J Mol Model (2010) 16:211–222
DOI 10.1007/s00894-009-0538-6

ORIGINAL PAPER

Superimposition of protein structures with dynamically weighted RMSD

Di Wu · Zhijun Wu

Received: 26 March 2009 / Accepted: 14 June 2009 / Published online: 1 July 2009

Abstract In protein modeling, one often needs to superimpose a group of structures for a protein. A common way to do this is to translate and rotate the structures so that the square root of the sum of squares of coordinate differences of the atoms in the structures, called the root-mean-square deviation (RMSD) of the structures, is minimized. While it has provided a general way of aligning a group of structures, this approach has not taken into account the fact that different atoms may have different properties and they should be compared differently. For this reason, when superimposed with RMSD, the coordinate differences of different atoms should be evaluated with different weights. The resulting RMSD is called the weighted RMSD (wRMSD). Here we investigate the use of a special wRMSD for superimposing a group of structures with weights assigned to the atoms according to certain thermal motions of the atoms. We call such an RMSD the dynamically weighted RMSD (dRMSD). We show that the thermal motions of the atoms can be obtained from several sources such as the mean-square fluctuations that can be estimated by Gaussian network model analysis. We show that the superimposition of structures with dRMSD can successfully identify protein domains and protein motions, and that it has important implications in practice, e.g., in aligning the ensemble of structures determined by nuclear magnetic resonance.

Keywords Superimposition · Root mean square deviation · Protein structure · Gaussian network model · Structural alignment

D. Wu (✉)
Department of Mathematics and Computer Science,
Bioinformatics and Information Sciences Center,
Western Kentucky University,
Bowling Green, KY 42101, USA
e-mail: di.wu@wku.edu

Z. Wu (✉)
Department of Mathematics, Program on Bioinformatics
and Computational Biology, Iowa State University,
Ames, IA 50011, USA
e-mail: zhijun@iastate.edu

Introduction

Proteins are important ingredients of biological systems; some are used to form the physical structures of the systems and others are responsible for the systems' biological activities. A protein is a polypeptide chain made of combinations of 20 different amino acids. The sequence of the amino acids in the chain determines the structure of the protein. The structure in turn determines the function of the protein. Therefore, it is always important to have some knowledge of the structure of a protein in order to study its function. X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are two major experimental techniques that can be used for protein structure determination [1, 2].

An important problem in protein structure determination and modeling is to position a given group of protein structures in three-dimensional space once they are determined, i.e., protein structure superimposition or alignment. The structures may be determined for the same protein but at different times, such as those obtained in NMR structure determination. It is then important to find the best superimposition for the structures, one that truly reflects the dynamic changes of the structures over time. The structures may also refer to different proteins, such as those obtained for mutated proteins or proteins from a specific

gene family. It is then critical to find the best alignment for the structures, in order to reveal some shared structural or functional motifs among the structures.

A conventional approach to superimposing a group of structures is to translate and rotate the structures so that the arithmetic average of the coordinate differences of the corresponding atoms in the structures, called the root-mean-square deviation of the structures, is minimized. Here, the best superimposition of the structures is obtained when the minimal possible root-mean-square deviation is reached. The latter is called the RMSD value of the structures and is used as a measure of the similarity of the structures. The RMSD can be calculated for all the atoms in the structures or a specifically selected subset of the atoms such as the set of all C_α atoms. The latter approach aligns only the specified subset of atoms in the structures, without counting all atoms equally in the calculations.

Even if only a subset of atoms is considered, the contribution of one atom to a meaningful superimposition of possible structures is in general different from that of another. Therefore, a more general way of computing the RMSD value is to assign a different weight for each different atom to be aligned or compared. The resulting RMSD value is called the weighted RMSD or wRMSD for short. We consider the simplest case when a pair of structures is to be superimposed. Let $x=[x_i=(x_{i,1}, x_{i,2}, x_{i,3})^T : i=1, \ldots, n]$ and $y=[y_i=(y_{i,1}, y_{i,2}, y_{i,3})^T : i=1, \ldots, n]$ be two sets of coordinates of the atoms selected to be aligned for two given structures, respectively. Assume that x and y have been translated so that their centers of geometry are both moved to the origin. Let Q be a rotation matrix. Then, the RMSD and wRMSD for the two structures are given by the following formulas.

$$RMSD(x,y) = \min_{Q} \sqrt{\frac{\sum_{i=1}^{n} \left(x_{i,1} - y'_{i,1}\right)^2 + \left(x_{i,2} - y'_{i,2}\right)^2 + \left(x_{i,3} - y'_{i,3}\right)^2}{n}} \quad (1)$$

$$wRMSD(x,y) = \min_{Q} \sqrt{\frac{\sum_{i=1}^{n} w_i \left[\left(x_{i,1} - y'_{i,1}\right)^2 + \left(x_{i,2} - y'_{i,2}\right)^2 + \left(x_{i,3} - y'_{i,3}\right)^2\right]}{n}} \quad (2)$$

where $y'_i = Qy_i$, and w_i are weights and $\sum_{1 \le i \le n} w_i = 1$.

Computing the RMSD value requires solving an optimization problem so that the superimposition of one structure on another is optimal. Diamond [3] investigated a non-orthogonal transformation method. Mclachlan [4] used an iterative method to find the optimal rotation between two structures. Kabsch [5] developed an eigenvalue method that requires the rotation matrix to be orthogonal. Determining an appropriate set of weights is not trivial, though. In general, more weights should be assigned to more "important" atoms, but the definition of "important" can be arbitrary. Kabsch [5] investigated the possibility of incorporating weight factors such as atomic masses. Damm et al. [6] developed a method with the weights determined by distance differences.

We are concerned with the effects of the thermal motions of the structures on their superimposition. Proteins have thermal motions, and some regions are flexible while others are relatively stable. Therefore, when superimposed, more stable regions of the structures should be assigned relatively larger weights, so that the resulting superimpositions can truly reflect the dynamic stabilities or flexibilities of the structures. The latter properties are of great importance in modeling practice such as in identifying active sites or the open and closed states of proteins [7–10]. We call such a weighted root-mean-square deviation the dynamically weighted RMSD or dRMSD for short.

Several research groups have tried to incorporate the dynamic properties of structures in RMSD calculations. Gerstein [11] tried to make an initial alignment to find stable regions of the structures and then further refine the alignment with larger weights assigned to those regions. Ye et al. [12] developed a knowledge-based method to compare flexible structures. Alexandrov [13] applied a Hidden Markov Model method to superimpose structures with a core subset of all the atoms. Schneider [14] used a genetic algorithm to identify the flexible regions in protein comparisons. Nichols et al. [15] developed an algorithm to identify rigid domains based on difference distance matrices.

In this paper, we propose a new algorithm for the calculation of the dRMSD of a group of protein structures. The weights are determined completely by the thermal motions of the atoms. We show that the thermal motions of the atoms can be obtained from several sources, such as the B-factors that can be determined from X-ray crystallography, or the mean-square fluctuations that can be estimated by Gaussian network model analysis and normal mode analysis. We show that the superimposition of the structures with dRMSD can successfully identify different domains of a protein and protein motions, and that it has important implications in practice such as aligning the ensemble of structures determined by NMR.

