Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 864135, 9 pages
http://dx.doi.org/10.1155/2014/864135

Research Article

acACS: Improving the Prediction Accuracy of Protein Subcellular Locations and Protein Classification by Incorporating the Average Chemical Shifts Composition

Guo-Liang Fan,[1] Yan-Ling Liu,[1] Yong-Chun Zuo,[2] Han-Xue Mei,[1] Yi Rang,[1] Bao-Yan Hou,[1] and Yan Zhao[1]

[1] *Department of Physics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China*
[2] *The Key Laboratory of Mammalian Reproductive Biology and Biotechnology of the Ministry of Education, College of Life Sciences, Inner Mongolia University, Hohhot 010021, China*

Correspondence should be addressed to Guo-Liang Fan; eeguoliangfan@sina.com and Yong-Chun Zuo; yczuo@imu.edu.cn

Received 19 May 2014; Revised 15 June 2014; Accepted 16 June 2014; Published 2 July 2014

Academic Editor: Hao Lin

The chemical shift is sensitive to changes in the local environments and can report the structural changes. The structure information of a protein can be represented by the average chemical shifts (ACS) composition, which has been broadly applied for enhancing the prediction accuracy in protein subcellular locations and protein classification. However, different kinds of ACS composition can solve different problems. We established an online web server named acACS, which can convert secondary structure into average chemical shift and then compose the vector for representing a protein by using the algorithm of auto covariance. Our solution is easy to use and can meet the needs of users.

1. Introduction

Knowledge of subcellular localization information of a protein may help to unravel its normal cellular function [1]. The proteins within the different compartments have different biological activity and functions; in turn, knowing the subcellular localization of a given protein helps in elucidating its functional role.

Recently, many computational approaches for subcellular localization predictions have been developed and plenty of methods for improving the accuracy of the prediction were applied. From two aspects the predictor can be described. One is the predicting algorithms, like support vector machine (SVM) [2–11], neural network [12], increment of diversity (ID) [13], random forest (RF) [14], K-nearest neighbor (K-NN) [15, 16], generating algorithm [17], and so on, or the combination of them [16, 18]. The other is the information source, such as widely used sequence-based information source, which are amino acid composition (AAC) and sorting signals [19–21], and textual descriptions of proteins [22, 23], which are protein physiochemical property [24], gene ontology (GO) [25], and so on. Actually, the structure information of a protein is very important, especially when it is used for representing the subcellular locations of a protein. However, the structure information of a protein cannot be easily described, and few methods using the structure information can be learned to our knowledge.

However, in NMR spectroscopy, as an important parameter, chemical shift, which is sensitive to changes in the local environments, can report the structural changes. Sibley et al. [26], Mielke and Krishnan [27], Spera and Bax [28], and Zhao et al. [29] have found that the ACS of a protein has intrinsic correlation with the protein's secondary structure and the function of this protein is determined by its structure. According to this point of view, there must be some relationship among the averaged chemical shift, protein structure, and functions [30, 31]. Wishart has developed a web server, namely, CS23D, for rapidly generating accurate 3D protein structures using only assigned NMR chemical shifts [32]. More than 100 proteins from BMRB [33] were tested

and found that the resulting structures generally exhibit good geometry and chemical shift agreement [32]. Also, there are some algorithms, which can predict the chemical shift from protein sequences and conformation [34–37]; few works have been done to determine a protein's functions by the chemical shifts [38, 39]. Therefore, how to use the chemical shift is still important and urgent.

In this paper, a benchmark data set of chemical shift was constructed, which consists of 1,552 proteins derived from BMRB website [33] and then extracted chemical shift values of ^{15}N, $^{13}C_\alpha$, $^{1}H_\alpha$, and $^{1}H_N$ for 20 amino acid residues. Then four types of average chemical shift for 20 amino acid residues were calculated and the autocovariance algorithm was used to convert the average chemical shift into the vector to describe the protein sample. The algorithm acACS (autocovariance of averaged chemical shifts) has been used to enhance the prediction accuracy in protein subcellular locations. The proposed acACS descriptor can be considered as a mode of generalized pseudoamino acid composition, which was summarized in [40]. Recently, the generalized pseudoamino acid composition methods have been systematically implemented by two powerful software, PseAAC-Builder [41] and PseAAC-General [42]. For the readers' convenience in using the current method, the acACS descriptor may be integrated into this software in future works. The details of how to deal with this calculation and how to use this method is shown as follows.

2. Material and Methods

2.1. Data Sets. When an electron moves around a proton, it will produce some magnetic field, which could affect proton's external electron field. Thus, the absorption frequencies of proton in different chemical environments would shift relatively to the absorption frequencies under standard magnetic fields. Chemical shift is the relative resonance frequencies shift of protons between different chemical environment and standard, which can be measured by NMR spectroscopy. Due to its sensitivity to local environments, such as the backbone dihedral angles and the secondary structure types [26, 27, 29], chemical shift can be an indicator for the changes of local conformations.

