2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Designing predictors of carbohydrate-binding proteins using informative physicochemical properties

Hui-Lin Huang, Hua-Chin Lee, Yi-Fan Liou, Ming-Che Li and Shinn-Ying Ho[+]

Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan

Abstract. Carbohydrate-binding proteins can interact with sugars but do not modify them, and play a pivotal role in a variety of important biological recognition processes. Most studies predict carbohydrate-binding sites, but not carbohydrate-binding proteins. The existing method of predicting carbohydrate-binding proteins with sequence identity 35% uses an amino acid based encoding method and support vector machine (SVM). The merits of this study are three-fold. First, we establish a large-scale data set of carbohydrate-binding proteins collected from three up-to-date databases, CAZy, CGF and Swiss-Prot. This data set CBPDS consists of 2380 positive and negative proteins with sequence identity 25% by removing sequence redundancy. Secondly, we propose an efficient SVM-based method for predicting carbohydrate-binding proteins using a small set of informative physicochemical properties obtained by using an inheritable bi-objective genetic algorithm. The prediction accuracy of independent test is 79.45% using 17 properties. Third, from the analysis of informative physicochemical properties, some interpretable knowledge of carbohydrate-binding and non-carbohydrate-binding proteins can be further discovered. The set of obtained physicochemical properties can also be used conveniently as the core for designing various predictors of carbohydrate-binding problems.

Keywords: Carbohydrate-binding, physicochemical properties, genetic algorithm, prediction, SVM.

1. Introduction

Carbohydrate-binding proteins can interact with sugars but do not modify them. Carbohydrates play a key role in a variety of important biological recognition processes like infection, immune response, cell differentiation, and neuronal development. All of these biological phenomena may be regulated by the interaction of these carbohydrates with proteins [1]. Carbohydrate-binding proteins are becoming extremely useful in curing various illnesses. Identifying carbohydrate-binding proteins using experimental work is costly and time consuming. Therefore, computational methods of predicting carbohydrate-binding proteins would be useful.

Developing an automated and efficient method for timely identification of novel carbohydrate-binding proteins is very important. Many researches mainly focused on prediction and analysis of carbohydrate-binding sites by using empirical rules [2] or a machine learning method [3]. They are designed for predicting binding sites of proteins that are already known as carbohydrate-binding proteins. In other words, they are not designed to predict non-carbohydrate-binding proteins. In this study, we are interested in protein sequence based classification of carbohydrate-binding and non-carbohydrate-binding proteins.

Someya *et al.* [4] first clarified the definition of carbohydrate-binding proteins and then constructed positive and negative datasets. They developed a prediction system of carbohydrate-binding proteins with sequence identity 35% using an amino acid based encoding method and support vector machine (SVM) [5], where the prediction of carbohydrate-binding proteins can be formulated as a binary classification problem.

Using both informative features and an appropriate classifier is essential to design an effective method for predicting carbohydrate-binding proteins using the primary sequence only. Someya *et al.* [4] trained the

[+] Corresponding author. Tel.: + 886-35712121-56910; fax: +886-35729288.
E-mail address: syho@mail.nctu.edu.tw

SVM with three different encoding methods: a direct encoding method (AA-20), and two grouping methods (Levitt-6 and Someya-7). Their method used the frequencies of triplets of group symbols as features which are based on the polarity and the propensity of the secondary structure. Kumar *et al.* [5] developed SVM modules for distinguishing cancer lectins from non-cancer lectins by investigating features of dipeptide composition, split composition, position specific scoring matrix (PSSM) profiles, and 14 PROSITE domains.

We investigate an optimal design of predictors for carbohydrate-binding proteins from sequences using both informative features and an appropriate classifier. Furthermore, we obtain a set of relevant physicochemical properties which can advance prediction performance. Physicochemical properties extracted from protein sequences were utilized as effective features in recent years. Our previous work Auto-IDPCPs [6] is an SVM-based classifier with automatic feature selection from a large set of physicochemical property features to predict DNA-binding domains/proteins. The POPI method used physicochemical properties as efficient features to predict peptide immunogenicity [7]. The prediction method UbiPred [8] mined informative physicochemical properties from protein sequences to identify promising ubiquitylation sites.

