6-8-2007

vGNM: a Better Model for Understanding the Dynamics of Proteins in Crystals

Guang Song
Iowa State University, gsong@iastate.edu

Robert L. Jernigan
Iowa State University, jernigan@iastate.edu

vGNM: a Better Model for Understanding the Dynamics of Proteins in Crystals

Abstract
The dynamics of proteins are important for understanding their functions. In recent years, the simple coarse-grained Gaussian Network Model (GNM) has been fairly successful in interpreting crystallographic B-factors. However, the model clearly ignores the contribution of the rigid body motions and the effect of crystal packing. The model cannot explain the fact that the same protein may have significantly different B-factors under different crystal packing conditions. In this work, we propose a new Gaussian network model, called vGNM, which takes into account both the contribution of the rigid body motions and the effect of crystal packing, by allowing the amplitude of the internal modes to be variables. It hypothesizes that the effect of crystal packing should cause some modes to be amplified, and others to become less feasible. In doing so, vGNM is able to resolve the apparent discrepancy in experimental B-factors among structures of the same protein but with different crystal packing conditions, which GNM cannot explain. With a small number of parameters, vGNM is able to reproduce experimental B-factors for a large set of proteins with significantly better correlations (having a mean value of 0.81 as compared to 0.59 by GNM). The results of applying vGNM also show that the rigid body motions account for nearly 60% of the total fluctuations, in good agreement with previous findings.

Keywords
Protein dynamics, crystal packing, space groups, Gaussian Network Model, B-factors

Comments
This is a manuscript of an article published as Song, Guang, and Robert L. Jernigan. "vGNM: a better model for understanding the dynamics of proteins in crystals." Journal of molecular biology 369, no. 3 (2007): 880-893. doi: 10.1016/j.jmb.2007.03.059. Posted with permission.

Published in final edited form as:
J Mol Biol. 2007 June 8; 369(3): 880–893.

vGNM: a Better Model for Understanding the Dynamics of Proteins in Crystals

Guang Song[*,†,§] and **Robert L Jernigan**[*,#,§]

[*]*Program of Bioinformatics and Computational Biology, Iowa State University, Ames, IA 50011, USA*

[†]*Department of Computer Science, Iowa State University, Ames, IA 50011, USA*

[#]*Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA*

[§]*L. H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA*

Abstract

The dynamics of proteins are important for understanding their functions. In recent years, the simple coarse-grained Gaussian Network Model (GNM) has been fairly successful in interpreting crystallographic B-factors. However, the model clearly ignores the contribution of the rigid body motions and the effect of crystal packing. The model cannot explain the fact that the same protein may have significantly different B-factors under different crystal packing conditions. In this work, we propose a new Gaussian network model, called vGNM, which takes into account both the contribution of the rigid body motions and the effect of crystal packing, by allowing the amplitude of the internal modes to be variables. It hypothesizes that the effect of crystal packing should cause some modes to be amplified, and others to become less feasible. In doing so, vGNM is able to resolve the apparent discrepancy in experimental B-factors among structures of the same protein but with different crystal packing conditions, which GNM cannot explain. With a small number of parameters, vGNM is able to reproduce experimental B-factors for a large set of proteins with significantly better correlations (having a mean value of 0.81 as compared to 0.59 by GNM). The results of applying vGNM also show that the rigid body motions account for nearly 60% of the total fluctuations, in good agreement with previous findings.

Keywords

protein dynamics; crystal packing; space groups; Gaussian Network Model; B-factors

1 Introduction

X-ray crystallography is one of the most powerful experimental tools for elucidating not only the static but also the dynamic structures of proteins and other macromolecules. X-ray diffraction data provides the average structure of a protein in a crystal as well as the scale of its atomic fluctuations normally expressed as isotropic Debye-Waller factors, or B-factors. In the refinement process, the normal practice is to consider the fluctuations of each atom to be independent of all others. Thus, for a protein with *n* atoms, *n* refinement parameters (for the individual B-factors) are needed to model the atomic fluctuations. However, it is widely conceived that the fluctuations of the atoms are concerted. Therefore, it is desirable to find some good collective variables to describe the fluctuations. There are several advantages in

Correspondence: Guang Song, Tel: 515-294-1696;Fax: 515-294-0258; E-mail: gsong@iastate.edu.

doing this. For one, by using the collective variables, the number of refinement parameters can be greatly reduced, which in turn reduces the data-to-parameter ratio, making possible more accurate refinements (e.g., anisotropic refinements). For another, these collective variables thus discovered are able to reveal useful information about protein dynamics. Such information can be further analyzed and applied to study protein conformation transitions [1] and the mechanisms of how proteins function. Indeed, results from previous work, using collective variables for either the external rigid body motion or the internal vibrational motion of proteins, have lent much support to such an idea. The earliest study of rigid body motion of molecules in crystals was by Cruickshank [2], who interpreted atom displacements with a translation tensor and a libration tensor. An extension of this method, which includes the correlation between translation and libration, is the translation, libration, and screw (TLS) model introduced by Schomaker and Trueblood [3]. TLS has been used not only in understanding the contribution of rigid body displacements to the total protein motion in crystals [4] but also in the refinement process itself [5]. Protein internal motions and their contribution to the total thermal fluctuations have also been investigated. In [6,7], Kidera and Go presented a ”normal mode refinement” method to explain the Debye-Waller factors. Their results showed that, by using a small number of variables, the normal mode refinement is able to give a better *R* factor in the refinement result and moreover, provide some information about the possible dynamics of proteins. In [8], Diamond presented a method which treated the amplitude coeffcients of normal modes as independent variables, and used them together with protein rigid body translation and rotation, to fit the thermal effects (fluctuations) in X-ray diffraction data. Their results showed that a small number of collective variables are able to mostly reproduce the atom fluctuations furnished by 892 isotropic temperature factors.

