Published in final edited form as:
Proteins. 2010 December ; 78(16): 3353–3362. doi:10.1002/prot.22842.

Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction

Tuo Zhang[1,2], **Eshel Faraggi**[1,2], and **Yaoqi Zhou**[1,2,*]

[1] School of Informatics, Indiana University Purdue University, Indianapolis, IN 46202

[2] Center for computational Biology and Bioinformatics, Indiana University School of Medicine, 719 Indiana Ave., Walker Plaza Building Suite 319, Indianapolis, IN 46202, USA

Abstract

Protein molecules exhibit varying degrees of flexibility throughout their three-dimensional structures. Protein structural flexibility is often characterized by fluctuations in the Cartesian coordinate space. On the other hand, the protein backbone can be mostly defined by two torsion angles φ and ψ only. We introduce a new flexibility descriptor, backbone torsion-angle fluctuation derived from the variation of backbone torsion angles from different NMR models. The torsion angle fluctuations correlate with mean-squared spatial fluctuations derived from the same collection of NMR models. We developed a neural-network based real-value predictor based on sequence information only. The predictor achieved ten-fold cross-validated correlation coefficients of 0.59 and 0.60, and mean absolute errors of 22.7° and 24.3° for the angle fluctuation of φ and ψ, respectively. This predictor is expected to be useful for function prediction and protein structure prediction when predicted torsion angles are used as restraints. Both sequence- and structure-based prediction of torsion-angle fluctuation will be available at http://sparks.informatics.iupui.edu within the SPINE-X package.

Proteins are dynamic and flexible macromolecules, which undergo constant thermal fluctuations and other types of dynamic and functional motions[1]. The structural flexibility that enables these motions is responsible for various biological activities, including molecular recognition, catalytic activity, allosteric regulation, antigen-antibody interactions, and protein-DNA interactions[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Many protein functions result from the flexible motion of the protein backbone.

Protein backbone flexibility can be measured by many different methods. For example, the temperature B-factor from X-ray structure determination reflects the degree of thermal motion and static disorder in a protein crystal structure. Given a protein structure, molecular dynamics simulations can provide the trajectories of protein motions. Here, we will describe backbone flexibility by the fluctuation of the backbone torsion angles because only two torsion angles are required for a nearly complete description of the backbone. Moreover, many functional motions result from significant change in the torsion angles of only a few amino-acid residues [13, 14, 15]. That is, potentially functional regions of proteins can be indicated by large torsion angle fluctuation. Furthermore, the conformational flexibilities of proteins described by normal modes can be better described in torsion angle space[16].

Our interest in torsion angle fluctuation is further enhanced because real-value prediction of torsion angles is considerably more useful than predicted three-state secondary structure as restraints for *ab initio* protein tertiary-structure prediction [17]. The former doubles the

[*] To whom correspondence should be addressed: Phone: (317) 278-7674, Fax: (317) 278-9201, yqzhou@iupui.edu.

success rate in sampling near native conformations within top ranked structures. Thus, if fluctuation of torsion angles can be predicted with reasonable accuracy, it will be useful for improving torsion angle restraints by providing allowable ranges of predicted angles and thus, have the potential to greatly enhance the efficiency of conformation sampling for protein structure prediction. This is a challenging problem in structural biology with little progress in *ab initio* template-free prediction in recent years[18].

Torsion angle fluctuation, however, cannot be obtained from the structures determined by the X-ray crystallographic method because it produces only one structure and measured temperature B-factors do not correlate strongly with fluctuation of torsion angles, as will be shown below. Here, we will estimate torsion angle fluctuation from angle variations in structural models determined by Nuclear Magnetic Resonance (NMR). The reason is that NMR-determined structures are typically made of an ensemble of model structures all of which are compatible with NOE restraints obtained from NMR experiments. The variations of those model structures are due in part to intrinsic fluctuations of proteins in solution [19,20].

We study the relationship between flexibility described by backbone torsion-angle fluctuation along the protein chain and the underlying physical characteristics such as secondary structure and solvent exposure. We will also establish a predictor for angle fluctuation with sequence information only. To our knowledge, this is the first method for sequence-based prediction of the fluctuation of backbone torsion angles. The sequence-based prediction of torsion angle fluctuation is motivated by the need for locating flexible (potentially functional) regions of a protein whose structure is unknown because a majority of proteins have unknown structures. Furthermore, it can assist protein structure prediction with predicted torsion angles and angle flexibility as restraints[17].

Our real-value prediction method is a two-layer neural network with guided learning technique developed previously by us for real-value prediction of backbone torsion angles (Real-SPINE,[21, 22, 17, 23]). Using a database of 997 non-redundant NMR structures, we achieve ten-fold cross-validated Pearson correlation coefficients (CC) of 0.598 and 0.602, and mean absolute errors (MAE) of 0.126 and 0.135 (22.7° and 24.3°, if we transform them back to real angles), for the torsion-angle fluctuations of φ and ψ angles, respectively. This predictor provides a new tool that will likely be useful for protein structure and function prediction.

1 Methods

1.1 Definitions

The torsion angle fluctuation, $\Delta\tau$, for a protein of length n is defined as the average difference of torsion angles ($\tau = \varphi$ or ψ) among different NMR models.

$$\Delta\tau_k=\frac{C_m}{m(m-1)/2}\sum_{i<j}\Delta(\tau_k^i,\tau_k^j) \tag{1}$$

with

$$\Delta(\tau_k^i,\tau_k^j)=\begin{cases} |\tau_k^i-\tau_k^j|/180, & if|\tau_k^i-\tau_k^j|<180,\\ 2-|\tau_k^i-\tau_k^j|/180, & Otherwise,\end{cases} \tag{2}$$

where $k = 1, 2, \cdots, n$ represents the k^{th} residue in the given structure, τ_k^i denotes the torsion angle (φ or ψ) of the k^{th} residue in the i^{th} model (i or $j = 1, 2, \cdots, m$ for a total of m NMR models), $\Delta(\tau_k^i, \tau_k^j)$ represents the normalized absolute minimum distance between angle τ_k^i and angle τ_k^j. We found that it is necessary to add an m-dependent factor C_m to ensure that range of $\Delta\tau_k$ is independent of m, because angle fluctuations calculated from different number of NMR models (m) have different upper limits as a result of angle periodicity. For example, the highest possible angle fluctuation is 180° for 2 NMR models, 120° for 3 NMR models, and 108° for 5 NMR models. In general,

$$C_m = \begin{cases} 2(m-1)/m, & if\, m\, is\, even, \\ 2m/(m+1), & if\, m\, is\, odd. \end{cases} \qquad (3)$$

ensures the maximum possible angle fluctuation is 1 from Eq. (1) regardless of the number of NMR models.

1.2 Dataset

The initial structural dataset was obtained from the precompiled CulledPDB lists by PISCES[24], which was generated on November 12, 2009 with a sequence identity threshold of 25%, including the structures from all experimental methods. It has 8027 protein chains of which 1268 chains were determined by NMR. The number of NMR structures was further reduced by removing the chains: 1) with less than 5 NMR models; 2) with chain size less than 25 amino acid residues; and 3) containing non-standard amino acid types. The final NMR dataset includes 997 chains (referred to as NMR997).

