PLOS | ONE

RESEARCH ARTICLE

Identify Beta-Hairpin Motifs with Quadratic Discriminant Algorithm Based on the Chemical Shifts

Feng YongE*, Kou GaoShan

College of Science, Inner Mongolia Agriculture University, Hohhot, PR China

* fengyonge@163.com

Citation: YongE F, GaoShan K (2015) Identify Beta-Hairpin Motifs with Quadratic Discriminant Algorithm Based on the Chemical Shifts. PLoS ONE 10(9): e0139280. doi:10.1371/journal.pone.0139280

Editor: Ayyalusamy Ramamoorthy, University of Michigan, UNITED STATES

Received: May 12, 2015

Accepted: September 9, 2015

Published: September 30, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The work was supported by the Inner Mongolia autonomous region higher school science and technology research projects (No. NJZY067) and Basic Science of Inner Mongolia Agriculture University Research Fund (No. JC2013004).

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: CSs, Chemical Shifts; Sn, Sensitivity; Sp, Specificity; MCC, Mathew's correlation coefficient; Acc, The overall accuracy; AAC, Amino

Abstract

Successful prediction of the beta-hairpin motif will be helpful for understanding the of the fold recognition. Some algorithms have been proposed for the prediction of beta-hairpin motifs. However, the parameters used by these methods were primarily based on the amino acid sequences. Here, we proposed a novel model for predicting beta-hairpin structure based on the chemical shift. Firstly, we analyzed the statistical distribution of chemical shifts of six nuclei in not beta-hairpin and beta-hairpin motifs. Secondly, we used these chemical shifts as features combined with three algorithms to predict beta-hairpin structure. Finally, we achieved the best prediction, namely sensitivity of 92%, the specificity of 94% with 0.85 of Mathew's correlation coefficient using quadratic discriminant analysis algorithm, which is clearly superior to the same method for the prediction of beta-hairpin structure from 20 amino acid compositions in the three-fold cross-validation. Our finding showed that the chemical shift is an effective parameter for beta-hairpin prediction, suggesting the quadratic discriminant analysis is a powerful algorithm for the prediction of beta-hairpin.

Introduction

Protein function is inherently correlated with its structure. So, the prediction of protein structure is an active research field in bioinformatics. At present, it is still difficult to predict the spatial structure directly from protein primary structure. However, the successful prediction of protein super-secondary structure is the key step in the spatial structure prediction. Protein super-secondary-structure motifs are composed of a few regular secondary structural elements connected by loops. These structural motifs play an important role in protein folding and stability because a large number of motifs exist in protein spatial structure. Generally speaking, the empirical prediction of protein super-secondary structure essentially consists of two parts: one is the prediction of different structural types from amino acid sequences [1–3]; another is the prediction of structural motifs [4–7]. In this article we concentrate on the latter. The prediction of beta-hairpin motif will be helpful to identify fold in the unknown structure. In the past decade, many researchers have focused on exploring methods for beta-hairpin prediction [6–10]. However, the features of these studies were mainly derived from the amino acid

acid compositions; QDA, Quadratic discriminant analysis; ANOVA, Analysis of variance.

compositions or dipeptide compositions. In this study, we introduced a novel feature, chemical shifts (CSs), to predict beta-hairpin motifs. Chemical shift describes the local chemical environment of nuclear spins in nuclear magnetic resonance [11]. Therefore, some researchers have utilized it for the determination of bimolecular structures and molecular dynamics studies [12–17]. Moreover, some works have studied on protein structure prediction [18–26] and protein backbone and side chain torsion angle prediction [27] by using chemical shifts, results showing that chemical shift is a powerful parameter for the determination of protein structure information.

In this paper, we would like to utilize CSs as parameters to predict beta-hairpin motifs combined with quadratic discriminant analysis. Using the benchmark dataset, we adopted three-fold cross-validation and achieved the sensitivity of 92% and specificity of 94% and the overall prediction accuracy of 87% by using CSs of six nuclei as features and combining with quadratic discriminant analysis (QDA) algorithm. At the same time, to compare with other parameter, we have performed the prediction by using 20 amino acid compositions (AAC) as inputs of the method of QDA. The results showed that the performance of CSs outperform that of 20 AAC in the prediction of beta-hairpin. At present, some machine learning algorithms were used in the prediction of beta-hairpin motifs [6–10]. Therefore, to test our method and facilitate comparison with other methods, we have performed the prediction by using the same six CSs as feature of the support vector machine (SVM) and Random forest (RF) algorithm in the same cross-validation. Compared results showed that QDA is better than the other two algorithms in terms of accuracies.