More recently, the coarse-grained elastic network model (ENM) introduced by Bahar et al [9] has gained popularity because of its simplicity and its ability to reproduce crystallographic B-factors [9]. It also has been applied to study protein conformational transitions and links between motions and to functions. By extending the idea proposed by Tirion for atoms [10], ENM, with its impressive simplicity, claims to be able to produce low frequency mode motions that match those gained by the original normal mode analysis (NMA), which on the other hand, requires sophisticated energy minimization and potential functions to derive the harmonic normal modes. The isotropic version of ENM, the GNM, models the protein by its alpha carbons (or all the atoms) and consider only the interaction of nearby alpha carbons (determined by a small cutoff distance, usually 7-8 Å) and the interactions are modelled with a single, uniform spring. As a result, the fluctuation of atoms can be analytically expressed as a summation of the contribution of each mode, weighted by the inverse of the corresponding eigenvalue. The low frequency modes contribute most of the fluctuations (see Methods section). The original GNM thus doesn’t take into account any solvent effect, nor crystal packing. Phillips and coworkers [11] showed that the correlation between GNM calculation and experimental B-factors can be marginally improved after considering neighbouring molecules in the GNM modeling.

However, although GNM [9] or GNM with neighbours [11] has demonstrated a reasonable success in reproducing B-factors, with the average correlation between GNM and experimental B-factor of 113 proteins studied in [11] being 0.59, and 0.66 after taking neighbouring molecules into modeling [11], there still remains much uncertainty and a significant room for improvement. In addition, GNM apparently leaves out some potentially important contributors to the mean-square fluctuations observed in X-ray diffraction data, such as rigid-body translation and rotation. As shown in [11], a simple version of the translation libration screw (TLS) model gives a mean correlation (of the same 113 proteins) as high as 0.52, which is fairly comparable with that from GNM (see Table 1). Thus, the internal motions that are inferred from the GNM are not so dependable.

Observations

The limitation of GNM in interpreting crystal temperature factors is also exposed by some experimental observations. GNM gives about the same atomic fluctuations for proteins with nearly identical structures, e.g., for two structures of the same protein crystallized under different conditions. However, the two structures, of the same protein but with different crystal packing, may have quite different B-factors. In such a case, GNM will not be able to account for the differences. The following case for myoglobin, shown in Figure 1, is an example. The two structures of the same protein (pdb id: 1AJG and 1ABS) are nearly identical with a root mean square distance (RMSD) of only 0.51 Å. However, their experimental B-factors look very different, and only have a correlation value of 0.61. Because of their structural similarity, the theoretical B-factors predicted by GNM are almost identical for the two structures (having a correlation of 0.98). Therefore, when applying GNM to explain these two sets of B-factor data which are very different from each other, we are literally trying to use one calculation to explain two sets of different experimental data, which is destined to fail. Indeed, the B-factors predicted by GNM have a correlation value of 0.58 with the experimental B-factors of structure 1ABS and only 0.44 with 1AJG. Therefore, we need a different model that can, (i) somehow take into account the differences between these two structures and, (ii) give good (and therefore necessarily different) B-factor predictions for either of them.

One notable difference between the two structures is their crystal packing conditions. 1A6G ispacked under space group $P2_1$ while 1ABS's space group is P6. One may reasonably think that GNM with packing neighbours proposed in [11] will meet the need. However, from Table 1 we know that the neighbouring effect as considered in [11] improves the correlation only by 0.07 on average and therefore is not likely to be able to account for the large differences existing between the experimental B-factors of 1ABS and 1AJG. Indeed, although GNM with neighbours [11] improves the correlations to some extent – from 0.58 to 0.64 for 1ABS, and from 0.44 to 0.45 for 1AJG, its cannot yield good B-factor predictions for both structures, and therefore cannot fully account for the B-factor differences between the two structures. But what should be considered to be 'good' B-factor predictions? From Table 1, we see that models with either internal motions or external rigid body motions alone can reproduce B-factors with correlations only at the level of 0.5 to 0.6. This implies that predictions having such a level of correlation cannot fully inform us about protein dynamics, or specifically, what contributes to the observed mean-square fluctuations. On the other hand, as we will see in what follows, the correlations between *experimental* B-factors of structures with similar crystal packing conditions can have correlations as high as 0.86 (see Table 2). We should expect a good theoretical prediction to have correlation values close to that.

More observations

There is a wealth of crystal structure data on sperm whale myoglobin. A collection of all myoglobin structures with no missing residues and having a resolution higher than 2 Å consists of 71 structures with space group P6 and 19 structures with space group $P2_1$. Table 2 lists the mean B-factor correlations and RMSD distances within the space groups and between the two space groups.

The data listed in Table 2 show convincingly that crystal packing reflected in the space groups strongly affects the mean-square fluctuations of atoms. The mean experimental B-factor correlation within the same space group is as high as 0.86 or 0.88, while the mean correlation between the structures of different space groups is only 0.51.

Hypothesis

To account for the vastly different mean-square fluctuations shown above, we take an empirical approach and postulate that the GNM modes may be so strongly affected by crystal packing

that some of the modes are excited, while others are suppressed, and that these excitation/suppression patterns are different under different crystal packing conditions. In other words, the contribution of each mode is no longer fixed to be in proportional to the inverse of its eigenvalue, as is in GNM. Another reason for not requiring the contribution of each mode to be proportional to the inverse of its eigenvalue is that, in the words of Hinsen et al. [12], "the precise identity of normal modes depends on the force field details that are beyond physical validity. For example, a different treatment of the long range electrostatic interaction will yield different sets of low frequency normal modes, and no physical argument can be given to decide which treatment is 'better.' However, the wider subspace of the multiple low frequency modes is unaffected by such details." That is to say, though the dynamics of a protein can be captured well by the total effect of the multiple low frequency modes, the exact contribution of any individual mode is less clear. It is also possible that the low frequency modes solved from one particular model are correlated in reality, as considered by Kidera and Go [6]. Here we assume the correlated terms contribute far less to the total fluctuations, and thus will not be considered. Lastly, the idea of having the amplitude coeffcients of normal modes as variables is not new, as for example, Diamond [8] applied a similar idea to study the thermal parameter refinement of Bovine Pancreatic Trypsin Inhibitor (BPTI).