In order to find out the correlation between chemical shift and the secondary structure of a protein, we construct a high-quality working data set, which started from the following steps: (1) the proteins star file with NMR spectroscopy data were downloaded from BMRB [33]; (2) the proteins less than 50 residues or not matched to PDB [43] entries were discarded; (3) the proteins with sequence identity higher than 40% were excluded by CD-HIT [44]. Finally, the benchmark data set has 1,552 proteins. The data set was available at our website. The data set contained 1,552 proteins sequences and BMRB star file, which was the original chemical shifts data file for all kinds of backbone atoms of each protein. We analyzed the averaged chemical shifts for every kind of amino acids type and secondary structure in order to find out the rules among averaged chemical shifts with every kind of amino acids type and secondary structure types and then used the autocovariance algorithm to calculate the feature vectors of the protein sequences from the statistic results. The feature vectors representing the protein sequences can be used in problems of subcellular location prediction or other protein classifications. Researchers may also develop better algorithms for protein representation using the data set.

2.2. Averaged Chemical Shift (ACS). In order to find the rule between the chemical shifts and structure information, the statistic about averaged chemical shift related to secondary structure and amino acids type was carried out.

Firstly, four types chemical shift values ω of ^{15}N, $^{13}C_\alpha$, $^{1}H_\alpha$, and $^{1}H_N$ from every amino acid residue were extracted from the BMRB star file for further calculation. In the BMRB star file, the amino acid residues, four kinds of protein backbone atoms of each amino acid residue, and matched PDB file were given. For example, the "bmr447.str" was extracted into four files: N_447.txt, Ca_447.txt, Ha_447.txt, and Hn_447.txt, which correspond to ^{15}N, $^{13}C_\alpha$, $^{1}H_\alpha$, and $^{1}H_N$ protein backbone atoms.

Secondly, the secondary structure information was extracted from PDB file which matched to BMRB star file. The secondary structure types of each amino acid residue are denoted by H, E, and C. Then the averaged chemical shifts for all the residues were calculated.

For protein backbone atoms "i" of amino acid type "j" with secondary structure type "k," the averaged chemical shift (ACS) is defined as

$$C_k^i(j) = \frac{1}{N}\sum_N \omega_k^i(j). \quad (1)$$

Here $i = {}^{15}$N, $^{13}C_\alpha$, $^{1}H_\alpha$, or $^{1}H_N$, j is one kind of 20 amino acids and k stands for the secondary structure types (H, E, or C) from DSSP [45] (H = helix, E = strand, and C = the rest). $\omega_k^i(j)$ is the chemical shift value extracted from the BMRB star file and N is the counts of $\omega_k^i(j)$ items.

By calculating the residues' ACS with (1) for 1552 proteins, we found that the ACS regularly varies with the secondary structure types and residues. The statistic results of averaged chemical shifts were listed in four tables, which can be accessed from our website. Take the $^{1}H_\alpha$ as an example, the ACS of $^{1}H_\alpha$ for each of 20 native amino acid residues with three types of secondary structure is shown in Figure 1. According to Figure 1, it can be concluded that we can use the ACS to represent the protein's secondary structure. In order to illustrate the algorithm, the flowchart of ACS is given in Figure 2.

2.3. Algorithm of Autocovariance of Average Chemical Shift (acACS). In order to obtain the correlation information between amino acids of a protein, the autocovariance of ACS was calculated. For a protein P,

$$P = [j_1, j_2 \cdots j_l \cdots j_L]. \quad (2)$$

Here, L is the sequence length and j_l is the amino acid in position l.

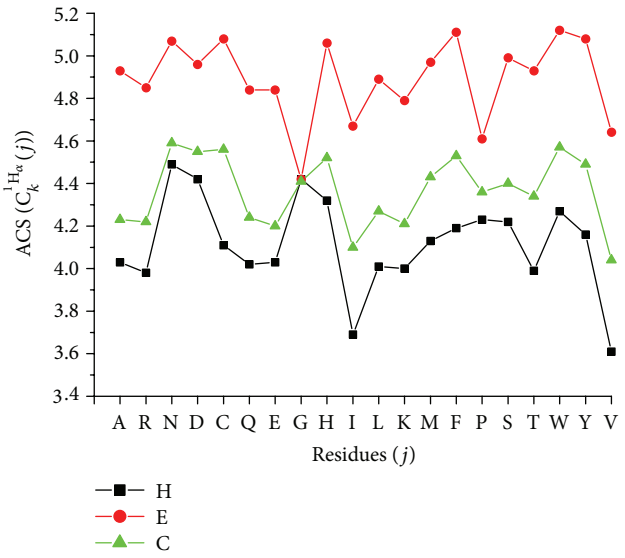


Figure 1: The average chemical shifts (ACS) of $^1H_\alpha$ for each of 20 native amino acid residues (j) with three types of secondary structure (k).

The secondary structure of protein P was predicted from Porter [46, 47] and then

$$P = [k_1, k_2 \cdots k_l \cdots k_L]. \quad (3)$$

Here k is the secondary structure types.

