

Protein Residue Contact Prediction using Support Vector Machine

Chan Weng Howe and Mohd Saberi Mohamad

Abstract—Protein residue contact map is a compact representation of secondary structure of protein. Due to the information hold in the contact map, attentions from researchers in related field were drawn and plenty of works have been done throughout the past decade. Artificial intelligence approaches have been widely adapted in related works such as neural networks, genetic programming, and Hidden Markov model as well as support vector machine. However, the performance of the prediction was not generalized which probably depends on the data used to train and generate the prediction model. This situation shown the importance of the features or information used in affecting the prediction performance. In this research, support vector machine was used to predict protein residue contact map on different combination of features in order to show and analyze the effectiveness of the features.

Keywords—contact map, protein residue contact, support vector machine, protein structure prediction

I. INTRODUCTION

BIOINFORMATICS is defined as a field of science that involve the application of statistics and computer science in the field of biology. It is an emerging field undergoing rapid growth in the past few decades. Bioinformatics at first is applied in the creation and maintenance of database of biological information and currently also applied in tasks like interpretation and analysis of biological data includes deoxyribonucleic acid (DNA) sequences, ribonucleic acid (RNA) sequences, protein structures, protein sequences and protein domains which referred as computational biology. The branch of bioinformatics that consists of the analysis and prediction of three dimensional structures of biological macromolecules such as DNA, RNA and proteins referred as structural bioinformatics. In structural bioinformatics, one of the challenges is the prediction of protein structure.

Protein is one of the most important compounds in human body. Function of a protein is defined by its structure. Protein structures are divided into few categories such as primary structure, secondary structure, tertiary structure, and quaternary structure. A protein consists of more than one linear chain of amino acids that further fold into polypeptides

Chan Weng Howe is with Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia. (email: stephenchanwh@gmail.com)

Mohd Saberi Mohamad is with Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia. (email: saberi@utm.my)

of different structures and features. Protein structure prediction has played an important role in protein design which is essential in several fields for instance medicine. In protein structure prediction, different information of protein has been used and one of them is protein contact map which is used in this research. Protein contact map is a compact representation of three-dimensional conformations of a protein. A contact map is a two-dimensional Boolean matrix representation of protein structure, each of the dimensions is represented by residue number, while the value is true when the corresponding residues are spatial neighbours and false otherwise [2]. Protein contact map is binary symmetric matrices where non-zero values represent the residue in contact [3] and this is illustrated in an example as in Fig. 1.

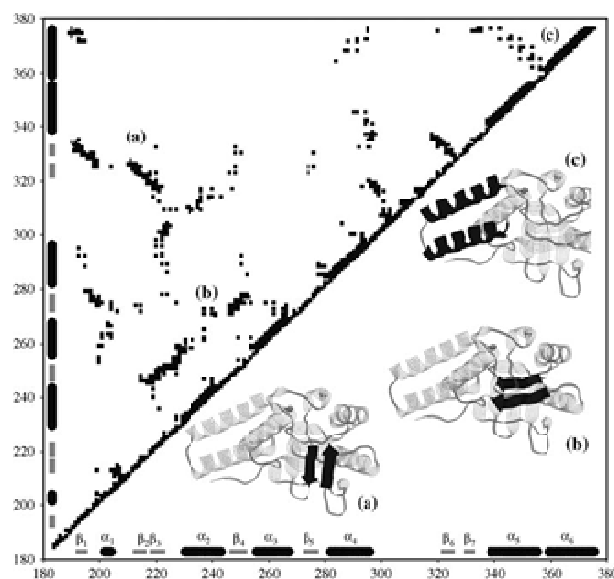


Fig. 1 Example of contact map of HSP-60 protein fragment

According to Fig. 1, Secondary structures are highlighted along the both axis. Both α -helices and β -strands represented by black and grey respectively. While on the left side, the structural protein features are shown: (a) Anti-parallel sheet contacts; (b) parallel sheet contacts; (c) contacts between helical regions. Generally, a residue pair considers as a contact when the distance between the residues within a pair is below a defined distance threshold. The distance threshold is calculated in angstrom (\AA) which measure as 0.1 nanometers or 1×10^{-10} metres. In CASP (Critical Assessment of Techniques for Protein Structure Prediction), default distance

threshold used for assessment is 8Å which is same as the threshold used in this research.

Based on the researches done in the past decade, many techniques and algorithms have been developed to predict contact map of a protein. Among those methods, machine learning algorithms have been widely used such as neural networks [5], [9], [10], support vector machine [1], [11], and genetic programming [7]. Machine learning is an artificial intelligent technique where it is a scientific discipline that is concern about design and development of algorithms that allows computer to learn and evolve behaviours based on the empirical data. It learns to recognize complex patterns thus make decision to gain useful output. In this research, support vector machine (SVM) has been used and implemented. SVM is a supervised learning method that analyses and recognizes pattern for classification and regression. SVM constructs a hyperplane or a set of hyperplanes in high or infinite dimensional space for classification and regression.

Since the prediction of protein contact map is significant in contributing the three-dimensional structure prediction of protein, refer to the state of art of protein contact map prediction, one of the main concerns is the performance issues of several predictors for protein contact map. According to the CASP (Critical Assessment of Techniques of Protein Structure Prediction), the accuracy and the coverage of the prediction of protein contact map are still low and performance varies with the type of structure of the tested protein. In fact, many of the predictors that had been developed tend to predict different correct contacts with implementation of different types of information obtained from protein such as protein profile, predicted secondary structure, solvent accessibility and so forth. Therefore consensus combination of predictors may lead to a better accuracy in protein contact map prediction. To date, researchers are still working on protein contact map prediction in order to enhance the predictor to obtain better and more accurate prediction. Besides, challenges also faced during the prediction of long sequence with many non-local contacts, non-local contacts which had appeared to be a problem because the global topology of the proteins is defined by non-local contacts (also known as long range-contacts) but the methods developed so far are more accurate on local contacts only [2].

This research concentrate in the performance related problem faced in the protein contact map prediction and thus with the use of support vector machine (SVM) method plus different combination of features, studies and experiments have been done in order to identify and determine the effectiveness of the features used in the prediction.

II. MATERIALS AND METHODS

A. Dataset

The dataset used in this research consist of 424 proteins and 48 proteins for both training and testing set respectively. This dataset had been used in previously done research [1] which consist of information regard to the particular protein such as

predicted secondary structure and predicted solvent accessibility generated from SSpro [4], protein sequence, beta partners as well as three-dimensional coordinates of alpha carbon for each residue in the protein. The dataset is redundancy reduced where the pairwise sequence identity of two sequences is less than 25%. Fig. 2 shows the example format of one of the data entry in the dataset.

```

1EJGA
46
T T A B P S I V A R S N F N V C R L P G
C C C C C C C C E E E C C C E E E C C C C C C H H H H C C C C E E E C C C C
0 34 33 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
50 50 10 10 50 50 50 50 10 50 50 10 10 10 10 10 10 50 50 50 50 10 10 10
16.9 12.8 4.2 13.9 11.5 5.9 13.6 10.7 9.6 10.7 8.9 11.3 9.5 9.1 14.9

```

Fig. 2 Example of dataset format

B. Input Features

This research is to construct different prediction models that consist of different combination of features in order to analyze and compare the effectiveness of the features implemented in the prediction model. There are total of five different kinds of features applied in this research which are local window feature, pairwise information feature, central segment window feature, segment average information feature, and protein information feature. These features are consisting of different information extracted from the dataset proteins including predicted secondary structure and solvent accessibility as well as the amino acid composition of the corresponding protein.

Local window feature is a 9-residue window feature which centered at each residue in each potential residue pair at which the distance of the residue in the pair is not less than the separation value set. In this features, each position within the window, there are 27 inputs which include 21 inputs for amino acids plus a gap, 3 inputs for predicted secondary structure (helix, coil, and sheet), 2 for the predicted solvent accessibility (exposed and buried) and 1 for the entropy. So this feature will have a size of 486.

In pairwise information feature, for each pair of position (i,j) in a multiple sequence alignment, 7 inputs are calculated, one input corresponds to the mutual information of the profiles of the two positions $\sum_{kl} p_{kl} \log(p_{kl}/(p_k p_l))$, where p_{kl} is the empirical probability of residues (or gap) k and l appearing at the two positions i and j simultaneously. While p_k and p_l refer to the probability of appearance of residues k and l respectively. Another two inputs are computed using cosine and correlation and one input for the amino acid type. Finally the last three inputs are regard to the pairwise potential values from three different pairwise potentials which are Levitt's pairwise potential [6], Jernigan's pairwise potential [8] and Braun's pairwise potential [12] for the residue pairs in the target sequence. This feature has a size of 16.