Materials and Methods

Database

All of the CSs data used in this paper were retrieved from the re-referenced protein chemical shift database RefDB [28]. The following steps were performed to construct our dataset. Firstly, only proteins in RefDB overlapping with the corresponding Protein Data Bank (PDB) file with sequence identity of 100% were considered. Secondly, only proteins with the beta-hairpin or beta-link (called not beta-hairpin) motifs information in ArchDB40 database [29] were considered. Thirdly, only proteins with six nuclei (C,C_α,C_β,H_N,H_α,N) assigned CSs were considered. Finally, we utilized the PISCES program [30] to remove the highly similarity sequences. After strictly following the aforementioned procedures, 123 proteins were obtained. Among 123 proteins, 87% (107 sequences) proteins have less than 25% sequence identity, and the sequence identity of the remains ranges from 25 to 30%. In 123 proteins, due to consider the six CSs information at the same time, finally we obtained 157 beta-hairpin fragments, in which the lengths are ranged from 7 to 38 amino acid residues. And 75 not beta-hairpin fragments, the lengths of these fragments are ranged from 8 to 40 amino acid residues. PDB IDs of 123 and CSs data of 157 beta-hairpin fragments and 75 not beta-hairpin fragments are listed in the Supplementary Materials S1–S3 files.

Feature parameter

In the two data subsets {beta-hairpin, not beta-hairpin}, we calculated the averaged CSs of six nuclei for a fragment of length l using following formula.

$$t_m = \frac{1}{l}\sum_{j=1}^{l} CS_m^j \tag{1}$$

Here $l = \left\{ \begin{array}{l} [7\tilde{}38] \;\; in\, beta - hairpin\, dataset \\ [8\tilde{}40] \;\; in\, not\, beta - hairpin\, dataset \end{array} \right\}$, $m = C,C_\alpha,C_\beta,H_N,H_\alpha,N$, and j represents

amino acid positions in the fragment. Therefore, a sequence fragment can be converted into a six-dimensional vector R:$\{t_m\}$.

Statistical distribution

Under the normal distribution, the analysis of variance (ANOVA) can be used to test whether there was a significant difference for two-group or multi-group samples [19, 31] in the database. In this paper, the ANOVA is defined by Eq (2)

$$MS_T = MS_B + MS_W \tag{2}$$

where MS_T, MS_B and MS_W denoted the square means of total, between groups and within a group, respectively. The statistical value, called F-value, is the ratio of MS_B and MS_W, which can be calculated by Eq (3)

$$F-\text{value} = MS_B/MS_W \tag{3}$$

From Eq (3), we can see that the MS_B becomes increasingly larger than MS_W, F-value will become larger. That is to say, there are significant differences between groups, otherwise, the lack of differences.

Quadratic discriminant analysis (QDA)

As mentioned above [6–10], various parameters such as amino acid compositions and dipeptide compositions have been employed in the prediction of beta-hairpin. Here, we used CSs as feature to predict beta-hairpin motifs.

The QDA [32–35] is an effective algorithm that has been widely applied in genomic and proteomic bioinformatics in recent years. Thus, we used it here to perform prediction.

For a sequence X to be classified, we calculated the averaged CSs of six nuclei using the Eq (1). So, the sequence is converted into a six-dimensional vector R:$\{t_m\}$

$$R : \{t_m\}\ (m = C, C_\alpha, C_\beta, H_N, H_\alpha, N) \tag{4}$$

Here we integrated six-dimensional vector by using QDA. Consider a sequence X is classified into two groups (beta-hairpin, not beta-hairpin). The discriminant analysis function between group i and group j is defined by

$$\xi_{ij} = \ln p(\omega_i|X) - \ln p(\omega_j|X) \tag{5}$$

According to Bayes' Theorem, we deduce

$$\begin{aligned}\xi_{ij} &= \ln\frac{p_i}{p_j} - \frac{\delta_i - \delta_j}{2} - \frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|} \\ &= (\ln p_i - \frac{1}{2}\delta_i - \frac{1}{2}\ln\left|\sum\nolimits_i\right|) - (\ln p_j - \frac{1}{2}\delta_j - \frac{1}{2}\ln\left|\sum\nolimits_j\right|)\end{aligned} \tag{6}$$