Methods

General RMSD

Given two proteins A and B with their structures represented by two coordinate matrices X and Y, the optimal superimposition of the two structures, in terms of their RMSD value, can be determined in the following two steps:

1. The two structures need to be translated so that the centers of geometry are located at the same place (e.g., the origin). Let $X=\{x_{ij}\}$ and $Y=\{y_{ij}\}$, $i=1, \ldots, n, j=1, 2, 3$. The centers of geometry can be computed by the formula,

$$x_c(j)=\sum_{i=1}^{n} x_{ij}/n, \quad y_c(j)=\sum_{i=1}^{n} y_{ij}/n, j=1,2,3 \tag{3}$$

Let X' and Y' be the coordinate matrices for the translated structures, $X'=\left\{x'_{ij}\right\}$ and $Y'=\left\{y'_{ij}\right\}$, $i=1, \ldots, n, j=1, 2, 3$. Then, $x'_{ij}=x_{ij}-x_c(j)$ and $y'_{ij}=y_{ij}-y_c(j)$.

2. A rotation matrix Q needs to be determined so that

$$\min_{Q} \|X'-Y'Q\|_F, \tag{4}$$

where $\|\cdot\|_F$ is the matrix Frobenius norm. Let $C=Y'^{\mathrm{T}}X'$ and the singular value decomposition of C be given by $C=U\Sigma V^{\mathrm{T}}$. Then, the optimal $Q=UV^{\mathrm{T}}$ and the RMSD of X and Y can be defined by the formula,

$$RMSD(X,Y)=\frac{\|X'-Y'Q\|_F}{\sqrt{n}} \tag{5}$$

Dynamically weighted RMSD

Let $X=\{x_{ij}\}$ and $Y=\{y_{ij}\}$, $i=1, \ldots, n$, $j=1, 2, 3$ be the coordinate matrices the structures of two proteins. Let X' and Y' be the translated coordinate matrices, as given in the above paragraphs. Then, the dRMSD of X and Y is defined as

$$dRMSD(X,Y)=\frac{\|D(X'-Y'Q)\|_F}{\sqrt{n}}, \tag{6}$$

where D is a diagonal matrix with the diagonal elements $D_{ii}=d_i$, $i=1, \ldots, n$, and d_i is the weight assigned to atom i and is defined to be inversely proportional to the fluctuation of atom i. Let the root-mean-square fluctuation of atom i be given by a value B_i. Then, we set $d_i=(B_i)^{-m}$, where m is an integer. In our calculations, we have used different m values for different groups of structures based on their regular RMSD values (see Table 1). This is obtained using a trial and error method.

Note that when there are more than two structures, we follow an algorithm similar to that developed in [16]: we first make an alignment for every pair of structures and find a representative structure for which the sum of its RMSD values from all the other structures is the minimum. We then re-align the structures iteratively until no changes can be made from the new alignments. Note that this iterative alignment method can be used for both regular and weighted superimpositions of multiple structures.

Table 1 Determination of the m value. *RMSD* Root mean squares deviation

Regular RMSD	m value
0–2.99 Å	2
3–5.99 Å	3
6–8.99 Å	4
9–11.99 Å	5
12–14.99 Å	6

Gaussian network model

The RMSDs of the atoms can be derived from their temperature factors or B-factors as determined by X-ray crystallography. If the B-factors are unknown from experiments, theoretical estimates may be used. The Gaussian network model (GNM) is a theoretical approach to obtaining the structural fluctuations of proteins around their equilibrium states at a residue level [17]. Let Γ be the contact matrix for a protein, and

$$\Gamma_{ij=}\begin{cases}-1 & if\, i\neq j \;\text{ and }\; d_{ij}\leq r_c\\ 0 & if\, i\neq j \;\text{ and }\; d_{ij}>r_c\\ -\sum_{i=1,i\neq j}^{n}\Gamma_{ij} & if\, i=j\end{cases} \tag{7}$$

where d_{ij} is the distance between ith and jth residues, and r_c is a cut-off distance. Let R_i be the equilibrium position of the ith residue, and ΔR_i the deviation of the residue from its equilibrium position. Then, using the GNM model, the root-mean-square fluctuations of the residues can be estimated by the formula,

Table 2 Protein data bank (PDB) IDs of proteins with two different crystal structures

Structure1	Structure2	RMSD (Å)
1AON	1OEL	12.38
1I2D	1M8P	4.07
1BNC	1DV2	3.9
1F3Y	1JKN	3.58
5CRO	6CRO	2.22
1A67	1CEW	4.64
1BMD	4MDH	2.22
1BYU	1RRP	14.4
1D5W	1DCM	1.8
1A32	1AB3	9.5
1G6O	1NLZ	0.93

Table 3 PDB IDs of proteins determined by nuclear magnetic resonance (NMR)

1BA4	1DVV	1HZL	1K8H	1RYK	3GCC
1AF8	1E8L	1I6F	1KUN	1SMG	3HSF
1AFI	1E17	1ICH	1LV3	1T0Y	3PHY
1BCN	1EIW	1IGL	1M94	1TIZ	
1BEG	1EO1	1IH0	1MZ5	1TNN	
1BEI	1EQ3	1IK0	1NE5	1UXC	
1BZK	1EZT	1IQO	1NRP	1VD4	
1C05	1F0Z	1IRH	1NYN	1WJ2	
1CN7	1FD6	1IRZ	1O8T	1XHJ	
1CRP	1FOW	1ITL	1P9K	1YEL	
1D8Z	1FUW	1J6Y	1PQX	1ZAC	
1D9A	1G92	1J56	1PV0	2CTN	
1DAX	1GB1	1JKZ	1RG6	2IGG	
1DKC	1GEA	1JOR	1RQ6	2SXL	
1DP3	1HFG	1K1V	1RWS	3CTN	

$$< \Delta R_i \cdot \Delta R_j >= \frac{\int (\Delta R_i \cdot \Delta R_j) e^{-V/k_b T} d\{\Delta R\}}{\int e^{-V/k_b T} d\{\Delta R\}} = (3k_b T/\gamma)\left[\Gamma^{-1}\right]_{ij}, \quad (8)$$

where k_b is the Boltzmann constant, γ is a spring constant and T is the absolute temperature. Note that if the singular value decomposition of Γ is $\Gamma = U\Lambda U^T$. Then, the ith residue fluctuation is obtained using the following,

$$< \Delta R_i \cdot \Delta R_i >= (3k_b T/\gamma) \sum_{k=1}^{n} U_{ki} \Lambda_{kk}^{+} U_{ki}, \quad (9)$$

where Λ^+ is the pseudo inverse of Λ.

Selected protein structures

A set of protein structures were selected as a set of test cases, based on their biological interest, conformational changes, sequential lengths, and modeling methods. Among them, proteins with two different crystal structures (Table 2) were used for pairwise structural alignment, while proteins with NMR-determined ensembles of structures (Table 3) were used for multiple structural alignments. The coordinate files of selected protein structures were downloaded from the Protein Data Bank [1] and Swiss-Prot [2].

Results

In this work, superimpositions of protein structures are studied at the residue level or coarse-grained level. For instance, only a C_α atom in each amino acid is taken into account in a protein structure. In addition, the weight values used in computational experiments of dRMSD are associated with predicted fluctuations of C_α atoms from GNM analysis. Note that a similar investigation can be conducted for superimpositions of proteins at the atomic level whenever the fluctuations of atoms in a protein is fully available; temperature factor values of crystal structures and order parameter values of NMR structures may also be used.