The informative physicochemical properties of amino acids selected in this study were used as features in designing SVM classifiers. An efficient algorithm inheritable bi-objective genetic algorithm (IBCGA) was used to select significant features which could discriminate the two classes of proteins. The feature sets selected by IBCGA were analyzed carefully to reveal the fundamental differences existed between carbohydrate-binding proteins and non-carbohydrate-binding proteins. In conclusion, we proposed a novel prediction method combining the informative physicochemical properties of amino acid with SVM to solve the prediction problem of carbohydrate-binding proteins.

2. Materials

2.1. Datasets

For advancing the study, we establish a large-scale data set of carbohydrate-binding proteins collected from three up-to-date databases, Carbohydrate-Active Enzyme (CAZy), Functional Glycomics (CFG) and Swiss-Prot. Carbohydrate-binding proteins are acquired from Consortium for the CFG and CAZy databases. All the records from those two databases are served as positive datasets which could bind carbohydrate. The gene ontology (GO) annotation terms about carbohydrate binding function are obtained from the GOA database. The GO term, carbohydrate binding function, and its child terms are collected. Finally, the number of GO terms which are defined as carbohydrate binding function is 778.

To obtain the negative datasets which are non-carbohydrate-binding proteins, the Swiss-Prot database, release 2011_06, is also used. The Swiss-Prot database is separated into positive and negative datasets using GO term mentioned above. The polypeptides in Swiss-Prot containing any GO terms defined as carbohydrate binding protein are classed into a positive dataset. The others are classed into a negative dataset.

The positive dataset is composed of 57330 polypeptides while the negative dataset is composed of 405046 polypeptides. USEARCH [9] is used to remove the sequence redundancy of the dataset. The threshold of USEARCH is set to 25%. After processing with USEARCH, the positive dataset contains 2380 polypeptides and the negative dataset contains 49647 polypeptides. The numbers of sequences in various stages are shown in Table 1.

Table 1. The numbers of sequences in various stages

stages	initial	identity threshold 25%	randomly chosen negative sequences	Final dataset CBPDS
positive	57330	2380	2380	1190
negative	405046	49647	2380	1190

2.2. Physicochemical Properties

AAindex is a database developed by Kanehisa et al. which collects numerical indices representing physicochemical and biochemical properties of amino acids [10]. The 544 properties retrieved from

AAIndex 9.0 were reduced to 531 after removing the features containing the value ‘NA’. The 531 properties from AAindex were used as initial features to construct SVM classifier for the discrimination between carbohydrate-binding proteins and non-carbohydrate-binding proteins. The original sequences of the datasets were transformed to the numerical indices according to the corresponding values of amino acids of each feature. The average values of each physicochemical property form a feature vector of the protein sequence. After the transformation of amino acids into numerical indices, these values were normalized to the scale between -1 and 1 for SVM.

3. Methods

3.1. Inheritable bi-objective genetic algorithm (IBCGA)

In order to select a minimal number of informative features while maximizing prediction accuracy, we used a previously developed inheritable bi-objective genetic algorithm to solve this problem [11]. IBCGA is an efficient algorithm consisting of an intelligent genetic algorithm IGA [12] which uses orthogonal array crossover to explore the search space efficiently. Moreover, the inheritable mechanism can efficiently improve the prediction accuracy during the searching process. Both feature selection and parameter tuning of SVM were encoded as binary genes in the chromosome of IBCGA. The gene and chromosome are commonly-used terms of genetic algorithm (GA), named GA-gene and GA-chromosome for discrimination in this paper. The GA-chromosome consists of n = 531 binary GA-genes b_i for selecting informative properties and two 4-bit GA-genes for tuning the parameters C and γ of SVM. If b_i=0, the i^{th} property is excluded from the SVM classifier; otherwise, the i^{th} property is included. This encoding method maps the 16 values of γ and C into $\{2^{-7}, 2^{-6}\ldots, 2^{8}\}$.

