HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1059.2019-0073
Natural Sciences 2019, Volume 64, Issue 10, pp. 61-69
This paper is available online at http://stdb.hnue.edu.vn

IMPROVING PREDICTED RESIDUE-RESIDUE CONTACTS BY FILTERING FALSE POSITIVE SAMPLES

Le Thi Tu Kien[1] and Nguyen Quynh Diep[2]

[1]*Faculty of Information Technology, Hanoi National University of Education*
[2]*Faculty of Computer Science and Engineering, Thuy Loi University*

Abstract. Research on the residue-residue contacts in interactive proteins is meaningful in determining the function and structure of proteins, structure-based drug design, and disease treatment. Previous methods showed good predicted results, however, the number of false positive non-residue-residue contacts (non-RRCs) is still much higher than the number of true positive residue-residue contacts (RRCs). In this research, we propose a method to eliminate false positive non-RRCs enhancing the predicted quality. The experimental results showed that our proposed method to increase the predicted quality in some cases.

Keywords: Protein, protein domain, protein-protein interactions, domain-domain interactions, residue-residue contacts.

1. Introduction

Proteins are macromolecules made up of one or more polypeptide chains, which are chains of amino acid residue. These chains can be coiled or folded in many ways to form different spatial structures of proteins.

Proteins form, maintain and replace cells in the body. Protein deficiency leads to malnutrition, slow growth, immunodeficiency, adversely affecting the function of organs in the body. It can be said that protein is related to all life functions of the body such as circulation, respiratory, genital, digestive, excretory, mental activity, etc...

To perform their functions, proteins interact with other proteins or other molecules in the cell. This interaction affects the activities of living in cells and the life processes of organisms. Therefore, the study of protein interactions is one of the most important issues in biology and bioinformatics.

The interaction of proteins is studied at three levels. At the first level, it is interested in whether two or more single proteins interact with each other. While in the second

Received July 16, 2019. Revised July 22, 2019. Accepted August 27, 2019.
Contact Le Thi Tu Kien, e-mail address: kienltt@hnue.edu.vn

level is interested in which domains of proteins interact. Many studies have demonstrated that in each protein there may be one or several protein domains. Each of these protein domains takes on one or more specific functions of the protein. When interacting with each other, depending on what biological functions need to be done, the protein domains that have the corresponding functions interact with each other to form interactive interfaces. The third level refers to how residues at the interactive surface contact together. Understanding the interactive surface in details will help to understand what the biological function is performed, supporting the process of predicting protein complexes and disease treatment. Biological experimental methods to perform the above problems often take a lot of time and cost. Therefore, many computational methods have been proposed to support solving them [1-10].

In recent years, residue-residue contacts (RRCs) prediction has yielded positive results. The Weigt *et al.* [4] developed the Direct-coupling analysis algorithm to find information RRCs of proteins. Then, Marks *et al.* [11] used this algorithm to predict the tertiary structure of proteins. In addition, González *et al.* [8] used Interaction profile Hidden Markov Model (ipHHM) and Support Vector Machine (SVM) to predict RRCs. Taking the advantages of the methods [4, 8, 12] into account, Le *et al*. [9] developed a RRCs prediction method that integrates formation about structure of proteins, coevolution relationship, and amino acid pairwise contact potentials.

Although experimental results have demonstrated that the proposed method in [9] gives better predictive results than previous methods, the number of misclass predicted non-RRCs (false positive samples) is still much higher than the number of true predicted RRCs (true positive samples). In this research, we propose a method to remove false positive samples to improve the quality of predicting results.

In the next section, we will present an overview of the RRCs prediction method in [9], the proposed method, experimental and results.

2. Content

2.1. Prediction residue-residue contacts by multiple interaction information

In [9], we developed the RRCs prediction method by integrating information of residue pairs from several sources. The general steps of the method are described as follows (Figure 1):

In the first step, data filtering, a subset of the pair of domain-domain interaction (DDIs) together with their residue-level information is filtered provided that the sequence distances between sequence domains within the query DDI and sequence domains in a filtered DDIs are less than a threshold t. In particular, the sequence distance is the smallest number of substitutions to perform a conversion of this protein domain sequence into another. The smaller the number of substitutions, the more is identical sequences.

In the second step, feature construction, the set of filtered DDIs is used to train two ipHMM models. Then, these ipHMMs are used to calculate the Fisher vector for each residue. ipHMM works to pass residue-level interactive information of domain protein to others in the same protein domain family which unknown interactive information.