We also employed a dataset[25] that contains 60 protein chains with both NMR- and X-ray-resolved structures without significant structural differences. Similar to the NMR997 dataset, we removed 12 chains with less than five NMR models and obtained a subset of 48 chains, referred to as NX48. This NX48 dataset was employed to study the relation between torsion-angle fluctuation, B-factor, and solvent accessibility.

The torsion angles, i.e., φ and ψ, were calculated by the DSSP program[26]. Four models from two chains (model 3, 4, 37 in 1OV2A; model 11 in 1LPVA) were removed since a few residues in those models did not contain the positions of necessary backbone atoms for torsion angle calculations.

1.3 Neural network

We employed a two hidden-layer neural network with a hyperbolic activation function and guided learning technique developed for Real-SPINE 3.0 for real-value torsion angle prediction[22]. Unlike Real-SPINE 3.0, our neural network employs a smaller number of hidden neurons (51 rather than 101) and one additional bias. This is because we found no improvement with a larger number of neurons. The back-propagation algorithm with momentum is applied to optimize the weights[27]. The learning rate and momentum are set to 0.001 and 0.4, respectively. To reduce random prediction errors caused by the randomly selected initial weights, we trained five independent predictors and the final prediction is based on their average.

1.4 Input for networks

A 34-dimensional vector was designed to characterize each residue. Its components include the 20-dimensional PSSM vector derived from the PSI-BLAST profiles[28] by searching a given sequence with three-iterations against the NCBI's non-redundant protein sequence database. Each dimension in the PSSM vector was divided by nine so that all values fell between −1.0 and 1.0. We further employed seven representative physical parameters, namely a steric parameter(graph shape index), hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability. Those parameters were identified by Meiler *et al* [29] and have proved helpful in protein secondary-structure prediction[21, 30]. In addition, we used predicted secondary structure (3 dimensions), solvent accessibility (1 dimension) and real-value torsion angles (2 dimensions), all from SPINE X[17]. The predicted secondary structure was encoded as a 3-dimensional probability vector, i.e., the probability of coil, strand and helix prediction. The predicted solvent accessibility was normalized by the solvent accessible surface area (ASA) of an extended conformation (Ala-X-Ala)[31, 32]. The two predicted torsion angles were normalized by 180°. Furthermore, we employed predictions of short disordered regions by IUpred[33], which yields a score between 0 and 1 for a current residue; scores above 0.5 indicate disorder. The majority of the dimensions in the input vector were in the range [−1, 1], for those that were not, we linearly transformed them between −1 and 1.

A sliding window, centered on the current residue, was introduced to include the information of its neighboring residues. The size of window was determined by optimization, as we will see later.

1.5 Training, test and evaluation

A ten-fold cross-validation test was performed on the NMR997 dataset. Specifically, we randomly divided the NMR997 dataset into ten subsets with roughly equal number of protein chains (7 with 100 chains and 3 with 99 chains). Each subset was in turn chosen as the testing set, while the remaining 9 subsets were merged to form the training set.

To measure the performance of torsion angle fluctuation predictions, we calculated the Pearson correlation coefficient and the mean absolute error between predicted and observed torsion-angle fluctuations, as given by

$$CC=\frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{[\sum_{i=1}^{N}(x_i-\overline{x})^2][\sum_{i=1}^{N}(y_i-\overline{y})^2]}} \tag{4}$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|x_i-y_i| \tag{5}$$

where x_i is the predicted torsion-angle fluctuation and y_i is the native torsion-angle fluctuation for the i^{th} residue in the sequence, and $\bar{x}$ and $\bar{y}$ are their corresponding sample means. CC is also used in measuring the correlation between torsion angle fluctuation, B-factor, and solvent accessibility. CC = 1 indicates that the variables are fully correlated, while CC = −1 means that the variables are fully anti-correlated. We note that the correlation can be calculated at the residue level[34] or at the chain level[35, 36]. In this paper, CC refers to correlation at the residue level, unless indicated otherwise.

2 Results

2.1 The distribution of torsion-angle fluctuation

The distribution of torsion-angle fluctuation is shown in Fig. 1. This distribution is based on 77421 residues in the NMR997 dataset with 50 bins. Similar distributions are observed for $\Delta\varphi$ and $\Delta\psi$ with small difference near 0 and 1. Both distributions are not uniform; the majority of residues (nearly 73% of residues) have angle fluctuations of no more than 0.2 and only about 15% of residues have large fluctuations (> 0.5). This reflects stable protein core structures with a small number of flexible residues.

2.2 Relationship between $\Delta\varphi$ and $\Delta\psi$

$\Delta\varphi$ and $\Delta\psi$ for the same residue represents the fluctuation of the neighboring rotational angles in the protein backbone. That is, chemical bond linkage will make them correlated because it is not possible to change one torsion angle without changing the other. Indeed, there is a significant correlation with the CC value of 0.75 as shown in Fig. 2A. The CC between $\Delta\varphi$ and $\Delta\psi$ is 0.66 if CC is calculated for each chain and then averaged over number of chains. The reduction of CC from residue-based to chain-based is probably due to the strong size dependence of correlation coefficients.

However, the correlation is far from perfect. The same residue can have a small torsion angle fluctuation in one angle but a large one in the other angle. We have analyzed 151 residues with $\Delta\varphi < 0.1$ and $\Delta\psi > 0.9$ (top left corner) and 34 residues with $\Delta\varphi > 0.9$ and $\Delta\psi < 0.1$ (bottom right corner). A majority (126/151) with $\Delta\varphi < 0.1$ and $\Delta\psi > 0.9$ are due to prolines with nearly fixed φ angles. There is no clear dominance of any other particular residue types at both off diagonal corners. All other residues (in both corners) are either all annotated as coil residues or mixed coil with either helix or strand in different NMR models. Mixed helix/strand assignments in different NMR models are rare in the entire NMR997 dataset, only 98 out of 77421 residues, possibly related to assignment inconsistency[37].

The imperfect correlation between $\Delta\varphi$ and $\Delta\psi$ is likely because the local packing and arrangement of neighboring residues may prohibit the fluctuation of one torsion angle but not the other. This effect likely disappears if we look at the overall flexibility of an entire protein chain. Indeed, the correlation coefficient between chain-averaged $\Delta\varphi$ and $\Delta\psi$ is 0.95 (as shown in Fig. 2B), significantly higher than 0.75 between $\Delta\varphi$ and $\Delta\psi$ at the residue level. This high correlation suggests that the average value of either $\Delta\varphi$ or $\Delta\psi$ can be used to represent the overall flexibility of a protein chain.

2.3 Relationship between torsion-angle fluctuation and secondary-structure types

We further investigate the relation between torsion-angle fluctuation and protein secondary-structure types. For convenience, we only use the first model to obtain secondary-structure types defined by the DSSP program[26]. Eight-state secondary-structure elements are reduced to three states by grouping α-helices (H), 3_{10}-helices (G), and π-helices (I) together as helices, β-bridges (B) and β-sheets (E) as strands, hydrogen-bonded turns (T), bends (S) and random coils or loops (−) as coils; the same grouping used in previous work[30, 32].