Third is the central segment window feature where this feature has a window size of 5 which locate at the position of $(i+j)/2$ which is the center of the potential residue pair. For each position of the window, 27 inputs are used same as in the

local window features which are 21 for amino acids plus a gap, 3 for predicted secondary structure, 2 for predicted solvent accessibility, 1 for the entropy. Therefore, central segment window feature has a size of 135. Another similar feature which is segment average information feature also using the information extracted from the segment between the residue pair. This feature has a size of 42 and it consist the information about the predicted secondary structure, solvent accessibility and the segment length information. Lastly is the protein information feature. This feature has a size of 30. In this feature, the global amino acid composition, secondary structure, and relative solvent accessibility of the target sequence are calculated.

C. Construction of Prediction Models

This research combines different features into several combinations and used to construct several prediction models. In order to compare and analyse the effectiveness of the features and with the availability of high performance computers from Centre of Information and Communication Technology (CICT) UTM Malaysia, a total of ten prediction models with different combination pairs of features are constructed. Table 1 shows the ten prediction models that constructed in this research.

The prediction models construction process consists of two major steps:

Step 1: Generation of SVM compatible input for all combination of features used in this research for training set proteins. In this step, corresponding necessary information are generated according to the feature involved and result in generation of a SVM readable format. This procedure continues until all the protein sequence in the training set is processed and all the information generated are appended to a single file that act as SVM input in later step for learning process.

Step 2: The generated SVM compatible files are then used as input for learning process using SVM Light to generate prediction models.

Throughout the learning process, steps done are written in Perl language with the jobs submitted in parallel to 4 of the nodes in the high performance computers provided. The overview of the process that generates the SVM compatible input is shown in Figure 3.

TABLE I
PREDICTION MODELS WITH CORRESPONDING FEATURES AND SIZE

Model	Features	Size
1	Pairwise Information + Local Window	502
2	Pairwise Information + Central Segment Window	151
3	Pairwise Information + Segment Average Information	58
4	Pairwise Information + Protein Information	46
5	Local Window + Central Segment Window	621
6	Local Window + Segment Average Information	528

7	Local Window + Protein Information	516
8	Central Segment Window + Segment Average Information	177
9	Central Segment Window + Protein Information	165
10	Segment Average Information + Protein Information	72

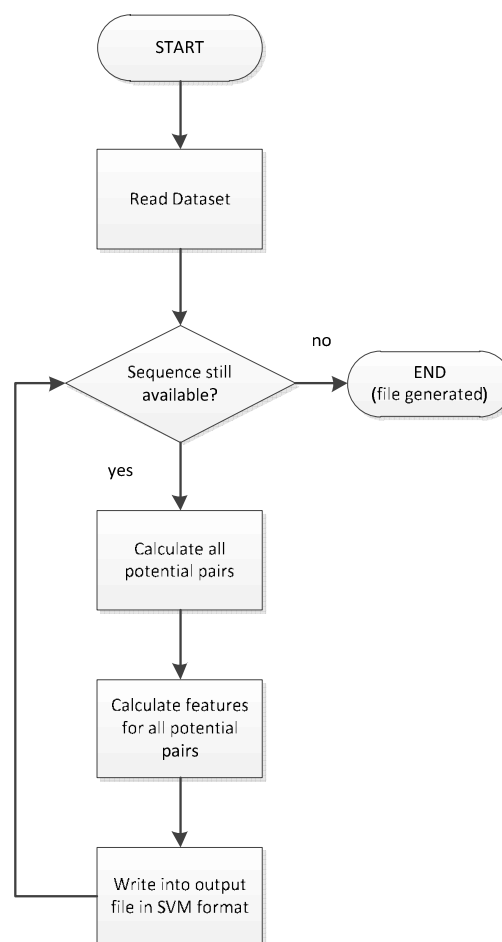


Fig. 3 Overview of generating SVM compatible input

D. Learning Using Support Vector Machine

Support vector machine (SVM) is used to predict an input feature vector which associated with a pair of residues to see if the two residues are in contact (positive) or not (negative). SVM provides several kinds of classification and regression methods that use linear to non-linear way to solve corresponding problem which control by the kernel methods. Kernel methods or kernel functions can re-map the data points into a higher dimensionality feature space solving the problem that are not solvable using linear method.

