Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 978503, 5 pages
http://dx.doi.org/10.1155/2014/978503

Research Article

Prediction of Four Kinds of Simple Supersecondary Structures in Protein by Using Chemical Shifts

Feng Yonge

College of Science, Inner Mongolia Agriculture University, Hohhot 010018, China

Correspondence should be addressed to Feng Yonge; fengyonge@163.com

Received 7 May 2014; Revised 3 June 2014; Accepted 4 June 2014; Published 18 June 2014

Academic Editor: Hao Lin

Knowledge of supersecondary structures can provide important information about its spatial structure of protein. Some approaches have been developed for the prediction of protein supersecondary structure. However, the feature used by these approaches is primarily based on amino acid sequences. In this study, a novel model is presented to predict protein supersecondary structure by use of chemical shifts (CSs) information derived from nuclear magnetic resonance (NMR) spectroscopy. Using these CSs as inputs of the method of quadratic discriminant analysis (QD), we achieve the overall prediction accuracy of 77.3%, which is competitive with the same method for predicting supersecondary structures from amino acid compositions in threefold cross-validation. Moreover, our finding suggests that the combined use of different chemical shifts will influence the accuracy of prediction.

1. Introduction

The prediction of protein structure is always one of the most important research topics in the field of bioinformatics. However, it is very difficult to predict the spatial structure directly from the protein sequence. Therefore, the prediction of supersecondary structure is an important step in the prediction of protein spatial structure. The supersecondary structural motifs are composed of a few secondary structural elements (namely, α or β) connected by loops. At present, there are four kinds of simple supersecondary structures, namely, α-loop-β, α-loop-α, β-loop-α, and β-loop-β. These motifs play an important role in protein folding and stability because a large number of motifs exist in protein spatial structure. Many researches have focused on exploring methods for protein supersecondary structure prediction [1, 2]. In 1995, Sun et al. predicted protein supersecondary structure and achieved an accuracy of between 70 and 80% by using neural networks [3]. Chou and Blinn presented a method for predicting beta turns [4–6], alpha turns [7], and all the tight turns [6]. Cruz et al. identified β-hairpin and non-β-hairpin [8]. Hu and Li identified four kinds of simple supersecondary structures in 2088 proteins and achieved an accuracy of 78~83 % [9]. Zou et al. also predicted four kinds of simple supersecondary structures from 3088 proteins by using support vector machine [10]. And the overall accuracy of 78% was achieved. The features of these studies were mainly derived from the amino acid compositions or dipeptide compositions.

Nuclear magnetic resonance (NMR) technique plays an important role in the determination of three-dimensional biological macromolecule structures. NMR chemical shifts encode subtle information about the local chemical environment of nuclear spins. For many years, there has been growing interest to access this information and utilize it for biomolecular structure determination [11, 12]. Recent progress was made by combining chemical shifts with protein structure prediction programs [13–20], showing that chemical shifts information is a power parameter for the determination of protein structure. In this paper, we utilized chemical shifts as parameters to predict four kinds of simple supersecondary structures in protein by the method of quadratic discriminant analysis. Using the benchmark dataset, we achieved the average of sensitivity of 76.3% and specificity of 74.3% and the overall prediction accuracy of 77.3% in threefold cross-validation by using six CSs ($C, C_\alpha, C_\beta, H, H_\alpha, N$) as features. Moreover, we have performed the prediction by combining the different chemical shifts as features. Results showed that the redundant information has great influence on the accuracy.

2. Materials and Methods

2.1. Database. The chemical shifts of all nuclei ($C, C_\alpha, C_\beta, H, H_\alpha, N$) in proteins were extracted from re-referenced protein chemical shift database (namely, RefDB [21]). The following steps were performed to construct the dataset. Firstly, only proteins with six nuclei assigned CSs were considered. Secondly, only proteins with the supersecondary structures information in ArchDB40 [22] were available. We finally utilized the PISCES program [23] to remove the highly similar sequences. After strictly following the aforementioned procedures, 114 proteins were obtained which have both CSs and supersecondary structures. Among 114 proteins, 92% (105 sequences) proteins have less than 25% sequence identity, and the sequence identity of the remains ranges from 25 to 30%. The appendix lists 114 proteins used in this study. Finally, we obtained 90 α-loop-α (HH), 89 α-loop-β (HE), 97 β-loop-α (EH), and 122 β-loop-β (EE) motifs, including the β-β link and β-β hairpin.