Each residue in the protein domain sequence is represented by a Fisher vector of size 20 corresponding to the number of amino acid such as:

$$\left\langle \frac{\partial}{\partial e_{M_i}^{A_1}} \log(x \mid \theta), \frac{\partial}{\partial e_{M_i}^{A_2}} \log(x \mid \theta), \ldots, \frac{\partial}{\partial e_{M_i}^{A_{20}}} \log(x \mid \theta) \right\rangle \quad (1)$$

In the expression (1), $\log(x \mid \theta)$ is the probability of the domain x given the model θ, which is a parameter of an ipHMM representing a domain family. $e_{M_i}^{A_k}, 1 \le k \le 20$ is the emission probability of amino acid A_k at the interacting or noninteracting match state M_i. The feature vector for a pair of residues is a concatenation of two Fisher vector. At the same time, coevolution scores and contact potential scores for residue pairs based on direct coupling analysis algorithm (mfDCA) [4] and amino acid pairwise contact potentials (AAPCPs) [12] are computed. All ipHMM, mfDCA, and AAPCPs features are combined to form the feature vector of each residue pair in the training data set and test set.

In the third step, classification, the training data set is used to train an SVM classification model. This model is then used to classify residue pairs in the test set into two classes RRC or non-RRC.

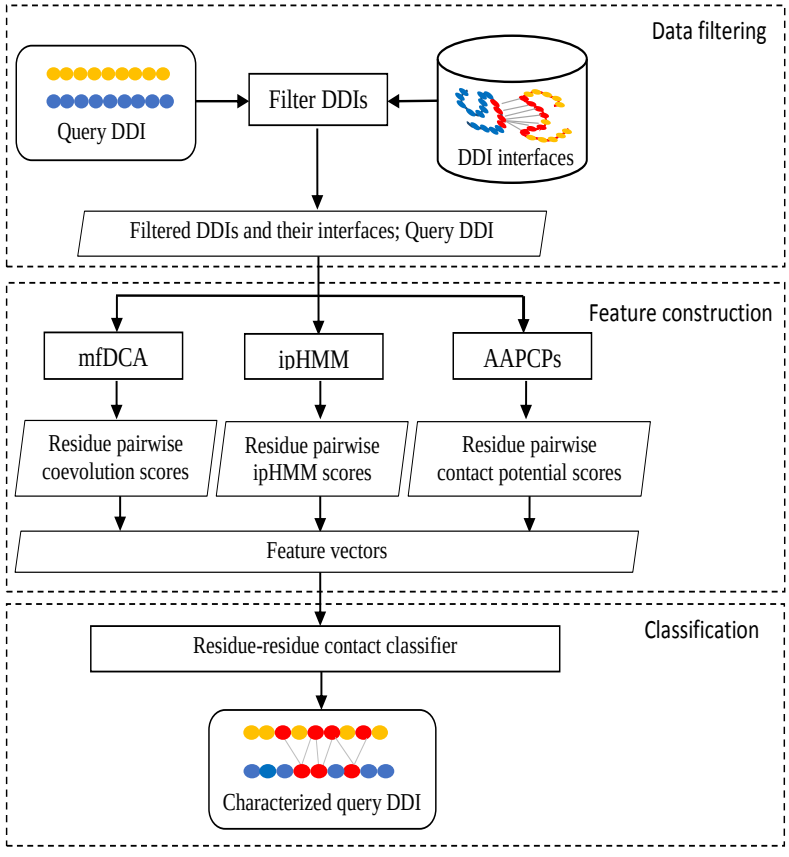


Figure 1. Steps to perform a prediction of residure-residure contacts in [9]

Experimental results in [9] proved that the predictor was high accurately and outperformed previous methods. However, this method still has some problems as follows:

Firstly, producing residue pairs by each residue in the first sequence sequentially matched with each other residues of the second sequence (i.e., if the protein domain pair has *m* and *n* residues, there will be $m \times n$ residue pairs generated) would cause a problem. Suppose a residue in the sequence M is predicted to contact with two residues in the sequence N, which are located very far from each other. In this case, it is likely that one of these two RRCs is false positive, i.e., one of the two RRCs does not contact but it is predicted to contact.