Set

$$\eta_v = \ln p_v - \frac{\delta_v}{2} - \frac{1}{2}\ln\left|\sum\nolimits_v\right| \tag{7}$$

where

$$\delta_v = (R - \mu_v)^T \sum\nolimits_v^{-1} (R - \mu_v) \tag{8}$$

where v = beta-hairpin, not beta-hairpin, and p_v denotes the number of samples in group v, δ_v is the square mahalanobis distance between R and μ_v with respect to Σv (notes: μ_v and $|\Sigma_v|$ are calculated in training set), and μ_v denotes chemical shift values of six nuclei R:$\{t_m\}$ averaged over group v, $|\Sigma_v|$ is the determinant of matrix Σ_v.

The six-dimensional vector μ_v can be written

$$\mu_m^{(v)} = \frac{1}{p_v}\sum_{n=1}^{p_v} t_m^n \tag{9}$$

here p_v denotes the number of samples in group v; t_m^n tdenotes the average CSs of m nuclei for n-th sequence in group v; v = beta-hairpin, not beta-hairpin; $m = C,C_\alpha,C_\beta,H_N,H_\alpha,N$.

The covariance matrix Σ_v is 6 × 6dimension, quantifying correlations between the chemical shifts of six nuclei.

$$\sum\nolimits_v = \begin{bmatrix} \sigma^v{}_{1,1} & \sigma^v{}_{1,2} & \cdots & \sigma^v{}_{1,6} \\ \sigma^v{}_{2,1} & \sigma^v{}_{2,2} & \cdots & \sigma^v{}_{2,6} \\ \vdots & \vdots & & \vdots \\ \sigma^v{}_{6,1} & \sigma^v{}_{6,2} & \cdots & \sigma^v{}_{6,6} \end{bmatrix}$$

where the element

$$\sigma_{i,j}^v = \frac{1}{p_v}\sum \left(t_i - \mu_i^{(v)}\right)\left(t_j - \mu_j^{(v)}\right) \tag{10}$$

here v = beta-hairpin, not beta-hairpin; i,j = $C,C_\alpha,C_\beta,H_N,H_\alpha,N$

From Eq (6) and Eq (7), we have concluded

$$\xi_{ij} = \eta_i - \eta_j \tag{11}$$

It can be easily proved that $p(w_k|X)$ is the maximum of $p(w_v|X)$, if η_k is the maximal one in η_v (v = beta-hairpin, not beta-hairpin). Then, we predict that X belongs to group k. In statistical results, fluctuation phenomenon inevitably exists. To correct predicted results, we define the coefficient of the error allowed scope as

$$R = \frac{\eta_{corr} - \eta_{wro}}{\eta_{corr}} \tag{12}$$

where η_{corr} denotes X belonging to itself class η, η_{wro} denotes X being predicted other class η. Set the appropriate R, the sequence X in the error allowed scope can be classified correctly by using Eq (12).

Performance evaluation

In statistical prediction, the jackknife test is considered to be the most rigorous test method [36] and has been widely used to evaluate the performance of various predictors [37–41]. However, considering the longer time needed for the jackknife test and because the goal of our paper concentrated on introducing a new model for beta-hairpin prediction, we adopted the three-fold cross-validation to evaluate the performance of our method. We randomly divided the training dataset into three parts, two of which are for training and the one for testing. The process is repeated three times. The final performance was calculated by averaging over all three datasets. The following parameters: the sensitivity (Sn), specificity (Sp), the overall accuracy (Acc) and Mathew's correlation coefficient (MCC) are used to evaluate the

predictive performance of our approach.