The digitized and normalized protein sequences of the training data set were used as the input for SVM. The fitness function of SVM is the overall accuracy of 5-fold cross validation. The feature selection algorithm of IBCGA is described as follows: Step 1) (Initialization) Randomly generate an initial population of individuals and $r=r_{start}$. Step 2) (Evaluation) Evaluate the fitness values of all individuals using fitness function. Step 3) (Selection) Select the winner from two randomly selected individuals to form a mating pool. Step 4) (Crossover) Select parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents. Step 5) (Mutation) Apply the swap mutation operator to the randomly selected individuals in the new population. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual. Step 6) (Termination test) If the stopping condition for obtaining the solution is satisfied, output the best individual. Otherwise, go to Step 2). In this study, the stopping condition is to perform 40 generations. Step 7) (Inheritance) If $r < r_{end}$, randomly change one bit in the binary GA-genes for each individual from 0 to 1; increase the number r by one, and go to Step 2). Otherwise, stop the algorithm.

In this study, the range of the size of candidate feature set selected by IBCGA is from r_{start}=4 and r_{end}= 30. 30 independent runs of IBCGA were performed to obtain a robust set of features that can reveal the differences between the two classes of proteins.

3.2. Prediction Method

The selected *m* physicochemical properties and the associated parameter set of SVM by using IBCGA are used to implement the computational system and analyze the physicochemical properties to further understand the carbohydrate-binding proteins. Since the IBCGA is a non-deterministic method, it should make more effort to identify an efficient and robust feature set of informative physicochemical properties. The procedure is described as the following steps:

Step 1 : Prepare the independent and training data sets for 5-CV.

Step 2 : IBCGA is performed *R* independent runs for each of independent data sets. In this study, *R* = 30. There are total 30 sets of *m* physicochemical properties for each of independent data sets.

Step 3 : Choose the set of selected physicochemical properties with a maximal accuracy.

IBCGA will automatically determine a set of informative physicochemical properties and an SVM model for predicting carbohydrate-binding and non- carbohydrate-binding proteins.

4. Results

4.1. Training results of SVM

The training data sets contain 595 positive and 595 negative samples. We performed 30 independent runs of carbohydrate-binding proteins to select a robust feature set which could improve the performance of SVM classifier on discriminating the two classes of proteins. The highest training accuracy of 30 IBCGA runs is 80.84%and its corresponding test accuracy was 79.45% (Table 2). This highest prediction accuracy for various numbers of selected features is given in Figure 1.

Table 2 Results of the training and independent test.

	Specificity (%)	Sensitivity (%)	MCC	Accuracy (%)
Training	0.77	0.77	0.54	80.84
Test	0.79	0.8	0.58	79.45

MCC: Matthews correlation coefficient

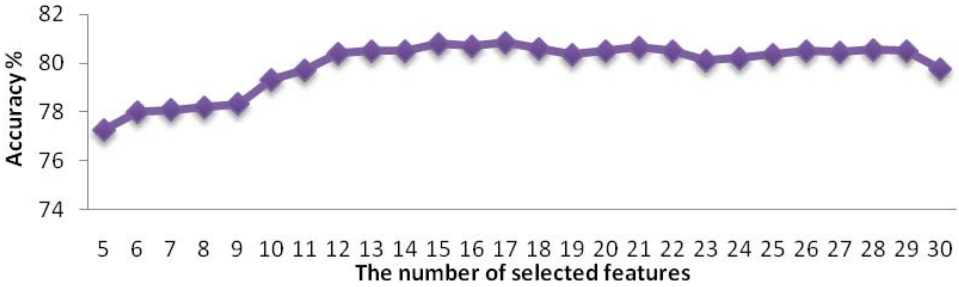

Fig. 1: Prediction accuracies for various numbers of selected properties.