Then, the amino acid j_l in protein P was replaced by its ACS "$C^i_{k_l}(j_l)$" according to its secondary structure type k_l. When $C^i_{k_l}(j_l)$ was redefined as S^i_l, P can be expressed as

$$P = \left[S^i_1, S^i_2 \cdots S^i_l \cdots S^i_L\right] \quad \left(i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^1\mathrm{H}_\alpha, {}^1\mathrm{H}_N\right). \quad (4)$$

Then, the autocovariance algorithm was used to calculate the correlation between amino acid l and $l+\lambda$ by the following equation:

$$\theta^i(\lambda) = \frac{1}{L-\lambda}\sum_{l=1}^{L-\lambda}\left[S^i_l - S^i_{(l+\lambda)}\right]^2, \quad \left(i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^1\mathrm{H}_\alpha, {}^1\mathrm{H}_N,\ 0 < \lambda < L\right). \quad (5)$$

After the above calculation, the protein P can be expressed as follows:

$$P_{\mathrm{acACS}} = \left[\theta^i(0), \theta^i(1), \theta^i(2), \theta^i(3), \ldots, \theta^i(\lambda); \ldots\right] \quad \left(i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^1\mathrm{H}_\alpha, {}^1\mathrm{H}_N,\ 0 < \lambda < L\right). \quad (6)$$

Here, $\theta^i(\lambda)$ is the correlation factor of average chemical shift S^i_l with average chemical shift $S^i_{l+\lambda}$. In particular, when $\lambda = 0$, with (5), $\theta^i(0) = 0$. In order to take use of ACS, the $\theta^i(0)$ was replaced by the average chemical shift S^i_l. The factor λ is a nonnegative integer and reflects the rank of correlation [40]. Based on different problems, in order to get a best result, a certain right number for factor λ should be given and so does i.

In order to give a pictorial representation of chemical shifting technique, a flow diagram is given in Figure 3, which shows how the acACS works.

Table 1: The comparison of the results with and without the acACS for predicting submitochondria locations and three membrane protein types with comparison to that without acACS.

	With acACS	Without acACS
Submitochondria locations	93.57%	91.46%
Three membrane protein types	97.79%	96.10%
Data set of Du [24]	94.95%	93.43%

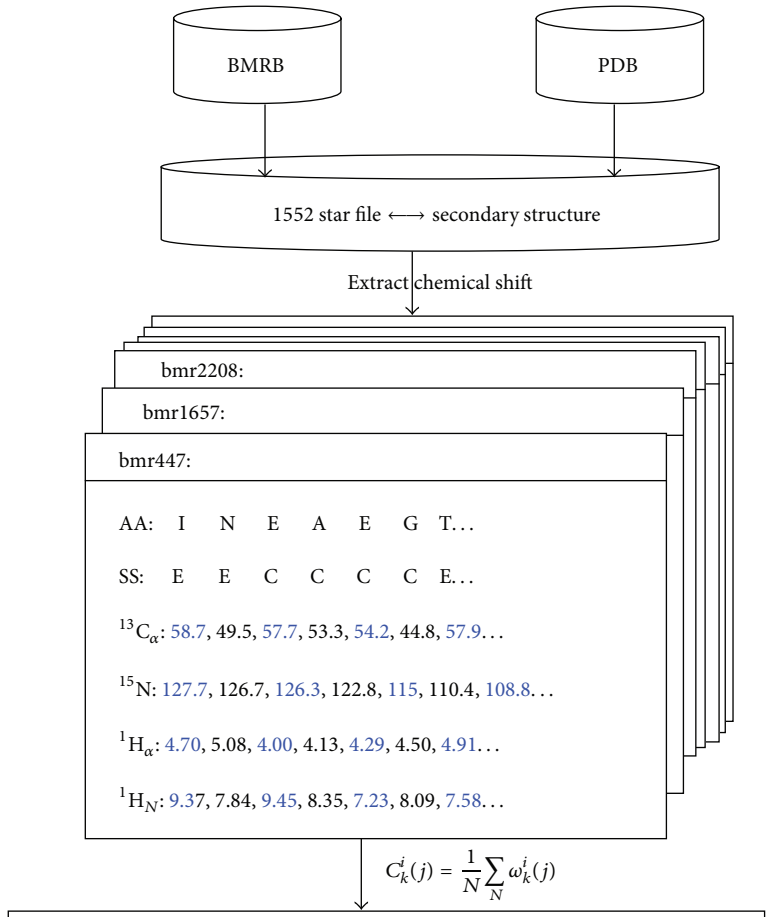
Table 2: The comparison of the results with and without the acACS for predicting mycobacterial subcellular localizations and three membrane protein types.

	With acACS	Without acACS
Mycobacterial subcellular localizations	87.77%	86.19%
Three membrane protein types	85.03%	83.71%
Data set of Rashid [53]	98.12%	96.85%

3. Results and Discussion

By using the acACS algorithm, we successfully represented the protein samples and accurately predicted submitochondria locations. We used the model to test the SML3-983 data set that was along with the SubMito-PSPCP [48]. The data set has 983 proteins sequences which were divided into three locations. Among the data set, there are 661 sequences from inner membrane, 177 sequences from matrix, and 145 sequences from outer membrane. We selected acACS combined with AAC, DC, PSSM, and GO and reduced physicochemical properties (Hn) as feature vectors for representing the proteins and then trained the model. Then 90.74% accuracy was obtained for SML3-983 data set with Jackknife cross-validation, which was 1.63% higher than SubMito-PSPCP. In order to compare the performance of acACS, the feature vector was recombined with AAC, DC, PSSM, GO, and Hn, without acACS. Then we trained the model and obtained the predicting accuracy of 89.52%, which was dropped about 1.2%.

The acACS algorithm has also been checked in our previous works [49–52]. In subcellular location prediction, we compared the results with and without the acACS in the submitochondria locations and mycobacterial proteins subcellular locations and got the better result which was listed in Tables 1 and 2. Actually, the acACS as a feature vector for representing the protein samples can also be used for other kinds of proteins prediction problem. In acidic and alkaline enzymes prediction and bioluminescent and nonbioluminescent proteins discrimination, we also improved the predicting accuracy by about 1.3%, which was listed in Table 3.



Average chemical shift for atoms $^{1}H_\alpha$

Secondary structure / Amino acids	H	E	C
A	4.03	4.93	4.23
R	3.98	4.85	4.22
N	4.49	5.07	4.59
D	4.42	4.96	4.55
C	4.11	5.08	4.56
Q	4.02	4.84	4.24
E	4.03	4.84	4.20
G	4.42	4.42	4.41
H	4.32	5.06	4.52
I	3.69	4.67	4.10
L	4.01	4.89	4.27
K	4.00	4.79	4.21
M	4.13	4.97	4.43
F	4.19	5.11	4.53
P	4.23	4.61	4.36
S	4.22	4.99	4.40
T	3.99	4.93	4.34
W	4.27	5.12	4.57
Y	4.16	5.08	4.49
V	3.61	4.64	4.04

Average chemical shift for atoms ^{15}N

Secondary structure / Amino acids	H	E	C
A	121.83	124.78	122.97
R	119.43	122.65	119.75
N	117.82	121.29	118.14
D	119.67	123.03	119.49
C	117.74	120.94	118.71
Q	118.70	122.13	118.87
E	119.55	122.57	119.98
G	107.82	109.81	110.24
H	118.51	121.68	118.97
I	119.80	123.06	119.65
L	119.97	124.75	120.61
K	119.45	123.14	120.16
M	118.67	122.23	119.46
F	119.71	121.54	118.69
P	119.85	119.40	119.76
S	115.32	117.45	116.02
T	115.08	117.59	113.11
W	120.77	122.80	119.33
Y	119.75	121.95	118.96
V	119.39	122.45	119.28

Average chemical shift for atoms $^{13}C_\alpha$

Secondary structure / Amino acids	H	E	C
A	54.70	50.92	52.94
R	58.92	54.88	56.50
N	55.35	52.46	53.65
D	56.74	53.45	54.64
C	61.74	56.14	57.57
Q	58.39	54.47	56.41
E	58.94	55.04	57.17
G	46.80	45.17	45.72
H	58.67	55.03	56.40
I	64.29	59.84	61.34
L	57.41	53.86	55.29
K	58.85	55.04	56.74
M	57.88	54.34	55.73
F	60.48	56.41	57.65
P	65.03	63.00	63.21
S	60.83	57.07	58.66
T	65.43	60.96	61.83
W	60.00	56.19	56.98
Y	60.63	56.58	57.90
V	65.83	60.66	62.43

Average chemical shift for atoms $^{1}H_N$

Secondary structure / Amino acids	H	E	C
A	8.11	8.63	8.16
R	8.10	8.65	8.18
N	8.54	8.60	8.38
D	8.22	8.53	8.34
C	8.15	8.67	8.12
Q	8.09	8.61	8.20
E	8.25	8.57	8.40
G	8.29	8.37	8.39
H	8.10	8.69	8.23
I	8.04	8.72	7.98
L	8.08	8.74	8.03
K	8.00	8.58	8.15
M	8.09	8.69	8.15
F	8.17	8.78	8.04
P	8.28	8.40	8.30
S	8.16	8.57	8.23
T	8.07	8.58	8.07
W	8.25	8.72	7.81
Y	8.12	8.76	8.07
V	7.99	8.68	8.00

Averaged chemical shifts

Figure 2: The flowchart of calculating the ACS. The AA denotes the amino acids and the SS denotes the secondary structure.

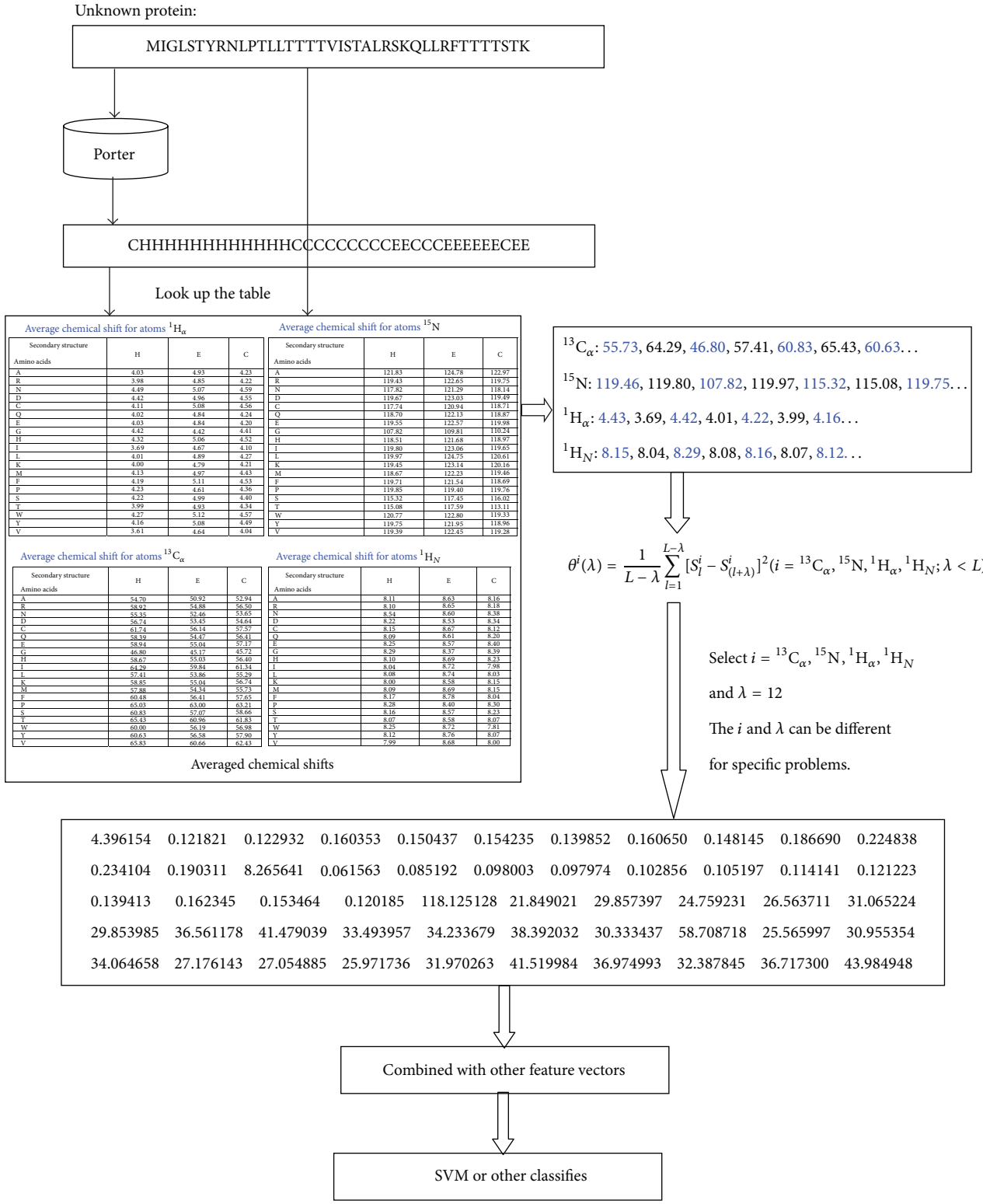


Figure 3: The flow diagram of the processing of the acACS.

Table 3: The comparison of the results with and without the acACS for other kinds of proteins prediction.

	With acACS	Without acACS
Acidic and alkaline enzymes	94.01%	92.52%
Bioluminescent and nonbioluminescent proteins	82.16%	80.90%