$$Sn = \frac{TP}{TP + FN} \times 100\% \tag{13}$$

$$Sp = \frac{TN}{TN + FP} \times 100\% \tag{14}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \tag{15}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TN + FN) \times (TP + FP) \times (TN + FP)}} \times 100\% \tag{16}$$

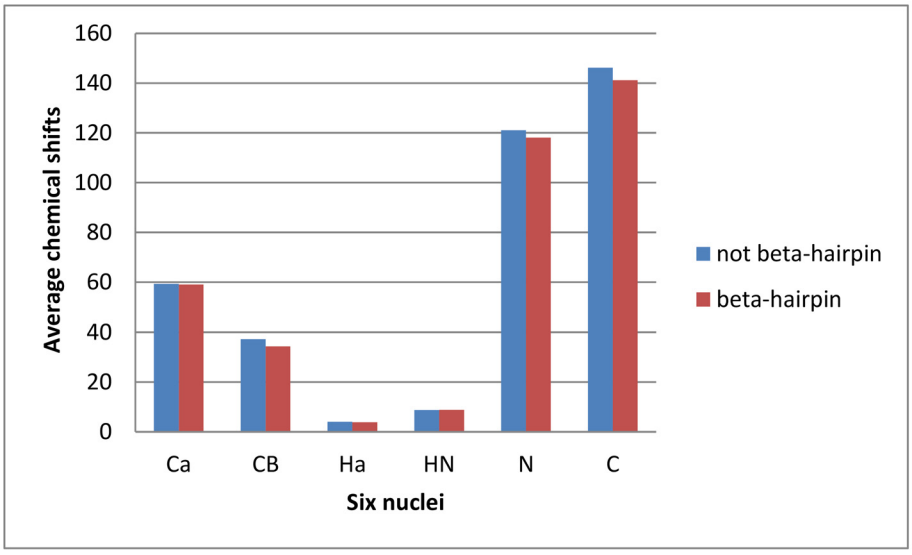
where true positive (TP) denotes the number of correctly predicted beta-hairpin motif, false negative (FN) denotes the number of the beta-hairpin misclassified as not beta-hairpin motif, false positive (FP) denotes the number of the not beta-hairpin misclassified as beta-hairpin motif, and true negative (TN) denotes the number of correctly predicted not beta-hairpin motif.

Results and Discussion

Statistical distribution of the average CSs of six nuclei

We analyzed the average chemical shifts of six nuclei in beta-hairpin and not beta-hairpin dataset. As showed in Fig 1, we found that the different distribution of the CSs six nuclei in beta-hairpin and not beta-hairpin dataset. The average chemical shift values of C,C_α,C_β,H_α,N nuclei are higher in not beta-hairpin dataset than beta-hairpin dataset. However, the average chemical shift value of H_N nuclei is lower in not beta-hairpin dataset than beta-hairpin dataset.

For further investigating whether the distribution of average CSs of six nuclei in two datasets are independent of one another, the analysis of variance (ANOVA) [19, 31] can be used for the



Notes: Red and blue lines represent the beta-hairpin and not beta-hairpin respectively.

Fig 1. Distribution chart of six-nuclei CSs in beta-hairpin and not beta-hairpin motifs.

doi:10.1371/journal.pone.0139280.g001

Table 1. The statistical test using ANOVA for CSs of six nuclei.

Nuclei	ANOVA (*p*–value)
C	4.42 (p<0.05)
C_a	4.01 (p<0.05)
C_B	4.44 (p<0.05)
H_N	4.13 (p<0.05)
H_a	4.36 (p<0.05)
N	4.12 (p<0.05)

doi:10.1371/journal.pone.0139280.t001

average CSs of six nuclei in beta-hairpin and not beta-hairpin statistical analysis under a normal distribution. Though we know that many test statistics are approximately normally distributed for large samples (generally>30 samples) under the central limit theorem. In order to strictly verify the validity of a normal distributional assumption, we implemented the statistical test. The Quantile-quantile (Q-Q) plot or Probability-probability (P-P) plot in statistics is often as a means to check the validity of a statistical distributional assumption for a dataset [42]. In term of P-P plot, if the data indeed follow the assumed normal distribution, then the points on the P-P plot will fall approximately on the diagonal line. The result demonstrated that the sampling distributions of six-nuclei CSs obey normal distribution (see supplementary material S4 file). Therefore, ANOVA can be implemented. Table 1 records the F-values of six nuclei and corresponding p-values. From Table 1 we observed that six p-values are less than 0.05 ($p<$ 0.05). This result shows that the average CSs of six nuclei have a significant difference between beta-hairpin and not beta-hairpin structures, suggesting that beta-hairpin motifs can be discriminated from not beta-hairpin sequences based on the CSs of six nuclei.

