Protein Interaction Prediction for Mouse PDZ Domains Using Dipeptide Composition Features

Songyot Nakariyakul*†, Zhi-Ping Liu* and Luonan Chen*‡
*Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for Pre-Diabetes,
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China
†Department of Electrical and Computer Engineering, Thammasat University,
Khlong Luang, Pathumthani 12120, Thailand
‡Collaborative Research Center for Innovative Mathematical Modelling,
Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan
Email: nsongyot@engr.tu.ac.th; zpliu@sibs.ac.cn; lnchen@sibs.ac.cn

***Abstract*—The PDZ domain is one of the largest families of protein domains that are involved in targeting and routing specific proteins in signaling pathways. PDZ domains mediate protein-protein interactions by binding the C-terminal peptides of their target proteins. Using the dipeptide feature encoding, we develop a PDZ domain interaction predictor using a support vector machine that achieves a high accuracy rate of 82.49%. Since most of the dipeptide compositions are redundant and irrelevant, we propose a new hybrid feature selection technique to select only a subset of these compositions that are useful for interaction prediction. Our experimental results show that only approximately 25% of dipeptide features are needed and that our method increases the accuracy by 3%. The selected dipeptide features are analyzed and shown to have important roles on specificity pattern of PDZ domains.**

***Index Terms*—Dipeptide compositions; feature selection; PDZ domain; protein interaction.**

I. Introduction

The PDZ (PSD-95/Discs-large/ZO-1) domain family is an important signaling protein that is involved in the development of multi-cellular organisms [1]. Many PDZ domains are key components in maintaining cell polarity, facilitating intercellular signaling system, and regulating synaptic development [2], [3]. They are composed of approximately 80 to 90 amino acid residues folded into six β strands (β1-β6) and two α helices (α1, α2). Prior studies showed that PDZ domains selectively bound C-terminal peptide sequences from voltage gated potassium channels and N-methyl-D-aspartate receptors [4], [5], specifically on residues up to -8 position of the peptide ligand (last residue numbered zero) [6]. Furthermore, many PDZ domains display promiscuity and bind to more than one ligand. However, experimental methods to determine the interaction specificity of the PDZ domains are time-consuming and expensive. Thus, a computational method that can provide accurate prediction is highly demanded.

Several computational methods have been proposed to predict interaction specificity of PDZ domains. Chen et al. [7] proposed an extension of a position-specific scoring matrix that predicted interactions between the 82 mouse PDZ domains and 93 peptides based on their primary sequences. They reported an area under the receiver operating characteristic (AUC) value of 0.87. Eo et al. [8] used amino acid contact matrices and physicochemical distance matrix to encode the protein complex into a feature vector. A support vector machine (SVM) classifier was employed to identify G protein-coupled receptors-binding PDZ domain proteins. Recently, Kalyoncu et al. [9] used trigram amino acid frequencies for feature encoding and a random forest classifier to build a model to predict the binding interactions of PDZ domains and peptide sequences. Resampling was used to address the problem of the imbalanced data set. They obtained an accuracy of 79.8% on the validation set of 27 binding and 62 non-binding interactions.

In this work, we propose to use dipeptide compositions as feature encoding to predict PDZ domain-peptide interactions and employing an SVM classifier to build our predictor. Dipeptide compositions have been shown to give useful information in prior protein-related work [10]–[12]. We compare our method with other feature encoding techniques based on primary sequences. Our experimental results demonstrate that our predictor can obtain a high prediction performance (accuracy of 82.49% and AUC of 0.8920). To further improve the prediction results, we develop a new hybrid feature selection algorithm named the mRMR_BIRS algorithm that is a combination of the minimal-redundancy-maximal-relevance (mRMR) algorithm [13] and the best incremental ranked subset (BIRS) algorithm [14]. We find that approximately 25% of dipeptide features are needed for interaction prediction and that our proposed method increases the accuracy by 3%. Analysis of selected dipeptide compositions is also given.

II. Materials and Methods

A. Data Set

We used the PDZ interaction data set provided in [9]. The data set was originally retrieved from the study of Stiffer et al. [15], which contained interaction data of 85 mouse PDZ domains and 181 mouse peptides. There are a total of 731 binding and 1361 non-binding interactions available for testing. The data set is imbalanced due to its nature. These interaction data were confirmed by fluorescence polarization experiments. The last 10 residues (up to -9 position) of each

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

peptide ligand are considered in our computational method due to the specificity of the PDZ domains. For more information about the data set, see [9].