Secondly, for each pair of protein domains, the number of RRCs is much less than the number of non-RRCs. This imbalance leads to a case such as even though the false positive rate of non-RRCs is low, (between 2 and 5 percent), but the number of misclass non-RRCs is still much more than the number of RRCs. For example, suppose that the sequence *M* has m=101 residues and the sequence *N* has n=100 residues. Hence, there is $m \times n$=101x100=10100 residue pairs and 100 pairs are RRCs while the number of non-RRCs is 10000 pairs. If the trained SVM model has true positive rate *TPR*= 80% and false positive rate *FP*R=3%, the number of true predicted RRCs is 80 pairs (over 100 pairs) while the number of false predicted non-RRCs is 300 pairs (over 10000 pairs). Thus, the number of false predicted non-RRCs is three times more than the true predicted RRCs. Therefore, it is necessary to filter these non-RRCs false positive.

Based on the above analysis, in the next section, we propose a solution to increase the quality of the predicted results of the method [9] .

2.2. Filtering non-RRCs false positive samples

Our main idea to filter false positive non-RRCs is that if one residue in the first sequence contacts to two residues in the second sequence, and if these two residues in the later locate far from each other, we will keep the first RRC and remove the RRC, that has a higher order of the residue in the second sequence. The following algorithm explicitly describes in details of this idea.

Input*:*

- A list *P* consist of predicted RRCs which have the first residue belong to the protein domain sequence *M* and the second residue belong to the protein domain sequence *N*.

- The orders of residues in the sequences

Output:

- *Q* is a list of remaining RRCs after filtering out the false positives samples

Method:

Step 0: Assign an empty list Q.

Step 1: Choosing one RRCs (x,y) in the list P and assign it to the list T.

Step 2: Finding other RRCs in the P that the first residue is x, then assign them to the list *T*.

Step 3: Sort the list *T* in ascending order by the order of residues belongs to the sequence *N*.

Step 4: Choosing the first RRC *(s, z)* in the list *T*, then assign it to the list *Q*. For each RRCs *(s, i)* from second RRCs in the *T*, calculate the distance between the residue *z* and the residue *y* based on the order of residues in the sequence. If the distance is greater than a threshold *d*, remove the RRC *(s, i)* from the *T*. Otherwise, assign *(s, i)* to the *Q*.

Step 5: Update the list *P* by removing all RRCs that exist in list *T*. Then, empty the list *T*.

Step 6: If the list *P* is empty, go to step 7. If *P* remains only one RRC, assign that RRC to the *Q* then go to step 7.

Step 7: End.

2.3. Experiments and results

2.3.1. Experimental data

To evaluate the effectiveness of the method proposed in section 2.2, we perform experiments on four datasets listed in Table 1. The first column is the sequence number of data sets, the second and third columns are the names of the Pfam protein domain families, the fourth column is the number of DDIs. Each set of data is built based on the following process: For each DDI, information about domain protein sequences is obtained from the Pfam database. In the Pfam database, domain protein sequences are grouped into Pfam domain protein families. Then, the interaction information at residue level of DDIs is extracted from the 3D Interacting Domain database (3DID). After that, we mapped Pfam domain information organized in 3did to PDB database to retrieve domain sequences for DDIs. Figure 2 shows the information of the Pfam domain family C1-set. In addition, the information of amino acid pairwise contact potentials is also collected from the AAindex database [12].

Table 1. The list of four experimental data sets

ID	DomainM	DomainN	#DDIs
1	C1-set	C1-set	**482**
2	Fib_alpha	Fib_alpha	**101**
3	Insulin	Insulin	**103**
4	Rhv	Rhv	**101**

2.3.2. Measures

We use the measure Matthew correlation coefficient (*MCC*) in the expression (2) to evaluate the performance of our proposed method. If the value of *MCC* is higher, it is better. *MCC* is also a good measure for imbalanced data sets.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

In the expression (2), *TP* (True Positive) and *TN* (True Negative) denote the number of positive and negative samples correctly classified, while *FN* (False Negative) and *FP* (False Positive) denote the numbers of positive and negative samples are misclassified.

2.3.2. Results

For each data set as shown in Table 1 and for each threshold value *t* (*t* = 0.1, 0.2, 0.3, 0.5, 0.7, 0.9), we perform odd one out five times to evaluation method. For each time, randomly select a pair of DDI as the query DDI and the remaining DDIs are training set. After predicting label 1 or 0 (RRC or non-RRC) for residue pairs of the query DDI, we apply the algorithm proposed in section 2.3 to remove residue pairs that are considered as false positive samples. The value of the threshold *d* is 10.

