ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 4, No.16 , July 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Classifying Protein-Protein Interaction Type based on Association Pattern with Adjusted Support

Huang-Cheng Kuo and Ming-Yi Tai

Department of Computer Science and Information Engineering
National Chiayi University
Chia-Yi City 600, Taiwan
hckuo@mail.ncyu.edu.tw

Abstract

Proteins carry out their functions by means of interaction. There are two major types of protein-protein interaction (PPI): obligate interaction and transient interaction. In this paper, residues with geographical information on the binding sites are used to discover association patterns for classifying protein interaction type. We use the support of a frequent pattern as its inference power. However, due to the number of transient examples are much less than the number of obligate examples, therefore there needs adjustment on the imbalance. Three methods of applying association pattern to classify PPI type are designed. In the experiment, there are almost same results for three methods. And we reduce effect which is correct rate decreased by data type imbalance.

Keywords: Protein-Protein Interaction, Association Pattern Based Classification, Type Imbalance

1. Introduction

Protein-protein interaction refers to an event generated on the change in physical contact between two or more proteins. Protein-protein interaction occurs when a number of proteins combine into an obligate protein complex or a transient protein complex. An obligate complex will continue to maintain its quaternary structure and its function will continue to take effect. A transient complex will not maintain its structure. It will separate at the end of its function. Protein-protein interaction occurs mainly on the binding surface of the proteins. The residues on the binding surface play an important role for deciding the type of protein-protein interaction. The residue distribution affects the contacting orientation and thus determines the binding energy which is important in interaction type.

In this paper, an association pattern method is proposed for classifying protein-protein interaction type. A transaction is instead of considering all the residues on the binding surface of a protein complex. We generate some small transactions from a protein complex, and each small transaction contains the residues which are geographically close to each other. The binding surface of a protein complex is usually curve. So, there have some residues of a protein with concave shape binding site and others are on the other protein with convex shape binding site. A transaction is a tuple <R, L>, where R is a set of residues of a protein, L is a set of residues of the other protein. Residues of a transaction are geographical close to each other. Patterns from obligate protein complexes and from transient protein complexes are mined separately [1].

In this paper, we assume proteins are in complex form. However, with the association patterns, proteins can be indexed under the patterns. So that biologists can quickly screen proteins that interact with the certain type of interaction.

2. Related Works

The ultimate goal of this paper is user input transient protein binding proteins, and then quickly screened out an experimental biological experimenter direction from the data library. As for how to predict protein interaction type, researchers have proposed method using machine learning classification methods to design the system module.

Mintseris *et al* for protein complexes can identify whether their prediction classification of information depends only limited participation in Pair of two proteins interact in order for the quantity of various atoms, called Atomic Contact Vector. There is a good accuracy, but there are two drawbacks. (1) Feature vector (171 dimensions), there will curse of dimensionality problem. (2) Focus only on contact with the atom, it did not consider the shape of the contact surface. The shape of the contact surface of the protein affects the contacting area and the types of atom contact

[2,3].
It is a popular field of study on protein complexes in addition to the recognition as a research outside the pharmacy. In pharmaceutical drug design research, the main goal is analyzing protein-protein interaction[4,5]. Pharmaceutical drug design research intends to find the protein that is located in a position gap, and the gap is mainly protein or protein-bound docking. In the protein binding site, there is information, such as shapes, notch depth and electrical distribution. The protein binding site is the main location of the occurrence of disease organisms, and the location is a place where compound produced protein chemistry and mutual bonding place. Therefore, when researchers design a drug, they look for existing molecules or the synthesis of new compounds. When a compound is placed in the protein gap, we must try to put a variety of different angles to constantly rotate. Looking for as much as possible fill the gap and produce good binding force between molecules. So, we can find a compound capable of proteins that have the highest degree of matching notches. This is called docking.
Protein-protein interaction network can help us to understand the function of the protein[6,7]. You can understand the basic experiment to determine the role of the presence of the protein, but the protein due to the huge amount of data.
We are not likely to do it one by one experiment. So predicting interactions between proteins has become a very important issue. In order to predict whether there has interaction between the protein situations more accurately [8]. So Biologists proposed using protein combination to increase the accuracy. The joint surface between the protein and the protein interacting surface is called protein domain.
Domain Protein binding protein is the role of surface functional units. Usually a combination of surface has more than one domain presence, and combined with the presence of surface property which is divided into the following categories: hydrophobicity, electrical resistance, residual Kitt, shape, curvature, retained residues[9, 10, 11]. We use information of residues.
Park et al also use classification association rules prediction interaction. The interaction into Enzyme-inhibitors, Non Enzyme-inhibitors, Hetero-obligomers, Homo-obligomers and other four categories[12], the total 147. Association rule is used in conjunction face value of 14 features, as well as domain, numerical characteristics such as average hydrophobicity, residue propensity, number of amino acids, and number of atoms.
There is about 90 percent correction rate, but the information has the same way. Other non-association rules, such as SVM, has 99% correct rate, and k-nearest-neighbor method also has about 93% accuracy[13]. Lukman *et al* divided interaction into three categories: crystal packing, transient and obligate. 4,661 transient and 7,985 obligate protein complexes are in order bipartite graph pattern mining complexes of each category, and find style single binding surface dipeptide. Find out which style or Patches, locate the joint surface, and can bring good accuracy whether protein interactions. Some people use the opposite operation, a collection of known protein acting plane. We want to find those proteins which have similar effects because the first know what is the role of surface interacted with each other exist. But it is possible to be able to know relationship exists for protein to protein.