2.2. Feature Parameter. In the four data subsets $\{HH, HE, EH, EE\}$, we calculated the averaged CSs of six nuclei for a sequence of length l using the following formula:

$$t_i = \frac{1}{l}\sum_{i=1}^{l} \mathrm{CS}_i, \tag{1}$$

where $i = C, C_\alpha, C_\beta, H, H_\alpha, N$. Therefore, a sequence can be converted into a six-dimensional vector $R : \{t_i\}$.

2.3. Prediction Algorithm. To design an efficient and accurate predicted algorithm the key step is in protein supersecondary structure prediction. The quadratic discriminant analysis [24] is a power algorithm that has been widely applied in genomic and proteomic bioinformatics. Thus, we used it here to perform prediction.

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

$$R = \{t_i\} \quad \left(i = C, C_\alpha, C_\beta, H, H_\alpha, N\right). \tag{2}$$

Here we integrated six-dimensional vector by using quadratic discriminant analysis function. Consider a sequence X is classified into four groups (HH, HE, EH, EE). The discriminant analysis function between group i and group j is defined by

$$\xi_{ij} = \ln p\left(\omega_i \mid X\right) - \ln p\left(\omega_j \mid X\right). \tag{3}$$

According to Bayes' Theorem, we deduce

$$\begin{aligned}\xi_{ij} &= \ln\frac{p_i}{p_j} - \frac{\delta_i - \delta_j}{2} - \frac{1}{2}\ln\frac{\left|\Sigma_i\right|}{\left|\Sigma_j\right|} \\ &= \left(\ln p_i - \frac{1}{2}\delta_i - \frac{1}{2}\ln\left|\Sigma_i\right|\right) \\ &\quad - \left(\ln p_j - \frac{1}{2}\delta_j - \frac{1}{2}\ln\left|\Sigma_j\right|\right).\end{aligned} \tag{4}$$

The result can be generalized to *four* groups directly and described as follows.

Set

$$\begin{aligned}\eta_v &= \ln p_v - \frac{\delta_v}{2} - \frac{1}{2}\ln\left|\Sigma_v\right| \\ &\left(v = HH, EH, HE, EE\right),\end{aligned} \tag{5}$$

where

$$\delta_v = \left(R - \mu_v\right)^T \Sigma_v^{-1}\left(R - \mu_v\right), \tag{6}$$

where 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 Σ_v (note: μ_v and $|\Sigma_v|$ are calculated in training set), and μ_v denotes chemical shift values of six nuclei $R : \{t_i\}$ averaged over group v; $|\Sigma_v|$ is the determinant of matrix Σ_v.

The six-dimensional vector μ_v can be written as

$$\mu_v^{(i)} = \frac{1}{p_v}\sum_{i=1}^{p_v} t_i, \tag{7}$$

where $v = HH, EH, HE, EE$; $i = C, C_\alpha, C_\beta, H_\alpha, H, N$; Σ_v is the covariance matrix of 6×6 dimension, quantifying correlations between the chemical shifts of six nuclei:

$$\Sigma_v = \begin{bmatrix} \sigma^v{}_{1,1} & \sigma^v{}_{1,2} & \cdots & \sigma^v{}_{1,6} \\ \sigma^v{}_{2,1} & \sigma^v{}_{2,2} & \cdots & \sigma^v{}_{2,6} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma^v{}_{6,1} & \sigma^v{}_{6,2} & \cdots & \sigma^v{}_{6,6} \end{bmatrix}, \tag{8}$$

where the element

$$\sigma_{i,j}^v = \frac{1}{p_v}\sum\left(t_i - \mu_v^{(i)}\right)\left(t_j - \mu_v^{(j)}\right). \tag{9}$$

Here $v = HH, EH, HE, EE$; $i, j = C, C_\alpha, C_\beta, H_\alpha, H, N$.

From (4) and (5), we have concluded

$$\xi_{ij} = \eta_i - \eta_j. \tag{10}$$

It can be easily proved that $p(\omega_k \mid X)$ is the maximum of $p(\omega_v \mid X)$, if η_k is the maximal one in η_v ($v = HH, EH, HE, EE$). Then, we predict that X belongs to group k.

2.5. Correction in the Error Allowed Scope. A sequence X is predicted for four kinds of supersecondary structures by using (1)~(10). If η_i is the maximal one in η_k ($k = HH, EH, HE, EE$), then we predict that X belongs to group i. However, there are slight differences among

η_k ($k = HH, EH, HE, EE$). To correct predicted results, we define the coefficient of the error allowed scope as

$$R = \frac{\eta_{\text{corr}} - \eta_{\text{wro}}}{\eta_{\text{corr}}}, \quad (11)$$

where η_{corr} denotes X belonging to itself class η, η_{wro} denotes X being predicted another class η. For example, if X is the super-secondary structure of HH, then η_{corr} is η_{HH} and η_{wro} is the maximum among $\eta_{EH}, \eta_{HE}, \eta_{EE}$.