B. Feature Encoding

The 400 dipeptide compositions of each protein sequence are computed using the following expression (1)

$$\text{Comp}_{\text{dipeptide}}(i,j) = \frac{n_{ij}}{L-1}, 1 \leq i,j \leq 20 \qquad (1)$$

where i, j stand for the distribution of amino acid i followed by amino acid j, n_{ij} is the number of residues of amino acid i followed by amino acid j, and L is the total number of residues in the protein sequence. For each binding/non-binding interaction, a PDZ domain and a peptide ligand are encoded into two vectors, each with 400 dipeptide compositions. We then concatenate two vectors into an 800-feature vector to represent each interaction.

C. Feature Selection

Feature selection refers to search algorithms that select a subset of features from an initial set of n features, where a criterion function J is used to evaluate the quality of each candidate subset. It is mainly used for identifying important features and improving classification results. Depending on the criterion function J used, feature selection methods can be categorized as filter or wrapper. Filter methods rely on the intrinsic properties of the data such as distance, dependency, and consistency and select subsets without any knowledge of the learning algorithm. Wrapper methods use the performance of a predetermined learning algorithm as the criterion function to select a subset. The wrapper method generally achieves better performance than the filter method, but it is also more computationally expensive. Since there are a total of 800 dipeptide features (400 for the PDZ domains and another 400 for the peptide ligands) for our work, we are interested in a wrapper method that is highly effective and computationally efficient.

In this work, a modified version of the best incremental ranked subset (BIRS) algorithm for feature selection is proposed. We first discuss the original BIRS algorithm [14] and then present its modification. The BIRS algorithm is a wrapper method that contains two phases; in the first phase, all of the n features in the set are ranked according to some evaluation measure. In the second phase, the search proceeds from the best to the worst ranked feature, and a feature is selected if adding it to the currently selected feature subset improves the accuracy significantly. That is, the algorithm starts by selecting the best ranked feature from the list. It then considers adding the second best ranked feature to the best one if and only if the resultant subset increases the accuracy rate significantly. If the accuracy obtained by adding the second best ranked feature to the set is not significantly better, the feature is discarded, and the third best ranked feature is considered next, and so on. A Student's paired two-tailed t-test is conducted to determine the statistical significance degree of difference between the accuracies of each subset using a fivefold cross-validation. The algorithm terminates when it reaches the worst ranked feature. Thus, BIRS runs in linear time and selects only relevant and irredundant features.

In the original BIRS work [14], the authors ordered the features according to their individual accuracy rates (the performance of a pre-defined classifier built with a single feature). Since ranking of all features in the first stage plays an important role on the performance of the algorithm, we thus propose using the minimal-redundancy-maximal-relevance (mRMR) algorithm [13] to order the feature set. The mRMR algorithm is a well-known filter search technique that selects feature subsets based mutual information. It is fast and shown to perform well in many applications. We thus expect it to give a better list of ranked features than that ranked by individual accuracy rates as done in prior work [14]. Moreover, to measure the significance of adding a feature to the current subset, we compute the difference between the AUC values of each subset rather than the accuracy rates, since our data set is imbalanced. We name this modified algorithm the mRMR_BIRS algorithm.

D. Support Vector Machines and Performance Evaluation

We choose the SVM classifier with a radial basis function to perform the classification, since it provides high classification results and is fairly resistant to feature selection. The software LIBSVM [16] version 3.0 is employed in this work. The regularization parameter C and kernel parameter γ in the SVM are selected by using a grid search approach. The SVM classifier is trained by using a fivefold cross-validation to maximize an area under the receiver operating characteristic (AUC), since our data set is imbalanced. The receiver operating characteristic (ROC) is a plot of the true positive rate (TPR) versus false positive rate (FPR). We also provide the accuracy (ACC) rate to measure the performance of our method. TPR, FPR, and ACC are defined as follows.

$$\text{TPR} = \frac{\text{TP}}{\text{TP + FN}} \qquad (2)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP + TN}} \qquad (3)$$

$$\text{ACC} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}} \qquad (4)$$

where true positive (TP) is the number of binding interactions correctly classified. True negative (TN) is the number of non-binding interactions correctly classified. False positive (FP) is the number non-binding interactions misclassified as binding interactions. False negative (FN) is the number of binding interactions misclassified as non-binding interactions.