In general, the RMSD value of a regular RMSD alignment is smaller than other modified weighted RMSD alignment and can be considered as the lower boundary. Evaluating the performance of a superimposition method is not always trivial. However, in this work, it is important to compare the correlation coefficients between residue fluctuations and residue RMS values of alignment results. A larger correlation value usually indicates better agreement

Table 4 A set of proteins with two different conformations was tested

ID	RMSD	dRMSD	Correlation1		Correlation2	
1AON VS 1OEL	12.38	15.5	0.59	0.7	0.78	0.86
1I2D VS 1M8P	4.07	5.6	0.42	0.39	0.79	0.79
1BNC VS 1DV2	3.9	4.51	0.83	0.69	0.82	0.69
1F3Y VS 1JKN	3.58	3.84	0.74	0.75	0.7	0.75
5CRO VS 6CRO	2.22	2.86	0.9	0.9	0.97	0.98
1A67 VS 1CEW	4.64	4.82	0.4	0.65	0.43	0.7
1BMD VS 4MDH	2.22	2.25	0.56	0.66	0.53	0.62
1BYU VS 1RRP	14.4	16.9	0.47	0.88	0.52	0.98
1D5W VS 1DCM	1.8	2.05	0.12	0.4	0.17	0.45
1A32 VS 1AB3	9.5	12.77	0.66	0.88	0.65	0.97
1G6O VS 1NLZ	0.93	1.03	0.6	0.59	0.66	0.67
Average	–	–	0.63		0.71	

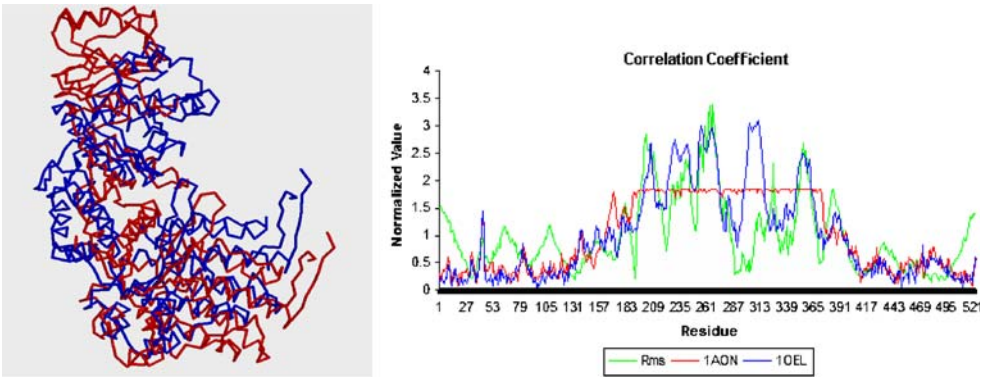

Fig. 1 Regular root mean squares deviation (RMSD) superimposition of 1AON and 1OEL. The *left picture* is the superimposition of proteins 1AON (*red*) and 1OEL (*blue*) visualized by Rasmol. The *right graph* plots Gaussian network model (GNM)-predicted fluctuation values of 1AON (*red*), GNM-predicted fluctuation values of 1OEL (*blue*) and RMS values (*green*) against residues. The linear coefficient between GNM-predicted fluctuation values of 1AON and RMS value is 0.59, and the linear coefficient between GNM-predicted fluctuation values of 1OEL and RMS value is 0.70

between RMS values of that alignment and protein dynamics.

The second two columns in Table 4 show the standard RMSD and dRMSD values. The last two columns show correlation coefficients between GNM-predicted residue fluctuations and residue RMS values after the standard RMSD and dRMSD alignments, respectively. Note that each pair has two structures, and two correlation coefficients are shown.

Pairwise alignment

A set of proteins with two different crystal structures was analyzed and aligned using the standard RMSD method and the dRMSD method, respectively. The proteins were selected based on availability of data, previous studies, and community interests. For each protein with two different structures, we first compute the alignment, and then calculate the correlation coefficient between the residue RMS values of alignment, and GNM-predicted residue fluctuations. Since each protein has two structures, in the dRMSD method, we apply GNM calculation to compute residue fluctuations of each structure and dRMSD alignment seperately. For instance, in Table 4, correlation 1 and correlation 2 show correlation coefficients between GNM-predicted residue fluctuations and residue RMS values for standard RMSD and dynamically weighted RMSD results, respectively. Within correlation 1 and correlation 2, we use both conformations and produce GNM-predicted fluctuation values of each conformation. Then, in either standard RMSD or dynamically weighted RMSD, we use GNM-predicted fluctuation values of each conformation separately and obtain correlation coefficients.

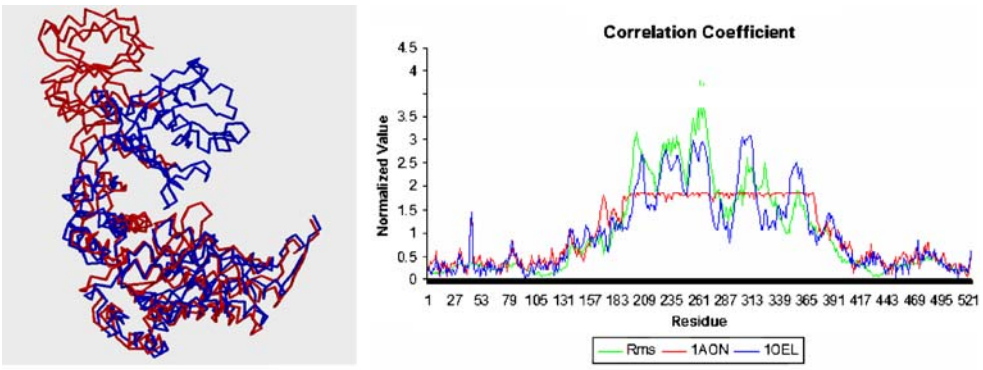

Fig. 2 Dynamically weighted RMSD (dRMSD) superimposition of 1AON and 1OEL. The *left picture* is the superimposition of proteins 1AON (*red*) and 1OEL (*blue*) visualized by Rasmol. The right graph plots GNM predicted fluctuation values of 1AON (*red*), GNM predicted fluctuation values of 1OEL (*blue*) and RMS values (*green*) against residues. The linear coefficient between GNM predicted fluctuation values of 1AON and RMS value is 0.78, and the linear coefficient between GNM predicted fluctuation values of 1OEL and RMS value is 0.86. Both are larger than previous results in Fig. 1 From the graph, two structural domains and the hinge motion can be identified

