Bioinformatics, 32, 2016, i314–i321
doi: 10.1093/bioinformatics/btw262
ISMB 2016

OXFORD

Ensemble-based evaluation for protein structure models

Michal Jamroz[1], Andrzej Kolinski[1] and Daisuke Kihara[2,3,*]

[1]Department of Chemistry, University of Warsaw, Warsaw, 02-093, Poland, [2]Department of Biological Sciences and [3]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Comparing protein tertiary structures is a fundamental procedure in structural biology and protein bioinformatics. Structure comparison is important particularly for evaluating computational protein structure models. Most of the model structure evaluation methods perform rigid body superimposition of a structure model to its crystal structure and measure the difference of the corresponding residue or atom positions between them. However, these methods neglect intrinsic flexibility of proteins by treating the native structure as a rigid molecule. Because different parts of proteins have different levels of flexibility, for example, exposed loop regions are usually more flexible than the core region of a protein structure, disagreement of a model to the native needs to be evaluated differently depending on the flexibility of residues in a protein.

Results: We propose a score named FlexScore for comparing protein structures that consider flexibility of each residue in the native state of proteins. Flexibility information may be extracted from experiments such as NMR or molecular dynamics simulation. FlexScore considers an ensemble of conformations of a protein described as a multivariate Gaussian distribution of atomic displacements and compares a query computational model with the ensemble. We compare FlexScore with other commonly used structure similarity scores over various examples. FlexScore agrees with experts' intuitive assessment of computational models and provides information of practical usefulness of models.

Availability and implementation: https://bitbucket.org/mjamroz/flexscore

Contact: dkihara@purdue.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins are intrinsically flexible molecules. Flexibility is essential in understanding protein activity where conformational entropy has a major role, including protein-ligand (Tzeng and Kalodimos, 2012), protein–protein interactions (Betts and Sternberg, 1999), protein allostery (Popovych *et al.*, 2006), and protein folding.

Although it is conventional to consider a single structure for a protein particularly when the structure was solved by X-ray crystallography, a protein can change its conformation in different experimental conditions (Andrec *et al.*, 2007; Garbuzynskiy *et al.*, 2005; Kosloff and Kolodny, 2008). Furthermore, there is compelling evidence that indicates even X-ray diffraction data contains flexibility information and that diffraction data can be interpreted into different conformations (DePristo *et al.*, 2004; Kuzmanic *et al.*, 2011). Indeed, a group of structural biologists proposed to deposit an ensemble of structures even for X-ray crystallography to public databases so that data represent structural heterogeneity and dynamics more properly (Furnham *et al.*, 2006). Thus, in principle it is more appropriate to represent a protein structure as an ensemble of alternative conformations (Fenwick *et al.*, 2011; Olsson *et al.*, 2014). In practice, flexibility information of a protein can be obtained either from experiments, such as NMR, or computational analysis including molecular dynamics (MD) simulations and normal mode analysis. By bringing flexibility into the picture of proteins, the protein sequence-to-structure-to-function paradigm (Fetrow and Skolnick, 1998) must be revised to sequence-to-structure-to-dynamics-to-function.

Reflecting the current situation that a protein is usually represented with a single conformation, most of the commonly used protein structure model evaluation methods compare a model against a single structure of the protein ('the native structure'). For example,

the most commonly used dissimilarity measure, the root mean square deviation (RMSD), quantifies difference of corresponding positions of two rigid structures after optimal superimposition (Kabsch, 1978):

$$rmsd(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|x_i^A - x_i^B\right\|^2} \quad (1)$$

where A and B are the Cartesian coordinates of atoms of two proteins, x_i^A is the coordinates of atom i in protein A and N is the number of atom pairs to be compared. In the protein structure prediction field, measures that are based on RMSD, such as GDT-TS (Zemla, 2003), TM-score (Zhang and Skolnick, 2005), are also used, which all compare two rigid structures. There are other methods that compare different representations of protein structures (Hasegawa and Holm, 2009), but they all belong to the rigid structure comparison realm and suffer from the same problem of neglecting protein flexibility. In evaluating a computational model, neglecting flexibility in the native structure results in overestimating displacement of residues in the model at flexible regions and underestimating displacement at rigid regions of the protein. For example, structure difference in a model at a flexible tail of the protein does not need to be penalized as long as the tail is modeled within the range of reasonable tail motion. On the other hand, a displacement of residues in a model at the core of a protein may be considered as a severe error of the model. Simply superimposing a rigid model to the single rigid conformation of the native structure cannot distinguish these two cases.

Although considering the flexibility is still not common for evaluating protein structure prediction models, there have been several structure comparison methods that consider protein flexibility. In the simplest form, flexibility can be indirectly represented as weights assigned to each residue in a pair of protein structures to be compared, where weights are computed from the B-factors in crystallography (Wu and Wu, 2010) or theoretical estimates of fluctuation. Weights can be also computed from the distance of corresponding residues in the previous round of alignment in iterative computations of RMSD (Damm and Carlson, 2006). FlexE uses a residue-level elastic network model and evaluates the difference of two structures in terms of energetic cost for deforming one structure into the other (Perez *et al.*, 2012).

An alternative strategy is to consider ensembles of protein structures. Brüschweiler proposed a fast computational method for averaging pairwise RMSD of two ensembles (Brüschweiler, 2003). The Kullback-Leibler divergence was used to quantify the similarity of two structure ensembles (Lindorff-Larsen and Ferkinghoff-Borg, 2009). An algorithm was developed for specifically aligning structure ensembles in a MD trajectory, which minimizes the variance of structures as well as pairwise displacement of structures from adjacent time frames (Gapsys and de Groot, 2013). THESEUS (Theobald and Wuttke, 2006) and bFit (Mechelke and Habeck, 2010) use maximum likelihood to superimpose flexible protein structures where atom positions are assumed to have Gaussian distribution. An advantage of these two methods is that they are free of *ad hoc* parameters.

