. [PubMed: 21070939]
59. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. *J Comput Chem*. 2005; 26:1668–1688. [PubMed: 16200636]
60. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly Efficient, load-balanced, and scalable molecular simulation. *J Chem Theory and Computation*. 2008; 4:435–447.
61. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. *J Comput Chem*. 2005; 26:1781–1802. [PubMed: 16222654]
62. Meyer T, Ferrer-Costa C, Perez A, Rueda M, Bidon-Chanal A, Luque FJ, Laughton CA, Orozco M. Essential dynamics: A tool for efficient trajectory compression and management. *J Chem Theory and Computation*. 2006; 2:251–258.
63. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics*. 2003; 19:1589–1591. [PubMed: 12912846]
64. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. *Structure*. 1997; 5:1093–1108. [PubMed: 9309224]
65. Lin CP, Huang SW, Lai YL, Yen SC, Shih CH, Lu CH, Huang CC, Hwang JK. Deriving protein dynamical properties from weighted protein contact number. *Proteins*. 2008; 72:929–935. [PubMed: 18300253]
66. Halle B. Flexibility and packing in proteins. *Proc Natl Acad Sci U S A*. 2002; 99:1274–1279. [PubMed: 11818549]
67. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982; 157:105–132. [PubMed: 7108955]

68. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983; 22:2577–2637. [PubMed: 6667333]
69. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. *J Mol Biol*. 1987; 196:641–656. [PubMed: 3681970]
70. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure*. 1999; 7:723–732. [PubMed: 10425675]
71. Sanner, M.; Olson, AJ.; Spehner, JC. Fast and robust computation of molecular surfaces. *Proceedings of 11th ACM Symposium on Computational Geometry*; 1995. p. C6-C7.
72. Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins*. 2005; 59:38–48. [PubMed: 15688434]
73. Kundu S, Melton JS, Sorensen DC, Phillips GN Jr. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys J*. 2002; 83:723–732. [PubMed: 12124259]
74. Chang C-C, Jin C-J. LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. 2001; 2:27:1–27:27.
75. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973; 181:223–230. [PubMed: 4124164]