1.1 Outline and Contributions of this Work

In this work, we thus present a method that treats the amplitude coeffcients of normal modes of vibration as independent variables and then apply it to understand protein dynamics in crystals. In addition, we take into account the contribution of protein rigid body motions as well, which is not presently included in GNM. In doing this and using only a small number of parameters (around 10) to be determined through least squares fitting, we are able to resolve the apparent discrepancy of B-factors in structures that are of the same protein but have different crystal packing conditions. Secondly, it conveniently allows us to determine the contributions of protein rigid body translation and rotation to the B-factors. Our results indicate that rigid body motions account for nearly 60% of the total atomic fluctuations, agreeing with results previously found by [4,8] and others. Thirdly, using this approach enables us to reproduce protein crystallographic B-factors much better than before (mean correlation is about 0.81 as compared to 0.59 from GNM). Our results thus suggest that crystal packing may have a much stronger effect than was previously thought [11], so much so that some internal modes are suppressed and others are excited and amplified. In the end, we show that the low frequency normal modes from GNM indeed form a significant vector basis (or subspace) for understanding atomic fluctuations (specifically the internal motions). A test with other bases such as random vectors shows that they are unable to reproduce B-factors, let alone protein dynamics.

2 Results and Discussion

Proteins studied

To show the overall improvement of reproducing B-factors by using the method described here, we use the large set of proteins studied by the Phillips' group [11]. The set includes 113 proteins. All these structures were solved by X-ray diffraction and have a resolution better than or equal to 2.0 Å (except 1ACC, 2.1 Å) and have only one chain in the asymmetric unit.

For the study of the effect of crystal packing conditions on B-factors, we used the myoglobin structures found in the PDB database [13] that have a resolution better than 2.0 Å and have only one chain in the asymmetric unit. There are 71 such structures with space group P6 and 19 structures with space group $P\ 2_1$. There are also 5 structures with space group $P2_12_12_1$ and a few singletons with their own space groups. These have not been included in this study since their numbers are small. Myoglobin is chosen for this study since it is a well studied protein

and there are many structures of it solved by X-ray, nearly all of which have a single chain in the asymmetric unit.

2.1 Differences in B-factors Explained

In the beginning of this article we showed an example of how the temperature factors of structures belonging to different space groups of the same protein can differ from one another (see Table 2). We also showed that even though GNM with packing neighbors [11] takes into account some effects of crystal packing on the modes, the effect is too small to account for the apparent large discrepancy between the B-factors of the two example structures 1ABS and 1AJG, and more generally, between the structures in space groups P6 and $P2_1$. As shown in Table 2, the average experimental B-factor correlation is only 0.51.

However, by using our new model vGNM, such discrepancies in B-factors are resolved. Our results suggest, due to different crystal packing conditions – especially the different space groups, that the internal modes should be excited and suppressed in different patterns. Table 3 shows, for 1ABS and 1AJG, along with 18 other proteins from the two space groups, which internal modes are selected and their contributions to the final B-factors. It is interesting to see that structures 1ABS and 1AJG have almost the opposite sets of modes selected, which possibly explains the significant differences between their B-factors.

What is also seen in Table 3 is that the proteins within the same space group tend to select similar sets of internal modes. This is especially true for the proteins in space group P6, where there is a clear pattern that modes 1,2,5,9,12, and 20 are consistently favored over the others.While for the proteins in space group $P2_1$, the pattern is less clear. But still, we see that for all the proteins in that set, mode 2, rather than mode 1, clearly has the dominant contribution to the total mean-square fluctuations. Mode 3 is also favored consistently.

Besides providing a feasible explanation for the significant B-factor differences between structures of different space groups, vGNM is also able to reproduce experimental B-factors with significantly better correlations, not only for myolgobin structures of different space groups as shown in the 'correlation' columns of Table 3, but also for a large set of various proteins, as we will see in the next section.

2.2 Significant Improvements Seen in Computed B-factors for a Large Set of Proteins

In the previous section, we have successfully applied our approach to resolve the apparent differences in B-factors of structures having different crystal packing conditions (different space groups) of the *same* protein. Now we will extend this approach to study the B-factors of the same large set of proteins studied in Ref. [11]. We want to demonstrate that, with this new method and a small number of collective variables, we are better able to reproduce the experimental B-factors for a large set of proteins with significantly improved accuracy.

Figure 2 shows the correlation of 113 proteins between experimental B-factors with calculated B-factors from normal GNM and our approach, vGNM. We see significant improvement in the B-factor correlations for all proteins. The mean correlation for all proteins is 0.81 for vGNM and 0.59 for GNM, a gain of 0.22 on average.

In Figure 3(a), taking Calmodulin (1osa.pdb) as an example, we show the B-factor distribution from X-ray diffraction data and various models (GNM and vGNM). It is remarkable to see that by using as few as 12 internal modes in addition to external rigid body motions we are able to reproduce B-factors extremely well (B-factor correlation improves to 0.87; was only 0.41 with GNM). Figure 3(b) shows the percentage contribution from each individual mode. It is interesting to see that the lowest frequency mode is not selected at all, and that the largest contribution comes from the third lowest frequency mode.

2.3 Contributions of Rigid Body and Internal Motions

Another advantage of this approach is that it allows us to determine the contributions of rigid body translation and rotation as well as the internal motions to the total mean-square fluctuations of proteins around their native states.

Figure 4 shows the contribution percentage from the rigid body translation and rotation (see Equation 7) for all 113 proteins. The mean rigid body contribution is about 59% (44% for translation and 15% for rotation), which agrees with results found by [4] and [8], implying the external rigid body motion are the major source of the fluctuations.

Table 4 shows, in addition, the mean contribution of each individual mode or rigid body motion, how often they are selected, and the mean amplitude (the ratio of contribution over frequency). It is seen from the table (the **frq** row) that, rigid body translation and rotation are almost always selected, modes 1 to 10 are selected about half of the time, modes 11 to 20 about a third of the time. In addition, the sum of all these frequencies gives an estimate of the number of parameters that is needed in vGNM. It is about 10 in this case, which means that although we consider 20 lowest modes in reproducing the B-factors with the least squares fit, we end up needing only about half as many parameters. These parameters indicate which modes or rigid body motions are selected as well as their weights/contributions. Surprisingly, vGNM also suggests that the lowest frequency mode is not the one selected most often (again see the **frq** row in Table 4). Instead, it is the second lowest mode. Lastly, even though vGNM removes the requirement that the contribution of each mode to be inversely proportional to its eigenvalue (therefore decrease as mode index increases), it is interesting to see that the amplitude of the internal modes decreases almost monotonically.

2.4 Validity of GNM Modes: A Comparison with Random Vectors

The improved B-factor predictions for all the proteins (Figure 2) is impressive. It is obtained by the least squares fit using about 10 parameters on average (see Table 4 and related discussion in the text). Yet, our ultimate goal is not just to reproduce the B-factors, but rather, through doing so, to learn as much as possible about the internal motions of proteins. A good understanding of protein dynamics may have many significant implications, from structure predictions [14] to decoding the mechanism of how some proteins realize their functions [15].

