Front. Biol. 2013, 8(3): 353–361
DOI 10.1007/s11515-013-1261-3

RESEARCH ARTICLE

Accurate prediction of protein dihedral angles through conditional random field

Shesheng ZHANG[1], Shengping JIN[1], Bin XUE (✉)[2]

[1] Department of Statistics, Wuhan University of Technology, Wuhan 430070, China
[2] Department of Molecular Medicine, College of Medicine, University of South Florida, Tampa, FL 33612, USA

Abstract **Identifying local conformational changes induced by subtle differences on amino acid sequences is critical in exploring the functional variations of the proteins. In this study, we designed a computational scheme to predict the dihedral angle variations for different amino acid sequences by using conditional random field. This computational tool achieved an accuracy of 87% and 84% in 10-fold cross validation in a large data set for φ and ψ, respectively. The prediction accuracies of φ and ψ are positively correlated to each other for most of the 20 types of amino acids. Helical amino acids can achieve higher prediction accuracy in general, while amino acids in beet sheet have higher accuracy at specific angular regions. The prediction accuracy of φ is negatively correlated with amino acid flexibility represented by Vihinen Index. The prediction accuracy of φ can also be negatively correlated with angle distribution dispersion.**

Keywords conditional random field, flexibility, angle distribution dispersion

Introduction

Subtle variations on amino acid sequences have been extensively observed in the studies on protein molecular biology. From the genetic point of view, mutation on nucleotides in coding region will result in a different codon. In the case of mis-sense mutation, this new codon will code for a different amino acid. Substitution of amino acids with similar properties may not severely affect the local and global structures of the protein and hence the function of the protein. However, substation to an amino acid with completely different physicochemical properties may disrupt the local and global 3-dimensional structures and eventually the function of the protein. Although it could be deleterious, amino acid substitution can also bring in novel functions to proteins and protein interaction networks (Ellegren, 2008; Tokuriki and Tawfik, 2009). That is the reason why the substitutions composed of the most magnificent pictures of competition and cooperation in the development of biological world.

Studying the influence of sequence variations on three-dimensional structures is still a formidable task. Traditional experimental methods for characterizing protein structures are X-ray and NMR. Although being very accurate, these two methods are both time and cost consuming, in addition to many critical technical issues in preparing protein samples. Another powerful experimental technique to study the local variance of amino acid on function is alanine-scanning mutagenesis (Cunningham and Wells, 1989; Ashkenazi et al., 1990; Gibbs and Zoller, 1991). This technique applies systematic alanine substitution to test various functional interactions. Since alanine has only backbone and C_β atoms, amino acid side-chain over C_β atoms will be removed upon the substitution. Through systematic substitutions and measurement of functionality, the functional epitopes and interaction sites can be identified. Nonetheless, alanine-scanning is laborious. All the substituted proteins have to be constructed, expressed, and re-folded in the experiments (Morrison and Weiss, 2001).

Hence, it is necessary to develop other efficient methods. Computational techniques are thus introduced into protein structure biology and are playing more and more important roles. Many computational tools are developed to predict three-dimensional structures, such as homology modeling, fold recognition, and *ab initio* modeling. Another category of

Received November 23, 2012; accepted February 16, 2013
Correspondence: Bin XUE
E-mail: bxue@health.usf.edu

computational methods are designed for various structural properties, such as secondary structure (Green et al., 2009), dihedral angles (Helles and Fonseca, 2009), accessible surface area (Chang et al., 2008), contact map (Xue et al., 2009), structural classes (Ahmadi et al., 2012), etc. The prediction of various structural properties is able to provide useful information on three-dimensional structure, active site, and function. The methods for the prediction of structural properties are mainly composed of various machine learning techniques. In all these machine learning methods, a consecutive segment of sequence is normally applied as the input. This unique scheme ensures that the influence of neighboring residues is taken into consideration properly. Besides, in all machine learning methods, only sequence related information was required for the prediction; none of the local structural information was needed in advance.

Here in this manuscript, we applied linear-chain Conditional Random Field (CRF) method to predict backbone dihedral angles. CRF is a very powerful undirected graphical technique similar to simple Hidden Markov Models (HMM). However, while HMM has very specific feature functions using constant probabilities, CRF can have any number and formality of feature functions. These characteristics make CRF very flexible and fitful for many purposes. Recently, CRF has been successfully applied in protein fold recognition (Liu et al., 2006), conformation sampling (Zhao et al., 2008), identification of phosphorylation site (Dang et al., 2008), and prediction of intrinsic disorder (Wang and Sauer, 2008). CRF can also be combined with artificial neural networks to sample protein conformation ensembles (Zhao et al., 2010). Therefore, we applied CRF function in a large data set of proteins to predict protein dihedral angles in this work. The 10-fold cross-validated prediction accuracy reached above 80%.

Method

Data set

Due to inadequate number of structures for single amino acid substitution in PDB, we came up with a compromised method to build up the data set. Since a sliding-window of amino acids will be used as inputs, we may only collect a group of non-homologous proteins as the data set. The proteins applied in this study are extracted from protein sequence culling server PISCES (Wang and Dunbrack, 2003). By choosing X-ray structure with resolution higher than 3 Å and sequence identity lower than 25%, 2531 sequences were finally selected. This set of proteins has length variance from 60 to 1244 residues, with the average length of 232 residues.

Protein backbone representation

The protein backbone structure can be represented by $\Omega = \{(\mathrm{R}_k,\psi_k)|k = 1,2,...,N\}$. Here, Ω represents a specific backbone conformation; R_k is the amino acid type of k-th residue; $\Psi_k = (\varphi_k, \psi_k)$ is the dihedral angles of the k-th residue. Dihedral angles have initial values from −180 to 180 degrees. N is the total number of residues in the protein.