2.6. Performance Evaluation. In statistical prediction, independent dataset test, cross-validation test, and jackknife test can be used to examine a predictor for its effectiveness in practical application. Among the three test methods, the jackknife test is deemed to be the least arbitrary that can always yield a unique result for a given benchmark dataset [25] and has been widely used to examine the performance of various predictors [26–37]. However, in this study we have used the threefold cross-validation to examine the performance of our method; in order to reduce the computational time, we randomly divided the training set into three parts, two of which are for training and the rest for testing. The process is repeated three times. The following three parameters: sensitivity (SN_i), specificity (SP_i), and overall accuracy (Q_{total}), are used to evaluate the predictive performance of our approach:

$$\text{SN}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \times 100\%, \quad (12)$$

$$\text{SP}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \times 100\%, \quad (13)$$

$$Q_{\text{total}} = \frac{\sum_i \text{TP}_i}{N} \times 100\%, \quad (14)$$

where $i = HH, HE, EH, EE$ and TP, FN, TN, and FP denote, respectively, true positives, false positives, true negatives, and false positives. N is total number of sequences in four data subsets.

3. Results and Discussion

Under the benchmark dataset, we calculated the average chemical shift values using (1). The sequences from four data subsets are converted, respectively, into six-dimensional vectors, which are derived from chemical shift values of six nuclei; then μ is also a six-dimensional mean vector, which is calculated in each of the datasets. In the training sets, determinant and inverse matrix of covariance matrix Σ_v are calculated. Given a sequence of the testing sets, we may calculate η_v by using (4)~(10) and compare the results. Then the class of sequence X was determined by the maximum of η_v ($v = HH, HE, EH, EE$). Moreover, the coefficient R given in (11) is used to correct predicted results. The current study utilized $R < 0.4$. The results of threefold cross-validation are listed in Table 1.

From Table 1, we can see that the averaged sensitivity, specificity, and overall accuracy of four kinds of supersecondary structures are 76.3%, 74.3%, and 77.3%, respectively, indicating that CSs are highly informative with regard to supersecondary structures.

Table 1: The predicted accuracies by using six CSs as features (3-fold cross-validation).

Class structure	SN (%)	SP (%)	Average SN (%)	Average SP (%)	Q_{total} (%)
	$R < 0.4$				
HH	73.0	71.0	76.3	74.3	77.3
EH	75.8	78.1			
HE	69.0	66.7			
EE	87.5	81.4			

Generally speaking, chemical shift measurements can be incomplete for a multitude of reasons. Often, chemical shifts can only be assigned partially or are missing. To assess the impact of incomplete chemical shift assignments, we performed the prediction by using the combination of the different chemical shifts as features. The results are shown in Table 2.

From Table 2, we found that omission of some CSs can result in radically different accuracy. Theoretically, incomplete chemical shifts provide relatively less information, so the predicted accuracy is also declined. But it actually did not in prediction. We used CSs of H, H_α, C as features and achieved the highest accuracy of prediction, indicating that the results are affected by the redundant data. According to the performances, we concluded that CSs of N, C_α, C_β are the most informative features in the prediction of four kinds of protein supersecondary structures. In addition, the information of C, H_α, N is commonly provided in protein database; we achieved the prediction accuracy of 79.1% by using CSs of C, H_α, N as the only inputs.

To test the method and facilitate comparison with other features, we used amino acid compositions (AAC) as inputs of the method of quadratic discriminant analysis. The compared results are recorded in Table 2. Compared results show that the performances of CSs are superior to that of AAC for supersecondary structures prediction, except *HE* structure (compared with six CSs).

4. Conclusions

In this paper, we have introduced a prediction model for supersecondary structures from protein chemical shifts. Our model is both simple and easy to perform. However, owing to the limitation of both information of supersecondary structures and corresponding chemical shifts of six nuclei that should be considered, only 114 proteins have been selected in this study. Based on the benchmark dataset, we investigated the relationship between supersecondary structures and chemical shifts. We achieved the overall accuracy of 77.3% by using six CSs as features and the maximum overall accuracy of 89.2% by using the combination of CSs of N, C_α, C_β. Results show that chemical shift is a good parameter for the prediction of four kinds of protein supersecondary structures. In summary, the chemical shifts

Table 2: Predicted results of different feature combinations ($R < 0.4$).