Therefore, we are interested in finding out how much we can learn about the internal motions through the observed B-factors. We want to know how much the basis we use to represent the internal motions matters in reproducing the B-factors. Is it possible to reproduce the B-factors equally well using other combinations of internal motions (or modes) and external motions, or even random motions? If the answer were yes, then the 'impressive' results we get would lose their meaning and usefulness in revealing protein dynamics, but be a mere consequence of the least squares fit.

In order to answer this question, we experiment with a variety of other bases to see how well they can reproduce experimental B-factors. Should they all reproduce B-factors similarly well, then we could conclude nothing about the internal motions. Should the basis we choose reproduce B-factors much better than others, we would have much greater confidence that the internal motions they represent are the actual motions.

The sets of bases we experiment with, besides the one we have been using (which is the 20lowest frequency modes of GNM with cutoff distance R_{cut} equal to 7.3 Å), are: the second set of 20 eigenvectors (i.e., eigenvectors 22-41), the 20 eigenvectors in the middle range of the spectrum (for different protein, the range will be different), the last 20 eigenvectors (high frequency end), the first 20 eigenvectors of GNM with $R_{cut} = 5\,\mathring{A}$, the first 20 eigenvectors of

GNM with R_{cut} = 15 $\mathring{A}$, and 20 random vectors (which means the motions of atoms are completely uncorrelated).

To make a fair comparison and determine how the basis for internal motions alone affects our ability to reproduce B-factors, we set the rigid body contribution (both translation and rotation) to be fixed and identical for all sets of basis and use the values we have determined earlier when the lowest 20 modes are used as basis.

Table 5 lists the mean correlations with the experimental B-factors of all 113 proteins when using these various bases. The mean correlation from plain GNM is listed for comparison purposes. As we presented earlier, vGNM with the lowest 20 modes (modes 2 to 21) as the basis yields significantly better results than GNM. When using the second set of 20 modes (modes 22 to 41), the 20 modes in the middle of the spectrum, and the last 20 modes (high frequency end of the spectrum), the results get increasingly worse. It is especially noticeable that the mean correlation using the last 20 modes as the basis is only 0.19. We know the high frequency modes represent highly localized motions, therefore it is quite understandable that with 20 such modes it is almost impossible to capture the fluctuations of a whole protein. As we can also see, the mean correlation from using the last 20 modes is even worse than that from using 20 random (orthogonal) vectors as basis. We also notice that the results for vGNM (using the default lowest 20 modes as basis) with different cutoff distances show large correlations too - 0.75 when using 5 Å as cutoff distance, 0.69 when using 15 Å However, the mean correlation is still the highest when using a cutoff distance around 7 Å

Our results indicate the space spanned by the 20 lowest modes of the GNM model captures well the potential dynamics of proteins in crystals. However, due to different crystal packing conditions, i.e., the space groups, the contribution of each mode is not necessarily inversely proportional to the eigenvalues given by the GNM. Instead, some are activated and amplified, while the others are suppressed.

3 Conclusions

In this work, we propose a new Gaussian network model, called vGNM, for understanding protein dynamics in crystals. vGNM includes two important contributors to protein motions in crystals that are missing in GNM. One is the rigid body motions and the other is the effect of crystal packing. vGNM takes into account the effect of crystal packing by allowing the amplitude of the internal modes to be variables. In doing so, vGNM is able to resolve the apparent discrepancy in experimental B-factors among structures of the same protein but with different crystal packing conditions, which GNM cannot explain. With a small number of parameters (around 10), vGNM is able to reproduce experimental B-factors for a large set of proteins with significantly better correlations (having a mean value of 0.81 as compared to 0.59 by GNM). The results of applying vGNM also show that the rigid body motions account for nearly 60% of the total fluctuations, in good agreement with previous findings.

vGNM does not disprove GNM. Rather, the success of vGNM relies on the strength of GNM, especially the quality of its low frequency modes. Our results indicate the space spanned by the 20 lowest modes of the GNM model indeed captures well the potential dynamics of proteins in crystals, while the spaces constructed using other bases cannot. However, different from GNM, in vGNM the contribution of each mode to the final mean-square fluctuations is not necessarily inversely proportional to its eigenvalue, presumably due to the effect of crystal packing. Instead, some modes are activated and amplified, while others become less feasible. Analysis using myoglobin structures as an example suggests that different space groups may activate quite different sets of internal modes.

The reason that Phillips' crystal packing calculations [11] were not better is probably because they underestimated the strengths of packing interactions. It is likely that symmetry over the macroscopic scale of the crystal makes the protein motions extremely restraint to some directions. The important motions of a structure are determined by the whole structure (in this case the whole crytal) and they can not be computed usually from partical structures (in this case one single chain or one single chain plus its packing neighbours). Results from our empirical model strongly suggest that this may be the case. Looking ahead, what is needed is a model that can give more direct physical explanations to what are observed here.

4 Materials and Methods

Gaussian Network Model (GNM)

Given a protein structure, GNM simplifies the system by modeling it with its alpha carbons only and attaching springs with uniform constants to all contacting alpha carbon pairs. Alpha carbon pairs are considered to be in contact when their separation distance is smaller than a preset cutoff distance, usually 7 to 8 Å All springs are set at their equilibrium for the input structure. One beauty of this approach is that the fluctuations of each point around its equilibrium position and their cross-correlations can be elegantly expressed in analytical forms. To determine the atomic fluctuations, we first write down the Kirchhoff matrix based on the contact information,

$$\Gamma = \begin{cases} -1 & if\ i \neq j \text{ and } R_{ij} \leq r_c \\ 0 & if\ i \neq j \text{ and } R_{ij} > r_c \\ -\Sigma_{i,i\neq j}\Gamma_{ij} & if\ i = j \end{cases} \quad (1)$$

where R_{ij} is the distance between atom i and j, and r_c is the cutoff distance. The mean square fluctuations of each atom and the theoretical B-factors can be conveniently expressed as:

$$< u_i^2 > = (3K_B T/\gamma)\,[\Gamma^{-1}]ii \quad (2)$$

$$B_i = 8\pi^2 < u_i^2 > /3 \quad (3)$$

The mean square fluctuations can also be expressed as the summation of the contributions from each mode, i.e.,

$$< u_i^2 > = (3k_B T/\gamma) \sum_{k=2}^{n} (\lambda_k^{-1} u_{ki}^2) \quad (4)$$

where u_{ki} is proportional to the amplitude of motion of atom i in mode j. We can see that the contribution from each mode is weighted by the inverse of its corresponding eigenvalue. The low frequency modes contribute most to the total fluctuations.