In the following analysis, while keeping the original values of φ, all the values of ψ were shifted by −50 degrees (Xue et al., 2008). Upon the angle-shifting, the populations of ψ near two extremes were reduced to zero and all the values of ψ have a very nice two-peak distribution. This angle-shifting technique improved the prediction accuracy of protein dihedral angles remarkably (Xue et al., 2008). After the angle-shifting, both φ and shifted ψ were further split into 18 bins with each bin corresponding to an angle difference of 20 degree. Finally, each φ_k and ψ_k has only a bin value from 1 to 18. We also applied 72 bins in the prediction. However, the results from 72-bin based prediction were only used in the distribution analysis.

Two-peak prediction

As shown by recent studies, both φ and shifted ψ angles have two-peak distribution (Xue et al., 2008). This observation leads to an interesting application: predict on which peak the dihedral angles are. The two-peak based prediction can provide high accurate information on the rough range of the dihedral angles. The rough range of dihedral angle is very helpful not only for understanding the structure, but also for refining the real value predictions (Faraggi et al., 2009). For these purposes, a two-peak analysis upon the amino acid substitution was also included in this paper.

Prediction evaluation

The prediction results were evaluated by Q10% for 18-bin based prediction, and 2-state true positive rate for two-peak prediction. In 18-bin based prediction, each angle is assigned a predicted value same as the middle value of the bin. After this assignment, Q10% is calculated as the true positive rate when predicted value is within 36 degrees of experimental value.

Conditional random field

Our main objective in this study is to predict the value of dihedral angles for a substituted residue under the condition of knowing the conformation of neighboring residues. For this purpose, the conditional probability in the CRF is defined as (Wallach, 2004; Sutton and McCallum, - Lafferty et al., 2007; Faraggi et al., 2009):

$$P(\boldsymbol{s}/\boldsymbol{o}) = \Pi_i \exp[\Sigma_k\ \mu_k f_k(\boldsymbol{o},s_i,i)]/Z_0.$$

Here, $P(\boldsymbol{s}/\boldsymbol{o})$ is the conditional probability. In which, "***s***" is the states of angles of the predicted residues; "***o***" is the input over a sliding window of 5 residues, including: the amino acid type of predicted residue, amino acid types of two neighboring residues at both side of predicted residue, and corresponding (φ, ψ) values of these four neighboring residues. f_k is one of the K feature functions. In this study, K is the product of number of residue types and number of angle bins. μ_k is the weight parameter of k-th feature function. Z_0 is the normalization factor defined by:

$$Z_0 = \Sigma_{\boldsymbol{s}}\ \Pi_i\ \exp[\Sigma_k\ \mu_k f_k(\boldsymbol{o}, s_i, i)].$$

The weight parameters of this CRF model can be obtained by maximizing the conditional log-likelihood $\boldsymbol{L}$ on the training data set:

$$\boldsymbol{L} = \ln\ P(\boldsymbol{s}/\boldsymbol{o}) = \Sigma_i \Sigma_k\ \mu_k f_k(\boldsymbol{o}, s_i, i) - \ln\ Z_0 - \Sigma_k\ \mu_k^2/\sigma^2.$$

In which the entire third item was introduced to prevent over-fitting and $\sigma^2 = 50$ was applied. This equation is convex and hence a global optimum can be expected. The equation was finally solved by a slightly modified Powell method for function maximization.

Angle distribution dispersion

To quantify the angle distribution, we calculated the angle distribution dispersion $D = \Sigma_k\ abs(A_k - A_{max})P_k$. Here, A_k and P_k are the middle value of k-th bin and percentage of angles distributed in the k-th bin. A_{max} is the middle value of the bin with the highest percentage of angles. Σ_k is the summation over all the bins with percentage higher than the threshold value. In this study, the threshold value is taken as 1%. “abs” indicates the absolute value. To take consideration of angular periodicity, values of $abs(A_k - A_{max})$ larger than 180 will be subtracted by 180.

Results and discussion

Table 1 shows the prediction accuracy of φ and ψ for 20 types of amino acids under two different prediction schemes: 18-bin based and two-peak based predictions. The results of comparison among them were further illustrated in Fig. 1. Clearly, several trends can be observed in the figure: (1) the prediction accuracy under two-peak scheme is always better than that under 18-bin based scheme. The averaged increment on φ and ψ are 10% and 6% as shown in Table 1. Although this result is not a surprise, it may be very useful in designing new prediction strategies. (2) Prediction accuracies of φ are always higher than that of ψ for almost all amino acid residues. The only exception is glysine. Under both prediction schemes, the prediction accuracy of ψ is 3% higher than that of φ for glysine. (3) There are outliers in the distribution of prediction accuracy. The prediction accuracy of ψ for glycine and asparagine are remarkably lower than others. In the prediction of φ, Aspatic Acid in addition to glysine and asparagine also has an obvious lower prediction accuracy than others. Meanwhile, the prediction of φ for proline goes to another extreme. Under both schemes, the prediction result for Proline is almost perfect. (4) For aliphatic

Table 1 Prediction accuracies of φ and ψ for 20 types of amino acids under 18-bin and two-peak schemes. Error is calculated as the standard error from 10 fold of predictions