Figure 3 shows average *MCC* values (vertical axis) on four sets of C1_set - C1_set, Fib_alpha - Fib_alpha, Rhv - Rhv, Insulin - Insulin values corresponding to the values of the threshold *t* (horizontal axis) from 0.1 to 0.9 of before and after filtering non-RRCs false possitive. In this figure, we made the following observation:

- Firstly, for the C1_set - C1_set data set, the average MCC values after filtering non-RRCs is higher at threshold values *t* of 0.1, 0.2, 0.7, and 0.9. On the other hand, *MCC* values are lower at threshold values t of 0.3 and 0.4.
- Secondly, for the Fib_alpha - Fib_alpha data set, the filtering gives better average *MCC* values at threshold values t from 0.1 to 0.5, but it is worse at the value of t from 0.7 to 0.9.
- Thirdly, for the Rhv family – Rhv data set, our method gives better average MCC values at all values of the threshold *t*.
- Finally, with the pair of Pfam Insulin – Insulin data set, our algorithm gives good average MCC results at the threshold value *t* of 0.1, 0.3, 0.9, and gives lower results at the remaining values.

Based on the above observations, we can conclude that our proposed method gives the average value of MCC better or worse depending on the value of t and each data set. Especially when *t* is equal to 0.1 or is equal to 0.2, all data sets give better *MCC* values. These results lead to some problems that we need to consider. Choosing the first RRC in the list T and then based on it to remove other RRCs might not suitable. Furthermore, two protein domains are often touching each other on some regions (Figure 4). It means that some adjacent residues of this sequence are in touch with some adjacent residues of another. This case does not include in our algorithm.

Fields	pdbid	chain	ORGANISM_SCIENTIFIC	ORGANISM_COMMON	Sequence	seqResNum	interResNum	interAA	interactions
1	'1ad0'	'C:118-205'	'HOMO SAPIENS'	'HUMAN'	'PPSDEQLKS...	*1x88 double*	*1x18 cell*	'KKQQKKK...	*1x88 double*
2	'1ad9'	'B:126-208'	'HOMO SAPIENS'	'HUMAN'	'PCSRSTSES...	*1x83 double*	*1x13 cell*	'EESSSSTTT...	*1x83 double*
3	'1adq'	'A:244-333'	'HOMO SAPIENS'	'HUMAN'	'PPKPKDTL...	*1x90 double*	*1x10 cell*	'LMIIISSSRR'	*1x90 double*
4	'1adq'	'A:352-437'	'HOMO SAPIENS'	'HUMAN'	'PPSQEEMT...	*1x86 double*	*1x16 cell*	'NGGGQM...	*1x86 double*
5	'1adq'	'A:352-437'	'HOMO SAPIENS'	'HUMAN'	'PPSQEEMT...	*1x86 double*	*1x6 cell*	'VVSHHN'	*1x86 double*
6	'1akj'	'A:188-271'	'HOMO SAPIENS'	'HUMAN'	'HMTHHAV...	*1x84 double*	*1x28 cell*	'TTTTTTQQ...	*1x84 double*
7	'1akj'	'A:188-271'	'HOMO SAPIENS'	'HUMAN'	'HMTHHAV...	*1x84 double*	*1x9 cell*	'EEEEQQDDV'	*1x84 double*
8	'1akj'	'B:11-92'	'HOMO SAPIENS'	'HUMAN'	'SRHPAENG...	*1x82 double*	*1x6 cell*	'KKKKDD'	*1x82 double*
9	'1bfo'	'G:119-205'	'RATTUS RATTUS'	'BLACK RAT'	'PPSTEQLAT...	*1x87 double*	*1x8 cell*	'VENNNGNN'	*1x87 double*
10	'1dee'	'A:119-206'	'HOMO SAPIENS'	'HUMAN'	'PPSDEQLKS...	*1x88 double*	*1x9 cell*	'EETHQQGLS'	*1x88 double*
11	'1e4x'	'H:123-204'	'MUS MUSCULUS'	'HOUSE MOUSE'	'PVCGDTTG...	*1x82 double*	*1x12 cell*	'NSSSGGSS...	*1x82 double*
12	'1etz'	'A:122-208'	'MUS MUSCULUS'	'HOUSE MOUSE'	'TPSSEELET...	*1x87 double*	*1x13 cell*	'NNNRRRR...	*1x87 double*
13	'1f3d'	'J:119-206'	'MUS MUSCULUS'	'HOUSE MOUSE'	'PPSSEQLTS...	*1x88 double*	*1x21 cell*	'KKDDDGG...	*1x88 double*
14	'1f3d'	'K:134-215'	'MUS MUSCULUS'	'HOUSE MOUSE'	'PGSAAQTN...	*1x82 double*	*1x9 cell*	'AAAQQQN...	*1x82 double*
15	'1fe8'	'M:119-206'	'HOMO SAPIENS'	'HUMAN'	'PPSSEQLTS...	*1x88 double*	*1x20 cell*	'DNNVKQQ...	*1x88 double*
16	'1fj1'	'A:119-206'	'MUS MUSCULUS'	'HOUSE MOUSE'	'PPSSEQLTS...	*1x88 double*	*1x18 cell*	'KDDNNEQ...	*1x88 double*
17	'1fl5'	'B:125-208'	'MUS MUSCULUS'	'HOUSE MOUSE'	'APSSKSTSG...	*1x84 double*	*1x10 cell*	'SSSNNNN...	*1x84 double*
18	'1fn4'	'D:119-200'	'RATTUS NORVEGICUS'	'NORWAY RAT'	'PGSAAQTN...	*1x82 double*	*1x12 cell*	'NSSGGGA...	*1x82 double*
19	'1ggi'	'L:119-206'	'MUS MUSCULUS'	'HOUSE MOUSE'	'PPSSEQLTS...	*1x88 double*	*1x17 cell*	'DEERRRRR...	*1x88 double*
20	'1hxm'	'B:139-222'	'HOMO SAPIENS'	'HUMAN'	'PSIAETKLQ...	*1x84 double*	*1x6 cell*	'KKKSQE'	*1x84 double*

