Protein Secondary Structure Online Server Predictive Evaluation

Shuping Zhu, Yihui Liu
School of computer science and technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
yxl@qlu.edu.cn

Abstract. The prediction of protein secondary structure is of great significance for studying the function of proteins and for making progress in the field of bioinformatics. Since 1951, the secondary structure of predicted proteins has been proposed. After 66 years of development, the prediction method has been continuously optimized, and the accuracy rate has exceeded 80%. Continuous Automated Model Evaluation (CAMEO) gives a prediction of protein tertiary structure evaluation for many current prediction methods, and secondary structure evaluation has not been achieved. In response to this problem, six servers were selected: PSRSM, MUFOLD, SPIDER, RAPTORX, JPRED and PSIPRED to evaluate the predicted secondary structure. The latest released protein from the Protein Data Bank (PDB) was applied just to ensure that the test set is not included in the training set. In the experiments of which protein homology was 30%, 50%, 70% and 90%, the obtained accuracy of PSRSM for Q3 was 91.44%, 88.12%, 90.17% and 87.39%, respectively. And the accuracy is higher than the best server among other prediction servers---MUFOLD, by 3.19%, 1.33%, 2.19% and 1.72%, correspondingly. It is proved that PSRSM has a better prediction quality than other servers for the same kind of homology data, the Sov and boundary accuracy as well. This paper focuses on analyzing the operating methods and corresponding results of various servers, thus, it is safe to say that the prediction of protein secondary structure should be studied on perspectives of big data, templates and deep learning.

1. Introduction
In order to study the function of proteins, people usually start from its structure, in which the research of secondary structure helps to discover three-dimensional structures and provide functional annotation of proteins. Therefore, most people are committed to the study of protein secondary structure and hope to contribute to the field of bioinformatics.

In 1951, Pauling and Corey first proposed the secondary structure of proteins\cite{1}, and the initial prediction method for protein secondary structure was mainly carried out by studying the amino acid sequence with an accuracy of about 60%. Rost\cite{2-3} used the PHD algorithm in the study to use the evolutionary information contained in the multi-sequence arrangement as the input of the neural network, and predicted that the secondary structure accuracy of the protein exceeded 70%. In the paper \cite{4}, the sparse algorithm of dynamic Bayesian classifier is adopted, and the prediction accuracy of protein secondary structure reaches 76.3%. Using SVM combined with position-specific scoring matrices (PSSM) and physicochemical characteristics of protein structures to predict, the accuracy rate is about 80\%\cite{5}. This paper \cite{6} solves the sequence-structure mapping relationship between input protein features and SS by combining PSSM and amino acid sequence information and using a
secondary structure recursive encoder-decoder network (SSREDN), and the accuracy of Q3 in the datasets CullPDB and CB513 is 84.2% and 82.9%. Protein secondary structure prediction methods continue to inject new vitality, and now many methods have realized the prediction of online servers. This paper selects six kinds of servers: PSRSM, MUFOLD, SPIDER, RAPTORX, JPRED and PSIPRED to illustrate the algorithm principle, and compares the prediction accuracy of each server through test data to give the evaluation of the secondary structure prediction server. Then proposes new ideas for the future development of protein research.

2. Online Server Principle

2.1. PSRSM
The server uses methods based on data partition and semi-random subspace method (PSRSM)\(^7\). In the traditional random subspace method, the low-dimensional subspace is generated by random sampling in high-dimensional space. The semi-random subspace method used by PSRSM can effectively ensure the accuracy and diversification of the basic classifier. The main steps of the method are as follows: First and foremost, the training data is divided into different subsets according to the length of the protein sequence, and then the subspace is generated using the semi-random subspace method, and the basic classifier is trained in the subspace. Finally combining them by majority vote rule on each subset. Support vector machine (SVM) was used as the most basic classifier in the experiment.

Specifically, PSSM generated for input using the PSI-BLAST program, and PSI-BLAST uses the BLOSUM62 Evolution Matrix to search for a reduced version of NCBI’s non-redundant (NR) database. The PSSM obtained according to the above principle is a 20*L matrix, 20 is the number of amino acids, and L is the length of each protein. In the experiment, 13 sliding windows were used to obtain the characteristic information of the center position of the protein sequence, and the secondary structure was predicted. Assuming a protein of length L, a 260*L (13*20*L) input matrix is generated. And 160 features are selected from 260 feature values as the main feature, 160*L is used as the network input. Finally, 12 classifiers were built for training. Then a new protein can be predicted based on the length and the appropriate classifier.

The experimental training set selected 2892 protein data from the ASTRAL data set and 12288 protein data from the CullPDB data set. After removing the higher similarity protein, and the training set included a total of 15,696 data. The experimental test set used 99 CASP10 data, 81 CASP11 data, 19 CASP12 data, 513 CB513 data, 1673 25PDB data, and 100 protein data (T100) released by PDB before February 2018. In the case of using 6 GTPC models, the Q3 prediction accuracy of protein secondary structure in the datasets 25PDB, CB513, CASP10, CASP11, CASP12 and T100 data were 86.38%, 84.53%, 85.51%, 85.89%, 85.55 and 85.09%, respectively. Forecast URL is: http://qilubio.qlu.edu.cn:82/protein_PSRSM/default.aspx.

2.2. MUFOLD
MUFOLD uses a new type of network called Deep inception-inside-inception (Deep 3I) to predict protein secondary structure and considers the input feature matrix carefully. The physicochemical properties of amino acids, PSI-Blast features and HHBlits characteristics are combined in the characteristic matrix\(^8\). Among them, for the physicochemical property matrix, 8 numbers selected from -1 to 1 are set to represent one amino acid; the first 7 digits represent the physicochemical properties of the amino acid; the last digit uses 1 or 0 to indicate whether the amino acid is input. For the characteristics of PSI-Blast, according to a similar principle, an amino acid is represented by a 22-bit number from 0 to 1, wherein the first 21 bits are set according to the obtained PSSM value, and the last bit is represented by 1 or 0 whether there is an input. The HHBlits function also selects a 31-digit number from 0 to 1 to represent an amino acid; the first 30 bits are set according to the HMM file; the last bit is also represented by 0 or 1. The above three feature sets combine the 58-bit total features as inputs to the MUFOLD network.
The Deep3I network consists of two Deep3I blocks, a series of convolutions and a fully interconnected dense layer. The Deep3I block is composed of the initial module recursive nesting. The initial module can effectively extract the non-local interaction between amino acid residues through the convolution operation. The Deep3I network predicts protein secondary structure by continuously training and experimenting with TensorFlow and Keras.

Protein sequences ranging from 50 to 700 in length were used in the MUFOLD experiment, and used five published protein databases from CullPDB, JPRED, CASP, CB513 and PDB. Specifically: 9581 data were selected from CullPDB, of which 9000 were randomly selected as the training set, and the remaining 581 were tested. The data also selected from JPRED data belong to different superfamilies. After the CASP data set was filtered, 98 data of CASP10, 83 data of CASP11, 40 data of CASP12 were used. CB513 and 385 PDB data were also applied to the experiment of MUFOLD. The MUFOLD test data range is 30 to 700, and the predicted URL is: http://mufold.org/mufold-ss-angle/.

2.3. SPIDER3
Reference [9] mentions that the study of protein secondary structure prediction and solvent contact surface area has been stagnant for many years because some amino acid residues are very close in the three-dimensional structure and far in the amino acid sequence. It is therefore more difficult to capture non-local interactions between amino acid residues. Existing machine learning methods basically use 10 to 20 sliding windows to obtain amino acid interactions. SPIDER does not use a sliding window, using a Long Short-Term Memory (LSTM) Bidirectional Recurrent Neural Network (BRNNs) machine learning model to achieve prediction, and it can capture the non-local interaction between amino acid residues.

The network structure of the LSTM-BRNN model consists of two hidden layers activated by the BRNN layer using LSTM elements and two tightly connected rectifying linear units (ReLUs), which are used in four iterations. For the input of the network, 7 representative protein physicochemical properties (PP), 20-dimensional PSSM from PSI-Blast and 30-dimensional hidden Markov model sequence spectra from HHBlits were selected. These data are entered into a network of LSTM-BRNNs for four iterations (one of which includes two LSTM-BRNNs), the final machine learning model. To prevent overfitting during training, a loss algorithm with a 50% loss rate was used, and Adam was used to optimize the training process, which was able to capture long and short distance interactions without using a sliding window.

The experiment used 5789 protein data, of which 4590 data were randomly selected for training, and the remaining 1199 data were used as independent test sets. The trained 4590 data is further divided into 10 10-fold cross-validated size-folded data, each containing training and validation data. In addition, based on the similarity below 30% before September 16, 2016, 115 protein data were selected as an additional test set[10]. The URL for the SPIDER3 test is: http://sparks-lab.org/server/SPIDER3/.

2.4. RAPTORX
RAPTORX uses Deep Convolutional Neural Fields (DCNF) combined with deep convolutional neural network (DCNN) and conditional random fields (CRF) to predict protein secondary structure, and the network is trained by a region under the ROC curve (AUC) maximization method, which can well solve the prediction problem of disordered sequence proteins[11]. In the literature [12], it is mentioned that RAPTORX can obtain about 84% Q3 accuracy and 72% Q8 accuracy, when using protein sequence files. When not using sequence files, it can achieve about 75% Q3 accuracy and 59% Q8 accuracy, respectively, and RAPTORX can effectively solve complex genetic structure relationship modeling and modeling between adjacent residues. Reference [13] pointed out that DCNF uses DCNN instead of the shallow neural network used in CNF to capture the complex relationship between input and output tags to capture remote sequence information.

The data used in the RAPTORX experiment included 6125 CullPDB data, CB513 data, 123 CASP10 data, 105 CASP11 data, and CAMEO data, as well as 1338 training data and 149 test data.
published by JPRED. The RAPTORX test data range is 26 to 4000 protein sequences, and the predicted URL is: http://raptorx.uchicago.edu/StructurePropertyPred/predict/.

2.5. JPRED
The JPRED server has been providing protein prediction since 1998 and has now grown to the JPRED4 version. The JPRED3 version provides a single protein sequence or multiple sequence alignment (MSA) prediction using the JNET algorithm, where JNET uses JNET v2.0. JNET v2.0 does not use frequency files, only uses PSI-BLAST's PSSM configuration file and HMMER's hidden Markov model to increase the neural network from 9 units to 100 units. The method was developed through a 7-fold cross-validation training of sequence and structural non-redundant data sets derived from Astral assembly of super-family-level SCOP data\cite{14}, finally, using 149 blind data tests gave 81.5% Q3 accuracy.

The JPRED4 version, like JPRED3, also uses the JNET algorithm and provides secondary prediction of protein sequences for single and multiple sequence alignments. The difference is that it selects one of the 1358 SCOPe/ASTRAL v.2.04 superfamilies and performs a 7-fold cross-validation experiment with JNET 2.3.1, a PSI-BLAST file was generated by looking for UniRef90 v.2014_07 and multiple sequence alignments were created for each protein sequence, finally, 82% accuracy was obtained on 150 training sets\cite{15}. The JPRED online server also provides predictions of solvent accessibility and coiled-coil zones at http://www.compbio.dundee.ac.uk/jpred4/index.html.

2.6. PSIPRED
The literature\cite{16} pointed out that the PSIPRED server combines three advanced technologies, namely PSIPRED, GenTHREADER and MEMSAT 2. Among them, PSIPRED uses a strict cross-validation process to evaluate performance, and uses two feed forward neural networks to analyze the output obtained from PSI-BLAST to obtain reliable secondary structure prediction results, GenTHREADER is used to infer the structure and topology of transmembrane proteins, MEMSAT2 is able to quickly identify protein folding information. The forecast URL is: http://bioinf.cs.ucl.ac.uk/psipred/.

3. Data Selection and Evaluation Criteria
Based on the premise that each server can be predicted, 200 protein data (T200) from April to August 2018 were selected for experimentation and evaluated using appropriate evaluation criteria.

3.1. Data selection
The data selection follows the following principles: First, the data is selected from the latest PDB data released in 2018, which ensures that the test set is not in the training set of the server; secondly, the data comes from different time periods and is more dispersed; then the data volume is larger, making the experimental results can be more convincing. Furthermore, data can represent inter-differentiation predictions of protein homology, the final selection of protein length allows each server to test and get predictions. Based on the above conditions, 50 data were selected for testing based on protein homology of 30%(DT1), 50%(DT2), 70%(DT3) and 90%(DT4), respectively. The data were selected as shown in Tab.1.

3.2. Evaluation criteria
In this paper, three methods for measuring the accuracy of protein secondary structure prediction are used: Q3, SOV and Boundary. The value of Q3 is mainly to measure the accuracy of individual residue allocation. The value of SOV is mainly to measure the prediction accuracy of all elements.

3.2.1. Q3
According to the DSSP, we usually divide the protein secondary structure into H, G, I, E, B, T, S and ‘-’, 8 states. In this state, an amino acid sequence is converted into H (helix), E (sheet), and C (coil) in
three states according to the manner of H/G/I → H, E/B → E, and others → C. Then Q3 represents the ratio of the number of amino acids in the three states correctly predicted to the entire amino acid sequence. The define of Q3 as:

\[ Q3 = \frac{S_H + S_E + S_C}{S} \quad (1) \]

| DATA | Protein name |
|------|--------------|
| DT1  | 5LOS 5LTL 5M6Y 5MCT 5MCU 5MCV 5MCW 5MF7 5MG5 5MH5 |
|      | 5MH6 5MIY 5MLP 5MNW 5MQX 5MV2 5MXB 5MXP 5MLQ 5NJ2 |
|      | 5N2P 5N9B 5SNAP 5NBC 5NCB 5NCM 5NDX 5NFX 5NK5 5NM3 |
|      | 5NPN 5NQ0 5SNUK 5NV9 5NWH 5NYH 5NZ4 5NZ5 5NZG 5NQ9X |
|      | 5OBA 5OC9 5OD3 5OOGX 5OOG 5OOG 5OOG 5OOGX 5OOGX |
| DT2  | 5MQX 5MXP 5NCB 5NDX 5NZ4 5O9X 5O9X 5O9X 5O9X |
|      | 5OQS 5OTU 5OW5 5OWB 5OWL 5OWO 5OQF 5UMP 5UMW 5VO3 |
|      | 5VX5 5W6K 5W6Q 5W7M 5W7M 5W8Z 5W8Z 5W8Z 5WCH 5WCQ |
|      | 5WCX 5WD6 5WDG 5WDK 5WEC 5WEW 5WFG 5WGH 5WGH 5WGH |
|      | 5WJM 5WKW 5WLIP 5WM9 5WQM 5X16 5X16 5X16 5X16 |
| DT3  | 5OG7 5OG9 5OGX 5OOG 5OGJ 5OGJ 5OGJ 5OGJ |
|      | 5OII 5OIJU 5OJ3 5OJ5 5OK2 5OK3 5OK6 5OLT 5OML 5OML |
|      | 5ONN 5ONQ 5OSO 5OP3 5OW4 5OWC 5OWK 5Q22 5ULY 5UQ9 |
|      | 5VIA 5VTL 5WAX 5WBD 5WMD 5WNB 5WOF 5WPO 5WPO |
|      | 5XJ5 5XFJ 5XNC 5XNE 5XQZ 5XWX 5X0 5XJ5 5XJ5 |
| DT4  | 5XML 5MXW 5NFX 5NM3 5NWN 5NYH 5NZG 5NZI 5NZI |
|      | 5NZK 5NZI 5NZM 5O6C 5OAO 5ODX 5OHU 5OH7 5OH7 5OH7 |
|      | 5OLN 5OQL 5TOS 5USB 5USN 5USO 5V80 5VAK 5VCK 5VFX |
|      | 5VFY 5VFZ 5VG5 5VG5 5VGN 5VHN 5VHU 5VJ3 5VJ3 5VJ3 |
|      | 5WBS 5WC7 5WDS 5NQ0 5O8M 5OAE 5OET 5OLM 5X9B 5YQ5 |

Among them: \( S_E \) is the number of E-class protein structures accurately predicted, \( S_H \) is the number of accurate predictions of class H protein structures, \( S_C \) is the number of accurate predictions of class C protein structures, and \( S \) is the total number of amino acids. Q3 refers to the accuracy of protein secondary structure prediction in three states.

3.2.2. Sov

Sov’s calculation is based on a measure of the ratio of overlapping segments, which treats the predicted results and the observed results equally. The protein secondary structure is also divided into three states of coil, sheet, and helix according to the above idea of Q3. If it is assumed that the observed sequence is denoted as \( S_1 \), the predicted sequence is denoted as \( S_2 \), and \( S_0 \) is the segment in which all states of \( S_1 \) and \( S_2 \) are the same. (Note that, like several predicted spirals and one observed spiral, each overlap produces a pair in \( S_0 \), each of which contains an observed spiral), next, the length of \( S_1 \) is length(\( S_1 \)), and the sum of the number of \( S_1 \) and \( S_2 \) sequences in each pair is recorded as \( \max(S_1, S_2) \), and the intersection number of \( S_1 \) and \( S_2 \) is recorded as \( \min(S_1, S_2) \). On the basis of the above, we define the calculation formula of Sov\[^{[17]}\] as:
\[ \text{Sov} = 100 \sum_{i} \left( \frac{\min(S_i, S_j) + \delta(S_i, S_j)}{\max(S_i, S_j)} \right) \times \frac{\text{length}(S)}{\text{length}(S_i, S_j)} \]  