Prediction of beta-hairpin based on the CSs of six nuclei

Results in Table 1 suggest that the CSs of six nuclei are capable of predicting beta-hairpin. Therefore, we examined the accuracy of six nuclei by using QDA algorithm. Under the benchmark dataset, we calculated the average chemical shift values using the Eq (1). The sequences from two data subsets are converted respectively into six-dimensional vectors. In the training sets, determinant and inverse matrix of covariance matrix Σ_v are calculated. And μ is a six-dimensional mean vector, which is calculated in each dataset. Given a sequence X in testing sets, we may calculate η_v by using Eqs (6–11) and compare the results. Then the class of sequence X was determined by the maximum of η_v (v = beta-hairpin and not beta-hairpin). Finally, the coefficient R given in Eq (12) is used to correct predicted results. The current study utilized R<0.2. The results of three-fold cross-validation are listed in Table 2.

From the Table 2, we can see that the sensitivity, specificity and total accuracy are 92%, 94% and 87%, respectively, indicating that chemical shift is a good parameter for the beta-hairpin prediction.

Chemical shift is an easily obtained experimental datum. However, Chemical shift values of a sequence are not always complete for a multitude of reasons. Often, chemical shifts can only

Table 2. Results of different parameters using QDA (R<0.2).

Parameters	Sn	Sp	Acc	MCC
Six CSs	92%	94%	87%	0.85
20 AAC	36%	87%	32%	0.26

doi:10.1371/journal.pone.0139280.t002

Table 3. Predicted results by using the CSs of five nuclei (R<0.2).

Parameters	Sn	Sp	Acc	MCC
C_α,C_β,H_N,H_α,N(omit C)	98%	52%	83%	0.61
C,C_β,H_N,H_α,N(omit C_a)	94%	48%	80%	0.50
C,C_α,H_N,H_α,N(omit C_B)	87%	76%	83%	0.62
C,C_α,C_β,H_α,N(omit H_N)	100%	48%	83%	0.61
C,C_α,C_β,H_N,N(omit H_a)	94%	32%	74%	0.35
C,C_α,C_β,H_N,H_α(omit N)	100%	14%	72%	0.29

doi:10.1371/journal.pone.0139280.t003

be assigned partially or are missing. To assess the impact of incomplete chemical shift assignment and determine the importance of chemical shift of each nucleus, we performed the prediction by removing any one of the CSs six nuclei. Then, the CSs of combination of five nuclei can be seen as features to predict the beta-hairpin. The results are listed in Table 3.

In table 3, we can see that all results are affected compared with using six CSs as features when a CSs feature is left out. If all six CSs are used, we reach a prediction overall accuracy of 87% (see Table 2). The absence of one CS leads to a significant decrease in prediction accuracy ranging from 4% for missing C or C_B or H_N shifts to 15% for missing N shifts. It is strange that the overall accuracy is worst when the CS of N nuclei is left out. This illustrates that N is the most important feature for prediction the beta-hairpin. According to the overall accuracy, we rank as the importance as: $N>H_a>C_a>C>H_N>C_B$ in this paper.

Comparison with other feature

To test our method and facilitate comparison with other feature, we used 20 amino acid compositions (AAC) as inputs of the method of QDA. Notes: Where μ is a twenty-dimensional mean vector, and Σ_v denotes the 20×20 dimensional covariance matrix. The results are also recorded in Table 2. Compared results show that the performance of CSs is more superior to that of 20 AAC for the beta-hairpin prediction.

Comparison with other approaches

Some approaches have been developed for predicting the beta-hairpin motifs [7–10]. However, due to differences in database, it is difficult to directly compare our results with other published results. Here we examined the predicted performance of other algorithms by use of the same CSs of six nuclei as features. At present, the support vector machine (SVM) and random forest (RF) are arguably the most widely used classification techniques in the Life Sciences [43–46]. In this paper, we implemented the SVM and RF algorithm based on R software package. The results are all listed in Table 4.