The distributions of torsion-angle fluctuations in the three-state secondary-structure types are shown in Fig. 3. Again, we observe similar distributions for $\Delta\varphi$ and $\Delta\psi$. The majority of residues with large torsion-angle fluctuation (> 0.5) are coil residues. Helix and strand residues are less flexible than coil residues; a majority (around 97%) have values less than or equal to 0.3. As expected, among the three secondary-structure types, the helix is the least flexible because a helix is a locally packed structures with the smallest allowable area in the Ramachandran diagram[38]. This agrees with previous observations[39, 40, 41].

2.4 Relationship between torsion-angle fluctuation and solvent accessibility

The flexibility of a residue should also be related to the degree of solvent exposure since exposed residues are more likely to be flexible as proteins can be classified as surface-molten solid[42]. We use the average ASA over all models to represent the solvent accessibility of a given residue. The absolute ASA values for each residue were obtained by running the DSSP program[26], and the absolute value was normalized by its ASA value in an Ala-X-Ala tripeptide in extended conformation to yield the relative solvent accessibility (RSA)[30, 32].

Figure 4 shows the distributions of torsion-angle fluctuations at different levels of solvent exposure. This figure was made by dividing residues into 3 RSA bins from fully buried (RSA = 0) to mostly exposed (RSA > 80%). The figure indicates that exposed residues indeed are more flexible than buried ones. Highly fluctuating regions (> 0.5) are dominated by exposed residues and residues with RSA ≤ 80% are unlikely to have large fluctuations for either φ or ψ. The CC between $\Delta\varphi$ ($\Delta\psi$) and RSA is 0.33 (0.36) if averaged over chains. This low CC comes about since exposed residues are not necessarily flexible.

2.5 Relationship between torsion-angle fluctuation and amino acid types

It is of interest to know how torsion-angle fluctuations differ for different amino acid types. The mean torsion-angle fluctuation values of each amino acid type is shown in Fig. 5. Glycine (G) and two hydrophilic residues Serine (S) and Histidine (H) are the top three most flexible residues while three hydrophobic residues tryptophan (W), valine (V), and isoleucine (I) are the least flexible residues. Glycine (G) is more flexible because its lack of side-chain allows for a great range of torsional angles. Hydrophobic residues are more likely buried and less flexible. To this end, we calculated the correlation between the mean torsion-angle fluctuation and the Fauchere-Pliska's hydrophobicity indices[43] for 20 amino acid types, and obtained negative CC values of −0.47 and −0.50 for $\Delta\varphi$ and $\Delta\psi$, respectively. This indicates that hydrophobic residues tend to have less fluctuations in torsion angles when compared to hydrophilic residues. We observe comparable $\Delta\varphi$ and $\Delta\psi$ for all amino acids except Proline (P), which is characterized by significantly low $\Delta\varphi$. This is due to the fact that the last atom of the Proline side chain is bonded to the main chain, forming a ring which restricts the available conformational space and results in a nearly fixed φ angle [44].

2.6 Relationship between torsion angle fluctuation and B-factor

In contrast to the angle fluctuations which characterize the mobility of backbone atoms in NMR-determined structures, B-factor reflects atomic thermal motion and static disorder of X-ray-determined structures. An interesting question is whether there is any relationship between these two measures. To find the answer, we used the NX48 dataset, which includes 48 proteins with both NMR- and X-ray-determined structures. Following previous work[45, 46, 32], the B-factor values of C_α atoms in each protein were extracted. Relative solvent accessibility (RSA) was included as a reference. We also calculated the mean-squared fluctuations ($< \Delta r^2 >$) of C_α atoms among different NMR models, which is directly related to the definition of B-factor. Results for the relationships amongst $\Delta\varphi$, $\Delta\psi$, RSA, $< \Delta r^2 >$, and B-factors in terms of CC are given in Table 1. All CC values reported in Table 1 are averaged over chains.

The highest correlation (0.53) is observed between $\Delta\varphi$ and $\Delta\psi$, as discussed above. This value is lower than the chain-averaged value of 0.66 for the significantly larger NMR997 dataset. This is followed by a CC of 0.5 between $< \Delta r^2 >$ and $\Delta\varphi$ (or $\Delta\psi$). This suggests that there is a correspondence between the fluctuation described in torsional space and that in Cartesian coordinate space. This correspondence is about the same as between $\Delta\varphi$ and $\Delta\psi$. The third highest correlation comes from the B-factor from X-ray structures and the mean-squared fluctuation $< \Delta r^2 >$ from NMR structures (CC= 0.43). That is, structural fluctuation observed in NMR models is somewhat consistent with the atomic fluctuation observed in X-ray structure. This result illustrates the usefulness of using different NMR models for calculating dynamic properties at the residue level. The consistency between two different experimental techniques gives further support to the validity of using an ensemble of NMR structures to probe the structural flexibility of proteins.

We also observe a weaker correlation (CC= 0.42) between the average RSA from NMR structure and the B-factor from X-ray structures. This result is consistent with the fact that solvent exposure and the flexibility of a given residue (expressed by either the B-factor or $< \Delta r^2 >$) are related[32]. However, there is a lack of significant correlation between the B-factors from X-ray structures and the two torsion angle fluctuations from NMR structures. The weak correlation (CC=0.50) between $< \Delta r^2 >$ and $\Delta\varphi$ (or $\Delta\psi$) and the weak correlation (CC= 0.43) between $< \Delta r^2 >$ and B-factor does not translate into a correlation of similar strength (CC= 0.29 or 0.27) between B-factor and $\Delta\varphi$ (or $\Delta\psi$).

2.7 Predicting real-value torsion-angle fluctuations

A neural-network-based predictor was built to predict real values of $\Delta\varphi$ and $\Delta\psi$. We optimized the sliding-window size by performing ten-fold cross-validation tests at different window sizes. As Fig. 6 shows, the value of the correlation coefficient between predicted and measured torsion angle fluctuation initially increases as the size of window increases. It saturates at a window size of 15, after which we observe small fluctuations (< 0.005) only. Thus, a sliding window of size 15 was chosen and the total number of inputs for each residue is 34×15 = 510.

To assess the consistency for the performance of our predictor, we performed five independent ten-fold cross-validation tests on the NMR997 dataset. The five independent predictors were started from different initial random weights so that different suboptimal weights and predictions were obtained. The final prediction was based on the average of the five independent predictors.

Table 2 shows the quality of the neural-network-based predictor. The results from five independent predictors are very consistent, with CC around 0.585 and MAE around 0.128 for $\Delta\varphi$, and CC around 0.59 and MAE around 0.137 for $\Delta\psi$. The final predictor based on the average of five predictors shows an improvement of 0.01 and 0.001 on CC and MAE, respectively. We performed statistical tests to verify significance of this improvement. Specifically, we first calculated CC and MAE per chain for the five independent predictors and the final predictor. Then we verified whether the values of CC and MAE for each predictor follow a normal distribution using the Shapiro-Wilk test[47] with 0.05 significance level. The tests have revealed that none of the quality measures is normal and therefore we used a non-parametric Wilcoxon rank sum test[48] with 0.05 significance level. We compared paired values of the quality measures computed per chain for each of the five independent predictors and the final predictor. The p-values are lower than 9×10^{-16}. That is, the final predictor is statistically significantly better (with higher CC and lower MAE) than each of the five independent predictors. As displayed in Table 2, our predictions are also significantly better than random ones. Moreover, our correlation coefficients are higher than the typical correlation coefficients of 0.5–0.55 between predicted and measured values of another flexibility descriptor: temperature B-factors[49, 34, 32].

Table 3 lists the MAE for the 20 amino acids, three secondary structure types and residues with different solvent exposures, as well as their mean torsion-angle fluctuations in the NMR997 dataset. The 20 amino acids are ordered in ascending order by the MAE values of $\Delta\psi$. The secondary-structure types were obtained from the first model of each structure. The

RSA values were averaged over different models of each structure, and residues were then divided into 3 bins according to their RSA values.

We observe a small MAE on $\Delta\varphi$ for proline (P). This is due to the ring structure formed by the last atom of its side chain and the main chain, which results in a nearly fixed φ angle. The table reveals that the MAE obtained for each amino acid type is strongly correlated with its mean torsion-angle fluctuations; the corresponding CC values equal 0.92 and 0.95 for $\Delta\psi$ and $\Delta\varphi$, respectively. This indicates that amino acid types with smaller mean torsion-angle fluctuations are easier to predict than those with larger values. Further investigation shows that the majority of amino acid types with smaller mean torsion angle fluctuations are hydrophobic residues, which prefer to be buried and thus are less flexible.

A simple relation between the MAE and the mean torsion angle fluctuations is also observed for three secondary-structure types and residues of different RSA. Larger mean torsion-angle fluctuations indicate more flexibility, thus result in larger prediction errors. The coil residues and the exposed residues are more difficult to predict.

2.8 Predicting overall flexibility of protein chains

It is of interest to assess the ability of our predictor to predict overall flexibility of a protein chain. In Fig. 7, we plot the native and predicted chain-averaged $\Delta\psi$ values. The CC between the native and predicted chain-averaged torsion-angle fluctuations are 0.416 and 0.423 for $\Delta\varphi$ and $\Delta\psi$, respectively. Results for $\Delta\varphi$ are nearly identical, not shown in Fig. 7 for clarity. From these results we can infer that some information on entire chain flexibility exists in our predictors, but to a lesser extent than at the residue-level. In particular, the highest predicted average $\Delta\psi$ is only 0.5 (Fig. 7), significantly smaller than the highest value of 0.75 (Fig. 2B) from NMR models.

2.9 Two-state prediction

By setting a cutoff we can use the real value prediction as a two-state prediction, i.e., predict whether a given residue has a flexible torsion angle or not. We arbitrarily assumed a cutoff of 0.5 to distinguish native flexible and rigid torsion angles. The two-state predictions were then evaluated by calculating the sensitivity (percentage of correctly predicted flexible φ or ψ), specificity (percentage of correctly predicted rigid φ or ψ) and Mathews Correlation Coefficient (MCC). We obtained two-state predictions on the NMR997 dataset with sensitivity of 29.3% and 33.1%, specificity of 97.8% and 97.9%, MCC of 0.40 and 0.44 for φ and ψ, respectively. However, the predictions were biased to less flexible residues. Hence, we further optimized the cutoff for the predicted values to yield more balanced predictions. A cutoff of 0.34 was selected to reach the highest MCC values for both φ (0.45) and ψ (0.47), and we obtained sensitivities of 50.4% and 51.3%, and the specificities of 93.0% and 93.3% for φ and ψ, respectively.

2.10 Predicting torsion angle fluctuation using structural information

We investigate whether the usage of the native secondary structure, relative solvent accessibility and torsion angles would further increase the quality of the angle fluctuation prediction. This is motivated by the fact that known structures provide templates for modeling a large fraction of proteins, and the usage of structural information may potentially lead to more reliable predictions. We performed ten-fold cross validation on the NMR997 dataset, using native values of all the three structural properties, and using native value of one structural property at a time, see Table 4. The native secondary structure, relative solvent accessibility and torsion angles were extracted from the first model of each protein chain. The inclusion of three structural properties resulted in improvement of 0.14 on CC to 0.74 for $\Delta\psi$ (0.75 for $\Delta\varphi$). This improvement is mainly due to the employment of native relative solvent accessibility, although each structural property improves the quality of the prediction.

3 Discussion

In this paper, we have presented a new structural descriptor, torsion-angle fluctuation, which describes the backbone flexibility derived from NMR-determined structures. The presented descriptor was then analyzed using two carefully constructed datasets, namely NMR997 and NX48. We observe that 1) only a small fraction of residues have high torsion-angle fluctuation while the majority of residues have a relatively rigid backbone structure; 2) $\Delta\varphi$ and $\Delta\psi$ are highly correlated; and 3) the residues with higher torsion-angle fluctuations are typically located in coil regions and at the protein surface. We have built the first sequence-based predictor for torsion-angle fluctuations based on neural networks. We achieved correlation coefficient of about 0.6 between predicted and observed torsion-angle fluctuations. This correlation coefficient is significantly higher than the highest correlation coefficient of 0.55 between predicted and measured temperature B-factors[49, 34, 32]. This suggests that our new flexibility indicator is more amenable to prediction.

Our default 34 inputs for the neural networks are based on a detailed assessment of relative contributions to the final accuracy of the predictor. The 34 inputs consists of six types of features, i.e. 20 PSSM, 7 physical parameters, 3 predicted secondary structures, 1 predicted relative solvent accessibility, 2 predicted torsion angles and 1 disordered prediction. We grouped them into three subsets according to their relevance, namely PSSM and physical parameters, predictions (secondary structures, relative solvent accessibility and torsion angles) from SPINE X, and disordered predictions from IUpred. To analyze the relative contributions of different inputs, one type of feature or one subset of inputs was removed at a time and the remaining inputs were used to build the predictor (see Table 5). Instead of performing the time-consuming ten-fold cross-validation test, we randomly selected one subset and performed all tests only on that subset. The final predictions were averaged over five independent predictors to reduce the effect of randomly generated initial weights. Among the six types of features, physical parameters contribute the most to the angle fluctuation prediction because their removal leads to the largest drop in CC and MAE, followed by predicted secondary structure and disorders and PSSM. The predictions by SPINE X seem to be complementary and prefer to work together, since removing all of them resulted in a large drop in both CC and MAE. Note that both the PSSM and the physical parameters are used as inputs for SPINE X.

During the process of designing our torsion angle fluctuation predictor, we also tried the following inputs. 1) One-dimensional input describing the terminal sequence positions. This was motivated by the observation that the residues at the termini are usually on or close to the surface, and thus are more flexible. 2) Flexibility derived from local fragment. We built a set of X-ray-determined structures in the precompiled CulledPDB list that have low sequence identity with the proteins in NMR997. We further calculated the variations of torsion angles of the center residue across all possible 8000 triplet peptides contained in the X-ray structures. 3) Local compositional complexity (variation of residue types)[50]. 4) Disordered predictions by two other predictors (VSL2 [51] and Disopred2[52]). We also built a filter predictor, i.e. using the initial predictions of $\Delta\varphi$ and $\Delta\psi$ as inputs to a second neural network, for refining the torsion-angle fluctuation predictions. However, using these additional inputs did not give any significant improvement to our predictor.

When we optimized the window size, we noted that our method reaches high correlation coefficient (0.570 for $\Delta\varphi$ and 0.561 for $\Delta\psi$) even at a window size of 1. Further increasing the window size resulted in only a small improvement in correlation coefficient. This observation does not indicate that torsion-angle fluctuations are wholly determined by local residues, but come about because our inputs include predictions from SPINE X and IUpred, both of which employ a large window size (21 and 25, respectively).

One should note that the structural variations in NMR structure ensembles are not necessarily caused by intrinsic fluctuations alone. They could be caused by, for example, a limited number of NOE restraints[53], thus leading to multiple solutions for an under-constrained optimization problem. An under-constrained problem in principle will lead to more solutions (or NMR models). Certainly, a direct relationship between the number of NOE restraints and the number of NMR models may not exist because the latter depends on specific experiments and the decision of the author(s) who published the structure. Nevertheless, it is of interest to examine if the number of NMR models has a systematic effect. We have evaluated the distribution of torsion angle fluctuation and the prediction accuracy as a function of the number of NMR models. We found no statistically significant difference between the distributions of angle fluctuations for protein structures with different number of NMR models. We further found that the correlation coefficient between prediction accuracy and number of NMR models is essentially zero (e.g. it is 0.04 between the number of NMR models and MAE of $\Delta\varphi$ and 0.01 between the number of NMR models and MAE of $\Delta\psi$). Thus, there is no visible systematic effect.

The feasibility of using NMR models to estimate dynamics of torsion angle fluctuation is also reflected from the following results. The torsion angle fluctuations obtained from NMR models are consistent with the known facts that surface, coil, and Gly are more flexible while helical and hydrophobic core residues are the least flexible (Figures 3, 4, and 5). Moreover, the mean-squared fluctuations obtained from NMR models have a positive correlation with a coefficient of 0.43 with X-ray B-factors, the commonly used dynamic indicator for protein structures solved by the X-ray crystallographic method (Table 1). In addition, they are more predictable (with a correlation coefficient of around 0.6) than temperature B-factor. This suggests that they are resulted from physical sources rather than random errors.

With the number of novel sequences being generated from genome projects rapidly increasing, the issue of how to determine their structure and function is among the most challenging problems in the post-genome era. Our predictor of torsion-angle fluctuation will be useful in assisting protein-structure prediction as restraints and for function prediction because flexible residues are often involved in functional motion. Work in this area is in progress.

Acknowledgments

This work is supported by National Institutes of Health (grants R01 GM 085003 and R01 GM 067168).

References

1. Karplus M, McCammon JA. The internal dynamics of globular proteins. CRC Crit Rev Biochem. 1981; 9:293–349. [PubMed: 7009056]
2. Bhalla J, Storchan GB, MacCarthy CM, Uversky VN, Tcherkasskaya O. Local flexibility in molecular function paradigm. Mol Cell Proteomics. 2006; 5:1212–1223. [PubMed: 16571897]
3. Carr PA, Erickson HP, Palmer AG. Backbone dynamics of homologous fibronectin type III cell adhesion domains from fibronectin and tenascin. Structure. 1997; 5:949–959. [PubMed: 9261088]
4. Daniel RM, Dunn RV, Finney JL, Smith JC. The role of dynamics in enzyme activity. Annu Rev Biophys Biomol Struct. 2003; 32:69–92. [PubMed: 12471064]
5. Demchenko AP. Recognition between flexible protein molecules: induced and assisted folding. J Mol Recognit. 2001; 14:42–61. [PubMed: 11180561]

6. Dodson G, Verma CS. Protein flexibility: its role in structure and mechanism revealed by molecular simulations. Cell Mol Life Sci. 2006; 63:207–219. [PubMed: 16389462]
7. Dunker AK, Obradovic Z. The protein trinity–linking function and disorder. Nat Biotechnol. 2001; 19:805–806. [PubMed: 11533628]
8. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D. Intrinsic dynamics of an enzyme underlies catalysis. Nature. 2005; 438:117–121. [PubMed: 16267559]
9. Kosloff M, Selinger Z. GTPase catalysis by Ras and other G-proteins: insights from Substrate Directed SuperImposition. J Mol Biol. 2003; 331:1157–1170. [PubMed: 12927549]
10. Tobi D, Bahar I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc Natl Acad Sci USA. 2005; 102:18908–18913. [PubMed: 16354836]
11. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999; 293:321–331. [PubMed: 10550212]
12. Yuan Z, Zhao J, Wang Z-X. Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng. 2003; 16:109–114. [PubMed: 12676979]
13. Alexandrov V, Lehnert U, Echols N, Milburn D, Engelman D, Gerstein M. Normal modes for predicting protein motions: a comprehensive database assessment and associated. Web tool Protein Sci. 2005; 14:633–643.
14. Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, Wells S, Yu EZ, Thorpe M, Gerstein M. The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res. 2006; 34:D296–301. [PubMed: 16381870]
15. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res. 1998; 26:4280–4290. [PubMed: 9722650]
16. Mendez R, Bastolla U. Torsional network model: Normal modes in torsion angle space better correlate with conformation changes in proteins. Phys Rev Lett. 2010; 104:228103. [PubMed: 20867208]
17. Faraggi E, Yang Y, Zhang S, Zhou Y. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure. 2009b; 17:1515–1527. [PubMed: 19913486]
18. Zhang Y. Protein structure prediction: when is it useful? Curr Opin Struct Biol. 2009; 19:145–155. [PubMed: 19327982]
19. Bonvin AM, Brunger AT. Local flexibility in molecular function paradigm. Mol Cell Proteomics. 1996; 5:1212–1223.
20. Chalaoux FR, O'Donoghue SI, Nilges M. Molecular dynamics and accuracy of NMR structures: effects of error bounds and data removal. Proteins. 1999; 34:453–463. [PubMed: 10081958]
21. Dor O, Zhou Y. Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins. 2007b; 68:76–81. [PubMed: 17397056]
22. Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins. 2009a; 74:847–856. [PubMed: 18704931]
23. Xue B, Dor O, Faraggi E, Zhou Y. Real-value prediction of backbone torsion angles. Proteins. 2008; 72:427–433. [PubMed: 18214956]
24. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19:1589–1591. [PubMed: 12912846]
25. Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray- and NMR-resolved protein structures? Proteins. 2005; 60:139–147. [PubMed: 15856480]
26. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]
27. Zupan J. Introduction to artificial neural network (ANN) methods: What they are and how to use them. Acta Chimica Slovenica. 1994; 41:327–354.

28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

29. Meiler J, Muller M, Zeidler A, Schmaschke F. Generation and evaluation of dimension reduced amino acid parameter representations by artificial neural networks. J Mol Model. 2001; 7:360–369.

30. Dor O, Zhou Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins. 2007a; 66:838–45. [PubMed: 17177203]

31. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins. 2003; 50:629–635. [PubMed: 12577269]

32. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. Proteins. 2009; 76:617–636. [PubMed: 19274736]

33. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005; 21:3433–3434. [PubMed: 15955779]

34. Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins. 2005; 58:905–912. [PubMed: 15645415]

35. Lin CP, Huang SW, Lai Y-L, Yen S-C, Shih C-H, Lu C-H, Huang C-C, Hwang J-K. Deriving protein dynamical properties from weighted protein contact number. Proteins. 2008; 72:929–935. [PubMed: 18300253]

36. Lu CH, Huang SW, Lai Y-L, Lin C-P, Shih C-H, Huang C-C, Hsu W-L, Hwang J-K. On the relationship between the protein structure and protein dynamics. Proteins. 2008; 72:625–634. [PubMed: 18247347]

37. Zhang W, Dunker AK, Zhou Y. Assessing secondary-structure assignment of protein structures by using pairwise sequence-alignment benchmarks. Proteins. 2008; 16:61–67. [PubMed: 17932927]

38. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963; 7:95–99. [PubMed: 13990617]

39. Chothia C, Lesk AM. Helix movements in proteins. Trends Biochem Sci. 1985; 10:116–118.

40. Flocco MM, Mowbray SL. C alpha-based torsion angles: a simple tool to analyze protein conformational changes. Protein Sci. 1995; 4:2118–2122. [PubMed: 8535248]

41. Li X, Jacobson MP, Friesner RA. High-resolution prediction of protein helix positions and orientations. Proteins. 2004; 55:368–382. [PubMed: 15048828]

42. Zhou Y, Vitkup D, Karplus M. Native proteins are surface-molten solids: application of the Lindemann criterion for the solid versus liquid state. J Mol Biol. 1999; 285:1371–1375. [PubMed: 9917381]

43. Fauchere JL, Pliska VE. Hydrophobic parameters pi of amino acid side chains from partitioning of N-acetyl-amino-acid amides. Eur J Med Chem. 1983; 18:369–375.

44. MacArthur MW, Thornton JM. Influence of proline residues on protein conformation. J Mol Biol. 1991; 218:397–412. [PubMed: 2010917]

45. Parthasarathy S, Murthy MR. Analysis of temperature factor distribution in high-resolution protein structures. Protein Sci. 1997; 6:2561–2567. [PubMed: 9416605]

46. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins. 2005; 61:115–126. [PubMed: 16080156]

47. Shapiro S, Wilk M. An analysis of variance test for normality (complete samples). Biometrika. 1965; 52:591–611.

48. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin. 1945; 1:80–83.

49. Pan XY, Shen HB. Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. Protein Pept Lett. 2009; 16:1447–1454. [PubMed: 20001907]

50. Wootton JC. Sequences with ’unusual’ amino acid compositions. Curr Opin Struct Biol. 1994; 4:413–421.

51. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins. 2005; 61 (Suppl 7):176–182. [PubMed: 16187360]

52. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Molec Biol. 2004; 337:635–645. [PubMed: 15019783]

53. Clore GM, Robien MA, Gronenborn AM. Exploring the limits of precision and accuracy of protein structures determined by Nuclear Magnetic Resonance Spectroscopy. J Molec Biol. 1993; 231:82–102. [PubMed: 8496968]

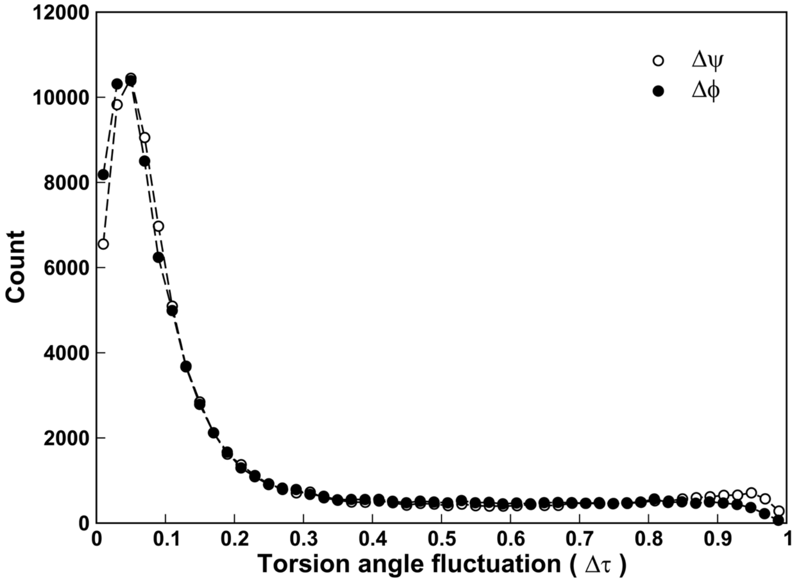

Figure 1.

Distributions of Δφ (filled circles) and Δψ (open circles) based on the NMR997 dataset.

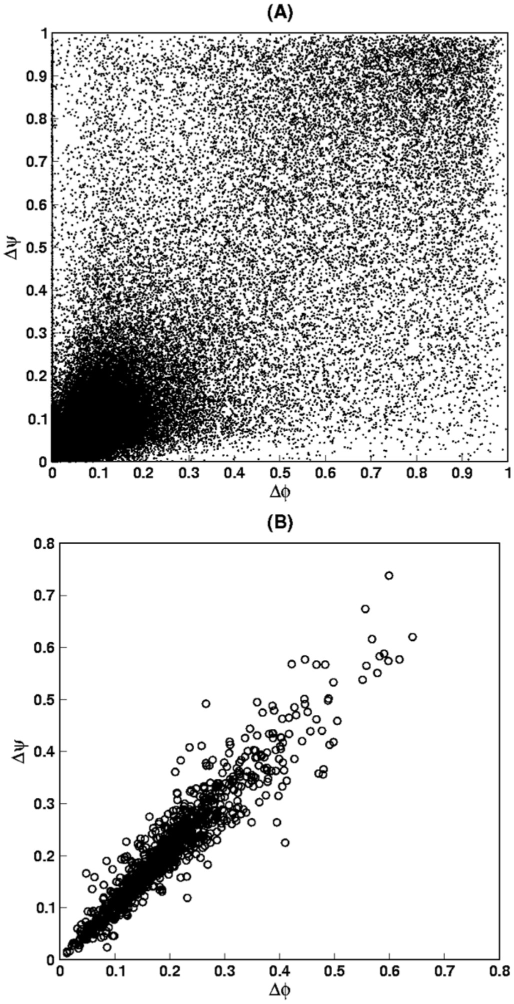

Figure 2.

(A) Fluctuations of *φ* vs. fluctuation of *ψ* on the NMR997 dataset; (B) Average fluctuation of *φ* vs. average fluctuation of *ψ* for each chain in the NMR997 dataset.

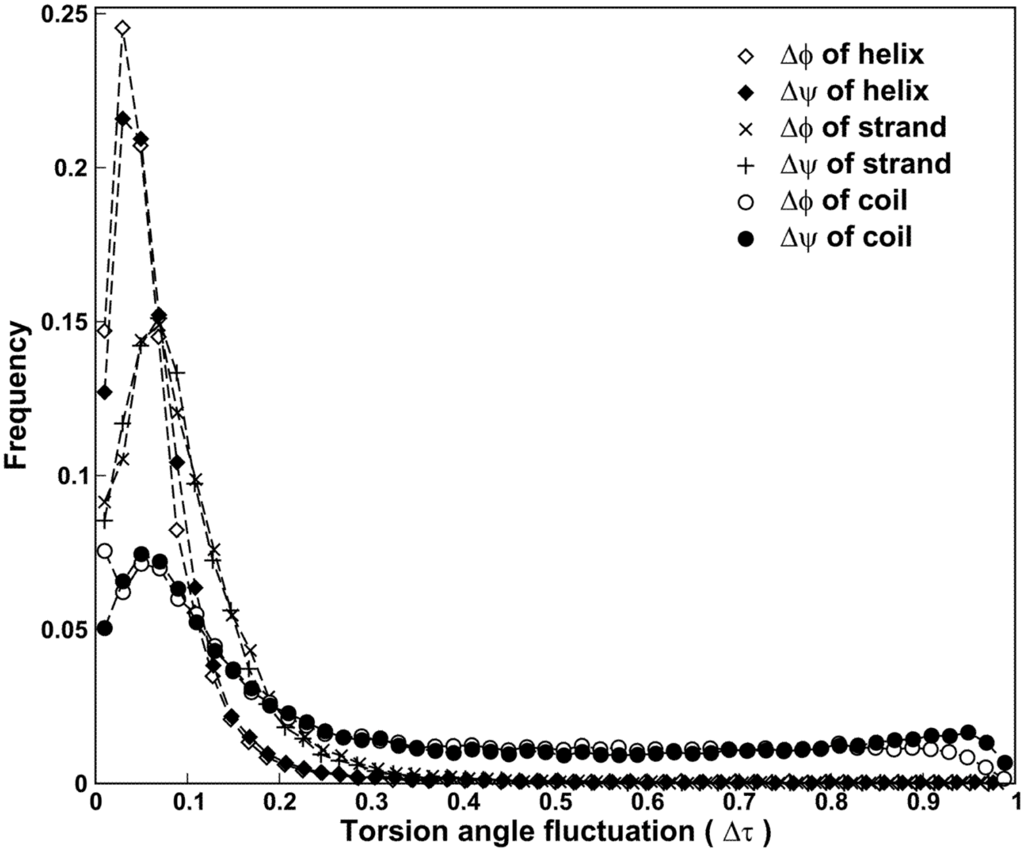

Figure 3.
Distributions of torsion angle fluctuations for the three major types of secondary structures on the NMR997 dataset.

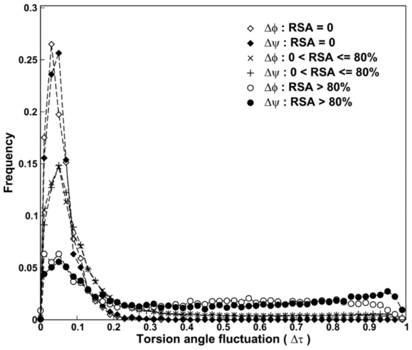

Figure 4.

Distribution of torsion angle fluctuations for residues binned according to their relative solvent accessibility (RSA) values.

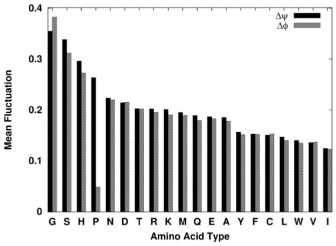

Figure 5.

Mean torsion-angle fluctuations of φ and ψ for the 20 amino acid types. Amino acid types are ordered by their mean angle fluctuation values of ψ.

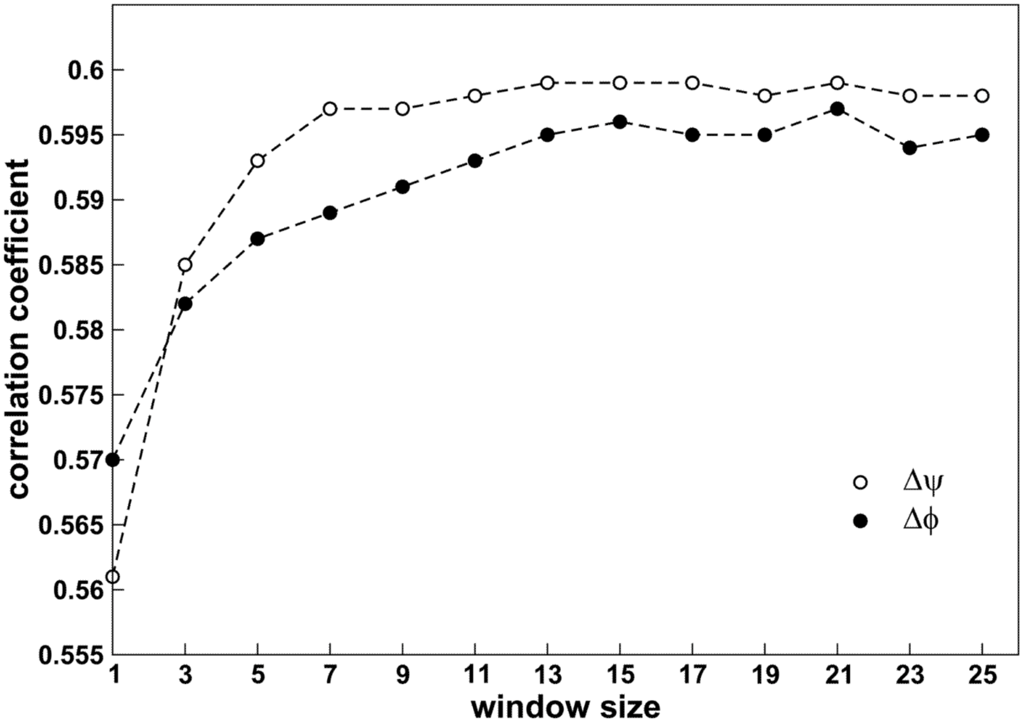

Figure 6.

The CC values (y-axis) of the neural network-based angle fluctuation predictors built using different window sizes (x-axis). These results are based on ten-fold cross-validations.