Figure 2. Information of their C1-set pfam

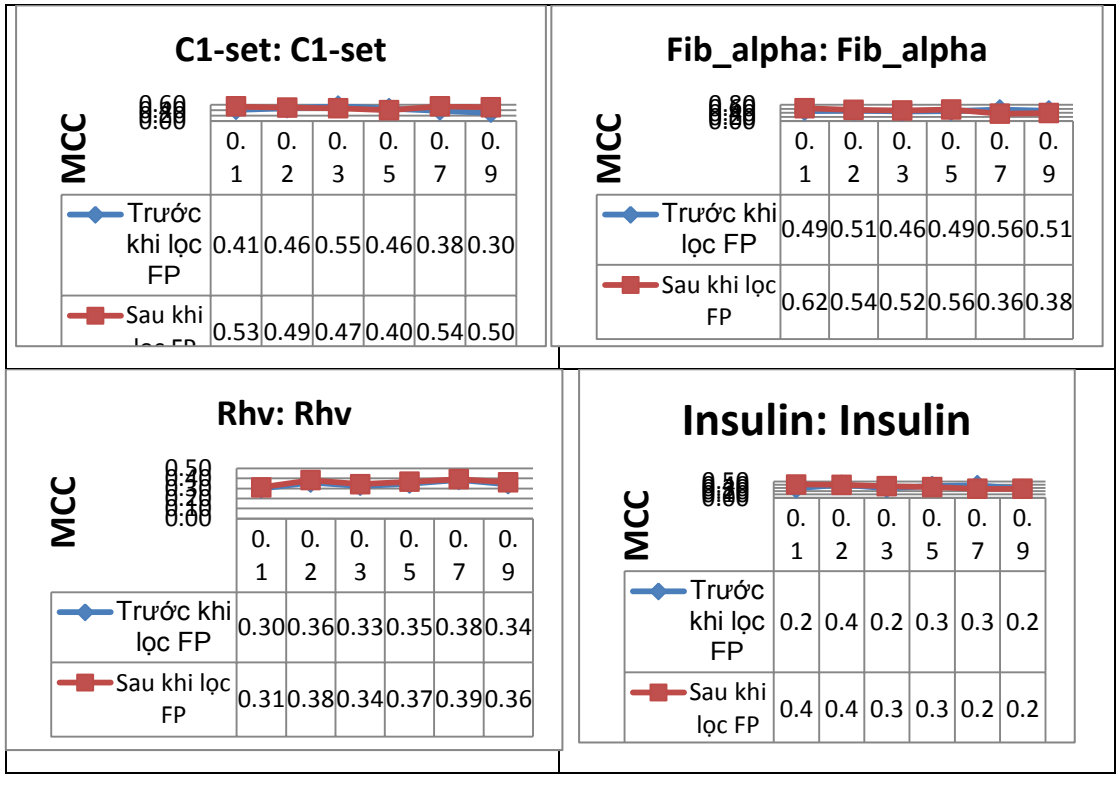
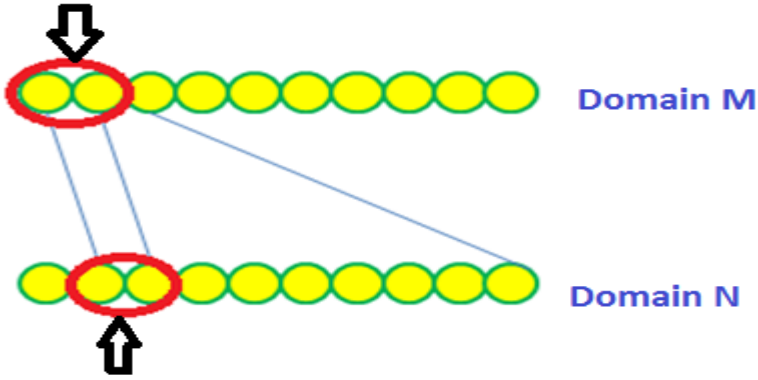


Figure 3. Comparison of MCC values of four data sets



3. Conclusions

Predicting RRCs from DDIs is significant in predicting the structure of proteins complexes, drug preparation, and disease treatment. In this study, we propose a solution to expect better the quality of predictive results. Although the proposed method is not effective in all cases, it leads to some further issues that need to be studied. Firstly, if each touch regions of a DDI contain several adjacent residues of two domains, and if one residue of the touch region is predicted to contact with a single residue that far away from it, this predicted RRC might be false positive non-RRC. Secondly, after predicting RRCs for the query DDI, we can compare the network of RRCs of the query DDI with the network of RCCs of the nearest DDI in the training set, and then based on it to remove false positive samples.

Acknowledgements: This work was supported by the Hanoi National University of Education (SPHN16-03TT).

REFERENCES

[1] T. M. W. Nye, C. Berzuini, W. R. Gilks, and M. M. Babu, 2006. *Predicting the Strongest Domain-Domain Contact in Interacting Protein Statistical*. Applications in Genetics and Molecular Biology, Vol. 5, No. 1.

[2] R. Jothi, P. F. Cherukuri, A. Tasneem, and T. M. Przytycka, 2006. *Co-evolutionary Analysis of Domains in Interacting Proteins Reveals Insights into Domain - Domain Interactions Mediating Protein - Protein Interactions*. J. Mol. Bio, pp. 861-875.

[3] H. Zhou and S. Qin, 2007. *Structural bioinformatics Interaction-site prediction for protein complexes: A critical assessment*. Bioinformatics, Vol. 23, No. 17, pp. 2203-2209.

[4] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, 2011. *Identification of direct residue contacts in protein - protein interaction by message passing*. PNAS, Vol. 106, No. 1, pp. 67-72.

[5] A. W. Ghoorah, M. Devignes, M. Smaïl-tabbone, and D. W. Ritchie, 2011. *Spatial clustering of protein binding sites for template based protein docking*. Bioinformatics, Vol. 27, No. 20, pp. 2820-2827.

[6] C. Chen *et al.*, 2012. *Protein-Protein Interaction Site Predictions with Three-Dimensional Probability Distributions of Interacting Atoms on Protein Surfaces*. PlosOne, Vol. 7, Iss. 6.

[7] R. A. Jordan, Y. El-manzalawy, D. Dobbs, and V. Honavar, 2012. *Predicting protein-protein interface residues using local surface structural similarity*. BMC Bioinformatics, Vol. 13, No. 1.

[8] A. J. González, L. Liao, and C. H. Wu, 2013. *Prediction of contact matrix for protein-protein interaction.* Bioinformatics, Vol. 29, Iss. 8, pp. 1018-1025.

[9] T. Kien T. Le *et al.*, 2014. *Predicting residue contacts for protein-protein interactions by integration of multiple information*. J. Biomed. Sci. Eng., Vol. 07, No. 01, pp. 28-37.

[10] T. Du, L. Liao, C. H. Wu, and B. Sun, 2016. *Prediction of residue-residue contact matrix for protein-protein interaction with Fisher score features and deep learning methods*. Methods of Elsevier, Vol. 110, pp. 97-105.

[11] D. S. Marks *et al.* 2011. *Protein 3D structure computed from evolutionary sequence variation*. PLoS One, Vol. 6, Iss. 12.

[12] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, 2008. *AAindex: Amino acid index database, progress report 2008*. Nucleic Acids Research, Vol. 36, Iss Database, pp. 202-205.