III. Results and Discussion

A. Feature Encoding Comparisons

We now compare our dipeptide composition model with other feature encoding proposed for predicting protein-protein interactions in the literature [11], [17], [18]. Amino acid composition (AAC) has been used in many prior protein-related work [11]. It is defined as the frequencies of 20 amino acids in a protein sequence. In the triad frequency model

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

[17], the 20 amino acids are grouped into 7 different classes according to their dipoles and volumes of the side chains. These classes contain [AGV], [ILFP], [YMTS], [HNQW], [RK], [DE], and [C] amino acids, respectively. Frequencies of three consecutive classes in each protein sequence are then computed and used as features. Pseudo amino acid composition (PseAAC) [18] incorporates both sequence-order information and protein properties to represent each protein sequence. The first 20 features of the PseAAC contain the AAC information, and the additional λ features represent the sequence-order information calculated by the hydrophobicity value, hydrophilicity value, and side-chain mass. We choose $\lambda = 5$, since this gives the highest AUC value.

Fig. 1 shows the ROC curves of the four feature encoding models. As seen, the AAC model gives the worst performance among the four models in many regions. This is expected, since it does not utilize the sequence-order information. The PseAAC model is performing slightly better than the triad model, but a clear difficulty with the PseAAC model lies in interpreting and understanding the model. The proposed dipeptide model is shown to outperform the other models in most regions. Table I summarizes the results of the four feature encoding models using a fivefold cross-validation test. Although the FPR of the PseAAC model (12.28%) is slightly lower than that of the dipeptide model (12.87%), its TPR (69.45%) is the worst one among the four algorithms. As seen in Table I, the ACC rate (82.49%) and AUC (0.8920) of the dipeptide model are highest. However, the TPR (73.84%) obtained using the dipeptide model is somewhat low due to the imbalance of the data set. We expect to use feature selection to combat this problem and improve the prediction results.

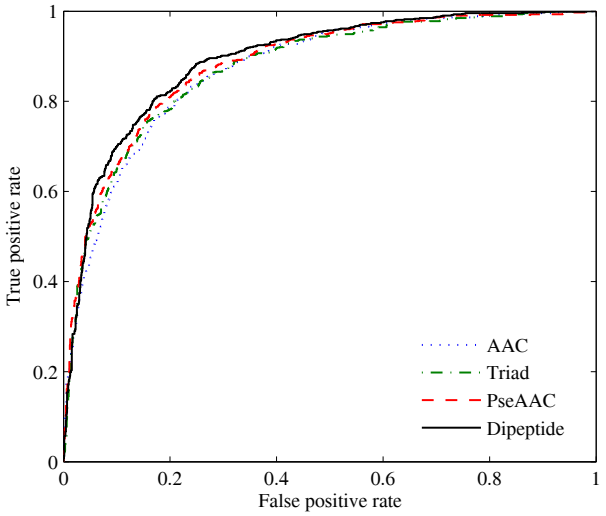

Fig. 1. ROC curves of four different feature encoding models.

B. Feature Selection Results

Since the lengths of PDZ domains and peptides are short, many encoded dipeptide features contains zeros. These features may not contribute to prediction results and can be deemed irrelevant. We propose the mRMR_BIRS algorithm to select only relevant and irredundant dipeptide features for interaction prediction. To determine the statistical significance degree of difference between the AUC values of each subset in the mRMR_BIRS algorithm, the confidence level is chosen to be $p < 0.5$ due to the small sample size. We thus expect many features to be selected by our algorithm. For comparison, we also apply the original BIRS feature selection algorithm to select dipeptide features.

TABLE I
FIVEFOLD CROSS-VALIDATION PREDICTION RESULTS FOR INTERACTION PREDICTION OF PDZ DOMAINS USING DIFFERENT FEATURE ENCODING MODELS.

Model	No. of features	TPR (%)	FPR (%)	ACC (%)	AUC
AAC	40	71.64	15.15	80.24	0.8726
Triad	686	71.78	13.90	81.10	0.8748
PseAAC	50	69.45	**12.28**	81.34	0.8837
Dipeptide	800	**73.84**	12.87	**82.49**	**0.8920**