4.2. Selecting a small set of physicochemical properties

The quantified effectiveness of individual physicochemical properties on prediction is useful to characterize the IBCGA mechanism by physicochemical properties. Orthogonal experimental design with factor analysis can be used to estimate the individual effects of physicochemical properties according to the value of main effect difference (MED) [7, 11]. The property with the largest value of MED is the most effective in predicting carbohydrate-binding proteins. According to MED, the 17 informative properties are ranked and their descriptions are shown in Table 3 and Figure 2. The most effective property with MED=33.067 is ARGP820102 denoting “Structural prediction of membrane-bound proteins”.

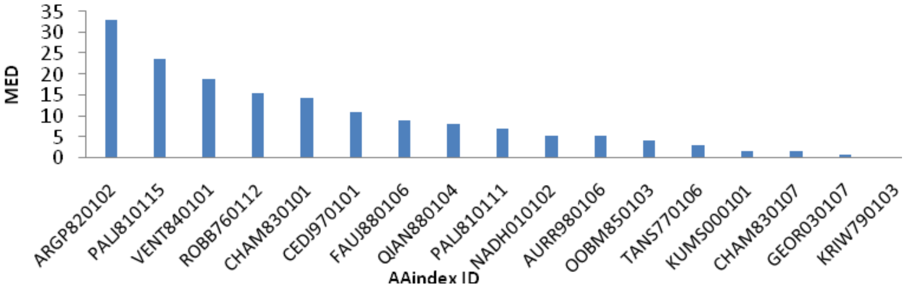

Fig. 2: The rank of the selected feature set with the highest training accuracy is analyzed by MED analysis.

Table 3.The highest accuracy with selected m = 17 feature set.

ID	AAindex ID	Description
3	ARGP820102	Signal sequence helical potential (Argos et al., 1982)
24	CHAM830101	The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983)
30	CHAM830107	A parameter of charge transfer capability (Charton-Charton, 1983)
83	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al., 1988)
150	KRIW790103	Side chain volume (Krigbaum-Komoriya, 1979)
220	OOBM850103	Optimized transfer energy parameter (Oobatake et al., 1985)
233	PALJ810111	Normalized frequency of beta-sheet in alpha+beta class (Palau et al., 1981)
237	PALJ810115	Normalized frequency of turn in alpha+beta class (Palau et al., 1981)
261	QIAN880104	Weights for alpha-helix at the window position of -3 (Qian-Sejnowski, 1988)
350	ROBB760112	Information measure for coil (Robson-Suzuki, 1976)
371	TANS770106	Normalized frequency of chain reversal D (Tanaka-Scheraga, 1977)

380	VENT840101	Bitterness (Venanzi, 1984)
408	AURR980106	Normalized positional residue frequency at helix termini N1 (Aurora-Rose, 1998)
440	KUMS000101	Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al., 2000)
447	NADH010102	Hydropathy scale based on self-information values in the two-state model (9% accessibility) (Naderi-Manesh et al., 2001)
456	CEDJ970101	Composition of amino acids in extracellular proteins (percent) (Cedano et al., 1997)
497	GEOR030107	Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa, 2003)

According to the vectors of amino acids for 531 properties, Hunag *et al.* [6] clustered them into 20 clusters by a fuzzy c-means algorithm based on normalized Euclidean distances. We analyzed the 30 feature sets of independent IBCGA experiments. The 30 sets of *m* properties belonging to the 20 clusters from the results of 30 runs are shown in Figure 3. From the statistic result, the clusters 7, 9, 10, 16, and 18 with very highly selected frequencies are more important for predicting carbohydrate-binding proteins.

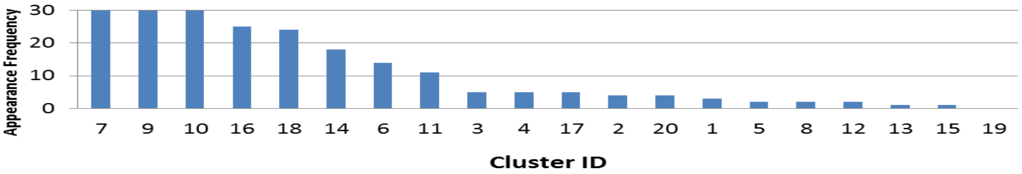

Fig. 3: The appearance frequency of each identified cluster in the 30 runs. The clusters 7, 9, 10, 16, and 18 are more informative.