One of the key property of kernel method is the embedding does not need to be given in explicit form. Given a set of training data points, $S = S^+ \cup S^-$ where S^+ represent the positive samples and S^- represent the negative samples,

using the theory of risk minimization, support vector machines learn a classification function $f(x)$ as follow where a_i are non-negative weights and b is the bias.

$$f(x) = \sum_{x_i \in S^+} a_i K(x, x_i) - \sum_{x_i \in S^-} a_i K(x, x_i) + b \quad (1)$$

$K(x, x_i)$ is the kernel method used, x_i is the training data points and x is the target data point that is predicted to be positive or negative by taking the sign of $f(x)$. This research used radial basis function (RBF) kernel to train the prediction models. The gamma parameter (γ) in the RBF kernel is set to 0.025 which similar to previous research to ease the results comparisons [1]. The gamma parameter (γ) determines the RBF width centered on the support vectors. The RBF kernel method can be represented as following equations.

$$K(x, y) = e^{-\gamma \|x-y\|^2} \quad (2)$$

$$K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \quad (3)$$

E. Performance Measurement

In order to justify the results obtained and the performance of the prediction models, data are compared among the prediction models in terms of prediction performance. In this research, the performance is measured by accuracy and coverage where accuracy is the number of correct predictions per total number of predictions; its value shows the ability of the prediction model to get correct prediction out of the total number of prediction. Higher accuracy implies that the models are able to get more correct prediction. Meanwhile for coverage is the number of correct predictions per total number of true contacts. This parameter is similar to the sensitivity, where it shows the ability of the prediction model to identify true contacts. Higher value of sensitivity implies that the percentage of the true contacts identified is high as well. In this research, both measurements are correlated to each other, if the accuracy of a model is high; the coverage also shows high value. This implies that the model is efficient in predicting true contact out of the prediction.

III. RESULTS AND DISCUSSIONS

The performance of the constructed prediction models are evaluated by the comparing the prediction to the true contacts information. Data are being tabulated into tables and the measurement is done in terms of accuracy and coverage where accuracy is defined as the total number of correct prediction per total number of predictions while coverage is defined as the total number of correct prediction per total number of true contacts in the protein. The overall accuracy results for all prediction models are shown in Table II while for the coverage results are shown in Table III.

Based on the results reviewed, and also based on the prediction performance data shown in Table II and Table III,

accuracy of the contact map prediction is directly correlated to the information or features integrated into the prediction model. This can be seen in this research, the prediction results of model 4, model 9 and model 10 which integrating protein information features as one of the information to predict contact map. However, the results obtained is very low in accuracy and performance is not balance and consistent on all types of proteins. While for model 1, model 5, model 6 and model 7, these models obtained good results among others. This can be clearly seen by observing the average accuracy and coverage obtained as shown in Fig. 4 and Fig. 5. These four models yielded overall consistent results throughout this research, four of this model have similarity where each of the models also implemented local window information as one of the feature. This further implies the effectiveness of the information of local window feature in distinguishing residue contact from protein sequence. Based on the findings from previous researches, performance of the prediction is affected by the reliability of the information used such as multiple sequence alignment, predicted secondary structure, predicted solvent accessibility and so forth.

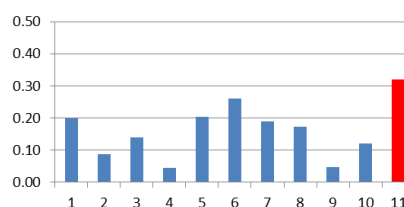


Fig. 4 Average prediction accuracy of all models (blue) and previous work model (red)

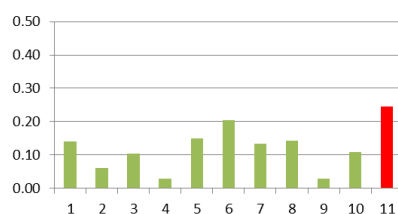


Fig. 5 Average prediction coverage of all models (blue) and previous work model (red)

Based on Fig. 4, prediction model 6 manages to get accuracy near the accuracy obtained by the previous work model [1]. This shows the significant of the features within the model especially the local window feature which shows significance on model 1, 5, and 7. Besides, based on Fig. 5, the coverage of model 6 is very near to the coverage obtained by previous work model. This shows that significance of the features used in model 6 has a high recall rate on the true contacts of the proteins.