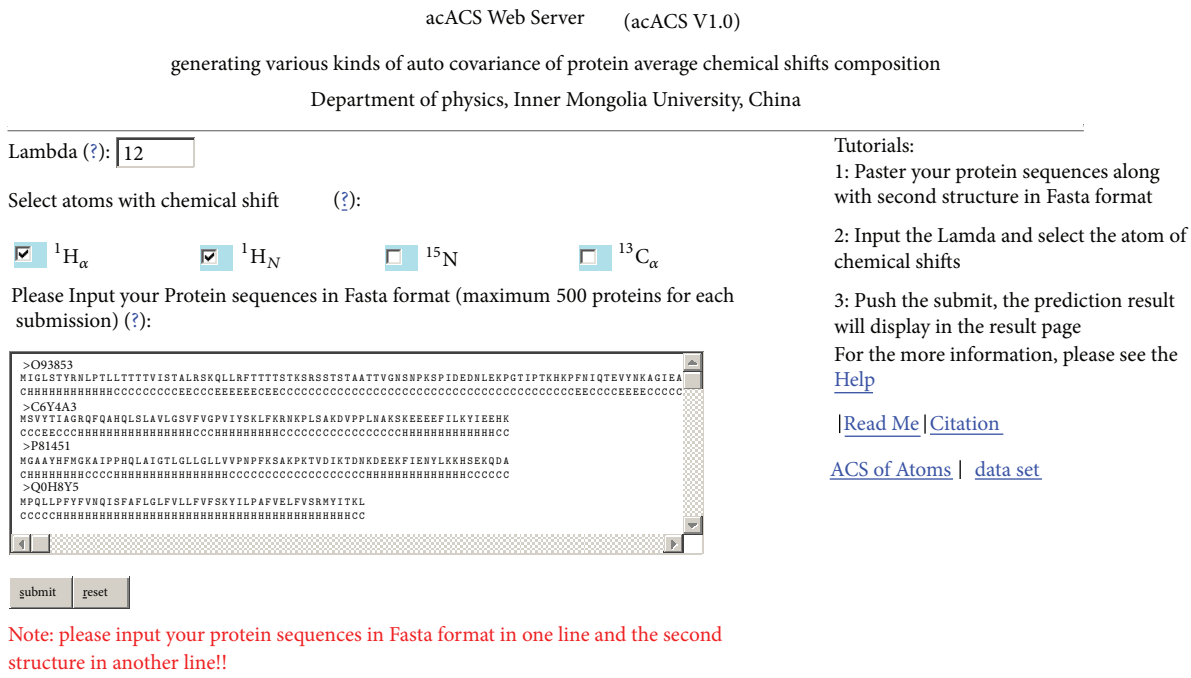


Figure 4: A screenshot for the top page of the web server acACS at http://wlxy.imu.edu.cn/college/biostation/fuwu/acACS/index.asp.

The protein functions, including its subcellular locations, are largely determined by its structure. Developing a novel method for improving the performance of predicting protein subcellular locations is urgent. However, the feature vectors in the methods were almost sequence-based in the past. Therefore, almost state-of-the-art methods tried to incorporate some other sequence-based information as its complement. Our method provides structure-based information and can be perfect complement to the sequence-based methods and can be used for other kinds of protein related problems. Actually, these methods can work side by side to help each other in a practical study.

For the chemical shift, it incorporates the structure information in the first place, so it can represent the protein sample better. What is the better ways to use of the chemical shift is still a hot topic for biologist and chemist. In this work, we used the autocovariance algorithm to process the averaged chemical shift and got the better results, but there are certainly some improvements that could be made for acACS. Actually other algorithms can be adopted to try to find the better method for representing the protein samples in the future. At present stage, it is not convenient for the user, for both the secondary structure information and the protein sequence that are used, to calculate the chemical shift. In the future, the secondary structure information will not be necessary, and it will be integrated into the algorithm.

4. Conclusions

In this work, the raw chemical shifts data set and averaged chemical shift data set were constructed. Then, the averaged chemical shift was calculated and the algorithm of acACS was presented. In order to check the performance of the acACS we got, proteins submitochondria locations, mycobacterial proteins subcellular locations, bioluminescent proteins discrimination, and acidic and alkaline enzymes classification were predicted. Based on the results we obtain, it can be concluded that the acACS can improve the accuracy of prediction at least 1%-2%, the performance of which is correlated with the correlation factor λ and the backbone atoms i. Some recent studies showed that the profile-based features [54–56], pseudoamino acid composition (PseAAC) [57], and features based on physicochemical proprieties of amino acids [58] were able to improve the performance of many computational predictors for protein remote homology detection, protein binding site identification, and so forth. Therefore, these features should be studied for protein subcellular location prediction in the future studies.

We have developed a web server acACS, which could automatically produce the vectors of proteins, when a custom submitted the protein sequences along with the secondary structure in batch mode. The data set can be a very useful addition to biomolecular NMR spectroscopists. The acACS will be of benefit to the proteomic research. The current work will become an important progress in the prediction of the protein subcellular locations and promote the study in the related areas.

5. Web Server and User Guide

To enhance the value of its practical applications, a web server for the acACS generator was established. Moreover, for the convenience of the user, here a step-to-step guide is provided for how to use the web server to get the desired results.

Step 1. Open the web server at http://wlxy.imu.edu.cn/college/biostation/fuwu/acACS/index.asp and you will see the top page of the acACS on your computer screen, as shown

in Figure 4. Click on the Read Me button to see a brief introduction about the acACS.

Step 2. Either type or copy/paste the query protein sequences into the input box at the center of Figure 4, and then copy/paste the secondary structure of the protein sequence in the next line. The input sequence should be in "ONE LINE" format. For the examples of sequences in ONE LINE format, click the "?" button above the input box.

Step 3. Input the Lambda value in the input box right of the Lambda label.

Step 4. Check atoms with chemical shift.

Step 5. Click on the Submit button to see the result page. For example, if you use the default example sequences, Lambda and atoms in the window, after clicking the Submit button, you will see the following message shown on the screen of your computer: "The lamda you have chosen is 12"; "The Atom of chemical shift you have chosen are $^1H_\alpha$, ^1H_N"; "The acACSs of the proteins you submitted are.....". Then the acACS of $^1H_\alpha$ atom was given and the acACS of ^1H_N atom followed for the first protein, then the acACS of second protein, the third, and so forth.

Step 6. Click the ACS of atoms and data set button to download the benchmark dataset used to calculate the ACS.

Step 7. Click the Citation button to find the relevant papers that document the detailed development and algorithm of acACS.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the reviewers for their helpful comments on their paper. This work was supported by a Grants from National Natural Science Foundation of China (61063016 and 31160188), The Scientific Research Program at Universities of Inner Mongolia Autonomous Region of China (NJZY13014), The Natural Science Foundation of Inner Mongolia Autonomous Region of China (2013MS0504 and 2013MS0503), and the Program of Higher-level Talents of Inner Mongolia University (135147).