Table II shows the number of selected dipeptide features and the prediction performances of the original BIRS algorithm and our mRMR_BIRS algorithm. As seen, the BIRS algorithm selects only 54 dipeptide features and yields poorer results than using all 800 dipeptide features (see Table I). The reason is that the BIRS algorithm does not employ a good feature ranking. Our proposed mRMR_BIRS algorithm, on the other hand, uses the mRMR method to provide an initial ranking, which is shown to be very effective. The prediction results of the mRMR_BIRS algorithm are much better than those of the BIRS algorithm and those using all 800 dipeptides; the ACC rate increases from 82.49% to 85.17%, and the AUC value increases from 0.8920 to 0.9110 using our feature selection method. The mRMR_BIRS algorithm selects 215 dipeptide features (approximately 25% of the original 800 features), which shows that many dipeptide features are redundant and irrelevant. Out of 215 features selected by mRMR_BIRS, 102 features are selected from PDZ domains and the other 113 features are chosen from peptides. In terms of computational complexities, the mRMR_BIRS algorithm takes only 10 minutes to perform the search, while the BIRS algorithm needs more than 20 minutes. Thus, our mRMR_BIRS algorithm is faster and more effective.

TABLE II
FIVEFOLD CROSS-VALIDATION PREDICTION RESULTS FOR INTERACTION PREDICTION OF PDZ DOMAINS USING DIPEPTIDE FEATURES AFTER FEATURE SELECTION.

Model	No. of features	TPR (%)	FPR (%)	ACC (%)	AUC
BIRS	54	61.10	11.40	79.00	0.8288
mRMR_BIRS	215	**76.85**	**10.37**	**85.17**	**0.9110**

We now analyze some important dipeptide features selected by our feature selection algorithm. For example, the most important (best ranked) dipeptide selected by our algorithm is 'Glu-Thr' of the peptide ligand, which is supported by prior finding [6] that many PDZ domains such as those of the Discs

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

Large Protein bind to the C-terminal motifs of the peptide ligand with the sequence of Glu-(Ser/Thr)-Xxx-(Val/Ile)-COOH, where Xxx represents any amino acid. Many 'Ser-Xxx' motifs such as 'Ser-Leu', 'Ser-Asn', 'Ser-Gln', and 'Ser-Ser' of the peptides are also selected by our mRMR_BIRS algorithm. 'Ile-Arg' of the PDZ domain chosen by our method is also found to play an important role in forming hydrophobic contact of the first PDZ domain of NHREF1 with amino acid Leu of the peptide ligand [19]. Furthermore, Bezprozvanny and Maximov [20] reported that eight PDZ domains in hINADL-5 bound to neurexin Ia, whose terminus is 'Glu-Tyr-Tyr-Val'. This is also in agreement with our prediction model that selects 'Tyr-Val' of the peptide as an important dipeptide motif. Thus, our selected dipeptides can be used as a guide for future study of prediction interaction specificity of PDZ domains.

IV. Conclusions

This study of PDZ domain-peptide interactions has two aims. First, we compared the prediction performance of the dipeptide composition model to those of three other feature encoding models based on primary sequences. We found that the SVM-based predictor based on the dipeptide model successfully achieved the fivefold cross-validation accuracy of 82.49%, which is slightly higher than those obtained using the other models. Second, we proposed a new mRMR_BIRS feature selection algorithm to further improve the prediction results and to identify important dipeptide motifs. The proposed method was shown to outperform the original BIRS algorithm and increased the prediction accuracy from 82.49% to 85.17%. Many important motifs of PDZ domains and peptides were identified by our method. These encouraging results could be used to facilitate future study on PDZ domain interactions. As a future topic, we will further consider to employ the network-based technique to improve the accuracy of the prediction [21], [22].

Acknowledgments

The authors would like to thank S. Kalyoncu, O. Keskin, and A. Gursoy for their helpful suggestions and for providing the data set used in this work. S. Nakariyakul was supported by the CAS Fellowship for Young International Scientist with Grant No. 2010Y1SB10 and NSFC with Grant No. 31050110435. This work was also supported by the Chief Scientist Program of SIBS of CAS with Grant No. 2009CSP002 (L. Chen) and by the Knowledge Innovation Program of SIBS of CAS with Grant No. 2011KIP203 (Z.P. Liu), and supported by NSFC under Grants No. 61072149 and No. 91029301 (L. Chen and Z.P. Liu), the Knowledge Innovation Program of CAS with Grant No. KSCX2-EW-R-01 (L. Chen and Z.P. Liu), and supported by the Key Project of Shanghai Education Committee (B.10-0412-08-001), Japan (JSPS) FIRST Program (L. Chen) and Shanghai NSF under Grant No. 11ZR1443100 (Z.P. Liu).