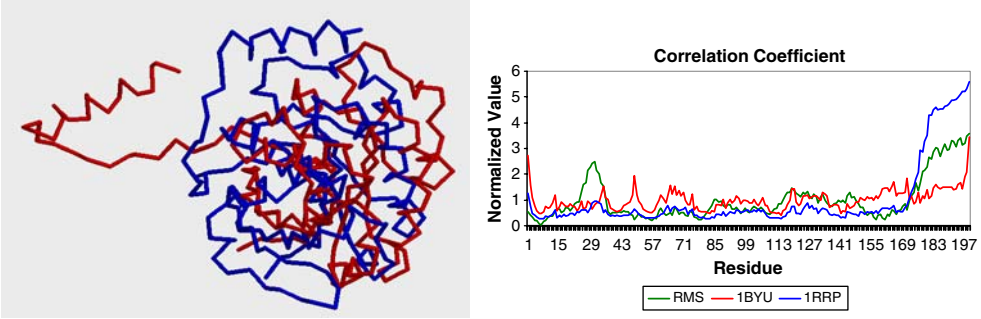

Fig. 3 RMSD superimposition of 1BYU vs 1RRP. The *left picture* is the superimposition of proteins 1BYU (*blue*) and 1RRP (*red*) visualized by Rasmol. The *right graph* plots GNM-predicted fluctuation values of 1BYU (*blue*), GNM-predicted fluctuation values of 1RRP (*red*) and RMS values (*green*) against residues. The linear coefficient between GNM-predicted fluctuation values of 1BYU and RMS value is 0.47, and the linear coefficient between GNM-predicted fluctuation values of 1RRP and RMS value is 0.88

After applying dRMSD alignment, improvement was seen in 8 of 11 proteins, in terms of the correlation coefficient between the residue RMS values of alignment and GNM-predicted residue fluctuations. On average, the correlation coefficient increased to 0.71 in dRMSD from 0.63 in RMSD. Note that the sample size is not very large, and the improvement varies case by case.

Identification of protein domains and motions

The standard RMSD alignment treats every atom or residue equally, while dRMSD alignment tends to give a biased alignment that agrees with the protein structural properties and functions well. The superimpositions could be severely affected by simply applying regular RMSD alignments, and hence, protein motifs, domains, functions and motions are very difficult to analyze and study in some cases. Therefore, for proteins with important functions, conformational changes and motions in biological systems, dRMSD alignment has the potential to be a powerful tool to study these issues. In our results, the dRMSD alignment successfully identifies protein domains and motions.

The function of the GroEL complex is to fold protein structures into their native states. Initially, the interior of this complex is highly hydrophobic and can hence easily bind unfolded proteins. After the protein is properly or nearly folded, the interior becomes hydrophilic. The folded protein is then released to the aqueous environment. The environmental change between hydrophilic and hydrophobic states is cyclic and corresponds to a conformational change in the GroEL complex that is driven by ATP hydrolysis [18]. Several issues related to GroEL dynamics and functions have been well studied both theoretically and experimentally [19, 20]. General RMSD and dRMSD were both performed on the superimposition of 1AON and 1OEL. In this test, we used a coarse-grained model for the protein. Note that it is also possible to use a full atomic model. Linear correlation coefficients can be used to determine or evaluate how the alignment results agree with protein dynamics. Our numerical results show that the linear correlation coefficient between the GNM-predicted

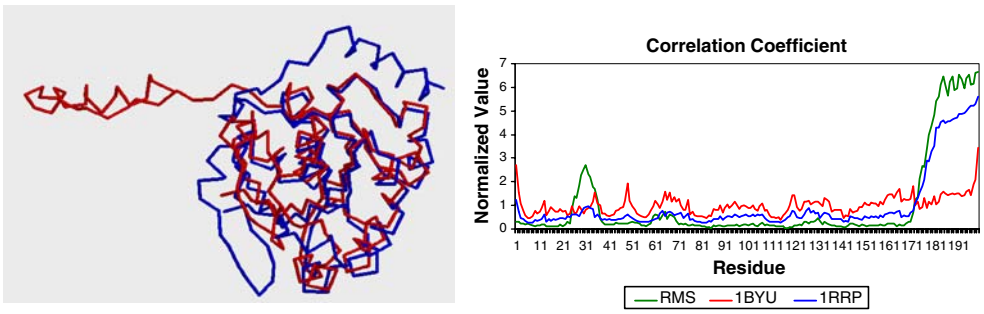

Fig. 4 dRMSD superimposition of 1BYU vs 1RRP. The *left picture* is the superimposition of proteins 1BYU (*blue*) and 1RRP (*red*) visualized by Rasmol. The *right graph* plots GNM-predicted fluctuation values of 1BYU (*blue*), GNM-predicted fluctuation values of 1RRP (*red*) and RMS values (*green*) against residues. The linear coefficient between GNM-predicted fluctuation values of 1BYU and RMS value is 0.52, and the linear coefficient between GNM-predicted fluctuation values of 1RRP and RMS value is 0.98. Both correlation values are larger than those in Fig. 3. From the graph, two structural domains and the hinge motion can be identified

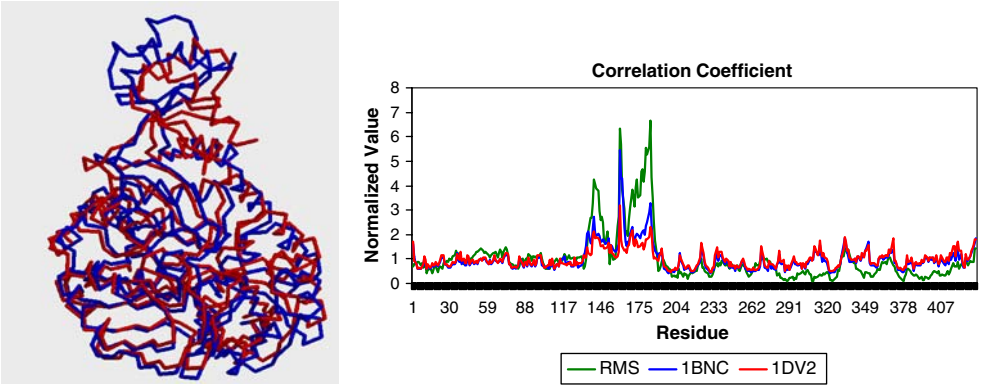

Fig. 5 RMSD superimposition of 1BNC vs 1DV2. The *left picture* is the superimposition of proteins 1BNC (*blue*) and 1DV2 (*red*) visualized by Rasmol. The *right graph* plots GNM-predicted fluctuation values of 1BNC (*blue*), GNM-predicted fluctuation values of 1DV2 (*red*) and RMS values (*green*) against residues. The linear coefficient between GNM-predicted fluctuation values of 1BNC and RMS value is 0.83, and the linear coefficient between GNM-predicted fluctuation values of 1DV2 and RMS value is 0.69

residue fluctuations of 1AON and residue RMS values of RMSD is 0.59, and that between GNM-predicted residue fluctuations of 1OEL and residue RMS values of RMSD is 0.70, while both coefficients obtained using dRMSD are 0.78 and 0.86, respectively (see Figs. 1, 2). This implies that, after incorporating residue fluctuations of protein dynamics, the superimposition of protein structures agrees with residue fluctuations better in this case. The superimposition of 1AON and 1OEL has also been visualized by Rasmol; Fig. 2 shows that a hinge motion and two structure domains of GroEL are identified by the dRMSD method, and the lower domain in the figure has clearly been well aligned also in dRMSD calculations, which is consistent with studies conducted previously by Damm et al. [6].

Similar results have also been found in 1BYU and 1RRP (see Figs. 3, 4). The protein is a trans-membrane protein responsible for importing proteins into the nucleus and also for exporting RNA molecules. The standard RMSD shows that the linear correlation coefficient between GNM-predicted residue fluctuations of 1BYU and residue RMS values of RMSD is 0.47, and that between GNM-predicted residue fluctuations of 1RRP and residue RMS values of RMSD is 0.88, while both coefficients obtained by using dRMSD are 0.52 and 0.98, respectively, which is greatly improved. The visualized alignment of dRMSD suggests that the protein may have two structural domains and hinge motion, and the central core region of 40–160 residues has been well aligned. The conformational change from 1BYU to 1RRP may illustrate how the protein functions in the transporting of molecules. The central core domain is very static and responsible for maintaining the structure, while the tail might be the functional domain. However, the standard RMSD alignment did not clearly show two structural domains and the motion.