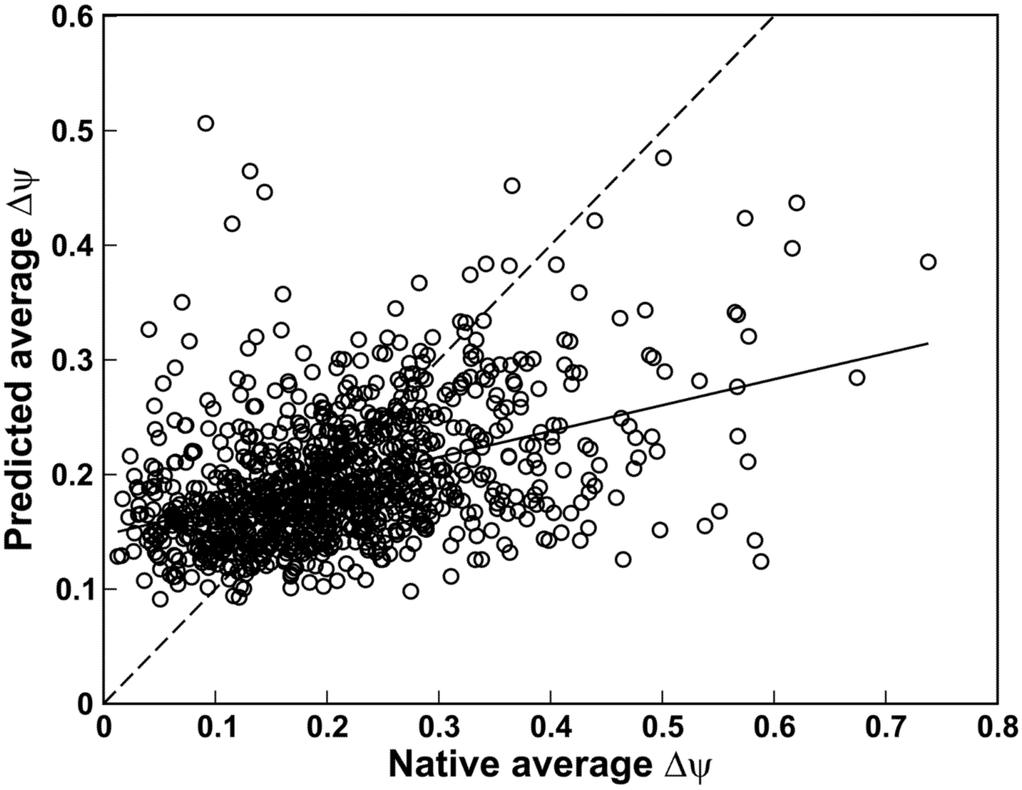

Figure 7.

Native Δψ averaged over entire chain versus predicted Δψ averaged over entire chains. Solid line from linear regression and dashed line when predicted and native values are equal.

Table 1

Correlation coefficient between two descriptors chosen from $\Delta\varphi$, $\Delta\psi$, RSA, $< \Delta r^2 >$ (calculated from NMR structures) and B-factor (derived from X-ray structures) on the NX48 dataset.

		NMR				X-Ray
		$\Delta\psi$	**$\Delta\varphi$**	**RSA**	**$< \Delta r^2 >$[a]**	**B-factor**
NMR	$\Delta\psi$	1.0	0.53	0.24	0.50	0.29
	$\Delta\varphi$		1.0	0.23	0.50	0.27
	RSA			1.0	0.40	0.42
	$< \Delta r^2 >$[a]				1.0	0.43
X-Ray	B-factor					1.0

[a] Mean-squared fluctuations of C_α atoms among different NMR models.

CC values are calculated for each chain, and then averaged over all chains.

Table 2

Ten-fold cross-validated results from five independent predictors and their consensus predictor by average.

		1	**2**	**3**	**4**	**5**	**Final**[a]	**Random**[b]
$\Delta\varphi$	CC	0.585	0.585	0.587	0.583	0.583	0.598	−0.001
	MAE	0.128	0.128	0.128	0.128	0.128	0.126	0.231
$\Delta\psi$	CC	0.590	0.589	0.590	0.591	0.590	0.602	−0.002
	MAE	0.137	0.136	0.136	0.137	0.137	0.135	0.246

[a] The final consensus predictor by averaging the predicted values from five predictors.

[b] Random predictions produced according to the distribution of angle fluctuations on φ or ψ.

Table 3

The MAE and mean torsion-angle fluctuations for the 20 amino acid types, three secondary-structure types, and three solvent accessibility types (buried, partially buried and exposed residues)

		MAE		Mean[c]		
		Δψ	**Δφ**	**Δψ**	**Δφ**	**Hydrophobicity[d]**
AA[a]	I	0.093	0.085	0.124	0.123	2.46
	V	0.097	0.094	0.136	0.137	1.66
	L	0.104	0.096	0.147	0.140	2.32
	W	0.106	0.110	0.140	0.135	3.07
	C	0.110	0.113	0.150	0.153	1.34
	F	0.111	0.110	0.153	0.152	2.44
	Y	0.115	0.108	0.157	0.151	1.31
	E	0.126	0.122	0.187	0.183	−0.87
	M	0.126	0.123	0.195	0.189	1.68
	A	0.127	0.125	0.185	0.178	0.42
	Q	0.133	0.125	0.189	0.179	−0.30
	R	0.138	0.135	0.202	0.196	−1.37
	K	0.139	0.130	0.201	0.190	−1.35
	T	0.141	0.135	0.202	0.202	0.35
	D	0.152	0.150	0.214	0.215	−1.05
	H	0.152	0.148	0.296	0.273	0.18
	N	0.157	0.156	0.223	0.220	−0.82
	S	0.162	0.158	0.338	0.312	−0.05
	P	0.180	0.047	0.263	0.048	0.98
	G	0.194	0.209	0.355	0.383	0.00
SS[b]	Helix	0.071	0.070	0.074	0.069	-
	Strand	0.074	0.074	0.097	0.100	-
	Coil	0.195	0.179	0.329	0.304	-
RSA	0.0	0.055	0.056	0.057	0.056	-
	(0.0, 0.8]	0.123	0.115	0.171	0.162	-
	(0.8, 1.0]	0.211	0.195	0.432	0.390	-

[a] Amino acid type

[b] Secondary-structure types

[c] Mean torsion-angle fluctuations averaged over the NMR997 dataset with respect to 20 amino acid types, three secondary-structure types and residues with RSA in specified range

[d] Fauchere-Pliska’s hydrophobicity index, larger value means more hydrophobic.

Table 4

Predicting torsion angle fluctuation using native structural information

	$\Delta\psi$		$\Delta\varphi$	
	CC	**MAE**	**CC**	**MAE**
origin	0.602	0.135	0.598	0.126
using native SS, RSA and TA	0.741	0.112	0.751	0.102
using native SS	0.705	0.121	0.700	0.113
using native RSA	0.742	0.112	0.751	0.102
using native TA	0.680	0.122	0.693	0.112

We performed ten-fold cross validation test on the NMR997 dataset, using predicted or native values of secondary structure (SS), relative solvent accessibility(RSA) and torsion angles(TA). Five rounds of tests were run, and predictions were averaged over the five rounds.

Table 5

Contributions of different subsets of inputs

	Δ*ψ*		**Δ*φ***	
	CC	**MAE**	**CC**	**MAE**
34 inputs	0.579	0.136	0.567	0.128
–PSSM –Physical parameters	0.556	0.140	0.538	0.132
–SPINE X[a]	0.567	0.139	0.556	0.129
–IUpred[b]	0.572	0.136	0.552	0.129
–PSSM	0.574	0.136	0.565	0.128
–Physical parameters	0.560	0.137	0.544	0.129
–SS[c]	0.572	0.136	0.567	0.127
–RSA[d]	0.579	0.136	0.568	0.127
–TA[e]	0.578	0.137	0.568	0.128

Tests were performed on one randomly selected subset. Five rounds of tests were run, and predictions were averaged over the five rounds.

[a] Predicted secondary structure, relative solvent accessibility and torsion angles (φ and ψ) by SPINE X.

[b] Disorder prediction by IUpred

[c] Predicted secondary structure by SPINE X.

[d] Predicted relative solvent accessibility by SPINE X.

[e] Predicted torsion angle (φ and ψ) by SPINE X.