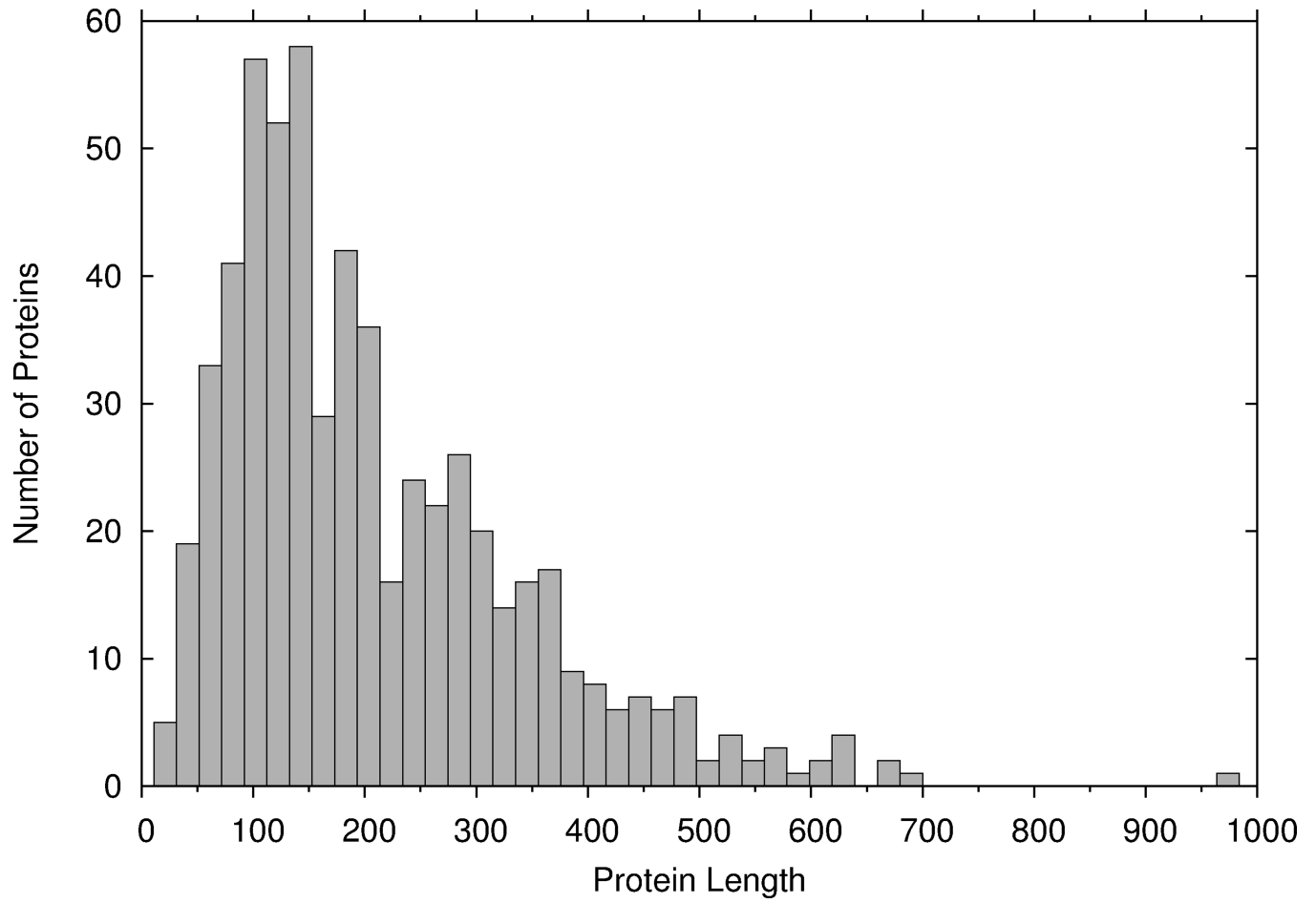


Figure 1.
Histogram of the length of proteins in the dataset. There are in total 592 proteins.