To measure how well the B-factors predicted by a model match the experimental data, the correlation coeffcient is normally used:

$$cof = \frac{\Sigma_{i=1}^{n}(B_i - <B>) \bullet (B_i^{\exp} - < B^{\exp} >)}{\| B - <B> \| \bullet \| B^{exp} - < B^{exp} > \|} \quad (5)$$

A perfect correlation between two vectors thus gives a value of 1 while perfect anti-correlation gives -1.

vGNM: GNM for proteins in crystal with amplitude as variables

In this work, we still use GNM to study the mean square fluctuations of proteins. The difference is that we allow the amplitude of each mode to be variable, in order to reflect the effects of different crystal packings on the protein's internal motions. We postulate that different crystal packing can cause different modes to be excited or suppressed. The effect of this on B-factor calculations is equivalent to allowing the weights in Eq. (4) to be variables. Since we still believe that the low frequency modes contribute most to the fluctuations, we limit the search for modes to those in the low frequency range, i.e., the first n_{low} modes. We will show how we determine the value of n_{low} later.

The fluctuations predicted by GNM modes represent the contribution of the internal motions. For proteins in crystals, it is widely believed that the rigid body motions of proteins probably contribute even more to the total fluctuations than is observed in X-ray diffraction data and modelled as isotropic temperature factors. For isotropic temperature factors, the contribution of translation can be simply represented by a uniform value, say w_{trans}. The contribution of rotation, for isotropic fluctuation, can be expressed as being proportional to the square of the distance between each alpha carbon to the protein's centroid [11], i.e., for atom i, the rotational mean-square fluctuation is

$$U_{rotate} = w_{rotate} * \| r_i - r_{centroid} \|^2, \tag{6}$$

where w_{rotate} is the weight that reflects the magnitude of contribution from rotational rigid body motion. Next, we will show how to determine the relative contributions to the final mean-square fluctuations from rigid body translation (i.e., w_{trans}), rigid body rotation (w_{rotate}), and the internal vibrations.

Least squares fit to experimental B-factors

As we described earlier, in this work we study how the internal motions and external rigid body motions (translation and rotation) influence the total mean-square fluctuations of atoms. For internal motions, we postulate that the contribution of each mode is *not* necessarily proportional to the inverse of its eigenvalue. Instead, under different packing conditions, some modes are excited, while some are suppressed. Some are simply more feasible in the context of a particular crystal than others. Therefore, the total mean square fluctuations can be expressed as (see Equations 4 and 6):

$$B_i^{calc} = w_{trans} + w_{rotate} * \| r_i - r_{centroid} \|^2 + \sum_{k=2}^{n_{low}} w_k * u_{ki}^2 \tag{7}$$

where u_{ki}, as in Equation 4 , is proportional to the amplitude of motion of atom i in mode k. Thus, the total fluctuation is expressed as the sum of all these terms weighted by parameters w_{trans}, w_{rotate}, and w_k's for each of the internal modes. To determine these parameters, we use the least squares fit between the calculated B-factor B_i^{calc} (see Equation 7) and experimental B-factors. i.e.to minimize,

$$\sum_{i=1}^{n} (B_i^{calc} - B_i^{exp})^2, \tag{8}$$

while requiring all weights to be non-negative. Another point worth pointing out is that the summation in Equation 7 does not run from 2 to n, the total number of modes, but instead to n_{low}. We restrict our search to the low frequency modes because we know they contribute most

to the internal fluctuations. This also ensures that we will not overfit the experimental B-factors with too many parameters.

4.1 Determine n_{low}

Before we apply least squares fitting to study how the different components contribute to the experimental observed temperature factors using Eqs. 7 and8, we first need to estimate what n_{low} should be. We know that n_{low} should be fairly small compared to the total number of modes since the low frequency modes contribute most of the internal fluctuation. Small n_{low} values will also help reduce overfitting the experimental B-factors that could be caused by the ”brute-force” nature of the least squares fit. Based on the mean correlation values found among experimental B-factors (see Table 2), we anticipate that there is a limit for the correlation value that would be unlikely to exceed about 0.85. Yet, we also need n_{low} to be suffciently large to represent well the important fluctuation patterns caused by the internal motions.

Figures 5 and 6 show the contribution of translation and rotation as a function of n_{low}, respectively, for six selected proteins of various sizes. One can see that when n_{low} is really small, the contributions of translation and rotation are unrealistically large, but they quickly reach a plateau as some further internal modes (starting from the low frequency end of the spectrum) are included in the least squares fit. From then on, adding more modes to the least squares fit has little effect on the magnitude of contribution of both translation and rotation, until n_{low} becomes so large that the contribution from translation (see Figure 5) starts to decrease and eventually drop to zero, which implies that overfitting has likely taken place - the inclusion of too many modes in the least squares fit causes the translation contribution to disappear completely. A similar trend is also observed for the rotation contribution plot in Figure 6.

From Figures 5 and 6 we also see n_{low} is only weakly related the size of protein. Therefore, for simplicity, we choose the same n_{low} value for all proteins. To have as few parameters as possible in our model and to minimize overfitting, we set n_{low} to be 20, about the smallest n_{low} at which the contributions from both translation and rotation seem to have stabilized (see Figures 5 and 6).

It is also worth noting that, as we use the least squares fit with Equation 7 to determine the parameters, only about half of the n_{low} modes have significant contributions (i.e., are excited). For the other modes, the weight w_k is simply zero or close to 0 after the least squares fit. This means than fewer parameters are actually required to reproduce the B-factors.

4.2 Summary of vGNM

In summary, vGNM differs from GNM in two major aspects: (i) it takes into account the contribution of rigid body translation and rotation; (ii) it takes into account the effect of crystal packing, by allowing the amplitude coeffcients of each mode to be variables. It hypothesizes that the effect of crystal packing should cause some modes to be amplified, and others to become less feasible.