Related works include structure alignment algorithms that explicitly consider local conformational change in structures caused by flexibility (Shatsky *et al.*, 2002; Ye and Godzik, 2003) and protein flexible functional site identification (Moll *et al.*, 2010; Sael and Kihara, 2012).

In this work, we propose a score for evaluating quality of protein structure models by taking flexibility of the native protein structure into account. The score named FlexScore (FS) quantifies how well each residue in the query model locates within the displacement range observed in structure ensemble. Among the existing methods mentioned above, only FlexE was designed for structure model evaluation. In contrast to FlexE, which needs parameter setting for an underlying elastic network model and outputs an energy value to a model that is not straightforward to interpret, our method does not need arbitrary parameters and provides an intuitive score that is easy to compare with conventional structure evaluation scores.

Below we first introduce FS and clarify the characteristics of the score using illustrative models. Then we compare FS with RMSD, GDT-TS (Zemla, 2003), and TM-score (Zhang and Skolnick, 2005) on several computational models. Lastly, FS was applied to evaluate computational models that were submitted to the Critical Assessment of Techniques for protein Structure Prediction (CASP; Moult *et al.*, 2014). It is shown that FS provides more reasonable and complete evaluation for structure models in comparison with other commonly used structure similarity scores.

2 Methods

We begin with outlining the algorithm of the structure superimposition method that was used for computing FS and then explain the dataset of protein models we used.

2.1 Maximum likelihood superimposition of protein ensembles

First, an ensemble of alternative conformations of a reference structure (native structure), against which computational models will be compared, is obtained from MD simulation (details given in the next section) or an experimental method, e.g. NMR. Then, the ensemble of conformations is represented as a probabilistic model, where a probability density function describes each residue Cα atom displacements from the ensemble mean. A probabilistic model is defined using structure superimposition by maximum likelihood estimation on the multivariate Gaussian model (Hirsch and Habeck, 2008; Mechelke and Habeck, 2010; Rother *et al.*, 2008). We used the framework of THESEUS for this ensemble superimposition, as it properly considers variance and correlations of atoms in the structures in an explicit fashion. Below we briefly outline the algorithm. For more details refer to the original paper (Theobald and Wuttke, 2006).

The coordinates of a structure X_i in an ensemble can be represented as

$$X_i = (M + E_i)R_i^T - 1_k t_i^T \quad (2)$$

using a mean structure M, a zero-mean Gaussian matrix displacement E_i that follows a Gaussian distribution of $N_{k,3}(0, \Sigma, I_3)$, where Σ is a $k \times k$ covariance matrix where k is the number of Cα atoms in the structure, a rotation matrix R_i, a 3×1 translational vector t_i, and a $k \times 1$ column vector of ones. T denotes transpose of a matrix. As commonly used in Bayesian analysis, eigenvalues of the covariance matrix are assumed to be distributed according to an inverse Gamma distribution, which is defined with a parameter α.

According to this multivariate Gaussian error model of the native state ensemble, the parameters can be iteratively estimated by optimizing the following log-likelihood function:

$$l_h = l(R,t,M,\alpha|X,\Sigma) = l(R,t,M,\Sigma|X) + l(\alpha|\lambda) \quad (3)$$

This is called a hierarchical model, which combines the log-likelihood that comes directly from Equation (2) (first term) and the log-likelihood of an inverse Gamma distribution, where λ are

eigenvalues of Σ. X is the ensemble that contains n structures. The parameters are estimated iteratively, given X and an initial value of estimated covariance, $\hat{\sum} = I$ and $\alpha = 0$.

First, t, M, and R are estimated iteratively. It is shown that the translation t can be estimated as

$$\hat{t}_i = -\frac{X_i^T \sum^{-1} 1_k}{1_k^T \sum^{-1} 1_k} \quad (4)$$

Hat (^) denotes an estimated value. The rotation matrix R is derived by single value decomposition (SVD) of $M^T \sum^{-1} \left(X_i + 1_k \hat{t}_i^T\right)$. The mean structure M is computed as the average of the translated/rotated ensemble structures coordinates. Then, the covariance matrix can be estimated as usual:

$$\hat{\sum}_s = \frac{1}{3n} \sum_{i=1}^{n} \left(\left(X_i + 1_k \hat{t}_i^T\right) R_i - \hat{M}\right)\left(\left(X_i + 1_k \hat{t}_i^T\right) R_i - \hat{M}\right)^T \quad (5)$$

Here $\hat{\sum}_s$ denotes the sample covariance matrix that is computed for the first term in the right hand side of Equation (3). Next, α of the inverse Gamma distribution and the eigenvalues λ are estimated iteratively by solving

$$\hat{\sum}_h = \frac{3n}{3n+3}\left(\frac{2\alpha}{3n} I + \hat{\sum}_s\right) \quad (6)$$

and

$$\hat{\alpha} = \frac{k}{2\left(mE\left(\lambda_{sm}^{-1} | \alpha, \gamma, c\right) + \sum_{i=1}^{k-m} \lambda_i^{-1}\right)} \quad (7)$$

where $\hat{\sum}_h$ is the estimated covariance matrix of the hierarchical model, the left side of Equation (3), m is the number of missing eigenvalues, $E\left(\lambda_{sm}^{-1} | \alpha, \gamma, c\right)$ is the expected value of the inverse of the m smallest missing eigenvalues conditional on the smallest observed eigenvalue c, γ is the shape parameter of the inverse Gamma distribution, which is set to 0.5. Once α and $\hat{\sum}_h$ are updated, t, M, and R are again updated and iterations continue until convergence.

2.2 FlexScore

Given the estimated covariance matrix $\hat{\sum}_h$ and the ensemble mean $\hat{M}$, it is possible to superimpose a protein structure model to the ensemble, which was not primarily included in the original ensemble. Consider a computational protein structure model Y. Y contains the amino acid sequence identical to the reference protein structures, X. Using $\hat{\sum}_h$ and $\hat{M}$ the translation vector and the rotation matrix for Y can be computed by applying Equation (4) for the translation vector and by SVD of $M^T \hat{\sum}_h^{-1} \left(Y + 1_k \hat{t}_i^T\right)$ for the rotation matrix.

After translation and rotation of Y, we define FlexScore, FS, as

$$FS(Y) = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{\sqrt{\lambda_i}} \| \hat{M}_i - Y_i^{\text{sup}} \| \quad (8)$$

where λ_i is the eigenvalue of the covariance matrix $\hat{\sum}_h$, thus variance of Cα atom i and Y_i^{sup} is the position of Cα atom i after superimposition. FS has a simple interpretation: how far, on average, residue positions of the computational model are to the native state ensemble in standard deviations units. The best score of 0 is obtained for a model if the model has the identical conformation as the mean. As a model deviates from the mean structure, the score will be larger.

2.3 Protein native-state ensembles

An ensemble of native structures of a target protein was constructed by a MD simulation. A 10-ns-long MD simulation of the native state of a target protein was performed with bound ligands and ions with explicit water representation using a structure deposited in the Protein Data Bank (PDB). 1 MD run is sufficient because as shown in Supplementary Figure S1, flexibilities observed in independent MD trajectories are usually consistent with each other. Selenium atoms in selenomethionine residues were replaced with Sulphur. NMR-solved protein structures were simulated using the first model in the PDB files. The force field used for simulations was AMBER-99SB* with a NVT system. For details of the simulations, see a Supplementary Material in a paper by Hospital *et al.* (2012). From a simulation trajectory, we extracted structure at each 10 picoseconds to form an ensemble of structures.

2.4 Computational protein models

We chose 10 CASP 10 and 18 CASP11 targets as follows, which meet the following criteria for applying and evaluating FS: (i) monomers; (ii) no missing residues in the middle of chains in the PDB file; (iii) does not contain ligands for which force-field are not developed yet. For these targets, we analyzed server models with a complete chain. The number of models for each target ranged from 201 to 257 (average: 224.9). For the CASP11 targets, 10 'template-based modeling' and 8 'free-modeling' where chosen. The number of models for CASP11 targets ranged from 163 to 191 (average 175.5).

2.5 Conventional scores

FS was compared with RMSD (Equation (1)) and two widely used model evaluation scores, GDT-TS (Zemla, 2003) and TM-Score (Zhang and Skolnick, 2005). GDT-TS is the average of fraction of residues in a model that are predicted within 1, 2, 4, and 8 Å after superimposing the model to a reference structure of the protein. TM-Score is also a fraction of residues in a model that are closer than a heuristic cutoff value to corresponding residue positions after superimposing the model to a reference structure. Thus, both scores range from 0 to 1, with 1 as the best score.

3 Results

3.1 FS for two example structures

To illustrate characteristics of FS, we made two structure models for the C-terminal domain (residue 529–577) of human CSTF-64 protein (PDB ID: 2j8p). From the first structure model of this protein in its PDB file (this protein was solved by NMR), two models were built by manual modification, which have identical RMSD, GDT-TS, and TM-Score values (1.47, 0.95, and 0.93, respectively) between each other. The first model (shown in green in Fig. 1A) was modified at a helical region (residues 550–554), while the second model (shown in blue) was modified at the C-terminal fragment (residue 573–577). Referring to the structure ensemble of this protein (all NMR models in the PDB file) shown in gray in the figure, the second model (blue) would be better than the first one in green because the deviation of the blue model at the C-terminus is still in the range of observed structure variation. In contrast to the existing scores that give the same quality evaluation to the two models, FS gave clear distinction between these models: the first model had FS = 1.96, and the second model had FS = 1.42. Figure 1B shows the variance of the NMR ensemble, and the deviation of the two models from the mean structure, and FS computed at each residue. It is clear that the deviated helical region in the first model was appropriately

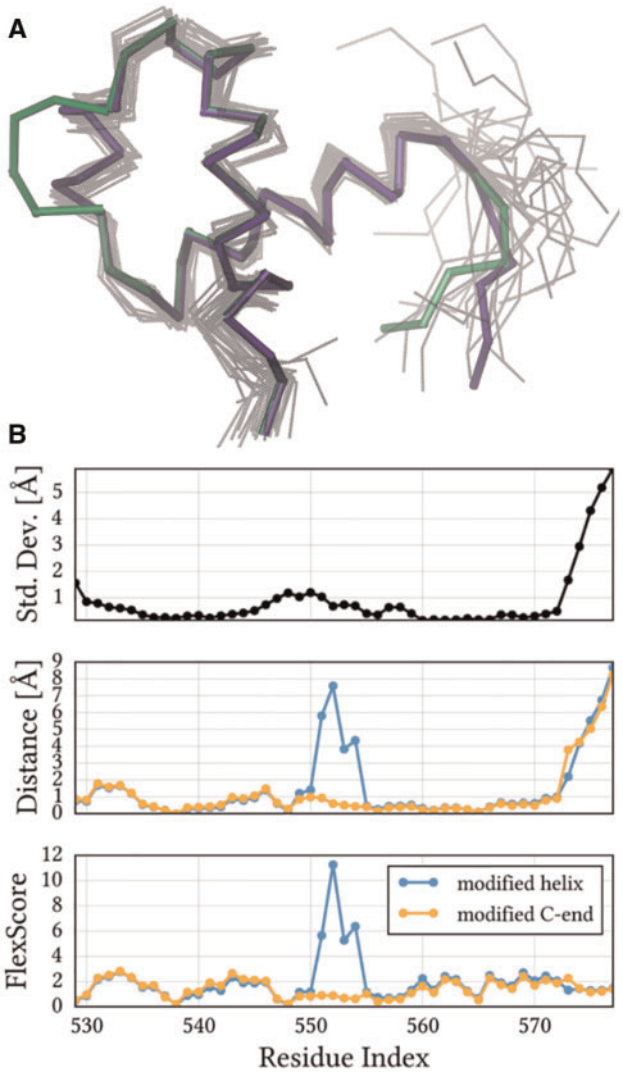

Fig. 1. Two artificially modified models showing fluctuation-dependence of FS. (**A**) Structure superimposition of native-ensemble of the protein, 2j8p, determined by NMR (gray) and two models (blue and green). Both models have identical RMSD, GDT-TS and TM-Score (1.47, 0.95, and 0.93, respectively) but distinct FS (green model: 1.96, blue: 1.42). (**B**) Structural variance of the NMR ensemble, the distance of residues in the two models to the mean structure of the NMR ensemble, and the FS of each residue of the models

penalized with a large value, while the C-terminal region of the second model is not.