References

[1] R. Casadio, P. L. Martelli, and A. Pierleoni, "The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation," *Briefings in Functional Genomics and Proteomics*, vol. 7, no. 1, pp. 63–73, 2008.

[2] J.-Y. Shi, S.-W. Zhang, Q. Pan, Y.-M. Cheng, and J. Xie, "Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition," *Amino Acids*, vol. 33, no. 1, pp. 69–74, 2007.

[3] M. R. Bakhtiarizadeh, M. Moradi-Shahrbabak, M. Ebrahimi, and E. Ebrahimie, "Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology," *Journal of Theoretical Biology*, vol. 356, pp. 213–222, 2014.

[4] T. D. Campos, N. D. Young, P. K. Korhonen et al., "Identification of G protein-coupled receptors in *Schistosoma haematobium* and *S. mansoni* by comparative genomics," *Parasites & Vectors*, vol. 7, no. 1, article 242, 2014.

[5] Y. L. Chen and Q. Z. Li, "Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 248, no. 2, pp. 377–381, 2007.

[6] H. Ding, S. Guo, E. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[7] S. Mondal and P. P. Pai, "Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction," *Journal of Theoretical Biology*, vol. 356, pp. 30–35, 2014.

[8] L. Zhang, B. Liao, D. Li, and W. Zhu, "A novel representation for apoptosis protein subcellular localization prediction using support vector machine," *Journal of Theoretical Biology*, vol. 259, no. 2, pp. 361–365, 2009.

[9] Y. C. Zuo, Y. Peng, L. Liu, W. Chen, L. Yang, and G. L. Fan, "Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns," *Analytical Biochemistry*, vol. 458, pp. 14–19, 2014.

[10] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis," *Molecular Biosystems*, 2014.

[11] H. Ding, E.-Z. Deng, L.-F. Yuan et al., "iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels," *BioMed Research International*, vol. 2014, Article ID 286419, 10 pages, 2014.

[12] Y. D. Cai and K. C. Chou, "Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins," *Molecular Cell Biology Research Communications*, vol. 4, no. 3, pp. 172–173, 2000.

[13] Q. Z. Li and Z. Q. Lu, "The prediction of the structural class of protein: application of the measure of diversity," *Journal of Theoretical Biology*, vol. 213, no. 3, pp. 493–502, 2001.

[14] Y. Jin, B. Niu, K. Feng, W. Lu, Y. Cai, and G. Li, "Predicting subcellular localization with AdaBoost learner," *Protein and Peptide Letters*, vol. 15, no. 3, pp. 286–289, 2008.

[15] Y. D. Cai and K. C. Chou, "Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition," *Biochemical and Biophysical Research Communications*, vol. 305, no. 2, pp. 407–411, 2003.

[16] K. C. Chou and H. B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers," *Journal of Proteome Research*, vol. 5, no. 8, pp. 1888–1897, 2006.

[17] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, 2008.

[18] K.-C. Chou and H.-B. Shen, "Predicting protein subcellular location by fusing multiple classifiers," *Journal of Cellular Biochemistry*, vol. 99, no. 2, pp. 517–527, 2006.

[19] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP and related tools," *Nature Protocols*, vol. 2, no. 4, pp. 953–971, 2007.

[20] A. Höglund, P. Dönnes, T. Blum, H. Adolph, and O. Kohlbacher, "MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition," *Bioinformatics*, vol. 22, no. 10, pp. 1158–1165, 2006.

[21] P. Horton, K. Park, T. Obayashi et al., "WoLF PSORT: protein localization predictor," *Nucleic Acids Research*, vol. 35, no. 2, pp. W585–W587, 2007.

[22] S. Brady and H. Shatkay, "EPILOC: a (working) text-based system for predicting protein subcellular location," in *Proceedings of the 13th Pacific Symposium on Biocomputing (PSB '08)*, pp. 604–615, January 2008.

[23] A. Fyshe, Y. Liu, D. Szafron, R. Greiner, and P. Lu, "Improving subcellular localization prediction using text classification and the gene ontology," *Bioinformatics*, vol. 24, no. 21, pp. 2512–2517, 2008.

[24] P. Du and Y. Li, "Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence," *BMC Bioinformatics*, vol. 7, article 518, 2006.

[25] K. C. Chou and Y. D. Cai, "Prediction of protein subcellular locations by GO-FunD-PseAA predictor," *Biochemical and Biophysical Research Communications*, vol. 320, no. 4, pp. 1236–1239, 2004.

[26] A. B. Sibley, M. Cosman, and V. V. Krishnan, "An empirical correlation between secondary structure content and averaged chemical shifts in proteins," *Biophysical Journal*, vol. 84, no. 2 I, pp. 1223–1227, 2003.

[27] S. P. Mielke and V. V. Krishnan, "Protein structural class identification directly from NMR spectra using averaged chemical shifts," *Bioinformatics*, vol. 19, no. 16, pp. 2054–2064, 2003.