TABLE II
RESULTS FROM DIFFERENT PREDICTION MODELS (ACCURACY)

Protein	length	TC	Type	ACCURACY									
				Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
1CTJA	89	95	alpha	0.023	0.011	0.180	0.023	0.023	0.124	0.023	0.135	0.000	0.214
1C75A	71	95	alpha	0.170	0.141	0.338	0.085	0.239	0.352	0.141	0.296	0.056	0.268
1CQYA	99	225	beta	0.172	0.283	0.131	0.051	0.192	0.131	0.182	0.253	0.212	0.121
1BMGA	98	220	beta	0.184	0.163	0.102	0.071	0.255	0.265	0.255	0.163	0.071	0.071
1MWPA	96	197	a+b	0.135	0.010	0.052	0.063	0.146	0.104	0.125	0.135	0.052	0.042
1G2RA	94	126	a+b	0.394	0.106	0.170	0.053	0.223	0.362	0.426	0.138	0.011	0.043
1CXQA	143	211	a/b	0.287	0.014	0.070	0.021	0.280	0.357	0.280	0.035	0.021	0.021
1F4PA	147	293	a/b	0.320	0.088	0.054	0.061	0.327	0.374	0.265	0.136	0.027	0.048
1A1HA	85	85	small	0.118	0.012	0.188	0.000	0.235	0.294	0.059	0.318	0.012	0.247
1EJGA	46	59	small	0.261	0.044	0.065	0.044	0.152	0.239	0.217	0.065	0.000	0.065
1AA0A	113	63	coil-coil	0.115	0.089	0.177	0.009	0.168	0.257	0.115	0.230	0.044	0.177

TABLE III
RESULTS FROM DIFFERENT PREDICTION MODELS (COVERAGE)

Protein	Length	TC	Type	COVERAGE									
				Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
1CTJA	89	95	alpha	0.0211	0.011	0.168	0.021	0.021	0.116	0.021	0.126	0.000	0.200
1C75A	71	95	alpha	0.126	0.105	0.253	0.063	0.179	0.263	0.105	0.221	0.042	0.200
1CQYA	99	225	beta	0.076	0.124	0.058	0.022	0.084	0.058	0.080	0.111	0.093	0.053
1BMGA	98	220	beta	0.082	0.073	0.046	0.032	0.114	0.118	0.114	0.073	0.032	0.032
1MWPA	96	197	a+b	0.066	0.005	0.025	0.031	0.071	0.051	0.061	0.066	0.025	0.020
1G2RA	94	126	a+b	0.294	0.079	0.127	0.040	0.167	0.270	0.318	0.103	0.0080	0.032
1CXQA	143	211	a/b	0.194	0.010	0.047	0.014	0.190	0.242	0.190	0.024	0.014	0.014
1F4PA	147	293	a/b	0.160	0.044	0.027	0.031	0.164	0.188	0.133	0.068	0.014	0.024
1A1HA	85	85	small	0.118	0.012	0.188	0.000	0.235	0.294	0.059	0.318	0.012	0.247
1EJGA	46	59	small	0.203	0.034	0.051	0.034	0.119	0.186	0.170	0.051	0.000	0.051
1AA0A	113	63	coil-coil	0.206	0.159	0.159	0.016	0.302	0.460	0.206	0.413	0.079	0.318

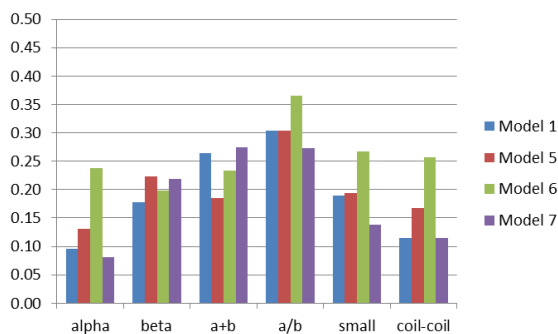


Fig. 6 Average accuracy on different protein structure

Meanwhile, the performance affection in terms of effectiveness of the features used to build the prediction models, the performance also affected by the type of the proteins that used for testing. Fig. 6 shows the average accuracy of model 1, 5, 6, 7 based on different types of protein structure of the tested proteins.

According to Fig. 6, clearly shown that the types of structure such as beta, a+b, and a/b tend to be predicted with higher accuracy. Refer to the research done previously [4], [7], the contacts that within beta-sheets are predicted with higher accuracy than contacts that between alpha helix and a beta strands or between alpha helix. This is probably because of the strong restraints between beta-strands such as hydrogen bond gives the increased accuracy. This are shown more clearly in Fig. 7 by average the accuracies obtained for all models based on different type of structure.