The dRMSD alignment successfully identifies the domains and motions, even if agreement between GNM-

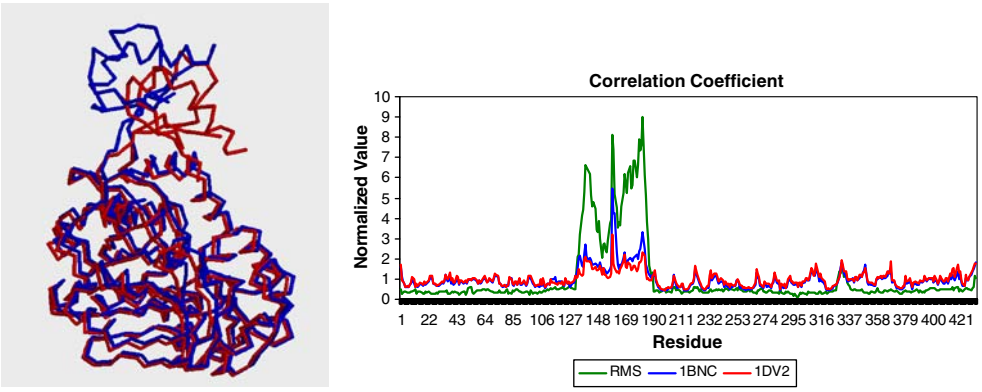

Fig. 6 dRMSD superimposition of 1BNC vs 1DV2. The *left picture* is the superimposition of proteins 1BNC (*blue*) and 1DV2 (*red*) visualized by Rasmol. The *right graph* plots GNM-predicted fluctuation values of 1BNC (*blue*), GNM-predicted fluctuation values of 1DV2 (*red*) and RMS values (*green*) against residues. The linear coefficient between GNM-predicted fluctuation values of 1BNC and RMS value is 0.82, and the linear coefficient between GNM-predicted fluctuation values of 1DV2 and RMS value is 0.69. Neither correlation value changes much. From the graph, the structural domains and hinge motion can be identified

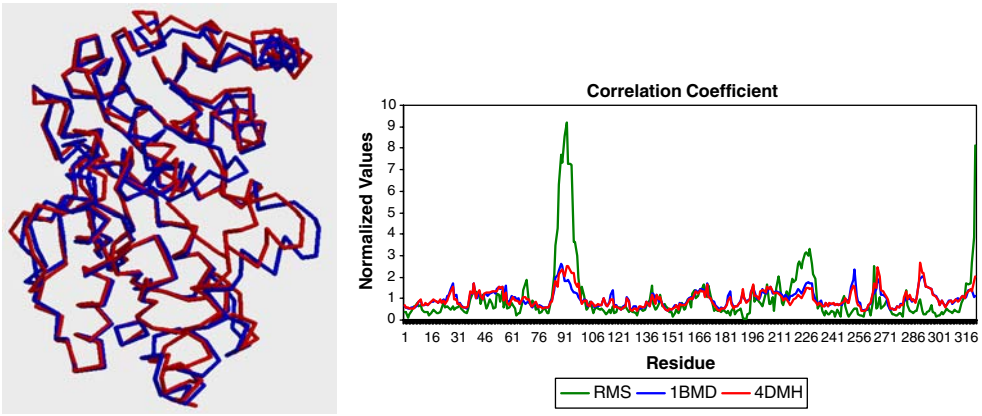

Fig. 7 RMSD superimposition of 1BMD vs 4MDH. The *left picture* is the superimposition of proteins 1BMD (*blue*) and 4MDH (*red*) visualized by Rasmol. The *right graph* plots GNM-predicted fluctuation values of 1BMD (*blue*), GNM-predicted fluctuation values of 4MDH (*red*) and RMS values (*green*) against residues. The linear coefficient between GNM-predicted fluctuation values of 1BMD and RMS value is 0.56, and the linear coefficient between GNM-predicted fluctuation values of 4MDH and RMS value is 0.66

predicted residue fluctuations and residue RMS values of alignment are not improved after using dRMSD. This can be seen in the test on 1BNC VS 1DV2 (see Figs. 5, 6). Biotin carboxylase has two structures (1BNC and 1DV2). The numerical results show that the linear correlation coefficient between GNM-predicted residue fluctuations of 1BNC and residue RMS values of RMSD is 0.83, and that between GNM-predicted residue fluctuations of 1DV2 and residue RMS values of RMSD is 0.69, and the coefficients obtained by using dRMSD are 0.82 and 0.69, respectively. The visualized alignment of dRMSD shows the two structural domains, and the motion of this protein seems very similar to that of GroEL.

However, it has also been found that the dRMSD alignment may not be applicable to a protein whose two conformations are very close. The test on 1BMD vs 4MDH illustrates exactly such a case (Figs. 7, 8). The standard RMSD value is 2.22, while the dRMSD value is just 2.25. The correlation coefficients of dRMSD between residue RMS values of dRMSD and residue fluctuations are a little worse compared to those of RMSD. Both alignments are very well produced. In most of our examples, the method can identify structural domains and motions, but may fail the test for proteins with two very close structures.

Multiple structure alignment

An NMR protein structure is usually determined with a set of models that form an ensemble. The multiple structure alignment of these models in the ensemble is critical to understanding the dynamic properties of the protein in solution, for instance, their motions and functions. The standard multiple structure RMSD calculation provides an average alignment, but the alignment can be affected by

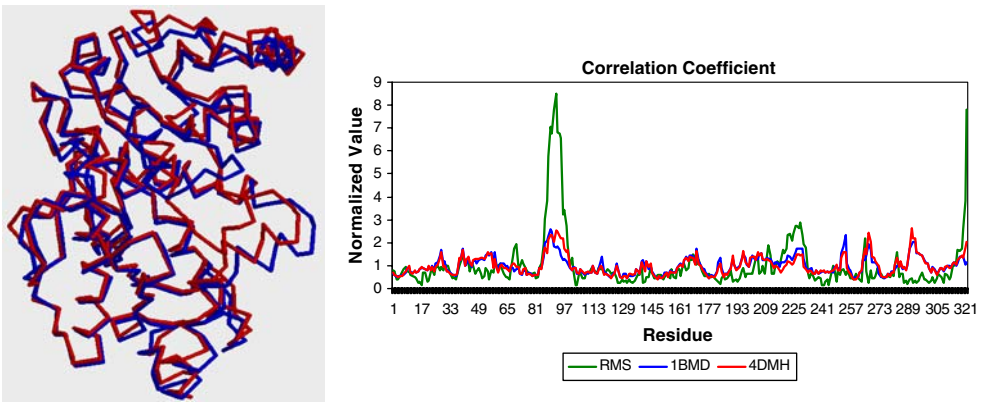

Fig. 8 dRMSD superimposition of 1BMD vs 4MDH. The *left picture* is the superimposition of proteins 1BMD (*blue*) and 4MDH (*red*) visualized by Rasmol. The *right graph* plots GNM-predicted fluctuation values of 1BMD (*blue*), GNM-predicted fluctuation values of 4MDH (*red*) and RMS values (*green*) against residues. The linear coefficient between GNM-predicted fluctuation values of 1BMD and RMS value is 0.53, and the linear coefficient between GNM-predicted fluctuation values of 4MDH and RMS value is 0.62. The correlation values are a little worse compared to those of RMSD in Fig. 7