The setting for \( \delta \) is to allow for changes in the fragments at the edges of the protein structure. The \( \delta(S_i, S_j) \) values are consistent with the following definitions:

\[ \delta(S_i, S_j) = \min \left( \frac{(\text{max}(S_i, S_j) - \text{min}(S_i, S_j))}{\text{length}(S_i)}, \text{int}[\text{length}(S_i)/2], \text{int}[\text{length}(S_i)/2] \right) \]

3.2.3. **Boundary**

In addition to the above two evaluation criteria, the evaluation of the boundary in protein secondary structure prediction was carried out.

4. **Experimental Result and Discussion**

Based on the difference in homology, the latest 200 protein data (T200) were downloaded from the PDB, and then uploaded to 7 prediction servers for testing. The Q3 and Sov values for each protein Tab.2 Experimental result

(a) Q3 average accuracy

| DATA NAME | PSRSM (%) | MUFOLD (%) | SPIDER (%) | RAPTORX (%) | JPRED (%) | PSIPRED (%) |
|-----------|-----------|------------|------------|-------------|-----------|-------------|
| DT1       | 91.44     | 88.25      | 87.44      | 84.20       | 80.88     | 82.02       |
| DT2       | 88.12     | 86.79      | 85.87      | 84.54       | 79.68     | 80.53       |
| DT3       | 90.17     | 87.98      | 86.19      | 83.45       | 78.62     | 80.15       |
| DT4       | 87.39     | 85.67      | 84.94      | 82.21       | 78.31     | 78.99       |
| T200      | 89.28     | 87.17      | 86.11      | 83.60       | 79.37     | 80.42       |

(b) Sov average accuracy

| DATA NAME | PSRSM (%) | MUFOLD (%) | SPIDER (%) | RAPTORX (%) | JPRED (%) | PSIPRED (%) |
|-----------|-----------|------------|------------|-------------|-----------|-------------|
| DT1       | 87.45     | 84.56      | 83.76      | 81.23       | 78.39     | 77.32       |
| DT2       | 81.95     | 81.67      | 80.23      | 78.74       | 75.30     | 76.05       |
| DT3       | 83.36     | 78.15      | 78.33      | 75.37       | 67.43     | 68.81       |
| DT4       | 81.61     | 79.67      | 81.35      | 76.39       | 72.91     | 72.25       |
| T200      | 83.59     | 80.01      | 80.92      | 77.93       | 73.51     | 73.61       |

(c) Boundary average accuracy

| DATA NAME | PSRSM (%) | MUFOLD (%) | SPIDER (%) | RAPTORX (%) | JPRED (%) | PSIPRED (%) |
|-----------|-----------|------------|------------|-------------|-----------|-------------|
| DT1       | 93.32     | 91.14      | 91.05      | 88.97       | 85.67     | 86.35       |
| DT2       | 90.22     | 89.26      | 89.29      | 87.14       | 83.94     | 85.23       |
| DT3       | 91.51     | 90.25      | 89.66      | 87.38       | 83.14     | 84.54       |
| DT4       | 86.69     | 88.33      | 88.82      | 86.63       | 83.19     | 83.16       |
| T200      | 90.44     | 89.75      | 89.70      | 87.53       | 83.98     | 84.82       |

were calculated by comparing the predicted results obtained from the server with the correct three-state DSSP results. Finally, the average prediction results of Q3, Sov and Boundary of the data DT1, DT2, DT3, DT4 and T200 are shown in Tab.2(a), Tab2(b) and Tab2(c) respectively.