	18-bin based prediction				Two-peak prediction			
	φ	Err-φ	ψ	Err-ψ	φ	Err-φ	ψ	Err-ψ
A	86.4	1.4	88.4	0.7	98.4	0.3	92.3	0.5
C	82.6	1.8	81.2	1.2	97.4	0.5	88.9	1.2
D	83.3	1.1	72.0	0.5	94.5	0.4	84.1	0.8
E	88.9	0.9	87.9	0.6	97.8	0.2	91.6	0.5
F	86.6	1.4	85.5	0.9	98.2	0.3	91.5	0.3
G	68.2	0.4	71.6	0.9	77.1	0.4	80.8	0.7
H	81.8	1.4	79.1	0.8	95.2	0.6	87.8	1.4
I	94.3	0.7	91.6	0.8	99.7	0.1	94.2	0.5
K	86.6	0.9	84.4	0.9	96.6	0.3	89.6	0.4
L	92.9	1.1	89.1	0.8	99.0	0.1	92.5	0.4
M	88.8	1.5	88.0	1.0	98.2	0.3	92.8	0.5
N	77.9	0.5	70.0	0.7	89.5	0.4	81.7	0.9
P	99.7	0.8	83.9	0.9	99.9	0.0	86.9	0.5
Q	87.4	0.9	85.8	0.9	96.6	0.4	90.6	0.4
R	86.3	1.1	85.0	0.9	97.0	0.4	90.0	0.6
S	81.3	1.1	80.1	0.7	97.5	0.3	85.3	0.5
T	88.4	0.7	83.1	0.8	99.3	0.2	87.5	0.4
V	93.2	0.5	90.8	0.7	99.7	0.2	93.8	0.2
W	85.3	0.9	87.5	1.4	98.1	0.4	92.5	0.7
Y	86.0	0.9	85.7	1.0	97.9	0.4	91.1	0.7
Average	86.5	1.1	83.7	0.9	96.0	0.3	89.2	0.6

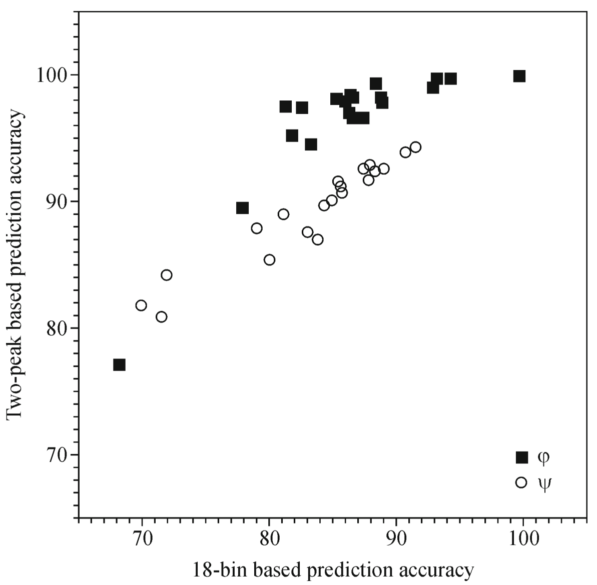

Figure 1 Comparison of prediction accuracy for φ and ψ under two different prediction schemes: 18-bin based and two-peak based. *x*-axis is the prediction accuracy from 18-bin based prediction, while *y*-axis is for the two-peak based prediction. Filled squares and blank circles are φ and ψ for 20 types of amino acids, respectively.

amino acids (isoleusine, valine, and leusine), the prediction accuracies are highly consistent between two schemes and are the highest among all 20 amino acids.

To understand the reasons why there is so much divergence in the prediction accuracy among amino acids, we presented further analysis in Fig. 2. The first factor in our consideration is amino acid composition. Biased distribution of samples will influence the results of various machine-learning methods very frequently. However, as indicated by all three panels in Fig. 2, there is no direct connection between prediction accuracy and amino acid composition. Actually, six amino acids have consistently lower prediction accuracy on both φ and ψ. These amino acids are: glysine; asparagine; aspatic acid; serine; histidine; and cysteine. In these six amino acids, only histidine and cysteine have much lower composition than the averaged value of 5%. The other four amino acids have the similar or higher compositions than the average. As a comparison, tryptophan and methionine have similar composition as histidine and cysteine but much higher prediction accuracy.

In Fig. 2A and 2B, we further compared the relation between prediction accuracy and amino acid flexibility for both φ and ψ. Here, Vihinen Index is applied to indicate the amino acid flexibility (Vihinen et al., 1994). In general, low flexibility amino acids have higher prediction accuracy, while high flexibility amino acids have lower prediction accuracy. Nonetheless, histidine and cysteine are abnormal examples in low flexibility amino acids. Lysine and glutamic acid are strangers in high flexibility amino acids. Exclusion of these four amino acids could result in a very nice correlation