5. Discussion

The merits of the proposed method are threefold: 1) establish a large-scale data set of carbohydrate-binding proteins CBPDS; 2) a small set of informative physicochemical properties is identified for predicting carbohydrate-binding proteins with promising accuracy, and 3) the small set of informative physicochemical properties can be more easily interpretable. The adopt IBCGA achieves a test accuracy of 79.45% using only 17 physicochemical properties for predicting carbohydrate-binding proteins.

The identified feature sets from 30 independent runs of IBCGA are very robust. The appearance frequency of each identified cluster in the 30 runs is shown in Fig. 3. From the statistic result, the clusters 7, 9, 10, 16, and 18 with very high selection frequencies are more informative for predicting carbohydrate-binding proteins. The selected clusters of the 30 runs are very similar in terms of cluster ID from the 20 clusters. The most effective property ARGP820102 belongs to the 10th cluster with Hydrophobicity propensity. IBCGA is an efficient approach to selecting informative physicochemical properties for designing SVM classifiers. This method can be also applied to other sequence-based prediction problems.

6. References

[1] A. Malik, A. Firoz, V. Jha, and S. Ahmad, “PROCARB: A database of known and modelled carbohydrate-binding protein structures with sequence-based prediction tools,” *Advances in Bioinformatics*, 2010, Article ID 436036, pp. 1-9.

[2] C. Shionyu-Mitsuyama, T. Shirai, H. Ishida, and T. Yamane, “An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins,” *Protein Engineering*, 2003, vol. 16, no. 7, pp. 467–478.

[3] A. Malik and S. Ahmad, “Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network,” *BMC Structural Biology*, 2007, vol. 7, article 1, pp. 1-14.

[4] S. Someya, M. Kakuta, M. Morita, K. Sumikoshi, W. Cao, Z. Ge, O. Hirose, S. Nakamura, T. Terada, and K. Shimizu, “Prediction of carbohydrate-binding proteins from sequences using support vector machines,” *Advances in Bioinformatics*. Vol. 2010, Article ID 289301, pp. 1-9.

[5] R. Kumar, B. Panwar, J. S Chauhan and G. PS Raghava, “Analysis and prediction of cancer lectins using evolutionary and domain information,” *BMC Research Notes* 2011, 4:237.

[6] H.-L. Huang, I.-C. Lin, Y.-F. Liou, C.-T. Tsai, K.-T. Hsu, W.-L. Huang, S.-J. Ho, and S.-Y. Ho, “Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties,” *BMC Bioinformatics* 2011, 12(Suppl 1):S47.

[7] C.-W. Tung and S.-Y. Ho, “POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties,” *Bioinformatics*, 2007, vol. 23, no. 8, pp. 942–949.

[8] C.-W. Tung and S.-Y. Ho, “Computational identification of ubiquitylation sites from protein sequences,” *BMC Bioinformatics*, July 2008, vol. 9:310.

[9] X. Yua, J. Caob, Y. Caic, T. Shia and Y. Lia, “Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines,” *Journal of Theoretical Biology*, 240 (2006), pp.175-184.

[10] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, “AAindex: amino acid index database,” *Nucleic Acids Res* 2008, 36(Database issue):D202-205.

[11] S.-Y. Ho, J.-H. Chen, and M.-H. Huang, “Inheritable genetic algorithm for bi-objective 0/1 combinatorial optimization problems and its applications,” *IEEE Trans. Syst. Man Cybern. Part B-Cybern.*, vol. 34, pp. 609-620, 2004a.

[12] S.-Y. Ho, L.-S. Shu, and J.-H Chen, “Intelligent evolutionary algorithms for large parameter optimization problems,” *IEEE Transactions on Evolutionary Computation*, Vol. 8, 2004, pp. 522–541.