Table 5 A set of NMR protein ensembles, each of which has 20 energy-minimized models. Columns 2–3 and 7–8 show the standard and dRMSD values, respectively. Columns 4–5 and 9–10 show the correlation coefficients between GNM B-factor values and RMS values after standard and dRSMD alignments, respectively

PDB ID	RMSD	dRMSD	Correlation 1[a]	Correlation 2[a]	PDB ID	RMSD	dRMSD	Correlation 1[a]	Correlation 2[a]
1BA4	5.07	6.46	0.46	0.77	1ITL	2.02	2.1	0.81	0.84
1AF8	2.34	2.66	0.63	0.7	1J6Y	2.53	2.78	0.54	0.62
1AFI	0.86	0.92	0.59	0.66	1J56	1.63	1.7	0.58	0.61
1BCN	2.31	2.54	0.82	0.84	1JKZ	1.01	1.26	0.7	0.82
1BEG	1.54	1.7	0.75	0.84	1JOR	1.97	2.11	0.76	0.77
1BEI	1.36	1.58	0.75	0.87	1K1V	1.76	2.02	0.63	0.73
1BZK	4.9	8.1	0.66	0.88	1K8H	2.38	2.53	0.7	0.74
1C05	1.42	1.63	0.87	0.84	1KUN	2.21	2.49	0.66	0.7
1CN7	1.56	1.64	0.86	0.86	1LV3	4.93	10	0.26	0.94
1CRP	1.17	1.24	0.72	0.76	1M94	1	1.07	0.69	0.72
1D8Z	1.89	2.09	0.67	0.68	1MZ5	1.47	1.56	0.82	0.84
1D9A	2.56	3	0.71	0.73	1NE5	1.4	1.73	0.96	0.98
1DAX	1.61	1.74	0.88	0.92	1NRP	1.51	1.74	0.91	0.9
1DKC	3.71	4.47	0.7	0.73	1NYN	3.35	5.07	0.67	0.93
1DP3	4.26	5.38	0.37	0.78	1O8T	5.38	7.36	0.15	0.77
1DVV	0.94	1.03	0.55	0.68	1P9K	2.86	3.12	0.63	0.71
1E8L	1.87	1.98	0.62	0.65	1PQX	2.2	2.47	0.66	0.64
1E17	2.07	2.23	0.81	0.84	1PV0	0.87	0.99	0.97	0.98
1EIW	2.09	2.34	0.73	0.81	1RG6	2.35	2.7	0.92	0.92
1EO1	2.33	2.57	0.62	0.72	1RQ6	2.3	2.63	0.81	0.85
1EQ3	2.32	2.73	0.64	0.8	1RWS	4.69	7.33	0.41	0.77
1EZT	1.27	1.36	0.55	0.69	1RYK	1.4	1.55	0.66	0.71
1F0Z	1.35	1.48	0.93	0.96	1SMG	1.6	1.77	0.84	0.91
1FD6	0.75	0.82	0.53	0.63	1T0Y	1.63	1.72	0.87	0.88
1FOW	2.51	2.89	0.92	0.95	1TIZ	0.96	1.04	0.78	0.82
1FUW	2.03	2.29	0.82	0.82	1TNN	1.46	1.52	0.66	0.71
1G92	4.06	5.78	0.21	0.94	1UXC	1.31	1.44	0.71	0.74
1GB1	0.43	0.47	0.61	0.66	1VD4	3.41	4.98	0.68	0.9
1GEA	2.36	3.03	0.94	0.95	1WJ2	1.96	2.22	0.85	0.88
1HFG	2.72	3.19	0.64	0.77	1XHJ	2.9	3.31	0.71	0.75
1HZL	1.29	1.35	0.73	0.74	1YEL	2.21	2.38	0.82	0.84
1I6F	1.47	1.66	0.91	0.89	1ZAC	1.27	1.38	0.54	0.62
1ICH	1.4	1.46	0.68	0.7	2CTN	1.72	1.84	0.8	0.84
1IGL	3.12	3.67	0.79	0.82	2IGG	2.4	2.68	0.84	0.89
1IH0	2.39	2.61	0.61	0.7	2SXL	2.45	2.58	0.85	0.87
1IK0	1.26	1.37	0.9	0.92	3CTN	2.07	2.47	0.72	0.84
1IQO	2.73	3.39	0.52	0.72	3GCC	1.45	1.62	0.92	0.92
1IRH	1.8	2.12	0.7	0.8	3HSF	1.45	1.62	0.92	0.92
1IRZ	3.1	3.84	0.76	0.84	3PHY	2.34	2.71	0.71	0.8

[a] Average correlation: correlation 1 (RMSD) : 0.71 Correlation 2 (dRMSD) : 0.80

mobile or flexible regions of the protein ensemble. The dRMSD could be a potential tool to address this issue, since, in the dRMSD method, the contributions of each residue in the alignment are based on their theoretically obtained fluctuations.

A set of NMR-determined proteins were selected with 20 models in each ensemble, as determined by the CNS 1.1 software package [21]. Generally, GNM computation requires a single structure in the computation. Since each NMR protein ensemble has 20 models, we compute the contact matrix for

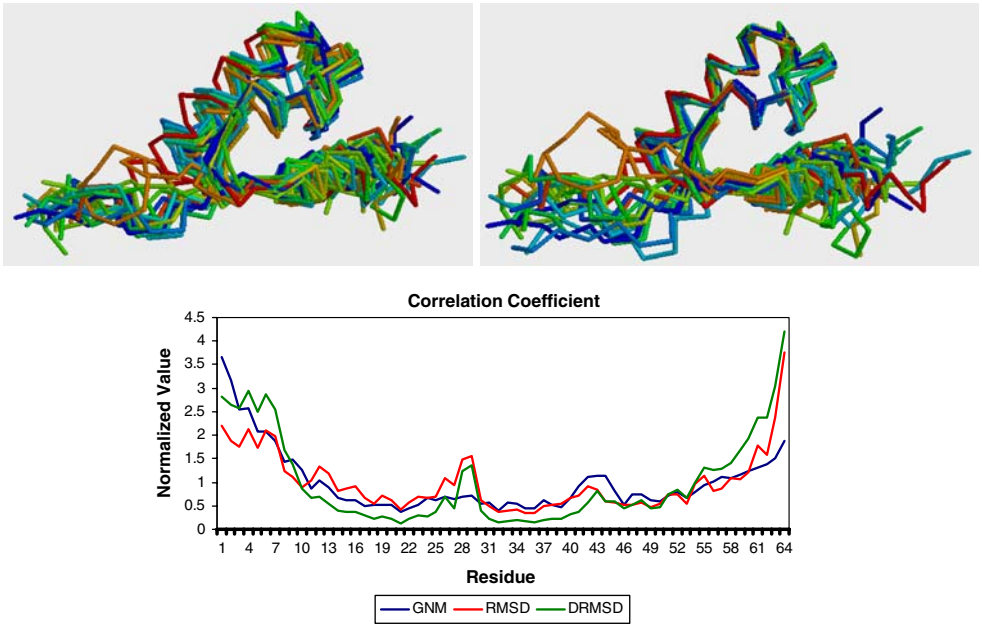

Fig. 9 RMSD vs dRMSD superimposition of 1IRZ. The *upper left picture* is the superimposition of proteins 1IRZ ensemble, using the standard RMSD visualized by Rasmol, while the *upper right picture* is the superimposition of proteins 1IRZ ensemble using dRMSD. The *bottom graph* plots GNM-predicted fluctuation values of 1IRZ (*blue*), RMS values of the standard RMSD (*red*) and RMS values of the dRMSD (*green*) against residues. The correlation coefficient between GNM-predicted fluctuation values of 1IRZ ensemble and RMS values of the standard RMSD is 0.76; the correlation coefficient between GNM-predicted fluctuation values of 1IRZ ensemble and RMS values of dRMSD is 0.84