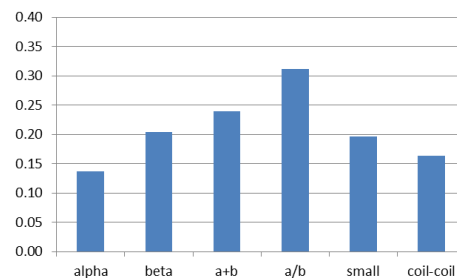


Fig. 7 Average accuracy on different types of structure

IV. CONCLUSION

Different with previous research, this research concentrate on the determination and analysis of the effectiveness of the features used in protein contact map prediction to contribute and improve protein contact map prediction which the main advantage of this research that is not so concentrated in previous research. Even though the accuracy achieved by the constructed prediction models is lower than previous research, however, based on the results that obtained by combination of two features, highest average results achieved is 81% of the average accuracy from previous research (5 features). This implies that the feature information (2 features) used in model 6 is efficient in predicting protein contact map, and indirectly implies the existence of unnecessary or inefficient features. This also shortens the execution time of the process with more experiments can be conducted. However, due to the time constraint, this research is done using combination of two features, more variety of combination can be made with implement of more features can be done in future. Therefore, in this research, with the construction of the multiple prediction models with different combination of features, effectiveness of the features that affect the performance of the prediction are identified, and further improve the knowledge

regard to the effective information to be used in protein residue contact prediction. We believe that, in order to further increase the accuracy of the predictions for all kind of proteins, a more informative feature of proteins is needed even combination of informative features that able to distinct the contacts among residues. This research had shown that the use of local window feature in the prediction model yield decent results among others, while on the other hand, this research also shows that combination of local window feature and segment average information (model 6) produce balance results among all structures. By the identification of these information, through combining others effective features with the one shown in this research, it is believed that this can help to improve the accuracy of the prediction.

ACKNOWLEDGMENT

Both authors thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grant (Vot number: Q.J130000.7107.01H29).

REFERENCES

- [1] J. Cheng, P. Baldi, "Improved Residue Contact Prediction Using Support Vector Machines and A Large Feature Set," *BMC Bioinformatics*, Vol. 8, no. 1, 2007.
- [2] X. Yuan, C. Bystroff, "Protein Contact Map Prediction," in *Computational Methods for Protein Structure Prediction and Modelling*, X. Ying, X. Dong, L. Jie, Ed. Heidelberg: Springer, 2007, pp. 255-277.
- [3] L. Bartoli, E. Capriotti, P. Fariselli, P. L. Martelli, R. Casadio, "The Pros and Cons of Predicting Protein Contact Maps," in *Protein Structure Prediction*, 2nd ed., M. Zaki, C. Bystroff, Ed. New Jersey: Humana Press, 2008, pp. 199-217.
- [4] J. Cheng, A. Randall, M. Sweredoski, P. Baldi, "SCRATCH: a protein structure and structural feature prediction server," *Nucleic Acids Research*, Vol. 33, pp. 72-76, 2005.
- [5] P. Fariselli, R. Casadio, "A Neural Network Based Predictor of Residue Contacts in Proteins," *Protein Engineering*, Vol. 12, pp. 15-21, 1999.
- [6] E. Huang, S. Subbiah, J. Tsai, M. Levitt, "Using a Hydrophobic Contact Potential to Evaluate Native and Near-Native Folds Generated by Molecular Dynamics Simulations," *J. Mol. Biol.*, Vol. 257, no. 3, pp. 716-725, 1996.
- [7] MacCallum, "Striped Sheets and Protein Contact Prediction," *Bioinformatics*, Vol. 20, no. 8, pp. 224-231, 2004.
- [8] S. Miyazawa, R. Jernigan, "An empirical energy potential with a reference state for protein fold and sequence recognition," *Proteins*, Vol. 36, pp. 357-369, 1999.
- [9] G. Pollastri, P. Baldi, "Improved Prediction of The Number of Residue Contacts in Proteins By Recurrent Neural Networks," *Bioinformatics*, Vol. 17, pp. 234-242, 2001.
- [10] A. N. Tegge, Z. Wang, J. Eickholt, J. Cheng, "NNcon: Improved Protein Contact Map Prediction Using 2D-Recursive Neural Networks," *Nucleic Acids Research*, Vol. 37, pp. 515-518, 2009.
- [11] Y. Zhao, G. Karypis, "Prediction of Protein Contact Maps Using Support Vector Machines," presented at *IEEE Symposium on Bioinformatics and Bioengineering*, Bethesda, MD, USA, March 10-12, 2003.
- [12] H. Zhu, W. Braun, "Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting," *Protein Sci.*, Vol. 8, pp. 326-342, 1999.