References

1. Song G, Jernigan RL. An enhanced elastic network model to represent the motions of domain-swapped proteins. Proteins 2006;63:197–209. [PubMed: 16447281]
2. Cruickshank DWJ. The analysis of the anisotropic thermal motion of molecules in crystals. Acta Cryst 1956;9:754–756.
3. Schomaker V, Trueblood KN. On the rigid-body motion of molecules in crystals. Acta Cryst 1968;B24:63–76.

4. Stec B, Zhou R, Teeter MM. Full-matrix refinement of the protein crambin at 0.83 a and 130 k. Acta Crystallogr D 1995;51:663–81. [PubMed: 15299796]

5. Winn MD, Isupov MN, Murshudov GN. Use of tls parameters to model anisotropic displacements in macromolecular refinement. Acta Cryst 2001;D57:122–33.

6. Kidera A, Go N. Refinement of protein dynamic structure: normal mode refinement. Proc Natl Acad Sci USA 1990;87:3718–22. [PubMed: 2339115]

7. Kidera A, Go N. Normal mode refinement: crystallographic refinement of protein dynamic structure I. theory and test by simulated diffraction data. J Mol Biol 1992;225:457–75. [PubMed: 1593630]

8. Diamond R. On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. Acta Crystallogr A 1990;46:625–35. [PubMed: 2206485]

9. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding Des 1997;2:173–81.

10. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 1996;77:1905–1908. [PubMed: 10063201]

11. Kundu S, Melton JS, Sorensen DC, Phillips GN. Dynamics of proteins in crystals:comparison of experiment with simple models. Biophys J 2002;83:723–32. [PubMed: 12124259]

12. Hinsen K, Thomas A, Field MJ. Analysis of domain motions in large proteins. Proteins 1999;34:369–382. [PubMed: 10024023]

13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucl Acids Res 2000;28:235–242. [PubMed: 10592235]

14. Qian B, Ortiz AR, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proc NatlAcad Sci USA 2004;101:15346–51.

15. Wang Y, Rader AJ, Bahar I, Jernigan RL. Global ribosome motions revealed with elastic network model. J Struct Biol 2004;147:302–14. [PubMed: 15450299]

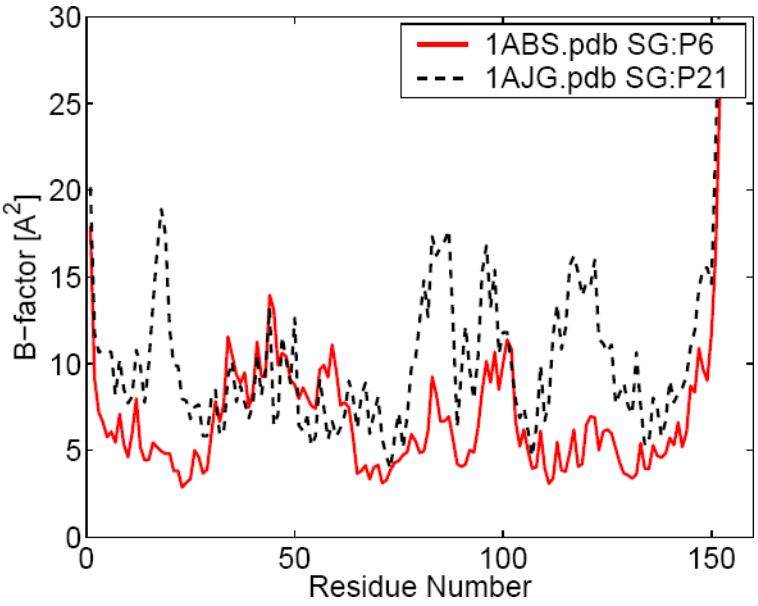

Figure 1.

Example of two different myoglobin crystal structures showing different temperature factors. The two structures, 1ABS.pdb (space group: P6) and 1AJG.pdb (space group: $P2_1$), of the same protein (sperm whale myoglobin), display rather different B-factors. The correlation between the two is only 0.61.

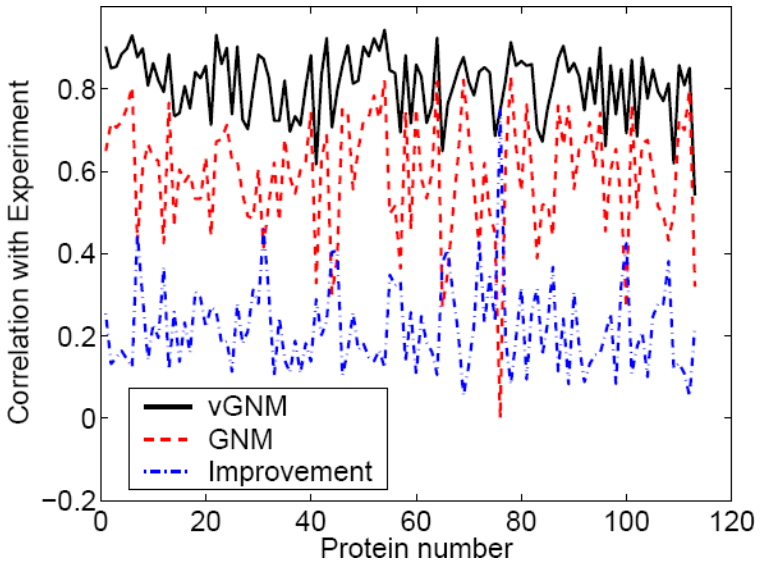

Figure 2.
Comparison of the correlations between experimental and calculated B-factors for 113 proteins from GNM and vGNM. The improvement is shown in dot dash line. vGNM produces significantly better correlations with a mean value of 0.81 (over all the proteins) as compared to 0.59 from normal GNM.

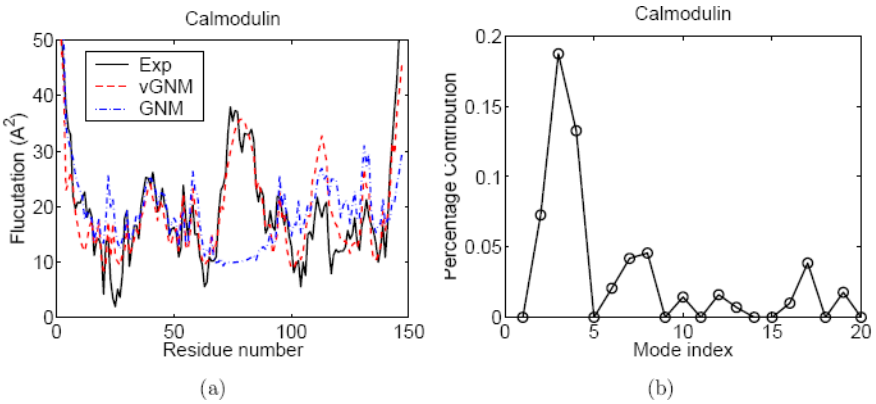

Figure 3.