each model, obtain the averaged contact matrix to represent the residue contacts of this protein, followed by GNM calculations, and then calculate the residue fluctuations of the ensemble. These residue fluctuations are then incorporated into our dRMSD multiple alignment. A total of 78 NMR ensembles were tested and 88% of NMR ensemble alignments were improved in terms of the correlation coefficients between GNM-predicted fluctuation values and residue RMS values of the ensemble. On average, the correlation increased to 0.80 in dRMSD from 0.71 in RMSD (see Table 5). The alignment of each ensemble can be visualized, and, for most models, the dRMSD can successfully identify structural domains and dynamics.

In 1IRZ, the correlation coefficient between GNM-predicted residue fluctuations and residue RMS values of the dRMSD alignment is 0.84, while that in the standard

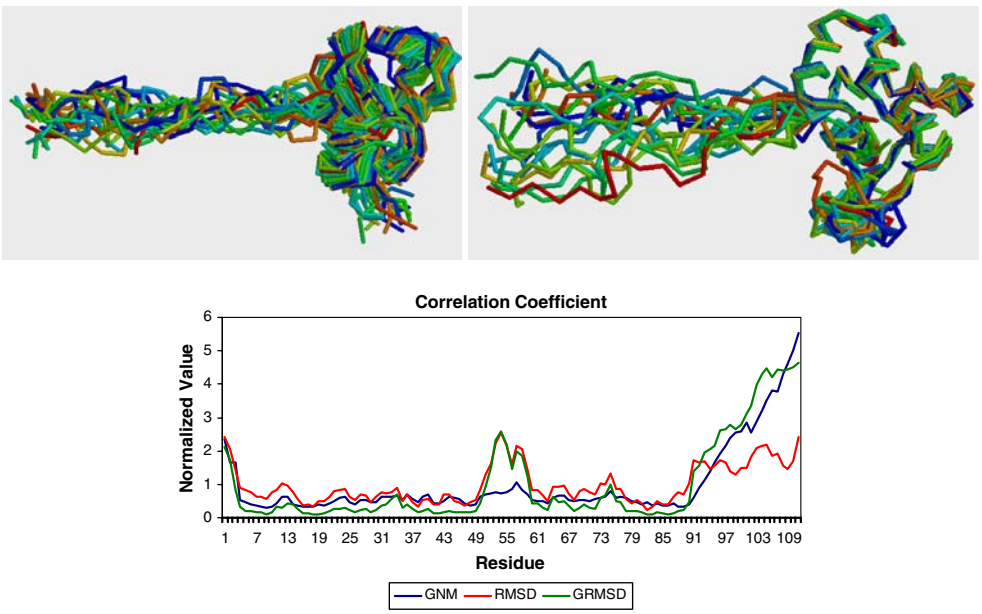

Fig. 10 RMSD vs dRMSD superimposition of 1NYN. The *upper left picture* is the superimposition of proteins 1NYN ensemble, using the standard RMSD visualized by Rasmol, while the *upper right picture* is the superimposition of proteins 1NYN ensemble, using the dRMSD. The *bottom graph* plots GNM-predicted fluctuation values of 1NYN (*blue*), RMS values of the standard RMSD (*red*) and RMS values of the dRMSD (*green*) against residues. The correlation coefficient between GNM-predicted fluctuation values of 1NYN ensemble and RMS values of the standard RMSD is 0.67; the correlation coefficient between GNM-predicted fluctuation values of 1NYN ensemble and RMS values of dRMSD is 0.93

RMSD method is only 0.76. The multiple structure alignment in the central helix region was improved in the dRMSD model (see Fig. 9). The N- and C-terminal regions both have flexible regions. However, the visualized alignment of RMSD does not clearly show this information. In a paper by Hosoda and colleagues [22], it was shown that side chains of several residues in the central helix region form hydrophobic interactions and stabilize the region, which is consistent with our result.

Similar results are also found in 1NYN. The correlation coefficient between GNM-predicted residue fluctuations and residue RMS values of RMSD alignment is 0.67, but the value is 0.93 in dRMSD. The protein structure has two structural domains, as seen clearly in Fig. 10. The central static region of residues 1–90 is mainly helix-turn-helix and aligns very well in the dRMSD method, and the tail region is more flexible. The standard multiple structure RMSD calculation fails to generate a clear alignment.

Relationships between RMSD and dRMSD

The fundamental goal of dRMSD is to incorporate dynamics into protein structure, and then produce the optimal alignment. However, several issues need to be clarified. How does the dRMSD value interpret the difference between two structures? Here, we used several examples to investigate the correlation between dRMSD and RMSD values of each protein in our sample.

First, we compared the RMSD value and dRMSD value of crystal proteins (see data in Table 4), each of which had two different conformations (see Fig. 11). The correlation was 0.99. Linear regression shows that the relationship between dRMSD and RMSD values follows the formula:

$$\mathrm{dRMSD} = 1.24 \times \mathrm{RMSD} - 0.18$$

We also compared the RMSD and dRMSD values of NMR ensembles (see data in Table 5). The correlation is 0.96 (see Fig. 12). The relationship between dRMSD and RMSD values of NMR ensemble alignments is expressed in the following formula, using linear regression,

$$\mathrm{dRMSD} = 1.59 \times \mathrm{RMSD} - 0.80$$

The coefficients of linear functions for pairwise and multiple structure alignments are different, and may vary for a number reasons, e.g., the number of structures in each alignment (each NMR ensemble has 20 structures), scaling factor values, and conditions in the calculations. Our analysis shows that dRMSD is correlated to regular RMSD.

Conclusions

Protein superimposition plays an important role in understanding protein structure models, protein functions and domains in biological activities and systems. Numerical methods developed based on the idea of RMSD have often been used and some of them have important applications in protein structural modeling-related research. In this paper, we proposed a new algorithm for the weighted RMSD calculations of a group of protein structures. The algorithm takes into account the information of protein fluctuations that can be obtained from either an experimental method, such as B-factor values in X-ray crystallography, or a theoretical method, such as pseudo B-factor values in GNM. Protein fluctuations are highly related to protein structure, function and dynamics. The method developed here incorporating protein dynamics hence has an interesting and strong biophysical background.