Feature combinations	*HH*		*EH*		*HE*		*EE*		Average SN (%)	Average SP (%)	Q_{total} (%)
	SN (%)	SP (%)	SN (%)	SP (%)	SN (%)	SP (%)	SN (%)	SP (%)			
$C, C_\alpha, C_\beta, H, H_\alpha$	63.3	77.0	84.5	45.6	34.8	100	71.3	77.0	63.4	74.9	64.6
$C, C_\alpha, C_\beta, H_\alpha, N$	90.0	85.3	66.0	97.0	85.4	86.4	93.4	75.5	83.7	86.1	84.2
C, C_α, C_β, N	55.6	87.7	61.9	80	44.9	93.0	95.1	52.5	64.4	78.3	66.8
C_α, C_β, N	90.0	87.1	94.8	83.6	79.8	93.4	91.0	91.7	88.9	89.0	89.2
C, H_α, N	90.0	73.6	75.3	82.0	79.8	81.6	73.8	80.4	79.7	79.4	79.1
AAC	73.3	73.6	73.0	77.8	72.4	71.3	77.5	75.8	74.1	74.6	75.8

Table 3: PDB 114 chains used in this work.

1a6g	1a6j	1a7g	1ail	1akh	1am7	1avs	1b2v
1b56	1bdo	1bed	1bgf	1bja	1by9	1byf	1c44
1cex	1cy5	1dfu	1dhn	1dqe	1dtl	1dyt	1e0c
1edh	1ejf	1ekg	1epf	1ew4	1f2l	1f35	1f3v
1f80	1F8H	1fdq	1ff3	1fil	1g6a	1g6h	1gaw
1gns	1gnu	1go4	1gwy	1gwy	1h4a	1h70	1hcb
1hfc	1hh8	1hrh	1hsl	1huu	1i4f	1ifo	1iho
1iko	1iw0	1iwm	1j1v	1j54	1j7d	1j97	1jr1
1jiw	1jr2	1jl3	1jrl	1jhf	1k82	1l0s	1l1d
1l6x	1lfo	1ljp	1lld	1m1f	1ml4	1mo1	1mxe
1naq	1ng2	1o15	1o5u	1oqr	1osp	1php	1ppf
1pz4	1q4r	1qav	1qfj	1qg7	1qog	1qst	1r5r
1rro	1rsy	1scj	1slm	1snc	1t15	1tkv	1tn3
1tph	1umu	1uoh	1uuh	1uv0	1vap	1vjh	1ycq
1ze3	256b						

will become a new parameter in prediction of the protein supersecondary structures in the near future.

Appendix

See Table 3.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The author is grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. The work was supported by Inner Mongolia Agriculture University PhD Research Fund (no. BJ08-30) and Basic Science of Inner Mongolia Agriculture University Research Fund (no. JC2013004).

References

[1] T. Blundell, D. Carney, S. Gardner et al., "Knowledge-based protein modelling and design," *European Journal of Biochemistry*, vol. 172, no. 3, pp. 513–520, 1988.

[2] H. J. Dyson and P. E. Wright, "Peptide conformation and protein folding," *Current Opinion in Structural Biology*, vol. 3, no. 1, pp. 60–65, 1993.

[3] Z. Sun, X. Rao, L. Peng, and D. Xu, "Prediction of protein supersecondary structures based on the artificial neural network method," *Protein Engineering*, vol. 10, no. 7, pp. 763–769, 1997.

[4] K. C. Chou, "Prediction of beta-turns in proteins," *Journal of Peptide Research*, vol. 49, pp. 120–144, 1997.

[5] K.-C. Chou and J. R. Blinn, "Classification and prediction of β-turn types," *Journal of Protein Chemistry* , vol. 16, no. 6, pp. 575–595, 1997.

[6] K.-C. Chou, "Prediction of tight turns and their types in proteins," *Analytical Biochemistry*, vol. 286, no. 1, pp. 1–16, 2000.

[7] K.-C. Chou, "Prediction and classification of α-turn types," *Biopolymers*, vol. 42, no. 7, pp. 837–853, 1997.

[8] X. de la Cruz, E. G. Hutchinson, A. Shepherd, and J. M. Thornton, "Toward predicting protein topology: an approach to identifying β hairpins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 17, pp. 11157–11162, 2002.