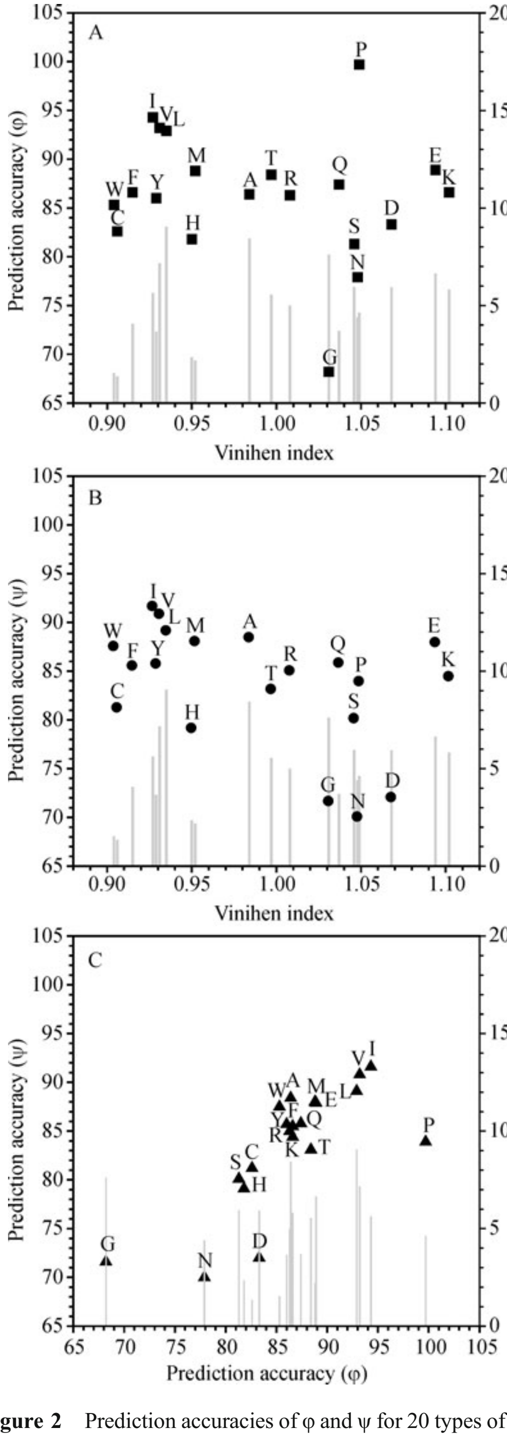

Figure 2 Prediction accuracies of φ and ψ for 20 types of amino acids, and their relations with amino acid composition and flexibility. The prediction accuracy is obtained from 18-bin based prediction. The amino acid composition is the percentage of each amino acid in the data set. Amino acid flexibility is represented by Vihinen Index. (A) and (B) are prediction accuracy of φ and ψ for 20 amino acids as a function of their Vihinen Indexes. Each gray bar is the composition of that amino acid overlapped with the gray bar on *x*-axis coordinates. The composition is measured by right-hand-side *y*-axis. (C) The correlation on prediction accuracy between φ (*x*-axis) and ψ (*y*-axis). Gray bars are also the composition overlapped with the amino acid on *x*-axis coordinates. Due to different order of amino acids, the pattern of gray bars in (C) is different from that in (A) and (B).

between prediction accuracy and flexibility. Proline is also a peculiar amino acid. The prediction of φ for Proline is almost completely correct in Fig. 2A. However, the accuracy of ψ for Proline is in line with the expectation that accuracy is correlated to Vihinen Index as shown in Fig. 2B.

We also presented the correlation on prediction accuracy between φ and ψ in Fig. 2C. Apparently, there are good correlations among all the amino acids, except Proline, aspatic acid, asparagine, and glysine. These four residues seem to have more divergence from the diagonal line in Fig. 2C. Part of the reasons could be the low prediction accuracies in either φ or ψ. From Fig. 2C, we can further identify several groups of residues in respect of their prediction accuracies on both φ and ψ: (1) Group I has the highest prediction accuracy and has three aliphatic amino acids: isoleusine, valine, and leusine; (2) Group II has higher accuracy and has ten residues (methionine; glutamic acid; alanine; tryptophan; phenylalanine; tryrosine; glutamine; arginine; lysine; threonine); (3) Group III has three residues: cysteine; histidine; and serine; (4) Group IV has four residues: proline; aspatic acid; asparagine; and glysine.

Since the distribution of ψ has an obvious two peaks with a separation of about 170 degrees (Xue et al., 2008), it is interesting to check if the prediction accuracy at two peaks is similar or not. Meanwhile, the distribution of φ has two extremely unbalanced peaks with the angular distance of 120 degrees. The second peak is much smaller than the first one. It is also necessary to examine the influence of biased distribution on the prediction accuracy. The results were shown in Table 2 and Fig. 3. In general, no matter for peak I or peak II, two-peak prediction scheme will have higher accuracy than 18-bin based prediction scheme. In peak I region, the overall improvement for φ and ψ is 2% and 3%, respectively. In peak II regions, the improvement for φ and ψ is 19% and 8%, accordingly. In addition, 18-bin based scheme has higher prediction accuracy in peak I range than in peak II region. The accuracy for φ and ψ was improved by 16% and 1%, accordingly. However, two-peak based scheme give more credits on peak II region. The improvement on φ and ψ was 0.5% and 4%, respectively.

We also analyzed various relations on prediction accuracy between different schemes and peaks in Fig. 3. As shown in panels A and C, the prediction accuracy of φ between peak I and II can be correlated to each other except several outliers in both 18-bin and two-peak based schemes. However, it is difficult to find such a correlation in panels B and D. Meanwhile, two-peak based scheme had more accurate prediction than 18-bin based scheme.

For the prediction of φ, almost all the prediction accuracies were improved in two-peak based scheme compared to that in 18-bin based scheme. In panel A, the accuracy for proline is almost 100% in both peak I and II. Three aliphatic residues

Table 2 Prediction accuracies of φ and ψ for 20 types of amino acids under 18-bin and two-peak schemes evaluated at angular ranges corresponding to two peaks. The original locations of two peaks for φ are −70 and 50 degree, and for ψ are −40 and 130 degree. After angle-shifting by −50 degree, the peak locations of ψ are −90 and 80 degree, respectively. Finally, peak I for (φ, ψ) are (−70, −90); peak II for (φ, ψ) are (50, 80)