3.2 FS for structure models

Next, we computed FS for structure models of the 28 CASP targets and compared the scores with RMSD, GDT-TS, and TM-Score. Table 1 summarizes correlation of the FS and the other three scores.

Overall, FS correlated well for most of the cases with GDT-TS, TM-Score, and RMSD. Because the smaller the better for FS, it has positive correlation with RMSD while negative with GDT-TS and TM-Score. However, there are exceptions, T0651, T780, T0808, T0814, and T0853, where FS has a strong correlation to RMSD while essentially no correlation with GDT-TS and TM-Score. In these cases, the quality of all the models is low as can be seen in the average score values in Table 1. Some models have long unphysical, totally stretched regions, which made their RMSD large. These bad regions of a model were severely penalized in RMSD and FS, which

Table 1. Correlation of FS with the other scores

Target	GDT-TS	TM-SCORE	RMSD	<GDT-TS>	<TM>	<RMSD>	<FS>
T0651*	−0.04	−0.13	1.00	0.27	0.36	24.02	62.76
T0655	−0.83	−0.88	0.77	0.49	0.58	13.95	15.41
T0657	−0.94	−0.95	0.92	0.63	0.68	7.69	9.64
T0662	−0.97	−0.96	0.99	0.67	0.67	3.87	5.24
T0667	−0.96	−0.98	0.98	0.57	0.69	6.73	13.34
T0669	−0.83	−0.84	0.96	0.46	0.50	9.21	16.70
T0673	−0.65	−0.58	0.95	0.33	0.27	11.85	22.87
T0675	−0.62	−0.56	0.74	0.37	0.33	11.14	6.96
T0714	−0.91	−0.92	0.98	0.78	0.79	2.67	5.24
T0716	−0.82	−0.79	0.88	0.65	0.62	7.55	5.62
T0763*	−0.30	−0.48	0.99	0.16	0.20	18.18	54.71
T0767*	−0.48	−0.69	1.00	0.11	0.19	33.84	94.69
T0769	−0.88	−0.87	0.80	0.50	0.53	11.58	13.22
T0773	−0.91	−0.89	0.85	0.52	0.49	9.45	12.04
T0777*	−0.63	−0.72	1.00	0.10	0.21	31.60	81.96
T0780	0.08	0.03	0.99	0.29	0.37	23.13	32.47
T0782	−0.88	−0.89	0.99	0.45	0.49	9.20	17.83
T0785*	−0.54	−0.59	0.97	0.18	0.20	16.40	37.16
T0790*	−0.28	−0.57	1.00	0.11	0.19	26.15	50.85
T0803	−0.27	−0.30	0.98	0.34	0.39	13.84	35.47
T0808*	−0.02	−0.15	0.99	0.11	0.21	26.47	70.98
T0814*	0.10	−0.43	0.98	0.10	0.19	27.14	75.96
T0829	−0.78	−0.72	0.95	0.47	0.42	9.63	22.38
T0832*	−0.41	−0.64	0.97	0.15	0.22	20.65	51.35
T0833	−0.94	−0.95	0.96	0.57	0.60	7.50	11.78
T0853	−0.27	−0.32	0.99	0.21	0.26	17.55	36.25
T0856	−0.89	−0.92	0.99	0.69	0.77	4.01	10.81
T0857	−0.89	−0.90	0.95	0.29	0.31	13.96	13.27

The four columns on the right side show the average score of the target proteins. A total of 18 CASP11 targets are shown in the bottom half. Stars (*) indicate free-modeling targets.

made their correlation high with each other but the correlation to GDT-TS and TM-Score low, because such regions were neglected by these two scores.

Table 1 shows that RMSD has the largest correlation among the three scores with FS. However, RMSD and FS have substantially lower correlation for T0655, T0675, T0716, T0769, and T0773 (Fig. 2A). This is owing to the flexible nature of these target proteins, which was shown in Figure 2B. Compared with a rigid target structure T0714 shown in Figure 2B for which FS has a correlation coefficient of 0.98 with RMSD, these three targets have one or two flexible regions. Because RMSD optimizes non-weighted (homoscedastic) mean deviation in comparing a model and a reference structure, it results in overestimating error for such flexible proteins, which makes a lower correlation with FS. Note that as shown in Supplementary Table S1, overall correlation between GDT-TS and TM-Score is slightly higher than that of between FS and RMSD. Correlation between the scores for invididual targets are provided in Supplementary Figure S2, where differences between FS and RMSD are observed in many targets.

We have also examined score correlation separately for template-based (TB) targets and free modeling (FM) targets (* in Table 1). Naturally, the quality of computational models are higher for TB targets (i.e. higher GDT-TS, TM-Score, and lower RMSD values) than FM targets, because the latter are more difficult to model. For the TB targets, FS showed high correlation to RMSD and had moderate correlation to GDT-TS and TM-Score (0.488 and 0.513,

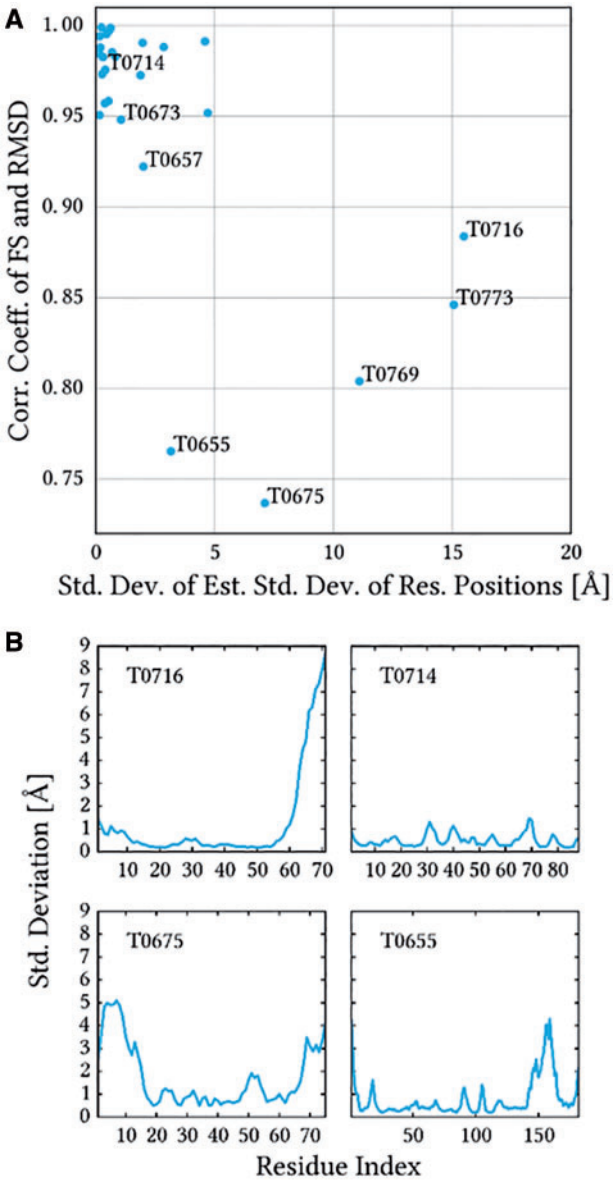

Fig. 2. Correlation of FS and RMSD. (**A**) Correlation coefficient between FS and RMSD relative to the average standard deviation of Cα atoms of ensemble structures. (**B**) The flexibility of four target proteins, T0716, T0714, T0675 and T0655. The standard deviation of each Cα atoms is shown. T0714 is an example of rigid structures for comparison

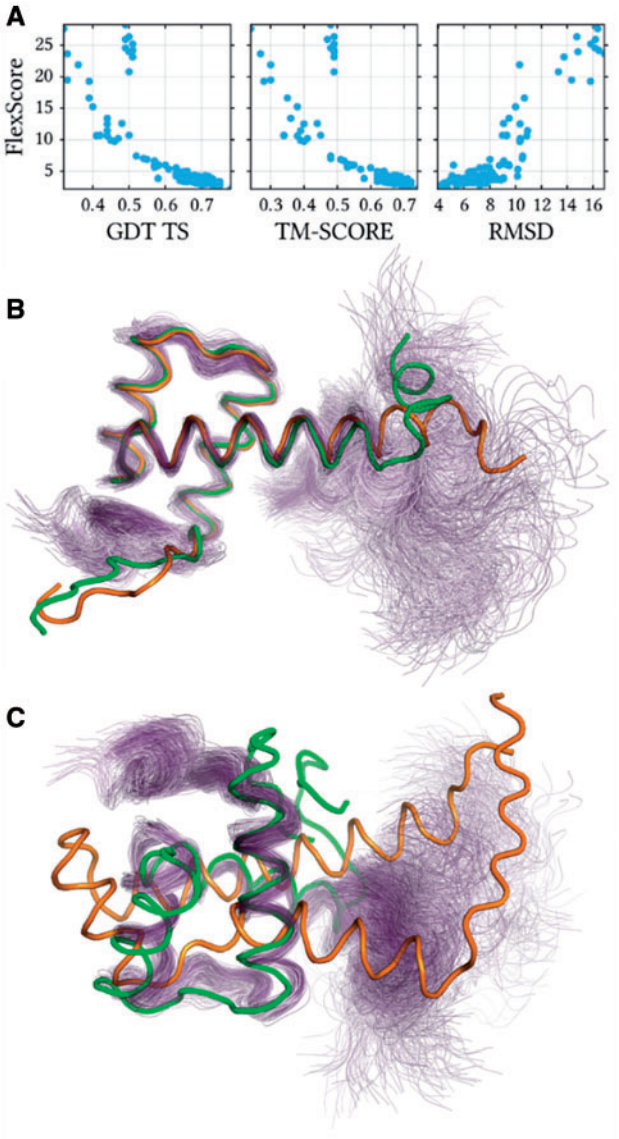

Fig. 3. Score comparison of models for T0716. (**A**) Score correlation of FS with GDT-TS, TM-Score and RMSD (left to right). The correlation coefficient between the two score was −0.82, −0.79, and 0.88, respectively. (**B**) Superimposition of two models (green and orange) for T0716 onto MD-derived structural ensemble (thin lines). Both models have a similar FS (2.75 and 2.71), GDT TS (0.75 and 0.73) and TM-Score (0.72 and 0.70), but distinguishable RMSD (3.93 Å and 5.40 Å, respectively). (**C**) Superimposition of another pair of models for T0716. They have similar GDT-TS (0.52 and 0.51 for green and orange) and TM-Score (0.48 and 0.49), but have substantially different FS of 7.4 and 24.1, respectively

respectively). For FM targets, the correlation to the GDT-TS and TM-Score dropped substantially to a low level (0.143 and 0.219) while the correlation to the RMSD became even higher for FM targets (0.93 and 0.99, respectively).

We now discuss some examples of models that have inconsistent relative scores by FS with RMSD, GDT-TS or TM-Score. The first examples are models of T0716 (Fig. 3). For this target, FS has overall sufficient correlation to the other three scores (Fig. 3A), 0.88 with RMSD, and −0.82 and −0.79 with GDT-TS and TM-Score. However, there are notable differences that deserve attention. There are models with similar FS between 2.5 and 4 but have diverse RMSD values that range between 4 and 8 Å (the right panel in Fig. 3A). Figure 3B shows superimposition of two such models relative to the ensemble structures generated by MD. These two models in green and orange have largely different RMSDs of 3.93 Å and 5.40 Å while with similar FS s of 2.75 and 2.71, respectively. As Figure 3B shows, the target protein is flexible in its N- and C-terminus. Considering the flexible regions of the protein, the quality of both models are essentially the same, because the core region of the two models is modeled correctly and the tail region is within the ensemble.

The second pair of models (Fig. 3C) has a similar GDT-TS score of 0.52 (green) and 0.51 (orange) but has different FS of 7.4 and 24.1, respectively (the left panel, Fig. 3A). As the figure shows, although the two models share common structures with the crystal structure at the middle part of the protein (the left side in the figure), the green model has a better agreement of the topology to the structure ensemble. The orange model has long-stretched helices in both

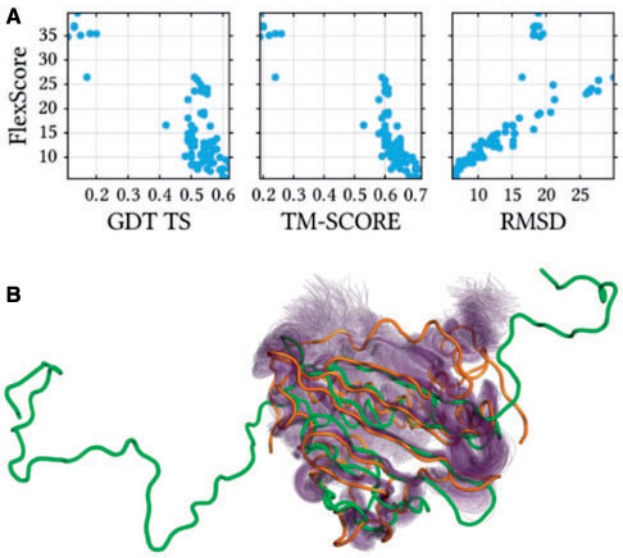

Fig. 4. Score comparison of models for T0655. (**A**) Score correlation of FS with GDT-TS, TM-Score and RMSD (left to right). The correlation coefficient between the two score was −0.83, −0.88 and 0.77, respectively. (**B**) Superimposition of two models (green and orange) for T0655 onto MD-derived structural ensemble (thin lines). Both models have a similar GDT-TS score (0.50 and 0.54), and TM-Score (0.61 and 0.66), but they have different FS (23.05 and 9.2, respectively)

chain termini, which largely disagree with the reference structure ensemble. Reflecting these characteristics of the two models, FS for the green model is substantially better than the orange model (7.4 for the chain in green and 24.1 for the chain in orange).

The next examples are for the target T0655 (Fig. 4). Looking at the score correlation in Figure 4A, there are a cluster of models that have a GDT-TS score around 0.5 or TM-Score of 0.6. FS, in contrast, distinguishes the quality of these models with a diverse range of score values. Figure 4B shows such examples. The two models in Figure 4B capture correct β-class fold in the middle of the protein. The green model has totally unstructured tails that are far off from the structure ensemble of its reference structure, which is reflected in a substantially worse FS of 23.05 than the orange one (9.2). This quality difference can be also detected by RMSD (25.93 Å and 7.28 Å, respectively, for the green and the orange model); however, the structure alignments computed for the green model by the FS and the RMSD computation are different (Supplementary Fig. S3). RMSD of the core region of the green model (residue 22–142) by FS computation was 4.01 Å, while the RMSD superimposition gave 17.22 Å. Thus, considering the ensemble helped making better alignments in the FS case.

The last example is from the target T0714 (Fig. 5). These two models have similar GDT-TS (0.84 and 0.83 for the green and the orange model, respectively), and similar TM-Score (0.83 and 0.86, respectively). We also computed GDT-HA score for these models, which was 0.64 and 0.61, respectively. In contrast, FS indicates that the orange model, which has a score of 2.69 has a better quality than the green one, which has a score of 4.42. This is a reasonable evaluation considering the larger number of incorrect regions in the green model that are off from the ensemble structures (indicated with blue circles in the figure).

3.3 FS using NMR ensemble

In Figure 6, we compared FS computed using a MD-generated ensemble and an NMR-derived ensemble. The two scores are consistent, particularly for models with higher accuracy (FS < 10), with an overall correlation coefficient of 0.994.

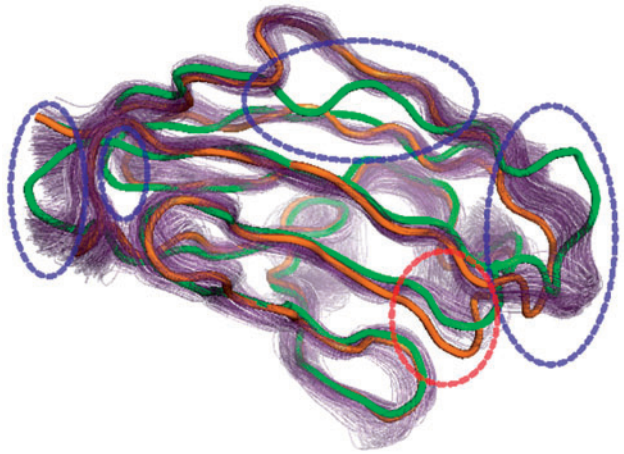

Fig. 5. Superimposition of two models of T0714 on to MD-derived structure ensemble (thin lines). The model in green and orange have similar GDT-TS of 0.84 and 0.83, and TM-Score of 0.83 and 0.86, respectively. FS of the two models are 4.42 and 2.69, for the green and the orange models, respectively

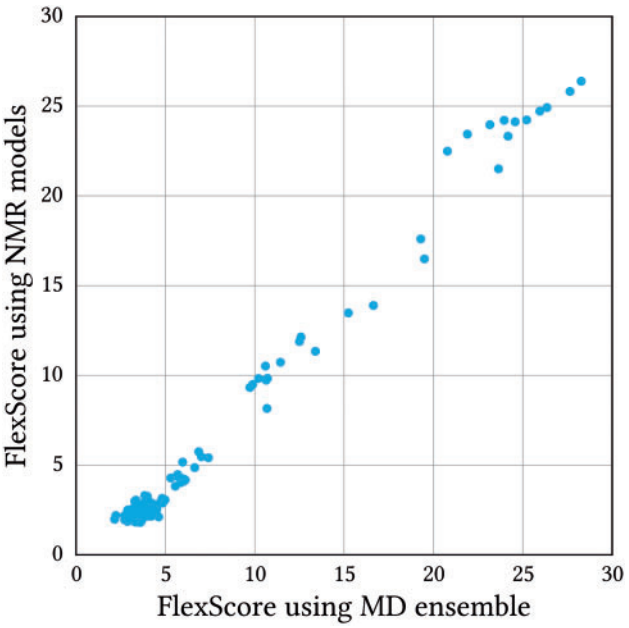

Fig. 6. Comparison of FS using a MD-generated ensemble and an NMR-derived ensemble for a protein solved by NMR (T0176). Scores of 235 models were plotted

3.4 Ranking prediction groups in CASP models

In the last section, we examine how performance of prediction methods are ranked among peers with FS. For this experiment, we used the 10 CASP10 targets selected in Section 2.4. There were 68 groups who submitted server models to at least one of these 10 targets in CASP10, who were subject to the analysis. In CASP, a group can submit up to five models for a target, but here only the first models (TS1 model) were evaluated. Ranking of the groups with FS was compared with those by three other scores, GDT-TS, TM-Score, and RMSD.

Table 2. Ranking of prediction groups in CASP10 with different scores

Rank	FS	FS-GDT	GDT-TS	TM	RMSD
1	A	A	A	A	A
2	B	D	B	B	B
3	C	B	F	C	C
4	D	C	C	F	F
5	E	F	D	D	E
6	F	E	I	I	G
7	G	O (14)	G	X (24)	J
8	H	J	J	L (12)	I
9	I	Q (17)	E	G	D
10	J	H	O (14)	Q (17)	H

FS = FlexScore; FS-GDT = FlexScore-GDT; TM = TM-Score. The alphabets denote the group IDs, which were assigned based on the ranking by FlexScore. If a group did not appear within top 10 by FlexScore, its ranking by FlexScore is shown in a parenthesis.

Table 3. Correlation of group ranking in CASP10 by the scores

	FS	FS-GDT	GDT-TS	TM	RMSD
A. Spearman's correlation coefficients or the rankings					
FS	–	0.931	0.930	0.913	0.940
FS-GDT	–	–	0.935	0.912	0.806
GDT-TS	–	–	–	0.984	0.901
TM	–	–	–	–	0.918
RMSD	–	–	–	–	–
B. Pearson's correlation coefficients of Sum of the Z-scores of groups					
FS	–	−0.929	−0.962	−0.963	0.976
FS-GDT	–	–	0.972	0.960	−0.898
GDT-TS	–	–	–	0.994	−0.943
TM	–	–	–	–	−0.952
RMSD	–	–	–	–	–

For this comparison, we devised another score named FS-GDT based on FS, which follows the concept of GDT-TS. Similar to the GDT-TS that counts the fraction of Cα positions that fall within 1, 2, 4 and 8 Å to the corresponding positions in the reference structure after superimposition, FS-GDT computes the average fraction of residues (Cα positions) that have FS within 1, 2, 4, and 8.

Ranking of the groups was determined by the accumulated Z-score in the same way as performed for the official CASP rankings: (i) For a target, Z-scores from a raw score were computed for all TS1 models. (ii) Then, bad models with a Z-score of −2.0 or lower were removed as outliers. (iii) Z-scores were recalculated without the outliers; (iv) and finally, each group accumulates the Z-score from the previous step from each target. Table 2 shows the group ranking results up to the top 10 groups. Because the purpose of ranking groups is not to decide who did well and who did not but to examine similarity of ranking by different scores, the group identities are denoted with alphabets.

Overall the group ranking by FS and FS-GDT correlated well with GDT-TS, TM-Score, and RMSD. In Table 2, Group A was consistently ranked the top by all the scores and every pair of scores shares at least four groups among the top 5 ranked groups by the scores. The Spearman's correlation coefficients between rankings by FS and FS-GDT against the other three existing scores are all high, all over 0.9 except for the FS-GDT and RMSD pair (0.806) (Table 3A). Correlation between FS and FS-GDT with the other scores is also high when the sum of the Z-scores used to rank the groups were compared (Table 3B). Thus, while FS provides alternative evaluation to structure models by reasonably considering protein flexibility (Section 3.1 and 3.2), evaluation is close and consistent with the other existing scores when it comes to ranking of groups.

Equivalent data to Tables 2 and 3 for the CASP11 targets are provided in Supplementary Material (Supplementary Tables S3 and S4), which consistently show strong correlation between group rankings by different scores.

4 Discussion

We proposed a new score named FS and its variant, FS-GDT, which evaluates the quality of computational protein structure models by considering flexibility of protein chains. FS effectively distinguishes deviations of a model from a reference structure at intrinsically flexible and rigid regions, and assigns more permissive scores to the former than the latter. This is reasonable from the biophysics perspective of protein structures.

The flexibility of a protein was measured from a MD simulation of 10 ns because it was long enough to observe large flexibility in protein terminal regions and to highlight incorrect regions of structure models that are beyond the range of chain flexibility. For a CASP11 model, T0733, we extended the MD run to 100 ns, but FS did not show meaningful change (Supplementary Table S2 and Fig. S4). The framework provided here is valid in principle with any data of protein flexibility, and FS would provide reasonable evaluation to these models under the provided flexibility information.

We point out that MD simulations are easier to perform than it used to be, for example by using web-based MD tools (Hospital *et al.*, 2012; Lee *et al.*, 2016). Previously, we developed a method named FlexPred, which predicts absolute values of residue fluctuation from a tertiary structure (i.e. a PDB file) of a target protein (Jamroz *et al.*, 2012), which can be also used to obtain flexibility.

At this juncture, it would be of interest to discuss difference of using MD simulations and using B-factors as the source of fluctuation. As mentioned in Introduction, the work by Wu & Wu introduced a B-factor weighted RMSD (2010). First of all, the level of fluctuation considered are different between them. B-factors indicate discrepancy of a solved structure to the X-ray diffraction pattern of the protein. Thus, although the B-factors were shown to have correlation to residue flexibility observed in computational simulation (Haliloglu and Bahar, 1998), it is flexibility in the crystal condition of proteins. It was shown that B-factors are appropriate measure of fluctuations for stable parts of proteins, but significantly underestimate motion in flexible regions (Eastman *et al.*, 1999). In contrast, MD used in FS is aimed at considering flexibility of proteins in solution, which is a more natural environment for proteins. From a technical stand point, to use B-factors for flexibility value, a weight parameter needs to be arbitrarily selected because B-factors do not provide absolute values of flexibility (Wu and Wu, 2010). Also in FS, flexibility is used more explicitly in the structure superimposition step than the B-factor weighted RMSD.

The results in the prediction group ranking shows that FS is in reasonable agreement with the other existing scores. In recent rounds of the CASP experiments, computational models are evaluated with a combination of several scoring terms that examine different structural aspects of models (Kim and Kihara, 2015; Moult *et al.*, 2014). Combined with other scores, FS will be able to provide a new viewpoint to the protein structure evaluation. The technique

of considering ensemble structures will be also useful for various other related problems including multiple protein structure alignment, flexible protein–protein or protein–ligand docking, flexible structure fitting to electron microscopy density maps.

Acknowledgements

The authors are grateful to Joshua McGraw for proofreading the manuscript.

Funding

This work was partially supported by the National Institutes of Health (R01GM097528) and the National Science Foundation (IIS1319551, DBI1262189, IOS1127027).

Conflict of Interest: none declared.

References

Andrec,M. *et al.* (2007) A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. *Proteins*, **69**, 449–465.

Betts,M.J. and Sternberg,M.J. (1999) An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng.*, **12**, 271–283.

Brüschweiler,R. (2003) Efficient RMSD measures for the comparison of two molecular ensembles. Root-mean-square deviation. *Proteins*, **50**, 26–34.

Damm,K.L. and Carlson,H.A. (2006) Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys. J.*, **90**, 4558–4573.

DePristo,M.A. *et al.* (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure*, **12**, 831–838.

Eastman,P. *et al.* (1999) Protein flexibility in solution and in crystals. *J. Chem. Phys.*, **110**, 10141–10152.

Fenwick,R.B. *et al.* (2011) Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles. *Eur. Biophys. J.*, **40**, 1339–1355.

Fetrow,J.S. and Skolnick,J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.*, **281**, 949–968.

Furnham,N. *et al.* (2006) Is one solution good enough? *Nat. Struct. Mol. Biol.*, **13**, 184–185; discussion 185.

Gapsys,V. and de Groot,B.L. (2013) Optimal superpositioning of flexible molecule ensembles. *Biophys. J.*, **104**, 196–207.

Garbuzynskiy,S.O. *et al.* (2005) Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? *Proteins*, **60**, 139–147.

Haliloglu,T. and Bahar,I. (1998) Coarse-grained simulations of conformational dynamics of proteins: application to apomyoglobin. *Proteins*, **31**, 271–281.

Hasegawa,H. and Holm,L. (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, **19**, 341–348.

Hirsch,M. and Habeck,M. (2008) Mixture models for protein structure ensembles. *Bioinformatics*, **24**, 2184–2192.

Hospital,A. *et al.* (2012) MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. *Bioinformatics*, **28**, 1278–1279.

Jamroz,M. *et al.* (2012) Structural features that predict real-value fluctuations of globular proteins. *Proteins*, **80**, 1425–1435.

Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, **A34**, 827.

Kim,H. and Kihara,D. (2015) Protein structure prediction using residue- and fragment-environment potentials in CASP11. *Proteins*, in press.

Kosloff,M. and Kolodny,R. (2008) Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins*, **71**, 891–902.

Kuzmanic,A. *et al.* (2011) Dynamics may significantly influence the estimation of interatomic distances in biomolecular X-ray structures. *J. Mol. Biol.*, **411**, 286–297.

Lee,J. *et al.* (2016) CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput.*, **12**, 405–413.

Lindorff-Larsen,K. and Ferkinghoff-Borg,J. (2009) Similarity measures for protein ensembles. *PloS One*, **4**, e4203.

Mechelke,M. and Habeck,M. (2010) Robust probabilistic superposition and comparison of protein structures. *BMC Bioinformatics*, **11**, 363.

Moll,M. *et al.* (2010) The LabelHash algorithm for substructure matching. *BMC Bioinformatics*, **11**, 555.

Moult,J. *et al.* (2014) Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins*, **82**(**Suppl 2**), 1–6.

Olsson,S. *et al.* (2014) Probabilistic determination of native state ensembles of proteins. *J. Chem. Theory Comput.*, **10**, 3484–3491.

Perez,A. *et al.* (2012) FlexE: Using elastic network models to compare models of protein structure. *J. Chem. Theory Comput.*, **8**, 3985–3991.

Popovych,N. *et al.* (2006) Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.*, **13**, 831–838.

Rother,D. *et al.* (2008) Statistical characterization of protein ensembles. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 42–55.

Sael,L. and Kihara,D. (2012) Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins*, **80**, 1177–1195.

Shatsky,M. *et al.* (2002) Flexible protein alignment and hinge detection. *Proteins*, **48**, 242–256.

Theobald,D.L. and Wuttke,D.S. (2006) THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*, **22**, 2171–2172.

Tzeng,S.R. and Kalodimos,C.G. (2012) Protein activity regulation by conformational entropy. *Nature*, **488**, 236–240.

Wu,D. and Wu,Z. (2010) Superimposition of protein structures with dynamically weighted RMSD. *J. Mol. Model.*, **16**, 211–222.

Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19**(**Suppl 2**), ii246–ii255.

Zemla,A. (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302.