[9] X. Z. Hu and Q. Z. Li, "Prediction of the β-hairpins in proteins using support vector machine," *Protein Journal*, vol. 27, no. 2, pp. 115–122, 2008.

[10] D. S. Zou, Z. S. He, J. Y. He, and Y. Xia, "Supersecondary structure prediction using Chou's pseudo amino acid composition," *Journal of Computational Chemistry*, vol. 32, no. 2, pp. 271–278, 2011.

[11] D. A. Case, "The use of chemical shifts and their anisotropies in biomolecular structure determination," *Current Opinion in Structural Biology*, vol. 8, no. 5, pp. 624–630, 1998.

[12] D. S. Wishart and D. A. Case, "Use of chemical shifts in macromolecular structure determination," *Methods in Enzymology*, vol. 338, pp. 3–34, 2001.

[13] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo, "Protein structure determination from NMR chemical shifts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 23, pp. 9615–9620, 2007.

[14] Y. Shen, O. Lange, F. Delaglio et al., "Consistent blind protein structure generation from NMR chemical shift data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 12, pp. 4685–4690, 2008.

[15] H. Lin, C. Ding, Q. Song et al., "The prediction of protein structural class using averaged chemical shifts," *Journal of Biomolecular Structure & Dynamics*, vol. 29, no. 6, pp. 643–649, 2012.

[16] M. Mechelke and M. Habeck, "A probabilistic model for secondary structure prediction from protein chemical shifts," *Proteins*, vol. 81, no. 6, pp. 984–993, 2013.

[17] 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.

[18] A. Pastore and V. Saudek, "The relationship between chemical shift and secondary structure in proteins," *Journal of Magnetic Resonance*, vol. 90, no. 1, pp. 165–176, 1990.

[19] Y. Wang, "Secondary structural effects on protein NMR chemical shifts," *Journal of Biomolecular NMR*, vol. 30, no. 3, pp. 233–244, 2004.

[20] W. S. Mao, P. S. Cong, Z. H. Wang, L. J. Lu, Z. L. Zhu, and T. H. Li, "NMRDSP: an accurate prediction of protein shape strings from NMR chemical shifts and sequence data," *PLoS ONE*, vol. 8, no. 12, Article ID e83532, 2013.

[21] H. Zhang, S. Neal, and D. S. Wishart, "RefDB: a database of uniformly referenced protein chemical shifts," *Journal of Biomolecular NMR*, vol. 25, no. 3, pp. 173–195, 2003.

[22] N. Fernandez-Fuentes, A. Hermoso, J. Espadaler, E. Querol, F. X. Aviles, and B. Oliva, "Classification of common functional loops of kinase super-families," *Proteins*, vol. 56, no. 3, pp. 539–555, 2004.

[23] G. Wang and R. L. Dunbrack Jr., "PISCES: recent improvements to a PDB sequence culling server," *Nucleic Acids Research*, vol. 33, no. 2, pp. W94–W98, 2005.

[24] Y. Feng and L. Luo, "Use of tetrapeptide signals for protein secondary-structure prediction," *Amino Acids*, vol. 35, no. 3, pp. 607–614, 2008.

[25] K.-C. Chou and H.-B. Shen, "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, vol. 3, no. 2, pp. 153–162, 2008.

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

[27] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[28] M. Hayat and A. Khan, "Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 411–421, 2012.

[29] C. Ding, L.-F. Yuan, S.-H. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.

[30] C. Chen, Z.-B. Shen, and X.-Y. Zou, "Dual-layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 422–429, 2012.

[31] K.-C. Chou and H.-B. Shen, "Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization," *PLoS ONE*, vol. 5, no. 6, Article ID e11335, 2010.

[32] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "IRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[33] W. Chen, H. Lin, P.-M. Feng, C. Ding, Y.-C. Zuo, and K.-C. Chou, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, no. 10, Article ID e47843, 2012.

[34] H. Lin, W. Chen, L.-F. Yuan, Z.-Q. Li, and H. Ding, "Using over-represented tetrapeptides to predict protein submitochondria locations," *Acta Biotheoretica*, vol. 61, no. 2, pp. 259–268, 2013.

[35] H. Lin, C. Ding, L.-F. Yuan et al., "Predicting subchloroplast locations of proteins based on the general form of Chou's pseudo amino acid composition: approached from optimal tripeptide composition," *International Journal of Biomathematics*, vol. 6, no. 2, Article ID 13500034, 2013.

[36] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "ILoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins," *Molecular BioSystems*, vol. 9, no. 4, pp. 634–644, 2013.

[37] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "IAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.