	Peak-I				Peak-II			
	18-bin-based		Two-peak		18-bin-based		Two-peak	
	φ	ψ	φ	ψ	φ	ψ	φ	ψ
A	96.7	90.5	98.2	92.3	67.6	84.4	98.7	92.2
C	92.0	78.7	95.3	82.5	75.1	83.2	99.0	94.0
D	90.4	79.4	93.0	84.8	73.5	61.6	96.4	83.0
E	96.1	89.6	97.6	91.9	74.1	84.3	98.1	91.1
F	94.3	84.2	97.0	88.1	79.5	86.7	99.4	94.6
G	83.8	77.5	85.7	83.5	45.0	63.0	64.3	76.8
H	89.4	78.8	93.0	84.2	73.2	79.3	97.8	91.8
I	98.2	88.7	99.6	90.4	91.2	93.9	99.8	97.3
K	93.7	85.4	95.9	89.5	75.2	82.9	97.8	89.9
L	97.5	88.8	98.7	90.7	86.3	89.5	99.4	95.1
M	96.2	88.7	97.7	91.4	78.9	87.0	98.8	94.8
N	83.6	77.5	86.3	83.7	70.2	59.9	93.9	78.9
P	99.6	74.6	99.8	75.5	99.8	90.7	99.9	95.1
Q	94.7	87.4	96.2	90.4	74.7	82.9	97.3	90.8
R	94.1	86.0	96.4	89.3	74.5	83.5	98.0	90.9
S	93.3	78.1	96.6	80.7	69.1	82.1	98.4	90.1
T	95.6	78.0	99.1	80.6	82.0	87.6	99.4	93.5
V	97.5	86.4	99.5	88.3	90.3	93.7	99.8	97.4
W	95.0	88.4	97.2	91.0	74.6	86.4	99.1	94.1
Y	93.7	83.6	96.6	87.3	78.6	87.7	99.2	94.7
Average	93.8	84.2	95.8	87.4	77.6	83.1	96.3	91.4

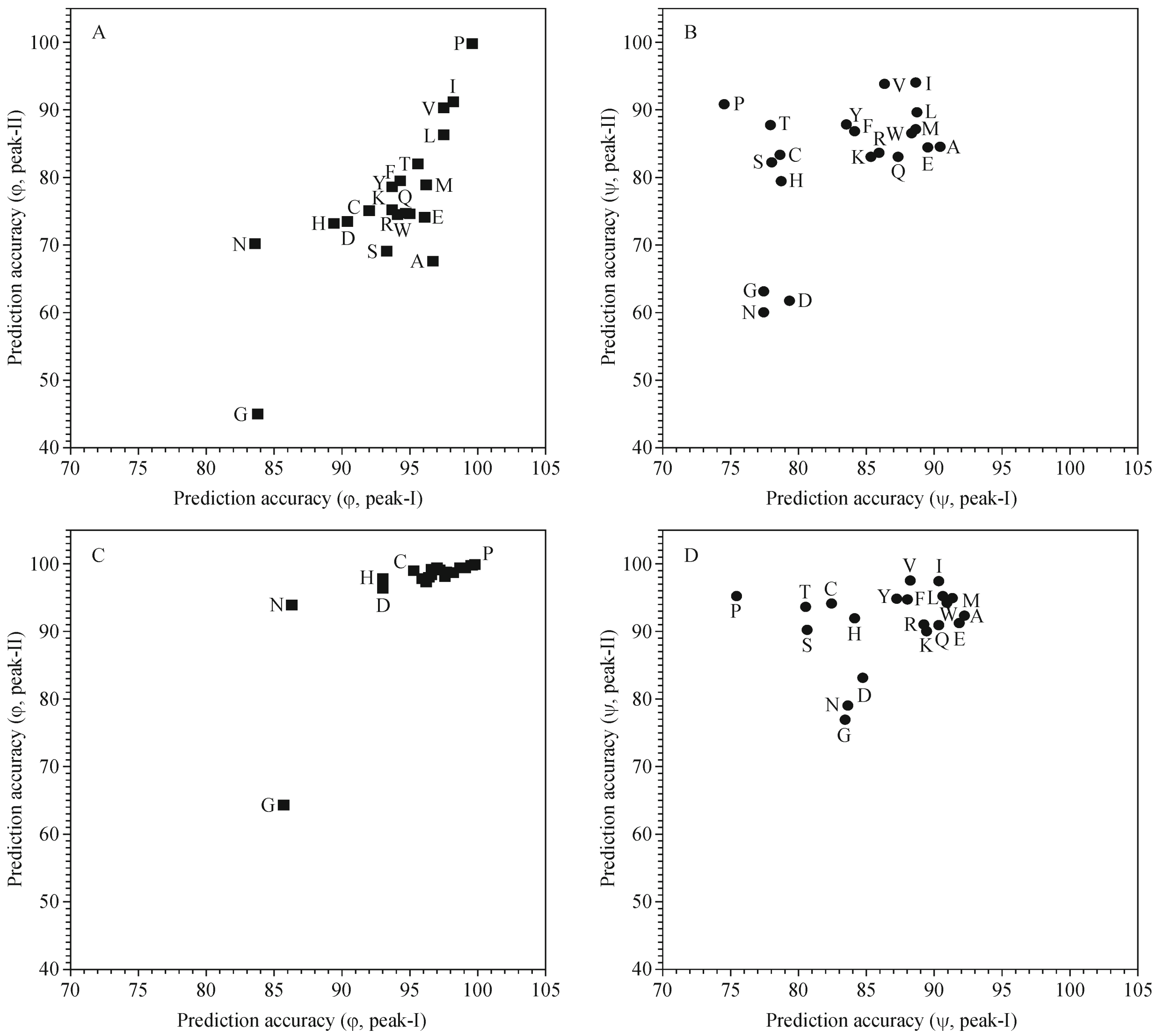

Figure 3 Comparison of prediction accuracy of φ and ψ for 20 amino acids between two-peak and 18-bin schemes. (A) and (B) show the results from 18-bin based predictions, while (C) and (D) present from two-peak scheme. Filled squares in (A) and (C) are accuracies of φ for 20 amino acids, while filled circles in (B) and (D) represent the accuracies of ψ for 20 amino acides. In panel (C), only several representative amino acids are labeled.