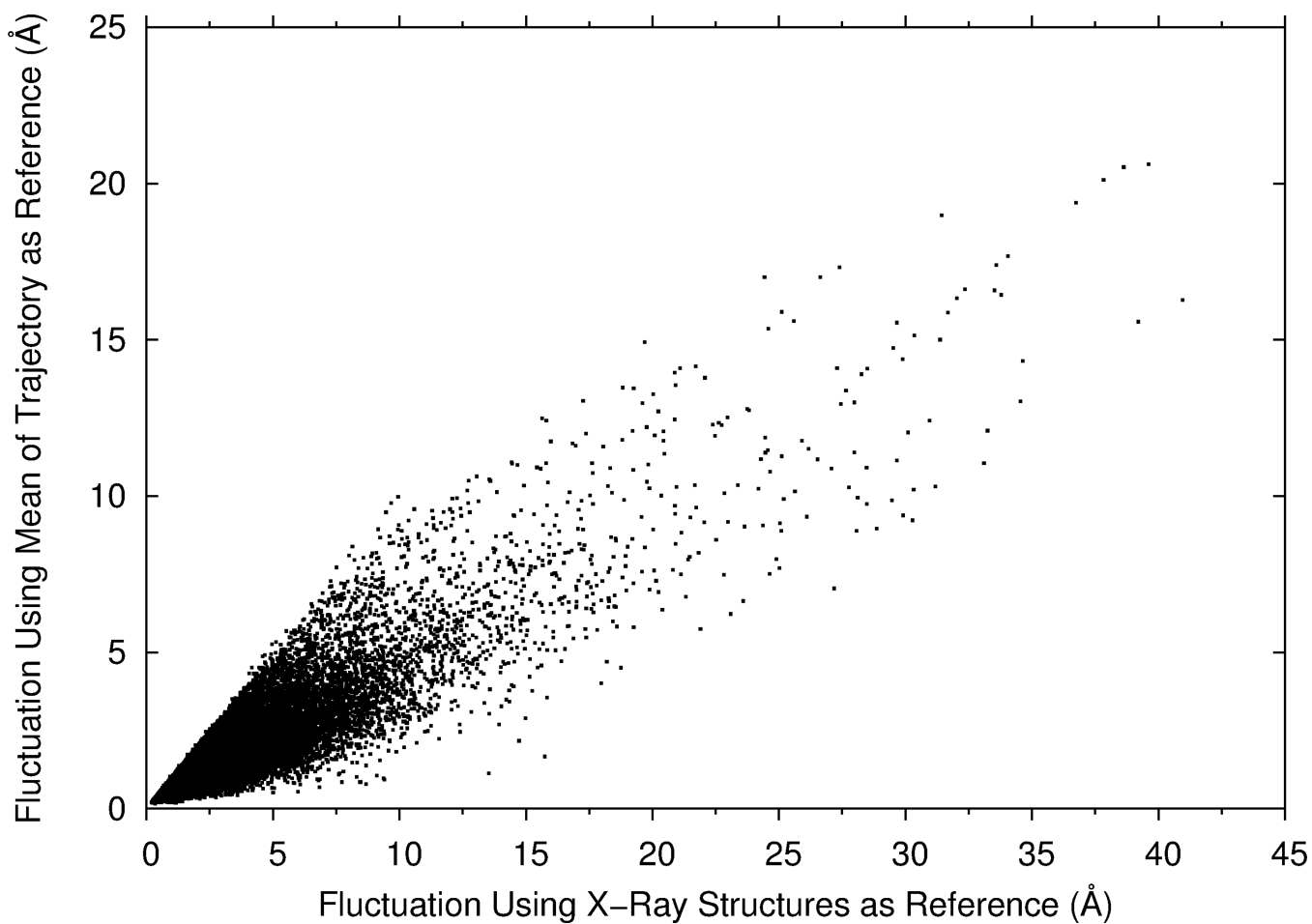


Figure 2. Average fluctuations of proteins in MD trajectories using two definitions. X values show fluctuations of residues relative to the crystal structures of proteins in the PDB (Eq. 2) while y values are fluctuations relative to the mean structure of each MD trajectory (Eq. 1).