Table 4 shows that QDA yields the best outcomes in using six CSs as feature. Therefore, we proposed using QDA to perform the beta-hairpin motifs prediction.

Table 4. The results of different approaches using the same six CSs information.

algorithm	Sn	Sp	Acc	MCC
QDA	**92%**	94%	**87%**	**0.85**
SVM	71%	**98%**	86%	0.75
RF	12%	**86%**	62%	0.28

doi:10.1371/journal.pone.0139280.t004

Conclusion

In this paper, we have introduced a model for predicting beta-hairpin motifs based on CSs. By the analysis of the statistical distributions of six-nuclei CSs in beta-hairpin and not beta-hairpin dataset, we found that the CSs of six nuclei are significantly different in beta-hairpin and not beta-hairpin motifs. Finally, we adopted three-fold cross-validation, and achieved the best prediction, namely the sensitivity (Sn) of 92%, the specificity (Sp) of 94%, the total accuracy (Acc) of 87% with 0.85 of Mathew's correlation coefficient (MCC) by using six CSs as features and the quadratic discriminant analysis. Results showed that chemical shift is indeed an effective parameter for the prediction of beta-hairpin motifs. Moreover, we have performed the prediction by combining the CSs of five different nuclei. Results showed that CSs of each nucleus has a different influence on the prediction of beta-hairpin structures. Our model is both simple and easy to perform. We hope this model will assist investigation the topology of protein structures in the near future [47–49]. As demonstrated in a series of recent publications [50–53] in developing new prediction methods, user-friendly and publicly accessible web-servers will significantly enhance their impacts [54], we shall make efforts in our future work to provide a web-server for the prediction method presented in this paper.

Supporting Information

S1 File. 123 proteins used in this paper.
(DOCX)

S2 File. CSs data of 157 beta-hairpin fragments.
(RAR)

S3 File. CSs data of 75 not beta-hairpin fragments.
(RAR)

S4 File. p-p plots of six nuclei.
(DOC)

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper.

Author Contributions

Conceived and designed the experiments: FYE. Performed the experiments: FYE KGS. Analyzed the data: FYE. Contributed reagents/materials/analysis tools: FYE KGS. Wrote the paper: FYE.

References

1. Bystro C, Thorsson V, Baker D. HMMSTR: a hidden markov model for local sequence structure correlations in proteins. J Mol Biol. 2000; 301(1): 173–90. PMID: 10926500
2. Burke DF, Deane CM. Improved protein loop prediction from sequence alone. Protein Eng. 2001; 14 (7):473–8. PMID: 11522920
3. Sun ZR, Rao X, Peng L, Xu D. Prediction of protein super secondary structures based on the artificial neural network method. Protein Eng. 1997; 10(7):763–9. PMID: 9342142
4. Chou KC. Prediction of beta-turns in proteins. J Pept Res. 1997; 49(2):120–44.
5. Chou KC, Blinn JR. Classification and prediction of beta-turn types. J Protein Chem. 1997; 16(6): 575–95. PMID: 9263121