The method was applied to test a set of proteins with two different conformations as determined by X-ray crystallography. Superimpositions and plots of residue fluctuations of proteins clearly show and identify protein domains and suggest possible protein motions. Some results are consistent with previous experimental and theoretical studies of these proteins. We checked the correlation between residue RMS and pseudo B-factor values. Using dRMSD algo-

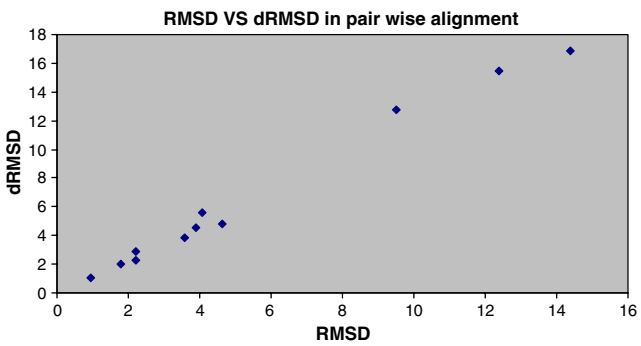

Fig. 11 Plot of RMSD and dRMSD values of proteins with two different conformations. Correlation = 0.99 and the linear regression model is dRMSD=1.24×RMSD−0.18

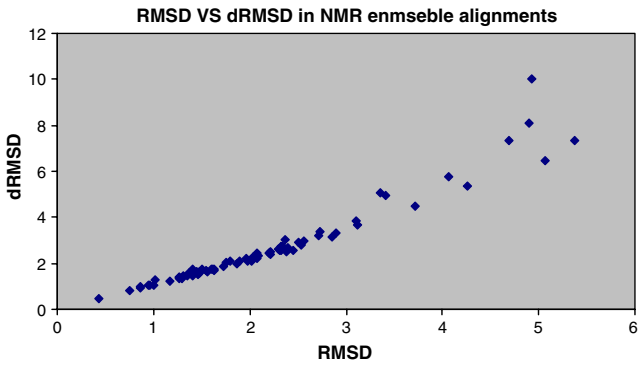

Fig. 12 Plot of RMSD and dRMSD values of NMR protein. Correlation = 0.96 and the linear regression model is dRMSD= 1.59×RMSD−0.80

rithms, the correlation value between the residue RMS values of the superimposition and residue fluctuation values predicted by GNM for a protein becomes larger, compared to the regular RMSD method. Though there is no theory strongly supporting the notion that a higher correlation must imply a more accurate superimposition, it was very interesting to note that a protein structure can be modeled to fluctuate or transform to a different conformation with a strong correlation between residue RMS values and protein fluctuations and dynamics. The superimposition suggested by dRMSD may also have the potential to become a starting point for researchers when studying protein conformational transformation and molecular dynamic simulation. However, when two structures of a protein are very close, dRMSD may not show any improvement in terms of the correlation between residue fluctuations and residue RMS.

A set of NMR-determined protein ensembles with 20 theoretical models was also tested by this method. An average contact matrix is used in the calculation of GNM and residue fluctuations are then determined. The superimposition of 20 structures in an ensemble is obtained through an iterative dRMSD algorithm. Superimposition of the results of dRMSD displayed by graphic software show very important implications in identifications of protein static and mobile domains and protein motions. Some results are consistent with experimental biochemical studies. The correlation value of dRMSD between averaged residue RMS and residue fluctuations predicted by GNM for an NMR ensemble increased substantially compared to the regular dRMSD method.

In summary, we have developed a weighted superimposition method, dRMSD, for protein structure alignment, with the weights determined completely from the thermal motion of the atoms. We show that the thermal motions of the atoms can be obtained from several sources, such as the mean-square fluctuations that can be estimated by GNM analysis. We show that the superimposition of the structures with dRMSD can successfully identify different protein domains and protein motions, and that it has important implications in practice, such as aligning an ensemble of structures determined by NMR. Therefore, the superimposition obtained by the dRMSd method may have important applications in many structure modeling areas, including protein structure transformation, identification of protein domains and motions, molecular dynamic simulation, quality assessment of protein structures, and NMR structure determination and refinement. Related applications and studies using dRMSD will be reported elsewhere in the future.

Acknowledgments The authors would like to thank the National Institutes of Health (NIH) and National Center for Research Resources (NCRR) Grant P20 RR16481 (Kentucky Biomedical Research Infrastructure Network) and National Science Foundation (NSF) Kentucky EPSCoR Research Enhancement Grant (REG) for support.

References

1. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542
2. Bairoch A, Boeckmann B (1991) The SWISS-PROT protein sequence data bank. Nucleic Acids Res 19(Suppl):2247–2249
3. Diamond R (1976) On the comparison of conformations using linear and quadratic transformations. Acta Cryst A 32:1–10
4. Mclachlan A (1972) A mathematical procedure for superimposing atomic coordinates of proteins. Acta Cryst A 28:656–657
5. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Cryst A 32:922–923
6. Damm K, Carlson H (2006) Gaussian-weighted RMSD superimposition of proteins: a structural comparison for flexible proteins and predicted protein structures. Biophys J 90:4558–4573
7. Tsai C, Xu D, Nussinov R (1998) Protein folding via binding, and vice versa. Folding Design 3:71–80
8. Dill K (1999) Polymer principles and protein folding. Protein Sci 8:1166–1180
9. Huber T, Botelho A, Beyer K, Brown M (2004) Membrane model for the G-Protein-Coupled RECEPTOR Rhodopsin: hydrophobic interface and dynamical structure. Biophys J 86 (4):2078–2100
10. Ben-Tal N, Honig B, Peitzsch RM, Denisov G, McLaughlin S (1996) Binding of small basic peptides to membranes containing acidic lipids: theoretical models and experimental results. Biophys J 71(2):561–575
11. Gerstein M, Echols N (2004) Exploring the range of protein flexibility, from a structural proteomics perspective. Curr Opin Chem Biol 8:14–19
12. Ye Y, Godzik A (2004) Database searching by flexible protein structure alignment. Protein Sci 13:1841–1850
13. Alexandrov V, Gerstein M (2004) Using 3D hidden Markov models that explicitly represent spatial coordinates to model and compare protein structures. BMC Bioinformatics 5:2–12
14. Schneider T (2004) Domain identification by iterative analysis of error-scaled difference distance matrices. Acta Cryst D 60:2269–2275
15. Nichols W, Zimm B, Ten Eyck L (1997) Conformation invariant structures of the a1b1 human hemoglobin dimer. J Mol Biol 270:598–615
16. Gower JC (1975) Generalized procrustes analysis. Psychometrika 40:33–51
17. Haliloglu T, Bahar I, Erman B (1997) Gaussian dynamics of folded proteins. Phys Rev Lett 79:3090–3093
18. Lin Z, Rye HS (2006) GroEL-mediated protein folding: making the impossible, possible. Crit Rev Biochem Mol Biol 41(4):211–39
19. Ellis J (1992) Protein folding: cytosolic chaperonin confirmed. Nature 358(6383):191–195
20. Ma J, Sigler PB, Xu Z, Karplus M (2000) A dynamic model for the allosteric mechanism of GroEL. J Mol Biol 302(2):303–13
21. Brunger A et al (1998) Crystallography and NMR System: a new software suite for macromolecular structure determination. Acta Cryst D 54:901–921
22. Hosoda K et al (2002) Molecular structure of the GARP family of plant Myb-related DNA binding motifs of the Arabidopsis response regulators. Plant Cell 14(9):2015–2029