(a) The B-factors calculated from vGNM and GNM for calmodulin (1osa.pdb). vGNM is able to reproduce extremely well the experimental B-factors. (b) the internal modes selected by vGNM, showing that the lowest frequency mode is not activated at all and the third lowest frequency mode makes the largest contribution.

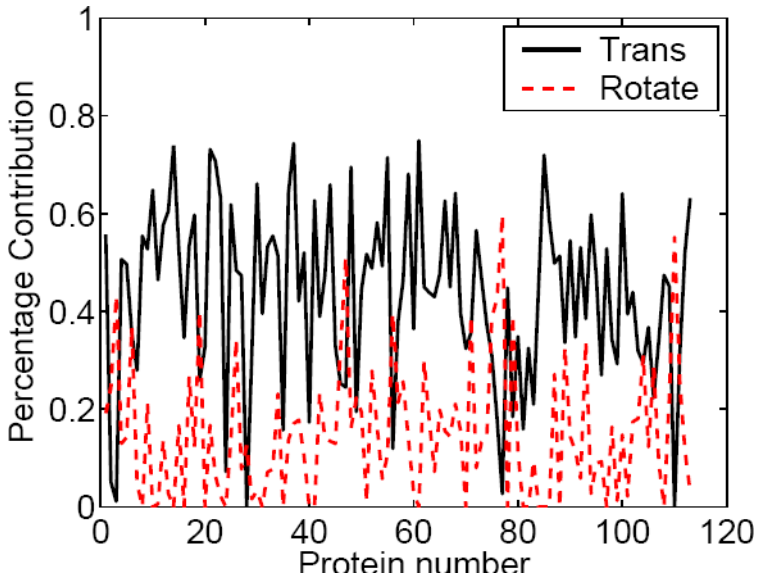

Figure 4.
The percentage contributions of rigid body translation and rotation to the total fluctuations for all proteins. The mean rigid body contribution is about 59%, with 44% from translational and 15% from rotational motions.

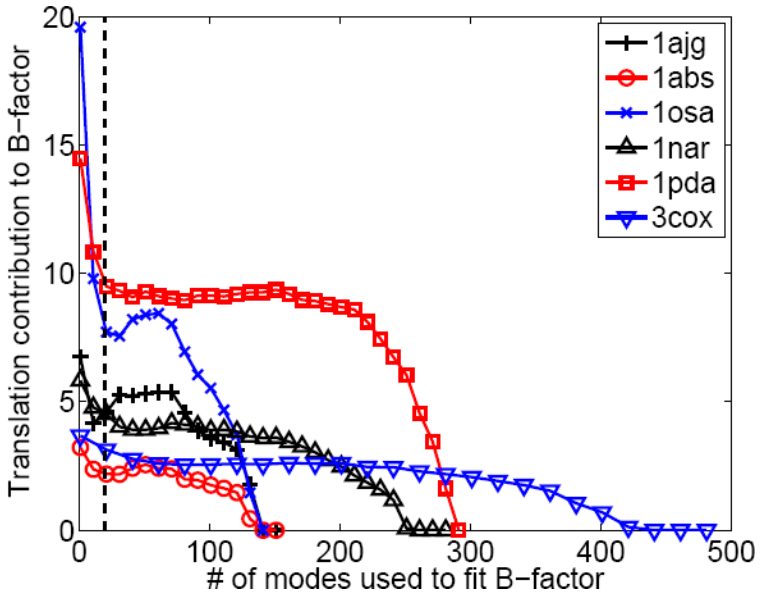

Figure 5.

The translation contribution to B-factors (w_{trans} inEq. 7) varies as the number of low frequency modes that are included in the least squares fit increases, for six example proteins of varies sizes.The contribution usually becomes stabilized after a small number of modes are included. The vertical dashed line marks where the number of modes is 20.

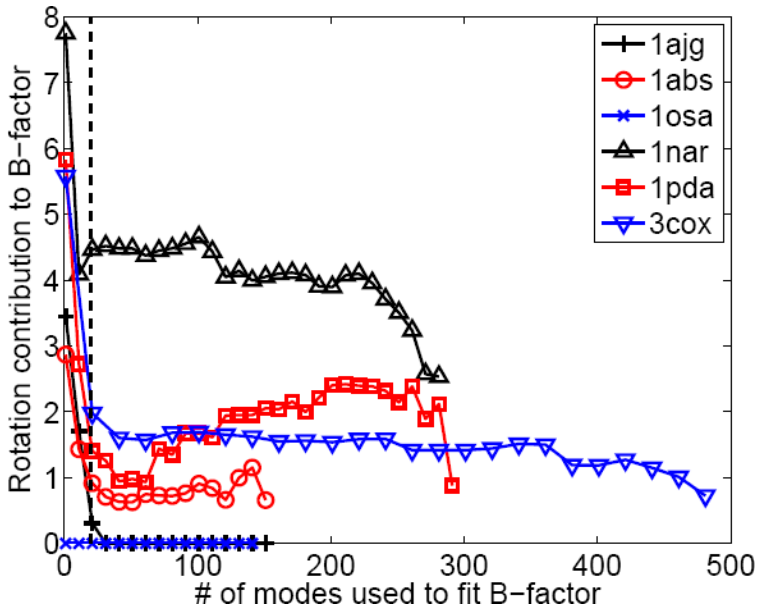

Figure 6.
The magnitude of the rotation contribution to B-factors (w_{rotate} in Eq. 7) varies as the number of low frequency modes that are included in the least squares fit increases, for six example proteins of varies sizes. The contribution becomes stabilized after a small number of modes are included. The vertical dashed line marks where the number of modes is 20.