Table 3 Prediction accuracy of φ and ψ for various secondary structures at two peaks and as a whole. H, E, and C stand for helix, sheet, and coil. The prediction accuracy is calculated from (a) 18-bin based predictions and (b) two-peak based prediction

	φ			Ψ		
	All	Peak-I	Peak-II	All	Peak-I	Peak-II
(a)						
H	97.3	96.2	71.5	94.1	94.7	57.7
E	85.1	77.1	87.9	92.6	53.9	94.9
C	67.0	69.5	66.0	61.2	49.9	70.5
(b)						
H	99.1	99.2	91.4	96.5	96.8	78.7
E	99.2	88.2	99.2	96.2	58.5	97.3
C	88.3	85.2	90.8	72.3	54.4	86.3

have the accuracy of more than 95% in peak I and around 90% in peak II. Another ten residues have an accuracy of around 95% at peak I and diversified range of accuracy from 65% to 85% at peak II. Histidine and Aspatic Acid are 90%

accurate on peak I and 70% accurate on peak II. Asparagine has 84% and 70% on peak I and II, respectively. Glysine achieved also 84% on peak I, but only 45% on peak II. After applying the two-peak based scheme as shown in panel C, the accuracy on peak I for every amino acid was increased by several percent. For example, asparagine and glysine are reaching 86%; histidine and aspatic acid achieved 93%; and all other residues are well above 95%. In the meantime, tremendous improvement was achieved in the prediction of peak II. The accuracy of glysine was improved to 65%; asparagine is almost 95%; hisditine and aspartic acid are over 95%; Other residues are approaching 100% in this binary prediction.

For the prediction of ψ, we can also observe the similar improvement by applying two-peak based scheme. Originally in panel B, the amino acids can be roughly divided into four clusters by their accuracies on peak I and II: proline has 75% on peak I and 90% on peak II; glysine, asparagine and aspartic acid have less than 80% on peak I and only 60% on peak II; serine, threonine, cysteine, and histidine have leass than 80% on peak I but over 80% on peak II; other residues have over 85% on peak I and over 80% on peak II. In panel D which is from two-peak based scheme, although four clusters can still be observed, the overall accuracies were improved significantly except Proline.

Table 3 shows the analysis of prediction accuracy on various secondary structures. Generally, helical residues have higher accuracy than residues in beta-sheet on both φ and ψ. And the accuracy for residues in beta-sheet is higher than in coils. After splitting amino acids into peak I and II, residues in peak I keep the same trend while residues in peak II are different. In peak II regions, beta-sheet residues have the highest accuracy for φ, followed by helical residues and then coiled residues. For the prediction of ψ in peak II regions, although residues in beta-sheet still have the highest accuracy, the accuracy of coiled residues exceeds that of helical residues. This observation is in accordance with the knowledge that residues in various secondary structures have different distribution on their dihedral angles as shown by their Ramanchandran plot.

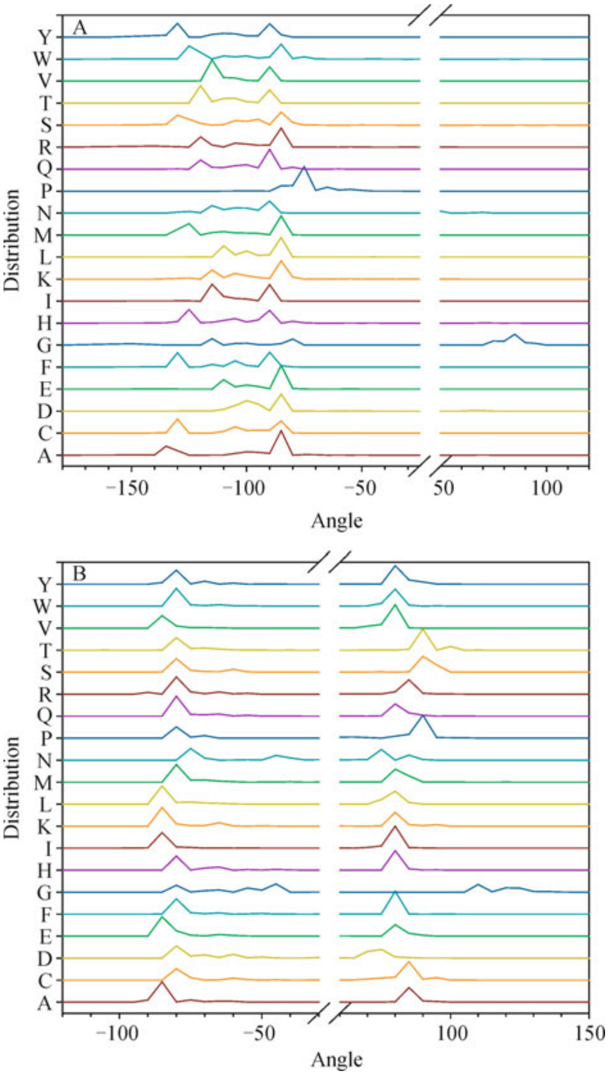

Figure 4 Distribution of predicted φ (a) and ψ (b) angles for 20 types of amino acids. The entire range of dihedral angles (from −180 to 180) was divided into 72 bins with each bin equivalent to 5 degrees. The statistics is based on the division of 72 bins. However, *x*-axis is still shown in degree. *y*-axis is the percentage at each bin for 20 types of amino acids. The values of percentage on *y*-axis are not listed. The maximal percentage is 56% for Proline.

Figure 4 presents the distribution of predicted φ and ψ angles for 20 types of amino acids. The original distribution of φ calculated from the data set statistics has a large peak and a tinny peak. The larger peak has a head and a shoulder separated by ~60 degrees. The distribution between the locations of head and shoulder is quite flat. Here in Fig. 4A, although the predicted φ still has a separation of about 60 degrees between its two major peaks, the flat distribution between them is missing. In addition, Proline has only one major peak at around −70 degree. Aspartic acid has a combined peak of head and shoulder. Glysine has three peaks with the major one at ~90 degree and far away from other amino acids. The original distribution of ψ from data set has also two peaks separated by 170 degree. In the distribution of predicted ψ, the location and separation of two peaks are maintained very well. Anyhow, we can see the shrunk separation for Asparagine and Aspartic Acid. Glysine also has three peaks with only one overlapped with other amino acids. The other two locate at −50 and 110 degree.

Figure 5 is the analysis on angle distribution dispersion. A nice correlation on angle distribution dispersion between φ and ψ can be observed for most amino acids in Fig. 5A. The outliers are: glysine; cysteine; tryptophan; leusine; aspartic acid. The first two amino acids have larger dispersion on φ, while the other three have larger dispersion on ψ. proline, Glutamic acid, valine, glutamine, alalnine, and isoleusine have smaller dispersion on both φ and ψ. Furthermore, we presented the dispersion for both φ and ψ as a function of

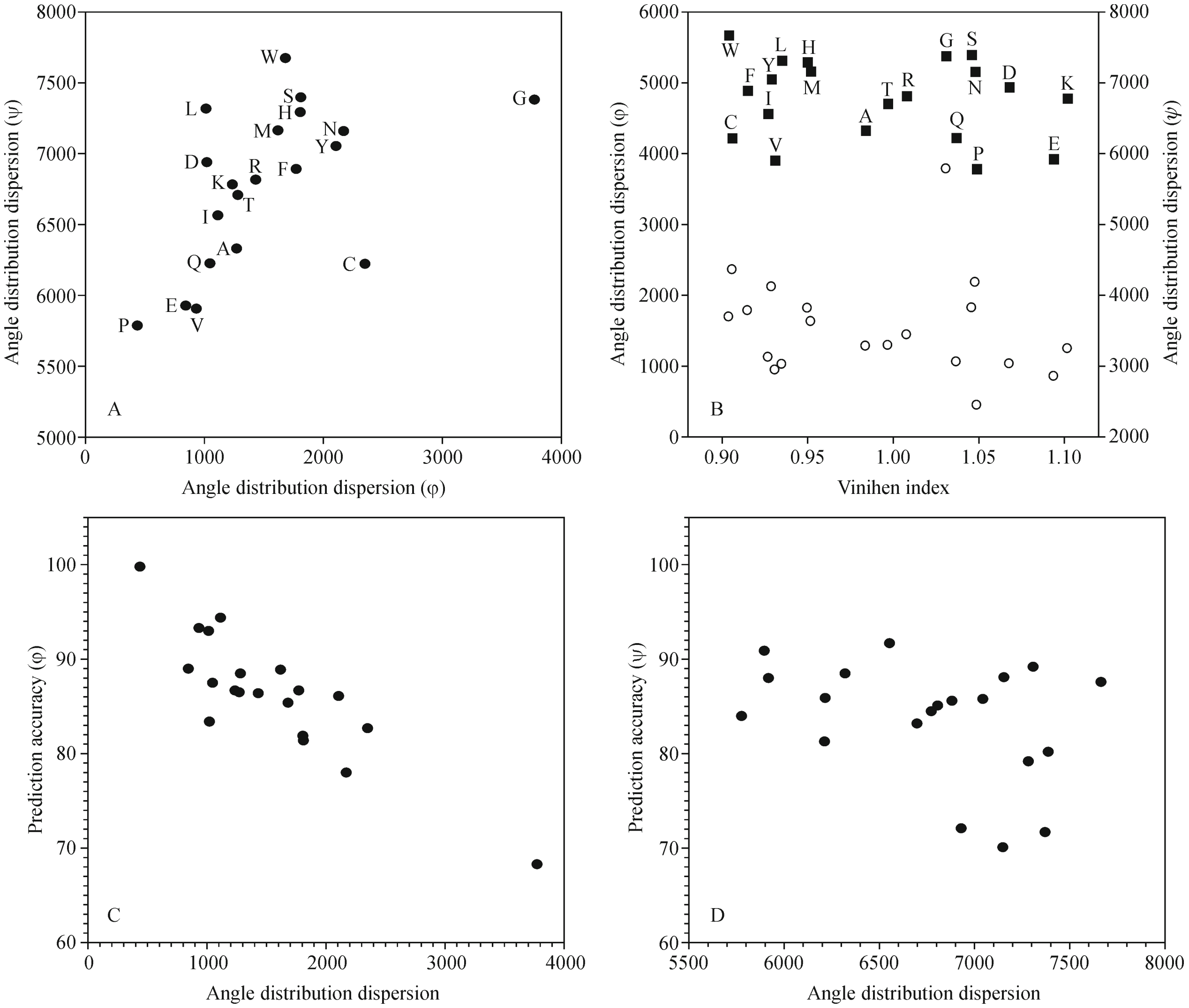

Figure 5 (A) Comparison of angle distribution dispersion between φ and ψ for 20 types of amino acids. (B) Variation of angle distribution between φ and ψ for 20 types of amino acids as a function of their Vihinen Indexes. Left-hand-side *y*-axis is angle distribution dispersion for φ (open circles), while right-hand-side *y*-axis is for ψ (filled squares). (C) Correlation between prediction accuracy and angle distribution dispersion for φ. (D) Relation between prediction accuracy and angle distribution dispersion for ψ.

amino acid flexibility in Fig. 5B. Clearly, there is an obvious negative correlation for φ and a marginal negative correlation for ψ. By taking into consideration that prediction accuracy of φ is also negatively correlated with Vihinen Index, we presented the relation between prediction accuracy and Vihinen Index for φ and ψ in Fig. 5C and 5D, accordingly. Obviously, the accuracy of φ can be linearly correlated with angle distribution dispersion.

Conclusion

We designed a computational model by applying conditional random field to predict dihedral angles. The overall accuracies by dividing dihedral angles into 18 bins are 86% for φ and 84% for ψ. In the case of two-peak prediction, the accuracies are improved to 96% for φ and 89% for ψ. Apparently, these high-accuracy predictions are very useful in understanding the structural variations of proteins.

In addition to the division of both dihedral angles into 18 equal bins, we also did the prediction by clustering angles into two peaks. By applying the two-peak based scheme, the prediction accuracy was further improved. The accuracy under two-peak scheme is positively correlated with the accuracy under 18-bin based scheme. Between two peaks, the prediction accuracy on φ showed correlations, but not for ψ. This is an interesting observation. As shown by previous study on the prediction of protein dihedral angles, the division of two peaks will be helpful in improving the overall prediction accuracy (Faraggi et al., 2009). Hence, the same strategy can be applied in the future on the study for amino acid substitution.

To understand the sources of the prediction error, we further analyzed the accuracies of 20 types of amino acids, as well as their relations with amino acid flexibility, composition, location, secondary structures, and angular distribution. As indicated by the analysis, glysine has the way low accuracy compared to other amino acids. Helical residues will usually have higher accuracy than other amino acids in other secondary structures. However, in specific angular regions, beta-sheet amino acids will have higher accuracy. The accuracy was also influenced by amino acid flexibility, but not amino acid composition.

Glycine is a very peculiar amino acid. From structural point of view, glycine has no side chains. Hence, glysine can be very flexible on its dihedral angles. This was reflected by several points in our study. First, glycine has a very large value of angle distribution dispersion on both dihedral angles. The dispersion of φ is almost twice larger than others, while the dispersion on ψ is the second largest among 20 amino acids. Second, the prediction accuracy for ψ of glycine is higher than that of φ. This is in contrary to all other amino acids. These properties make it very difficult to prediction dihedral angles of glycine. The similar situation is also observed for asparagine.

We further applied a quantity called angle distribution dispersion to describe the width of distribution for predicted angles. Other than above mentioned 18-bin and two-peak based schemes, we tested 72-bin based prediction and applied the results in distribution analysis. The angle distribution dispersion for φ is negatively correlated with amino acid flexibility. The prediction accuracy of φ can also be negatively correlated to the scale of angle distribution dispersion.

Compliance with ethics guidelines

Shesheng Zha, Shengping Jin, and Bin Xue declare that they have no conflict of interest.This article does not contain any studies with human or animal subjects by any of the authors.

References

Ahmadi Adl A, Nowzari-Dalini A, Xue B, Uversky V N, Qian X (2012). Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences. J Biomol Struct Dyn, 29(6): 623–633

Ashkenazi A, Presta L G, Marsters S A, Camerato T R, Rosenthal K A, Fendly B M, Capon D J (1990). Mapping the CD4 binding site for human immunodeficiency virus by alanine-scanning mutagenesis. Proc Natl Acad Sci USA, 87(18): 7150–7154

Chang D T, Huang H Y, Syu Y T, Wu C P (2008). Real value prediction of protein solvent accessibility using enhanced PSSM features. BMC Bioinformatics, 9(Suppl 12): S12

Cunningham B C, Wells J A (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science, 244(4908): 1081–1085

Dang T H, Van Leemput K, Verschoren A, Laukens K (2008). Prediction of kinase-specific phosphorylation sites using conditional random fields. Bioinformatics, 24(24): 2857–2864

Ellegren H (2008). Comparative genomics and the study of evolution by natural selection. Mol Ecol, 17(21): 4586–4596

Faraggi E, Xue B, Zhou Y (2009). 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, 74(4): 847–856

Faraggi E, Yang Y, Zhang S, Zhou Y (2009). Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure, 17(11): 1515–1527

Gibbs C S, Zoller M J (1991). Identification of electrostatic interactions that determine the phosphorylation site specificity of the cAMP-dependent protein kinase. Biochemistry, 30(22): 5329–5334

Green J R, Korenberg M J, Aboul-Magd M O (2009). PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. BMC Bioinformatics, 10(1): 222

Helles G, Fonseca R (2009). Predicting dihedral angle probability distributions for protein coil residues from primary sequence using neural networks. BMC Bioinformatics, 10(1): 338

Lafferty J, McCallum A, Pereira F(2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pages 282–289.

Liu Y, Carbonell J, Weigele P, Gopalakrishnan V (2006). Protein fold recognition using segmentation conditional random fields (SCRFs). J Comput Biol, 13(2): 394–406

Morrison K L, Weiss G A (2001). Combinatorial alanine-scanning. Curr Opin Chem Biol, 5(3): 302–307

Sutton C, McCallum A (2007) An Introduction to Conditional Random Fields for Relational Learning. In: Getoor L, Taskar B, ed. Introduction to Statistical Relational Learning. MIT Press

Tokuriki N, Tawfik D S (2009). Stability effects of mutations and protein evolvability. Curr Opin Struct Biol, 19(5): 596–604

Vihinen M, Torkkila E, Riikonen P (1994). Accuracy of protein flexibility predictions. Proteins, 19(2): 141–149

Wallach H (2004). Conditional Random Fields: An Introduction. University of Pennsylvania CIS Technical Report MS-CIS-04-21

Wang G, Dunbrack R L Jr (2003). PISCES: a protein sequence culling server. Bioinformatics, 19(12): 1589–1591

Wang L, Sauer U H (2008). OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields. Bioinformatics, 24(11): 1401–1402

Xue B, Dor O, Faraggi E, Zhou Y (2008). Real-value prediction of backbone torsion angles. Proteins, 72(1): 427–433

Xue B, Faraggi E, Zhou Y (2009). Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins, 76 (1): 176–183

Zhao F, Li S, Sterner B W, Xu J (2008). Discriminative learning for protein conformation sampling. Proteins, 73(1): 228–240

Zhao F, Peng J, Xu J (2010). Fragment-free approach to protein folding using conditional neural fields. Bioinformatics, 26(12): i310–i317