Among the six servers, PSRSM, SPIDER3 and RAPTORX can upload data and download experimental results in batches. The server JPRED and PSIPRED only allow one protein data to be uploaded at a time, and PSIPRED only allows 20 data to be submitted in the same time period. Then, the acquisition of the results of these two servers is also complicated. Although the website description of the MUFOLD server can upload less than 10 data in sequence, uploading more than 4 items in the experiment will cause some problems, so the prediction is performed by uploading 3 pieces of data at a time. For the six servers, PSRSM, MUFOLD, and SPIDER3 are the servers that have just appeared in this field in recent years. RAPPRORX, JPRED and PSIPRED are more classic predictive servers. From the perspective of the forecasting process, the six servers have little difference in prediction time, there differences lies in the requirements of uploading protein length and the acquisition of experimental results. From the perspective of experimental results, PSRSM's Q3, Sov and boundary are better than other servers, while MUFOLD and SPIDER's Q3 and Sov results are not much different. MUFOLD's
Q3 is slightly better, and SPIER3's Sov is sometimes better. The remaining servers RAPTORX, JPRED and PSIPRED differed greatly from the above three results. Among them, RAPTORX predicted higher results, followed by PSIPRED and finally JPRED. From the perspective of prediction methods, they all focus on the prediction problems of amino acid residues that are close in the three-dimensional structure and far in the sequence. Compared to other servers, PSRSM is able to achieve the best predictive results due to the use of its templates, big data and appropriate prediction methods. Through the experimental results and the respective methods, it can be seen that the superior results of the servers can be closely related to the size of the training data and the deep learning algorithms.

5. Conclusion
Based on the above experimental results and discussion, the conclusions are summarized as follows:

(1) PSRSM has the best protein secondary structure prediction effect in T200 data. The reason for the better PSRSM results from the use of its template and the huge amount of data.
(2) When the amount of data is not much different, the prediction results of other servers are closely related to the deep learning algorithm.
(3) In the future, the study of protein secondary structure can be broken by increasing the amount of training data, establishing classification templates and improving deep learning algorithms.

Acknowledgments
The work is supported by National Natural Science Foundation of China (81402538), National Natural Science Foundation of China (61375013), Shandong Provincial Natural Science Foundation, China (ZR2013FM20).

References
[1] Yang Y, Gao J, Wang J, et al. Sixty-five years of the long march in protein secondary structure prediction: the final stretch?[J]. Briefings in Bioinformatics, 2016:bbw129.
[2] Rost B, Sander C. Prediction of Protein Secondary Structure at Better than 70% Accuracy[J]. Journal of Molecular Biology, 1993, 232(2):0-599.
[3] Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure[J]. Proteins-structure Function & Bioinformatics, 1994, 19(1):55-72.
[4] Aydin Z, Singh A, Bilmes J, et al. Learning sparse models for a dynamic Bayesian network classifier of protein secondary structure[J]. Bmc Bioinformatics, 2011, 12(1):154.
[5] Kurniawan I, Haryanto T, Hasibuan L S, et al. Combining PSSM and physicochemical feature for protein structure prediction with support vector machine[J]. Journal of Physics: Conference Series, 2017,835:012006.
[6] Wang Y, Mao H, Yi Z. Protein Secondary Structure Prediction by using Deep Learning Method [J]. Knowledge-Based Systems, 2016,118.
[7] Ma Y, Liu Y, Cheng J. Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method[J]. Scientific Reports, 2018,8.
[8] Fang C, Shang Y, Xu D. MUFOLD-SS: New Deep Inception-Inside-Inception Networks for Protein Secondary Structure Prediction[J]. Proteins: Structure, Function, and Bioinformatics, 2018.
[9] Heffernan R, Yang Y, Paliwal K, et al. Capturing Non-Local Interactions by Long Short Term Memory Bidirectional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers, and Solvent Accessibility[J]. Bioinformatics, 2017.
[10] Heffernan R, Dehzangi A, Lyons J, et al. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins[J]. Bioinformatics, 2015:665.
[11] Wang S, Ma J, Xu J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields[J]. Bioinformatics, 2016,32(17):i672-i679.
[12] Wang S, Li W, Liu S, et al. RaptorX-Property: a web server for protein structure property prediction[J]. Nucleic Acids Research, 2016,44:w430-w435.

[13] Wang S, Peng J, Ma J, et al. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields[J]. Scientific Reports, 2016, 6.

[14] Cole C, Barber J D, Barton G J. The Jpred 3 secondary structure prediction server.[J]. Nucleic Acids Research, 2008,36:w197.

[15] Drozdetskiy A., Cole C, Procter J, et al, JPred4: a protein secondary structure prediction server[J], Nucleic Acids Research, 2015,43.

[16] McGuffin L J, Bryson K, Jones D T. The PSIPRED protein structure prediction server[J]. Bioinformatics, 2000, 16(4):404-405.

[17] Zevalbe M. Understanding Bioinformatics[M], Trans. Li Y X, Hao P, Beijing,Science Press, 2012.(in Chinese).