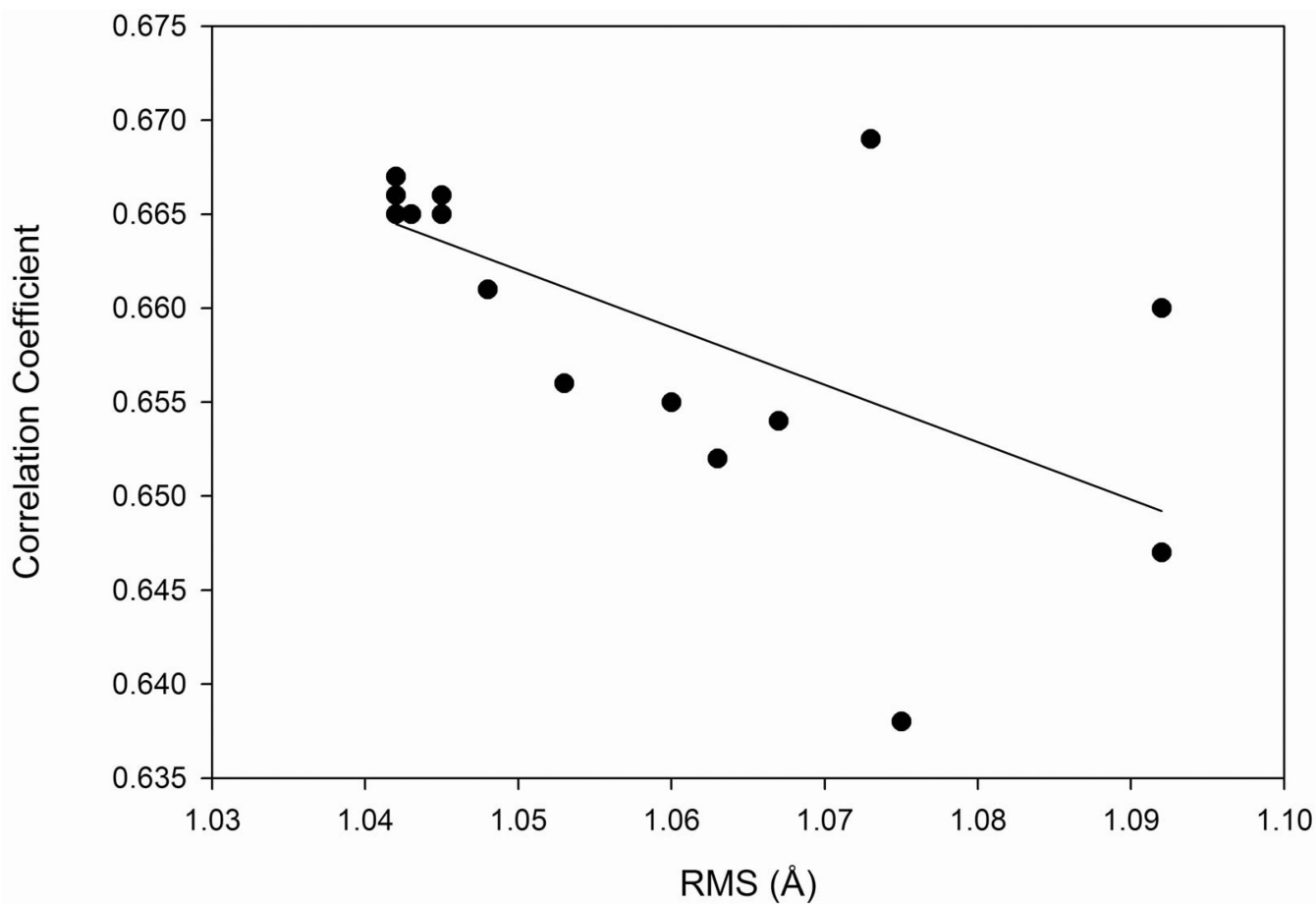


Figure 3.
The average correlation coefficient and RMS of predicted and actual fluctuations.
Predictions were made with SVR using seventeen different feature combinations (Table 2).

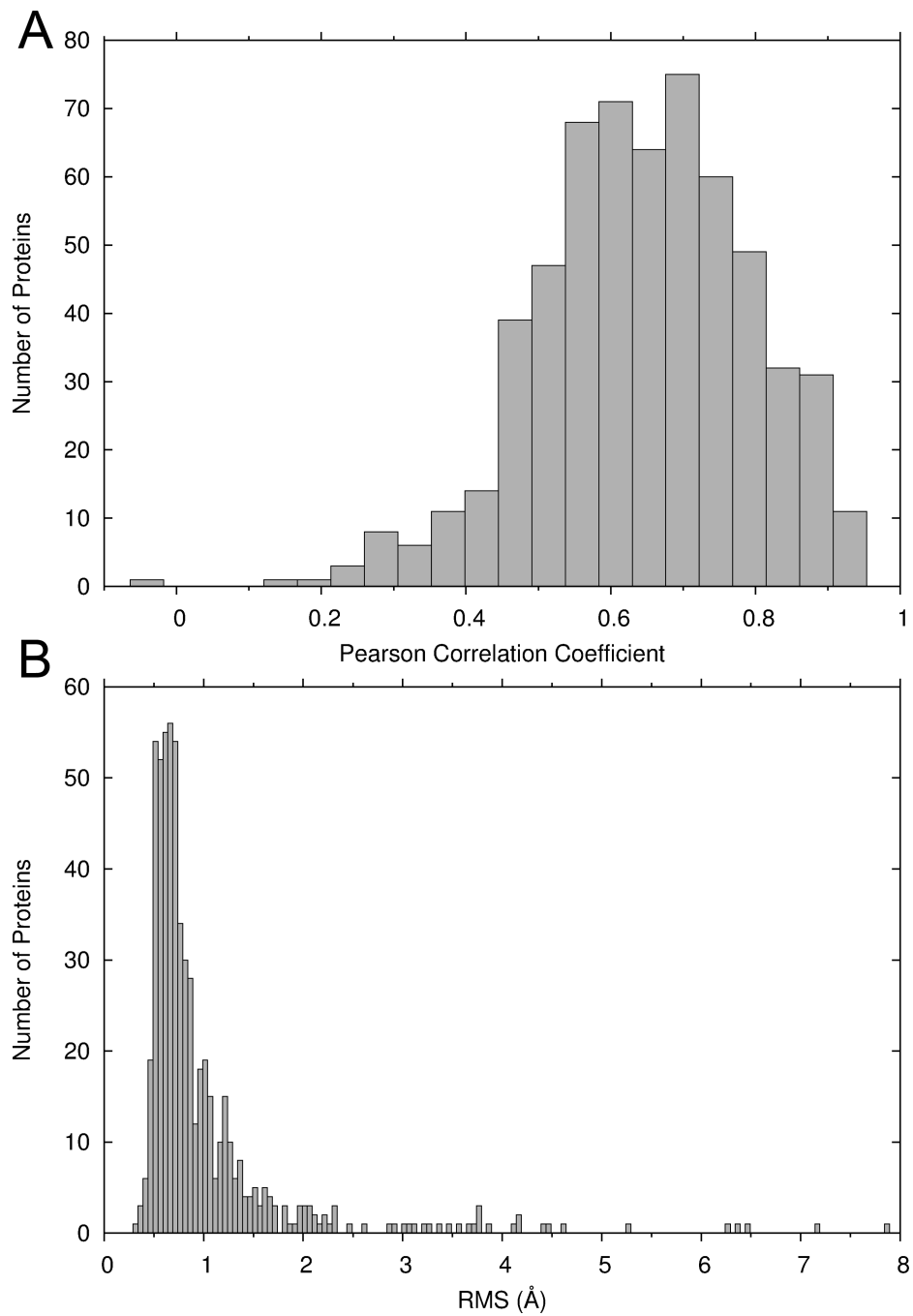


Figure 4. Distribution of **A**, correlation coefficients; **B**, RMS (Å); of predicted and actual fluctuations computed for 592 proteins in the dataset.

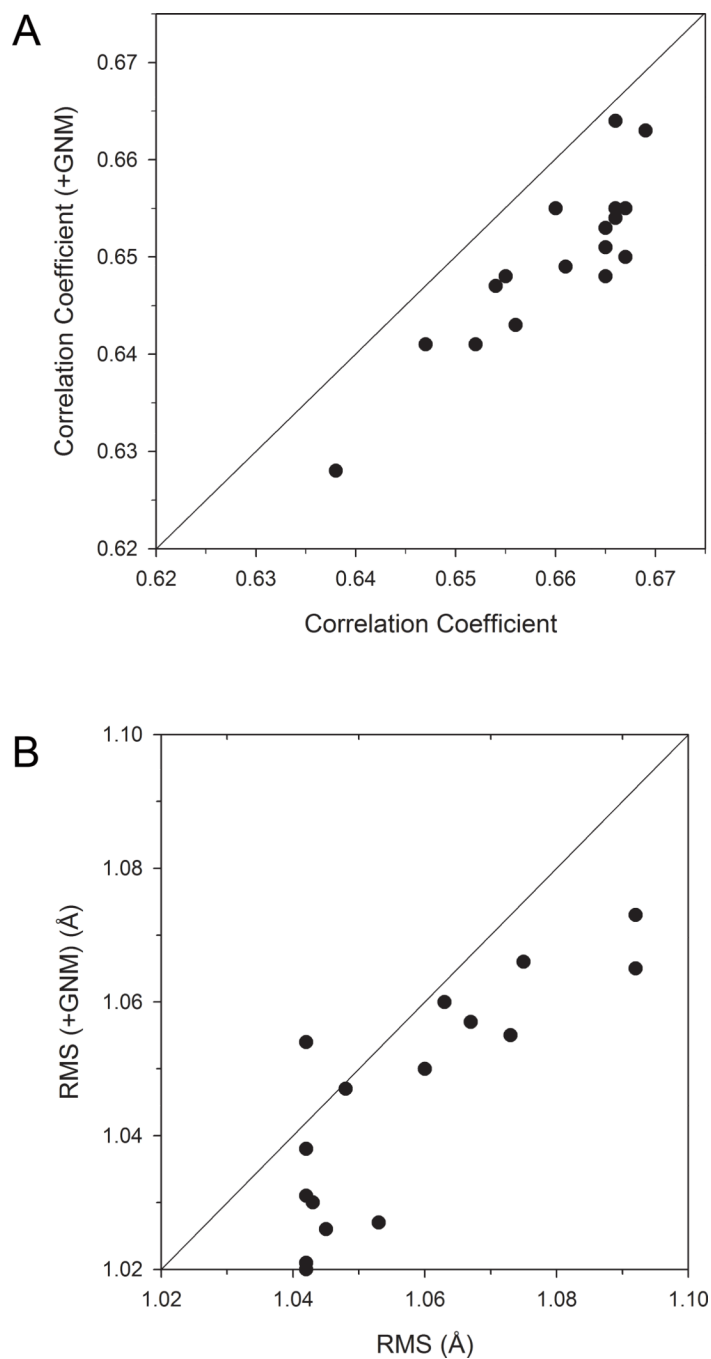


Figure 5. Comparison of the prediction performance with and without using GNM as a feature. $\langle(\Delta R_i)^2\rangle$ predicted by GNM was added to each SVR feature set listed in Table 2. **A**, Average correlation coefficient; **B**, average RMS predicted by SVR with and without $\langle(\Delta R_i)^2\rangle$ from GNM are plotted.

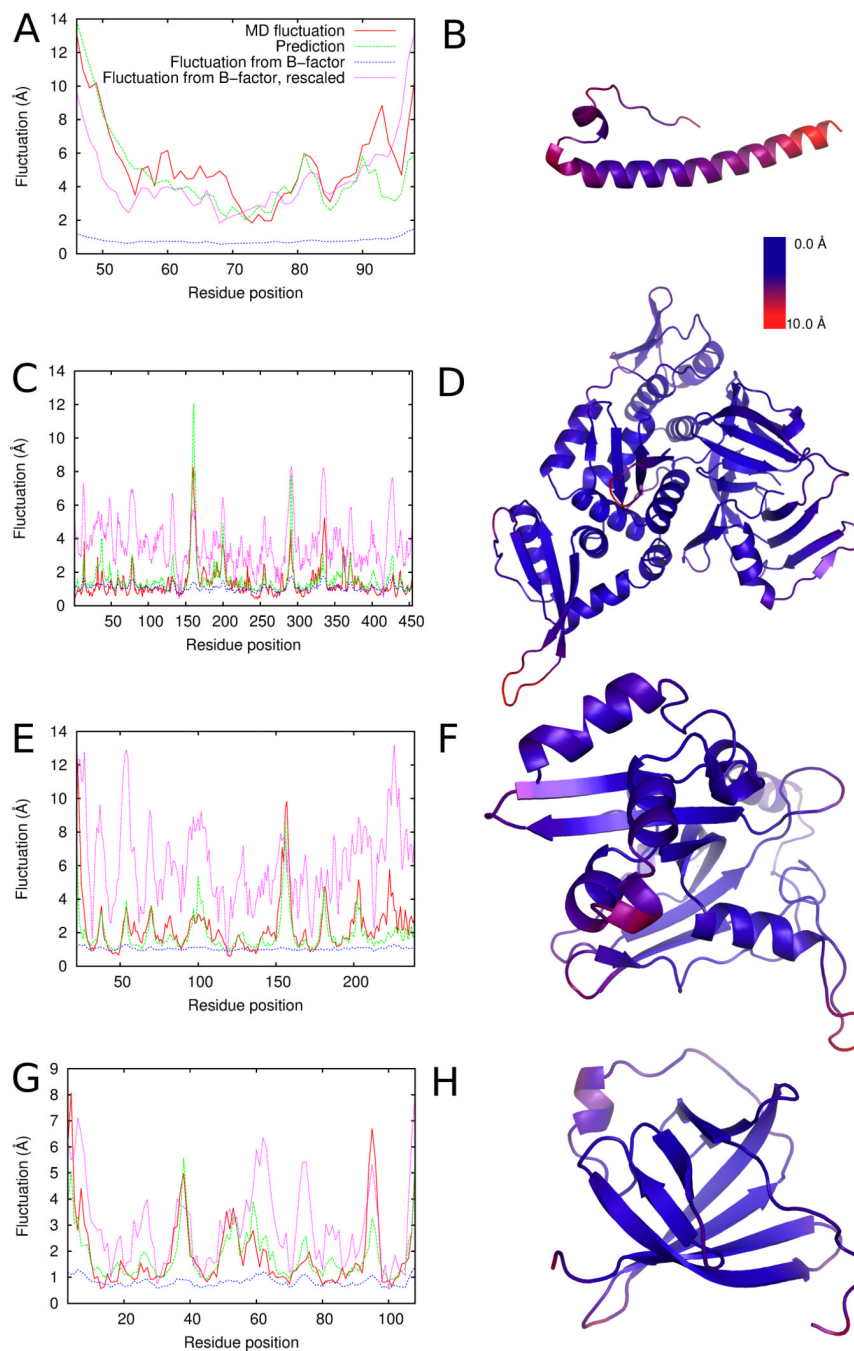


Figure 6. Examples of predicted fluctuations in comparison with B-factor derived fluctuations and MD simulation fluctuations. Left panels show the values of fluctuations: red, fluctuations observed in the MD trajectories; green, predicted fluctuations; dotted blue line, fluctuations computed from B-factors; dotted magenta line, rescaled fluctuations from B-factors ($\alpha = 1.0$). The correlation coefficients and RMS are summarized in Table 3. Right-hand panels show the magnitude of fluctuations in a color scale with blue indicating lower fluctuations and red for higher fluctuations. **A, B**, 1mof; **C, D**, 1dq3; **E, F**, 1gpc; **G, H**, 1a1x.

Table 1

Correlation coefficients between structural features and fluctuations.

Structural Features	Number of proteins with p-value < 0.05 (%) ^{a)}	Avg. corr. coeff. ^{b)}
B-Factor	565 (95.4)	0.484 (0.504)
Distance to center of mass	584 (98.6)	0.509 (0.514)
Square of distance to center of mass	586 (99.0)	0.545 (0.549)
Contact number (cutoff 6 Å)	571 (96.5)	-0.374 (-0.384)
Contact number (8 Å)	591 (99.8)	-0.480 (-0.481)
Contact number (12 Å)	590 (99.7)	-0.554 (-0.556)
Contact number (15 Å)	587 (99.2)	-0.568 (-0.571)
Contact number (16 Å)	571 (96.5)	-0.567 (-0.571)
Contact number (18 Å)	587 (99.2)	-0.562 (-0.565)
Contact number (20 Å)	585 (98.8)	-0.555 (-0.559)
Contact number (22 Å)	584 (98.6)	-0.545 (-0.551)
Accessible Surface Area (ASA) ^{c)}	580 (98.0)	0.404 (0.407)
ASA normalized	590 (99.7)	0.476 (0.477)
Residue depth (residue mean) ^{d)}	559 (94.4)	-0.352 (-0.371)
Residue depth (C α)	553 (93.4)	-0.339 (-0.359)
Half upper sphere exposure (C α) ^{e)}	568 (95.9)	-0.385 (-0.398)
Half lower sphere exposure (C α)	567 (95.8)	-0.389 (-0.402)
Half upper sphere exposure (C β)	537 (90.7)	-0.339 (-0.363)
Half lower sphere exposure (C β)	561 (94.8)	-0.383 (-0.399)
Prediction by GNM (cutoff 16 Å) ^{f)}	586 (99.0)	0.643 (0.648)
Prediction by GNM (no cutoff)	591 (99.8)	0.646 (0.646)

The largest correlation coefficients among the static structural features are highlighted in bold.

^{a)} The number of proteins that have significant correlation coefficient to the fluctuations (with p-value < 0.05) are counted. The total number of trajectories (proteins) is 592.

^{b)} The average value calculated only for the subset of proteins with p-value < 0.05 is shown in the parentheses.

^{c)} Accessible surface area (Å²) of amino residues without normalization. The next row is the correlation with the normalized Accessible surface area.

^{d)} The residue depth computed as the average distance for each atom in the residue and the distance for the C α atom (next row).

^{e)} The lower/upper half-sphere exposure of a residue using the C α or the C β atom to determine the position of the plane which cut the sphere to half.

^{f)} Fluctuations predicted by GNM (Eq. 6).

Table 2

Summary of fluctuation prediction using SVR models with different feature combinations.