References

[1] C.P. Ponting, "Evidence for PDZ domains in bacteria, yeast and plants," Protein Sci 6, pp. 464-468, 1997.

[2] T. Pawson, and P. Nash, "Assembly of cell regulatory systems through protein interaction domains," Science 300, pp. 445-452, 2003.

[3] K.K. Dev, "Making protein interactions druggable: Targeting PDZ domains," Nat Rev Drug Discov 3(12), pp. 1047-1056, 2004.

[4] H. Kornau, R.L. Schenke, M. Kennedy, and P. Seeburg, "Domain interaction between NMDA receptor subunits and the postsynaptic density protein PSD-95," Science 269, pp. 1737-1740, 1995.

[5] J. Schultz, U. Hoffmuller, G. Krause, J. Ashurst, M.J. Macias, P. Schmieder, J. Schneider-Mergener, and H. Oschkinat, "Specific interactions between the syntrophin PDZ domain and voltage-gated sodium channels," Nat Struct Biol. 5, pp. 19-24, 1998.

[6] Z. Songyang, A.S. Fanning, C. Fu, J. Xu, S.M. Marfatia, A.H. Chishi, A. Crompton, A.C. Chan, J.M. Anderson, L.C. Cantley, "Recognition of unique carboxyl-terminal motifs by distinct PDZ domains," Science 275, pp. 73-77, 1997.

[7] J.R. Chen, B.H. Chang, J.E. Allen, M.A. Stiffler, and G. MacBeath, "Predicting PDZ domain-peptide interactions from primary sequences," Nat Biotechnol 26(9), pp.1041-1045, 2008.

[8] H.S. Eo, S. Kim, H. Koo, and W. Kim, "A machine learning based method for the prediction of G protein-coupled receptor-binding PDZ domain proteins," Mol Cells 27, pp. 629-634, 2009.

[9] S. Kalyoncu, O. Keskin, and A. Gursoy, "Interaction prediction and classification of PDZ domains," BMC Bioinformatics 11, 357, 2010.

[10] Q.B. Gao, Z.C. Jin, X.F. Ye, C. Wu, J. Lu, and J. He, "Improving the classification of nuclear receptors with feature selection," Protein Pept Lett 16(7), pp. 823-829, 2009.

[11] M. Rashid, S. Ramasamy, and G.P. Raghava, "A simple approach for predicting protein-protein interactions," Curr Protein Pept Sci 11(7), pp. 589-600, 2010.

[12] S. Nakariyakul, Z.-P. Liu, and L. Chen, "Detecting thermophilic proteins through selecting amino acid and dipeptide composition features," Amino Acids, doi:10.1007/s00726-011-0923-1, 2011.

[13] H.C. Peng, F.H. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans Pattern Anal Mach Intel 27, pp. 1226-1238, 2005.

[14] R. Ruiz, J.C. Riquelme, and J.S. Aguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," Pattern Recog 39, pp. 2383-2392, 2006.

[15] M.A. Stiffler, J.R. Chen, V.P. Grantcharova, Y. Lei, D. Fuchs, J.E. Allen, L.A. Zaslavskaia, and G. MacBeath, "PDZ domain binding selectivity is optimized across the mouse proteome," Science 317, pp. 364-369, 2007.

[16] C.-C. Chang, and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM Trans Intell Syst Technol 2, pp. 27:1-27:27, 2011.

[17] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein-protein interactions based only on sequence information," Proc Natl Acad Sci USA 104(11), pp. 4337-4341, 2007.

[18] L. Liu, Y. Cai, W. Lu, K. Feng, C. Peng, and B. Nu, "Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection," Biochem Biophys Res Commun 380, pp. 318-322, 2009.

[19] S. Karthikeyan, T. Leung, and J.A. Ladias, "Structural basis of the Na+/H+ exchange regulatory factor PDZ1 interaction with the carboxyl-terminal region of the cystic fibrosis transmembrane conductance regulator," J Biol Chem 276(23), pp. 19683-19686, 2001.

[20] I. Bezprozvanny, and A. Maximov, "Classification of PDZ domains," FEBS Lett 509, pp. 457-462, 2001.

[21] L. Chen, R. Wang, and X. Zhang, Biomolecular Network: Methods and Applications in Systems Biology, Wiley, London, 2009.

[22] L. Chen, R. Wang, C. Li, and K. Aihara, Modelling Biomolecular Networks in Cells: Structures and Dynamics, Springer, Berlin, 2010.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE