Protein Interface Residues Prediction Based on Amino Acid Properties Only

Bing Wang[1,2], Peng Chen[3], and Jun Zhang[2]

[1] School of Electrical Engineering and Information, Anhui University of Technology,
Maanshan, Anhui 243002, China
[2] Department of Chemistry, University of Louisville, Louisville, KY 40292, USA
[3] Hefei Institute of Intelligent Machines, Chinese Academy of Sciences,
Hefei, Anhui 230031, China
wangbing@ustc.edu

Abstract. Protein-protein interactions play essential roles in protein function implementation. A computational model is introduced in this work for predicting protein interface residues based on amino acid chemicophysical properties only. 17 amino acid properties are selected from AAindex database and used as input features of a prediction model which is constructed by support vector machines method to infer protein interface residues in protein hetero-complexes. The results achieved in this work demonstrated the properties used in this work can actually capture up the difference between interface and noninterface residues.

Keywords: Amino Acid Property, Protein Interface Residues, Support Vector Machines, Hetero-complexes.

1 Introduction

The interactions between proteins play a very important role for the majority of biological functions, such as DNA replication, signal transduction, immunological recognition and protein synthesis [1]. In recent years, many methods have been developed to predict protein interaction sites or location of interface residues. Those approaches have addressed various aspects of protein structure and properties, such as amino acid composition [2-7], solvent accessibility [8], sequence entropy and secondary structure [9, 10], evolutionary profiles and conservation score [4].

In this paper, a method based on the amino acid chemical and physical properties is proposed for prediction of protein-protein interface residues using a a support vector machines (SVMs) predictor. The amino acid properties were extracted from the amino acid index (AAindex) database. Then has been constructed for differ protein interface residues from non-interface residues in protein chains. The results based on a non-redundant protein chains set show that the amino acid properties what we used in this work are effective to capture the difference between interface and non-interface residues.

D.-S. Huang et al. (Eds.): ICIC 2011, LNBI 6840, pp. 448–452, 2012.

2 Methods

2.1 Amino Acid Properties

The complexes used in this work were same as our previous study [4]. The amino acid properties are extracted from AAindex database, which is a database of numerical indices representing various physiochemical and biochemical properties of amino acids. Only the AAindex1 database was used in this work which comprises 544 sets of numerical indices for the 20 amino acids, and all of them are derived from published literature. Firstly we remove the 13 property items for there are some N/A values within it. Then all similar properties are eliminated for information-redundancy. The Pearson's correlation coefficient was used here to calculate similarity values among the properties and 0.3 is set up as the similarity threshold. As a result, 17 amino acid properties are used in this work for identifying protein interface reside

$$S(p_i,p_j)=\frac{\mathrm{cov}(p_i,p_j)}{\sigma_{p_i}\sigma_{p_j}} \tag{1}$$

where p_i and p_j denotes amino acid property which is a 20-dimensional vector where the value in each dimension is the index of one of 20 amino acids, and cov means covariance, σ means standard deviations.

Table 1. The selected amino acid properties

Property id	Property description	Property id	Property description
ANDN920101	alpha-CH chemical shifts	JOND920102	Relative mutability
ARGP820101	Hydrophobicity index	KHAG800101	The Kerr-constant increments
BEGF750101	Conformational parameter of inner helix	FAUJ880104	STERIMOL length of the side chain
BUNA790103	Spin-spin coupling constants 3JHalpha-NH	PALJ810107	Normalized frequency of alpha-helix in all-alpha class
BHAR880101	Average flexibility indices	RACS820114	Value of theta(i-1)
BURA740102	Normalized frequency of extended structure	WERD780103	Free energy change of alpha(Ri) to alpha(Rh)
GEOR030101	Linker propensity from all dataset	YUTK870102	Unfolding Gibbs energy in water, pH9.0
CHOP780204	Normalized frequency of N-terminal helix	CHAM830102	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet
CHOP780215	Frequency of the 4th residue in turn		

2.2 Predicting Model Construction

Similar to previous works, a sliding window technique is used in this study in order to involve the association among neighboring residues because protein interface is formed by some residues which closed to each other in spatial position. Therefore, the input vector of predicting model is fed with a window of 11 residues, centered on the target residue and including the five spatially neighboring residues on each side. As a result, each residues is represented by a 11×17=187 components vector.

A ten-fold cross-validation strategy was employed to conduct the subsequent experiments. In this strategy, proteins in the dataset are divided into 10 subsets which consist of roughly the same number of proteins, one of them is for the test process and the other ones are for the model training process. The number of positive samples or so-called interface residues is much smaller than that of negative samples or non-interface residues. Only 34.3% of the samples are interface residues in this work, which leads to a rather imbalanced data distribution. To overcome this problem, we will randomly select negative samples in the model training process to make sure the number negative and positive samples is same.

3 Results

3.1 Correlation of the Selected Properties

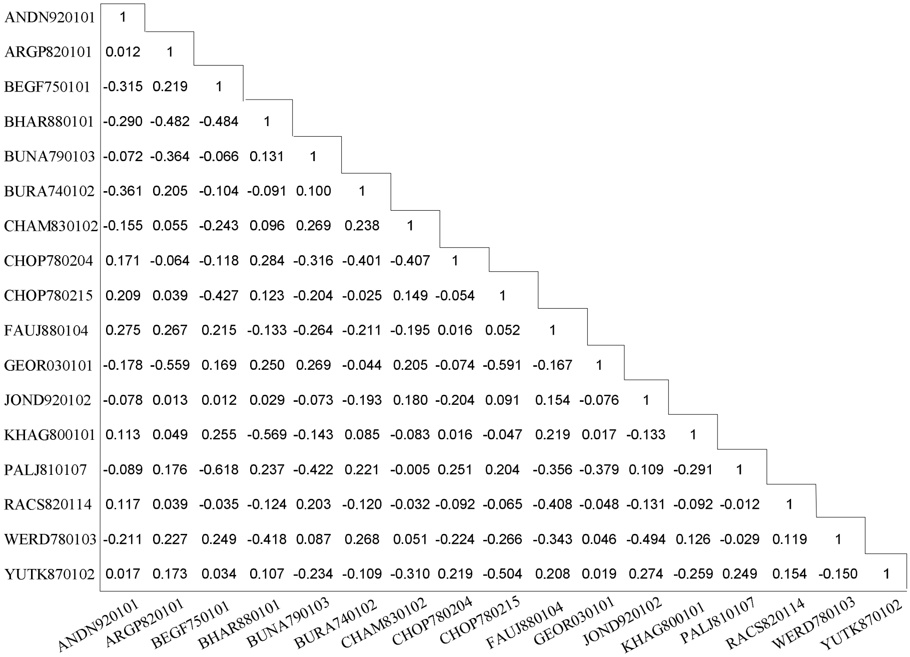
Obviously, More discriminative power what the property can differentiate interface and non-interface residues is, more successful the prediction model will be. The similarities among 17 amino acid properties can be seen in Table 2.

It can be seen that the maximum similarity among the selected amino acid properties is 0.28, the minimum value is -0.62, and the mean value is -0.05. The very low correlation among the selected amino acid properties means that there is no information-redundancy in the feature set what we used in this work. Reducing the number of amino acid properties from the original 544 to present 17 decreases the computational complexity drastically and speeds up the model learning process. Meanwhile, removing most irrelevant and redundant features from the data can enhance generalization capability of our proposed model.

3.2 Prediction Performance

Among 10329 protein surface residues, our prediction shows 4484 of them are assigned to +1 (interface residue), and 5845 of them are assigned to -1 (non-interface residue) by our proposed prediction model, respectively. Based on the definitions of performance measures, our proposed model can obtain a Sen of 57.9%, a Spec of 65.0%, a Prec of 46.9%. Furthermore, the value of MCC our model achieved in this work, 0.22, denotes that the selected properties can actually captures up the difference between the interface and non-interface residues.

In order to further evaluate our presented model, we compared it with two previous models: one is backpropagation (BP) neural network model, another is radial basis function (RBF) neural network model which is optimized by expectation maximum algorithm (EM), and these two models use amino acid residue sequence profile as input features for predicting protein interface residues. The comparison results can be found in Table 2. It can be seen that our proposed model can obtain best performance among this three computational models, especially for sensitivity measure. Furthermore, the MCC of 0.22 shows the importance of our selected amino acid properties in prediction of protein interface residues.



	ANDN920101	ARGP820101	BEGF750101	BHAR880101	BUNA790103	BURA740102	CHAM830102	CHOP780204	CHOP780215	FAUJ880104	GEOR030101	JOND920102	KHAG800101	PALJ810107	RACS820114	WERD780103	YUTK870102
ANDN920101	1																
ARGP820101	0.012	1															
BEGF750101	-0.315	0.219	1														
BHAR880101	-0.290	-0.482	-0.484	1													
BUNA790103	-0.072	-0.364	-0.066	0.131	1												
BURA740102	-0.361	0.205	-0.104	-0.091	0.100	1											
CHAM830102	-0.155	0.055	-0.243	0.096	0.269	0.238	1										
CHOP780204	0.171	-0.064	-0.118	0.284	-0.316	-0.401	-0.407	1									
CHOP780215	0.209	0.039	-0.427	0.123	-0.204	-0.025	0.149	-0.054	1								
FAUJ880104	0.275	0.267	0.215	-0.133	-0.264	-0.211	-0.195	0.016	0.052	1							
GEOR030101	-0.178	-0.559	0.169	0.250	0.269	-0.044	0.205	-0.074	-0.591	-0.167	1						
JOND920102	-0.078	0.013	0.012	0.029	-0.073	-0.193	0.180	-0.204	0.091	0.154	-0.076	1					
KHAG800101	0.113	0.049	0.255	-0.569	-0.143	0.085	-0.083	0.016	-0.047	0.219	0.017	-0.133	1				
PALJ810107	-0.089	0.176	-0.618	0.237	-0.422	0.221	-0.005	0.251	0.204	-0.356	-0.379	0.109	-0.291	1			
RACS820114	0.117	0.039	-0.035	-0.124	0.203	-0.120	-0.032	-0.092	-0.065	-0.408	-0.048	-0.131	-0.092	-0.012	1		
WERD780103	-0.211	0.227	0.249	-0.418	0.087	0.268	0.051	-0.224	-0.266	-0.343	0.046	-0.494	0.126	-0.029	0.119	1	
YUTK870102	0.017	0.173	0.034	0.107	-0.234	-0.109	-0.310	0.219	-0.504	0.208	0.019	0.274	-0.259	0.249	0.154	-0.150	1

Fig. 1. The pairwise correlation values among the 17 selected amino acid properties

Table 2. Comparisons of prediction performance among three models

Model[a]	Sen	Spec	Prec	Acc	F1	MCC
BP NN	44.9%	62.9%	49.2%	0.61	0.47	0.12
RBF_EM	42.5%	74.8%	57.5%	0.60	0.49	0.18
SVM_AA	57.9%	65.0%	46.9%	0.625	0.52	0.22

[a] BP NN denotes backpropagation neural network; RBF_EM is RBF neural network optimized by EM algorithm; SVM_AA is present model in this work.

4 Conclusions

This work proposed a promising method which can infer protein interface residues from protein surface in hetero-complexes. The results achieved here demonstrated the effectiveness of our proposed method by comparison of two computational models which address the same job. Pearson correlation approach has been employed to the selection of amino acid properties whose original number is 544 and current number is 17, which enhance the model generalization capability and in the same time deduce the computational complexity obviously. Furthermore, the MCC value we achieved in this work shows that the selected amino acid properties we used actually differentiate the interface and non-interface residues.

Acknowledgement. This work was supported by the National Science Foundation of China (Nos. 60803107).

References

1. Alberts, B.D., Lewis, J., Raff, M., Roberts, K., Watson, J.D.: Molecular Biology of the Cell, 2nd edn. Garland, New York (1989)
2. Fariselli, P., Pazos, F., Valencia, A., Casadio, R.: Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur. J. Biochem. 269(5), 1356–1361 (2006)
3. Ofran, Y., Rost, B.: Predicted protein-protein interaction sites from local sequence information. Febs Letters 544(1-3), 236–239 (2003)
4. Wang, B., Chen, P., Huang, D.S., Li, J.J., Lok, T.M., Lyu, M.R.: Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett. 580(2), 380–384 (2006)
5. Wang, B., Chen, P., Wang, P., Zhao, G., Zhang, X.: Radial basis function neural network ensemble for predicting protein-protein interaction sites in heterocomplexes. Protein Pept. Lett. 17(9), 1111–1116 (2010)
6. Wang, B., Wong, H.S., Chen, P., Wang, H.Q., Huang, D.S.: Predicting Protein-Protein Interaction Sites Using Radial Basis Function Neural Networks. In: International Joint Conference on Neural Networks, pp. 2325–2330 (2010)
7. Yan, C., Dobbs, D., Honavar, V.: A two-stage classifier for identification of protein-protein interface residues. Bioinformatics 20(suppl. 1), 371–378 (2004)
8. Porollo, A., Meller, J.: Prediction-based fingerprints of protein-protein interactions. Proteins 66(3), 630–645 (2007)
9. Yan, C., Wu, F., Jernigan, R.L., Dobbs, D., Honavar, V.: Characterization of protein-protein interfaces. Protein J. 27(1), 59–70 (2008)
10. Neuvirth, H., Raz, R., Schreiber, G.: ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J. Mol. Biol. 338(1), 181–199 (2004)