Feature Set	Features used ^{a)}	Number of proteins with p-value < 0.05 (%)	Average corr. coeff. ^{b)}	RMS (Å) ^{c)}
1	C(16), D ²	584 (98.6)	0.638 (0.644)	1.075
2	C(16), D ² , B	587 (99.2)	0.654 (0.658)	1.067
3	C(16), D ² , B, C(18)	587 (99.2)	0.655 (0.659)	1.060
4	C(16), D ² , B, C(18), Sec	589 (99.5)	0.661 (0.664)	1.048
5	C(16), D ² , B, C(18), Res-type	586 (99.0)	0.652 (0.657)	1.063
6	C(16), D ² , B, C(18), Sec, C(12)	589 (99.5)	0.665 (0.668)	1.042
7	C(16), D ² , B, C(18), Sec, C(12), C(8)	588 (99.3)	0.667 (0.668)	1.042
8	C(16), D ² , C(18), C(12), C(8), C(6)	588 (99.3)	0.656 (0.660)	1.053
9	C(16), D ² , B, C(18), C(12), C(8), C(6)	588 (99.3)	0.666 (0.669)	1.045
10	C(16), D ² , B, C(18), C(12), C(8), C(6), Sec	589 (99.5)	0.665 (0.667)	1.043
11	C(16), D ² , B, C(18), C(12), C(8), C(6), Acc	587 (99.2)	0.665 (0.669)	1.045
12	C(16), D ² , B, C(18), C(12), C(8), C(6), C(20)	588 (99.3)	0.666 (0.670)	1.042
13	C(16), D ² , B, C(18), C(12), C(8), C(6), C(20), C(22)	588 (99.3)	0.667 (0.670)	1.042
14	C(16), D ² , B, C(18), C(12), C(8), C(6), C(15), C(20), C(22)	588 (99.3)	0.666 (0.670)	1.042
15	C(16), B, C(18), C(12), C(8), C(6), C(20), C(22)	588 (99.3)	0.669 (0.673)	1.073
16	C(16), C(18), C(12), C(8), C(6), C(15), C(20), C(22)	587 (99.2)	0.660 (0.665)	1.092
17	C(16), B, C(18), C(12), C(8), C(6), C(20), C(22), HP	587 (99.2)	0.647 (0.651)	1.092

^{a)} C(x), the residue contact number with x Å distance cutoff; B, B-factor; D² - square of the distance between the C α atom to the protein center of mass; Sec, the secondary structure; Acc, normalized accessible surface area; HP, the number of hydrophilic/hydrophobic contacts, Res-Type, amino acid type of residues.

^{b)} The average correlation coefficients between predicted and actual fluctuations. Values calculated only for the subset of proteins that have significant correlation with p-value < 0.05 is shown in the parentheses.

^{c)} The RMS (Eq. 7) was averaged over all the proteins in the dataset.

Table 3

Correlation coefficients and RMS of the four example predictions.

PDB ID	Correlation Coefficient		RMS (Å)		
	B-factor	Prediction	B-factor, rescaled $\alpha = 1.0$ <i>a)</i>	B-factor, rescaled α <i>b)</i>	Prediction
1mof	0.69	0.80	4.92	1.91 (1.0)	1.55
1dq3	0.50	0.81	0.94	2.64	0.85 (0.4)
1gpc	0.55	0.78	1.93	4.32	1.42 (0.4)
1alx	0.61	0.82	1.60	1.72	1.09 (0.6)

The data correspond to plots at the left panels in Figure 6.

a) Fluctuations computed from B-factor were rescaled with $\alpha = 1.0$ in Equation 9. This value corresponds to the curve "Fluctuation from B-factor, rescaled" in Figure 6.

b) Fluctuations computed from B-factor were rescaled with the weight factor α (Eq. 9) ranging from 0.1 to 1.0 with an interval of 0.1. Then, the smallest RMS obtained is shown together with the used α value in the parentheses.

Table 4

Comparison of fluctuations of NMR models, MD, and our prediction.

Compared data	Number of proteins with p-value < 0.05 (%)	Corr. coeff.	RMS(Å)
NMR vs. MD	136 (97.1)	0.651 (0.667)	2.425
NMR vs. prediction	138 (98.6)	0.739 (0.747)	1.808
MD vs. prediction	138 (98.6)	0.686 (0.693)	2.165

140 non-redundant proteins in the MoDEL database were used whose structures were determined by NMR.