Table 1

The crystallographic B-factors of 113 proteins interpreted with elastic network model (GNM), GNM with crystal neighbours, and rigid body motions. The results are shown as the mean correlation between experimental B-factors and calculated ones. GNM and rigid body motions give nearly comparable results. It is therefore important to know how the internal motion *and* external motion, *together* contribute to the observed fluctuations.

Understanding B-factors and Protein Dynamics

	GNM (internal motion)	GNM w/ Neighbours [11]	Rigid Body (external motion)	Is a combination of internal + external motions better
Mean Correlation	0.59	0.66	0.52	

Table 2

Myoglobin structures in two space groups. The B-factor correlations are significantly higher for structures within the same space group than those between different space groups. The root mean square deviations (RMSD) among the structures, on the other hand, are all small, implying the GNM will inevitably produce nearly identical B-factor predictions for all of these structures.

Space group	P6	$P2_1$	between
# of structures	71	19	–
RMSD [Å]	0.34 ± 0.12	0.36 ± 0.18	0.58 ± 0.11
B-factor correlation	0.88 ± 0.07	0.86 ± 0.09	0.51 ± 0.11

Table 3

The impact of crystal packing: internal modes selected and their contributions to the final B-factors for 20 sperm whale myoglobin structures of two different space groups according to vGNM. Also listed are the correlation with experimental B-factors from GNM and vGNM. Proteins within the same space group tend to select similar sets of internal modes, especially for those in P6. The contribution I_k of each mode related to w_k in Eq. (7) by a constant factor, i.e., $I_k = < w_k * u_{ki}^2 > = w_k * \frac{\Sigma u_{ki}^2}{N} = w_k / N$,where N is the number of residues.

SG	PDB ID	Correlation		vGNM: modes selected and their contribution I_k																		
		vGNM	GNM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
P6	1ABS	.87	.63	.85	-	-	.71	-	-	-	.06	.34	.64	.15	.10	-	-	-	-	-	-	.55	
	2SPL	.86	.59	2.4	1.0	.55	-	.14	-	-	-	.17	.81	.19	.06	.25	-	-	-	-	-	-	1.2
	1D01	.75	.55	.51	.41	-	-	.51	-	-	-	.26	.17	.66	.26	-	-	-	-	-	-	.08	.15
	101M	.85	.63	3.9	1.9	.64	-	.4	-	-	-	.67	.4	1.1	.85	-	-	-	-	-	.11	.15	.94
	1CH2	.81	.64	2.7	.97	.68	.1	.41	-	-	-	.29	.86	-	-	.57	.25	-	-	-	.37	-	1.4
	1CIK	.76	.58	1.7	.79	.11	-	1.9	.28	.17	-	-	-	.89	-	.64	-	-	-	.2	.1	-	1.9
	1DXC	.89	.68	1.4	.12	.11	-	.48	-	-	-	.24	-	.58	1.0	.15	-	-	-	-	.06	-	.48
	1H1X	.83	.64	1.0	.41	-	-	.52	-	-	-	.14	-	.49	.73	.3	-	-	-	-	-	-	.36
	1JW8	.86	.66	.91	-	-	-	.35	-	-	-	.28	.5	-	.24	.39	-	-	-	.08	-	.11	.69
	1NAZ	.84	.64	.86	.53	-	-	.66	-	-	-	.28	-	.46	.82	.26	-	-	-	-	-	.15	.34
$P2_1$	1AJG	.70	.44	-	2.1	.85	-	-	.67	.51	.53	-	-	-	-	-	-	.09	-	.54	-	-	
	1BZP	.96	.69	.38	2.1	.44	.66	1.3	-	1.7	-	.34	.08	-	1.8	1.1	-	-	-	.11	.32	-	.06
	1BZ6	.86	.72	.14	1.1	.43	.53	-	.71	.39	.34	.06	-	.08	.44	.44	-	-	-	.13	-	-	-
	104M	.67	.41	0.2	1.8	.40	-	-	.86	.45	-	-	.14	.25	-	-	.06	-	-	-	.14	.70	-
	2MB5	.75	.47	-	1.8	.63	.52	-	-	-	.33	-	-	-	-	.08	-	.07	.26	.15	-	-	.14
	1HJT	.75	.30	1.2	2.9	-	-	.71	-	-	.17	.31	-	.25	-	.71	-	.15	.21	-	-	2.4	-
	1YOH	.79	.47	.32	2.7	1.1	.35	-	1.3	.12	-	-	-	.29	-	.09	-	-	-	-	-	.87	-
	1VXD	.83	.72	.66	2.7	1.2	-	.23	.74	.06	.71	.19	-	.15	-	.17	-	-	-	.39	.16	-	-
	1VXE	.81	.52	1.3	2.7	1.1	-	-	1.6	.21	-	-	.15	.18	.61	-	-	-	.34	-	-	.99	-
	1EBC	.71	.35	-	3.6	1.1	1.1	-	1.3	-	-	-	.27	-	-	.57	-	-	-	.61	-	-	1.1

Table 4

The frequency (frq), in percentage, of how often rigid body translation or rotation or each individual mode is selected by the least squares fit, the mean contribution percentage (cnt) and the mean amplitude (amp) of each component (amp = cnt/frq). The sum of all frequencies, which is about 10, gives an estimate of the number of parameters that is needed for vGNM. Note that the first mode is not selected most often among the internal modes and that the amplitude of the internal modes decreases almost monotonically.

	T	R	Internal modes																		
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
frq	98	81	47	64	58	52	51	51	47	50	44	34	38	29	29	32	26	27	35	30	21	30
cnt	44	15	4.9	5.6	4.5	3.3	2.9	2.5	2.3	2.1	2.0	1.0	1.6	1.1	0.9	1.1	0.7	1.0	1.0	1.1	0.7	0.9
amp	44	19	11	8.8	7.8	6.4	5.6	5.1	5.0	4.3	4.4	3.2	4.3	4.0	3.1	3.5	2.7	3.8	2.8	3.5	3.4	2.9

Table 5

The mean B-factor correlations of 113 proteins using plain GNM and vGNM with different bases,see the text. The results from using the 20 lowest modes as basis are significant better than the others.

Model	GNM Rc=7.3	vGNM (Rc=7.3)				VGNM Rc=5	VGNM Rc=15	Random vectors
		1st 20	2nd 20	mid 20	last 20			
Correlation	0.59	**0.81**	0.59	0.36	0.19	0.75	0.69	0.44