6. de la Cruz X, Hutchinson EG, Shepherd A, Thornton JM. Toward predicting protein topology: an approach to identifying beta hairpins. Proc Natl Acad Sci,USA. 2002; 99(17): 11157–62. PMID: 12177429
7. Hu XZ, Li QZ, Wang CL. Recognition of β-hairpin motifs in proteins by using the composite vector. Amino Acids. 2010; 38: 915–21. doi: 10.1007/s00726-009-0299-7 PMID: 19418016
8. Kuhn M, Meiler J, Baker D. Strand-loop-strand motifs: prediction of hairpins and diverging turns in proteins. Proteins. 2004; 54(2): 282–8. PMID: 14696190
9. Kumar M, Bhasin M. Bhairpred: prediction of B-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res. 2005; 33: 154–9.
10. Hu XZ, Li QZ. Prediction of the B-hairpins in proteins using support vector machine. The Protein Journal. 2008; 27(2):115–22. PMID: 18071887
11. Saitô H, Ando I, Ramamoorthy A. Chemical shift tensor-the heart of NMR: Insights into biological aspects of proteins. Prog Nucl Magn Reson Spectrosc. 2010; 57(2): 181–228. doi: 10.1016/j.pnmrs.2010.04.005 PMID: 20633363
12. Lee DK, Wittebort RJ, Ramamoorthy A. Characterization of ^{15}N Chemical Shift and $^{1}H–^{15}N$ Dipolar Coupling Interactions in a Peptide Bond of Uniaxially Oriented and Polycrystalline Samples by One-Dimensional Dipolar Chemical Shift Solid-State NMR Spectroscopy. J Am Chem Soc. 1998; 120: 8868–74. doi: 10.1021/ja981599u
13. Poon A, Birn J, Ramamoorthy A. How Does an Amide-N Chemical Shift Tensor Vary in Peptides? J Phys Chem B. 2004; 108(42): 16577–85. PMID: 18449362
14. Brender JR, Taylor DM, Ramamoorthy A. Orientation of Amide-Nitrogen-15 Chemical Shift Tensors in Peptides: A Quantum Chemical Study. J Am Chem Soc. 2001; 123: 914–22. doi: 10.1021/ja001980q PMID: 11456625
15. Birn J, Poon A, Mao Y, Ramamoorthy A. Ab initio study of ^{13}Ca chemical shift anisotropy tensors in peptides. J Am Chem Soc. 2004; 126(27): 8529–34. doi: 10.1021/ja049879z PMID: 15238010
16. Case DA. The use of chemical shifts and their anisotropies in biomolecular structure determination. Curr Opin Struct Biol. 1998; 8: 624–30. PMID: 9818268
17. Wishart DS, Case DA. Use of chemical shifts in macromolecular structure determination. Methods Enzymol. 2001; 338: 3–34. PMID: 11460554
18. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA. 2007; 104: 9615–20. PMID: 17535901
19. Lin H, Ding C, Song Q, Yang P, Ding H, Deng KJ, et al. The predictin of protein structural class using averaged chemical shifts. J Biomolecular Struc and Dynamics. 2012; 29(6): 643–9.
20. Mao WS, Cong PS, Wang ZH, Lu LJ, Zhu ZL, Li TH. NMRDSP: An accurate prediction of protein shape strings from NMR chemical shifts and sequence data. PLoS ONE. 2013; 8(12): e83532. doi: 10.1371/journal.pone.0083532 PMID: 24376713
21. Martin M, Michael H. A probabilistic model for secondary structure prediction from protein chemical shifts. Proteins. 2013; 81(6): 984–99. doi: 10.1002/prot.24249 PMID: 23292699
22. Mielke SP, Drishnan VV. Protein structural class identification directly from NMR spectra using average chemical shifts. Bioinformatics. 2003; 19(16): 2054–64. PMID: 14594710
23. Pastore A, Saudek V. The relationship between chemical shift and secondary structure in proteins. J Magn Reson. 1990; 90:165–76.
24. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, et al. Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA. 2008; 105(12): 4685–90. doi: 10.1073/pnas.0800256105 PMID: 18326625
25. Wang Y. Secondary structure effects on protein NMR chemical shifts. J Biomol NMR. 2004; 30(3): 233–44. PMID: 15754052
26. Shen Y, Bax A. Identification of helix capping and beta-turn motifs from NMR chemical shifts. J Biomol NMR. 2012; 52(3): 211–32. doi: 10.1007/s10858-012-9602-0 PMID: 22314702
27. Shen Y, Bax A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR. 2013; 56(3): 227–41. doi: 10.1007/s10858-013-9741-y PMID: 23728592
28. Zhang H, Neal S, Wishart DS. RefDB: A database of uniformly referenced protein chemical shifts. J Biomol NMR. 2003; 25(3):173–95. PMID: 12652131
29. Fernandez-Fuentes N, Hermoso A, Espadaler J, Querol E, Aviles FX, Oliva B. Classification of common functional loops of kinase super-families. Proteins. 2004; 56(3): 539–55. PMID: 15229886
30. Wang G, Dunbrack RJ. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res. 2005; 33: w94–98. PMID: 15980589