[28] S. Spera and A. Bax, "Empirical correlation between protein backbone conformation and C.alpha. and C.beta. 13C nuclear magnetic resonance chemical shifts," *Journal of the American Chemical Society*, vol. 113, no. 14, pp. 5490–5492, 1991.

[29] Y. Zhao, B. Alipanahi, S. C. Li, and M. Li, "Protein secondary structure prediction using NMR chemical shift data," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 5, pp. 867–884, 2010.

[30] P. Luginbühl, T. Szyperski, and K. Wüthrich, "Statistical basis for the use of $^{13}C^{\alpha}$ chemical shifts in protein structure determination," *Journal of Magnetic Resonance B*, vol. 109, no. 2, pp. 229–233, 1995.

[31] D. S. Wishart, B. D. Sykes, and F. M. Richards, "Relationship between nuclear magnetic resonance chemical shift and protein secondary structure," *Journal of Molecular Biology*, vol. 222, no. 2, pp. 311–333, 1991.

[32] D. S. Wishart, D. Arndt, M. Berjanskii, P. Tang, J. Zhou, and G. Lin, "CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data.," *Nucleic acids research*, vol. 36, pp. W496–502, 2008.

[33] B. R. Seavey, E. A. Farr, W. M. Westler, and J. L. Markley, "A relational database for sequence-specific protein NMR data," *Journal of Biomolecular NMR*, vol. 1, no. 3, pp. 217–236, 1991.

[34] Y. Shen and A. Bax, "Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology," *Journal of Biomolecular NMR*, vol. 38, no. 4, pp. 289–302, 2007.

[35] B. Han, Y. Liu, S. W. Ginzinger, and D. S. Wishart, "SHIFTX2: significantly improved protein chemical shift prediction," *Journal of Biomolecular NMR*, vol. 50, no. 1, pp. 43–57, 2011.

[36] Y. Tian, S. J. Opella, and F. M. Marassi, "Improved chemical shift prediction by Rosetta conformational sampling," *Journal of Biomolecular NMR*, vol. 54, no. 3, pp. 237–243, 2012.

[37] J. A. Vila, M. E. Villegas, H. A. Baldoni, and H. A. Scheraga, "Predicting 13Cα chemical shifts for validation of protein structures," *Journal of Biomolecular NMR*, vol. 38, no. 3, pp. 221–235, 2007.

[38] C. J. Markin and L. Spyracopoulos, "Accuracy and precision of protein-ligand interaction kinetics determined from chemical shift titrations," *Journal of Biomolecular NMR*, vol. 54, no. 4, pp. 355–376, 2012.

[39] C. J. Markin and L. Spyracopoulos, "Increased precision for analysis of protein-ligand dissociation constants determined from chemical shift titrations," *Journal of Biomolecular NMR*, vol. 53, no. 2, pp. 125–138, 2012.

[40] K. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[41] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[42] P. Du, S. Gu, and Y. Jiao, "PseAAC-general: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets," *International Journal of Molecular Sciences*, vol. 15, no. 3, pp. 3495–3506, 2014.

[43] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[44] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinformatics*, vol. 17, no. 3, pp. 282–283, 2001.

[45] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[46] G. Pollastri, A. J. M. Martin, C. Mooney, and A. Vullo, "Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information," *BMC Bioinformatics*, vol. 8, article 201, 2007.

[47] G. Pollastri and A. McLysaght, "Porter: a new, accurate server for protein secondary structure prediction," *Bioinformatics*, vol. 21, no. 8, pp. 1719–1720, 2005.

[48] P. Du and Y. Yu, "SubMito-PSPCP: predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions," *BioMed Research International*, vol. 2013, Article ID 263829, 7 pages, 2013.

[49] G. L. Fan and Q. Z. Li, "Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition," *Amino Acids*, vol. 43, no. 2, pp. 545–555, 2012.

[50] G. L. Fan and Q. Z. Li, "Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 304, pp. 88–95, 2012.

[51] G. L. Fan and Q. Z. Li, "Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 334, pp. 45–51, 2013.

[52] G. L. Fan, Q. Z. Li, and Y. C. Zuo, "Predicting acidic and alkaline enzymes by incorporating the average chemical shift and gene ontology informations into the general form of Chou's PseAAC," *Process Biochemistry*, vol. 48, no. 7, pp. 1048–1053, 2013.

[53] M. Rashid, S. Saha, and G. P. S. Raghava, "Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs," *BMC Bioinformatics*, vol. 8, article 337, 2007.

[54] B. Liu, X. Wang, L. Lin, and Q. Dong, "A discriminative method for protein remote homology detection and fold recognition combining Top-*n*-grams and latent semantic analysis," *BMC Bioinformatics*, vol. 9, article 510, 2008.

[55] B. Liu, J. Xu, and Q. Zou, "Using distances between Top-n-gram and residue pairs for protein remote homology detection," *Bmc Bioinformatics*, vol. 15, article S3, 2014.

[56] B. Liu, D. Zhang, R. Xu et al., " Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[57] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation," *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[58] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.