31. Sprinthall RC. Basic statistical analysis. 7th ed. Boston, MA: Pearson Education Group; 2003.
32. Feng YE, Lin H, Luo LF. Prediction of protein secondary structure using feature selection and analysis approach. Acta Biotheoretica. 2014; 62(1):1–14. doi: 10.1007/s10441-013-9203-7 PMID: 24052343
33. Feng YE, Luo LF. Use of tetrapeptide signals for protein secondary structure prediction. Amino acids. 2008; 35(3): 607–14. doi: 10.1007/s00726-008-0089-7 PMID: 18431531
34. Feng YE. Prediction of four kinds of simple super secondary structures in Protein by using chemical shifts. Scientific world journal, 2014, 978503. doi: 10.1155/2014/978503 PMID: 25050407
35. Kou GS, Feng YE. Identify five kinds of simple super secondary structures with quadratic discriminant algorithm based on the chemical shifts. J Theor Biol. 2015; 380: 392–8. doi: 10.1016/j.jtbi.2015.06.006 PMID: 26087283
36. Chou KC, Shen HB. Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms. Nat Protocol. 2008; 3(2): 153–62.
37. Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 2013; 41(6): e68. doi: 10.1093/nar/gks1450 PMID: 23303794
38. Esmaeili M, Mohabatkar H, Mohsenzadeh S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol. 2010; 263(2): 203–9. doi: 10.1016/j.jtbi.2009.11.016 PMID: 19961864
39. Hayat M, Khan A. Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC. Protein Pept Lett. 2012; 19(4): 411–21. PMID: 22185508
40. Lin H, Chen W, Yuan LF, Ding H. Using over-represented tetrapeptides to predict protein submitochondria locations. Acta biotheoretica. 2013; 61(2): 259–68. doi: 10.1007/s10441-013-9181-9 PMID: 23475502
41. Xiao X, Wang P, Lin WZ, Jia JH, Chou KC. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem. 2013; 436(2): 168–77. doi: 10.1016/j.ab.2013.01.019 PMID: 23395824
42. Liang JJ, Pan WS, Yang ZH. Characterization-based Q-Q plots for testing multinormality. Stat. Probabil. Lett. 2004; 70: 183–90.
43. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 2008; 9: 319. doi: 10.1186/1471-2105-9-319 PMID: 18647401
44. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res. 2007; 35: W339–44. PMID: 17553836
45. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012; 99: 323–29. doi: 10.1016/j.ygeno.2012.04.003 PMID: 22546560
46. Goldstein BA, Polley EC, Briggs FBS. Random forests for genetic association studies. Stat Appl Genet Mol Biol. 2011; 10: 1–34.
47. Ramamoorthy A, Wei Y, Lee DK. PISEMA solid-state NMR spectroscopy. Ann Rep NMR Spectrosc. 2004; 52: 1–52.
48. Dürr HN Ulrich, Yamamoto K, Im SC, Waskell L, Ramamoorthy A. Solid-State NMR Reveals Structural and Dynamical Properties of a Membrane-Anchored Electron-Carrier Protein, Cytochrome b5. J Am Chem Soc. 2007; 129: 6670–71. doi: 10.1021/ja069028m PMID: 17488074
49. Yang JS, Kim J, Park S, Jeon J, Shin YE, Kim S. Spatial and functional organization of mitochondrial protein network. Scientific Reports 3, 2013; Article number: 1403.
50. Xu Y, Wen X, Wen LS, Wu LY. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE. 2014; 9: e105018. doi: 10.1371/journal.pone.0105018 PMID: 25121969
51. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J. Theor. Biol. 2011; 273: 236–47.
52. Guo SH, Deng EZ, Xu LQ, Ding H. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics, 2014; 30: 1522–29. doi: 10.1093/bioinformatics/btu083 PMID: 24504871
53. Liu Z, Xiao X, Qiu WR. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem., 2015; 474: 69–77. doi: 10.1016/j.ab.2014.12.009 PMID: 25596338
54. Chou KC. Impacts of bioinformatics to medicinal chemistry. Medicinal Chemistry. 2015; 11: 218–34. PMID: 25548930