3. Data Preparation

3D coordinate position of the plan of proteins derived from the RCSB Protein Data Bank. Identification of protein complexes, there are several sources:
1. 209 protein complexes collected by Mintseris and Weng [3].
2. Protein complexes obtained from the PDB web site[14]. Then type of the complexes is determined by using NOXclass website. NOXclass uses SVM to classify the protein-protein interaction of three types into biological obligate, biological transient and crystal packing. We keep only the protein complexes which as classified as biological obligate and biological transient. The accuracy rate is claimed to be about 92%. So, we use the data classified by NOXclass as genuine data for experiment. We collected total 243 protein complexes [15] by this way.

3.1 The Binding Surface Residues

A protein complex is composed of two or more proteins, where in PDB[14] each protein is called a chain. The all the chains in a protein complex share a coordinate system, and the coordinates of each chain of each residue sequentially label in a number of the more important atoms. Because it is a complex of a common coordinate system, so that the relative position is found between the chains. If there are two chains bind together by the chain of residue position determination, but no residues are indicated at the bonding surface on[16]. It is therefore necessary to further inquiries by other repositories or algorithms judgment.
There have been many studies on protein sequence data to predict which residues are on the binding surface[17]. In our research, the atom-to-atom distance between two

residues is used to decide whether the two residues which are on the binding surfaces. They are binding surface residues if there exists an atom-to-atom distance which is less than a threshold. The distance threshold is 5Å in this paper[18,19,20] .

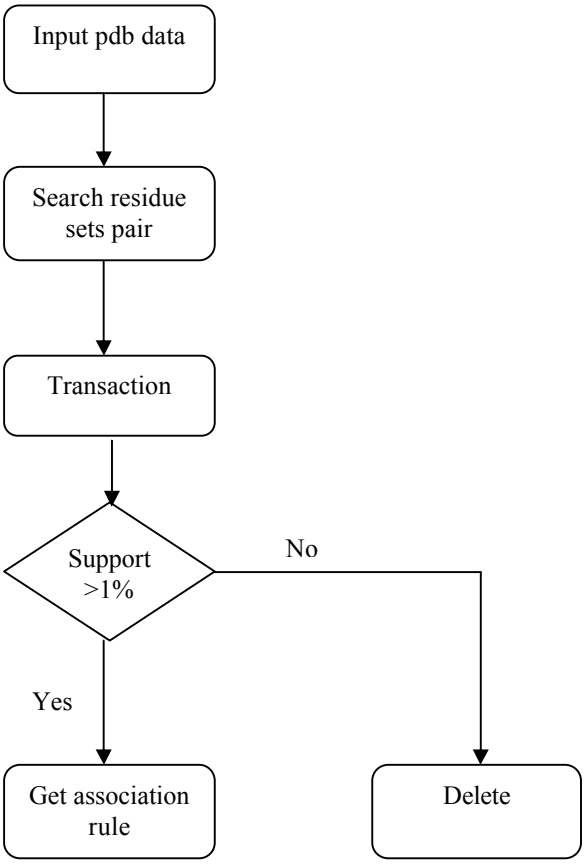

Fig 1. Associative Classification Mining

We input a dataset of PDB files[14], then find the residues on the binding site for each complex and get a pair of residue sets, one set on the convex side, the other one on the other side. And partition each of the pairs of residue sets into transactions for association rule mining in next box. Finally if the transaction's value of supper is less than 1%, then delete this transaction.

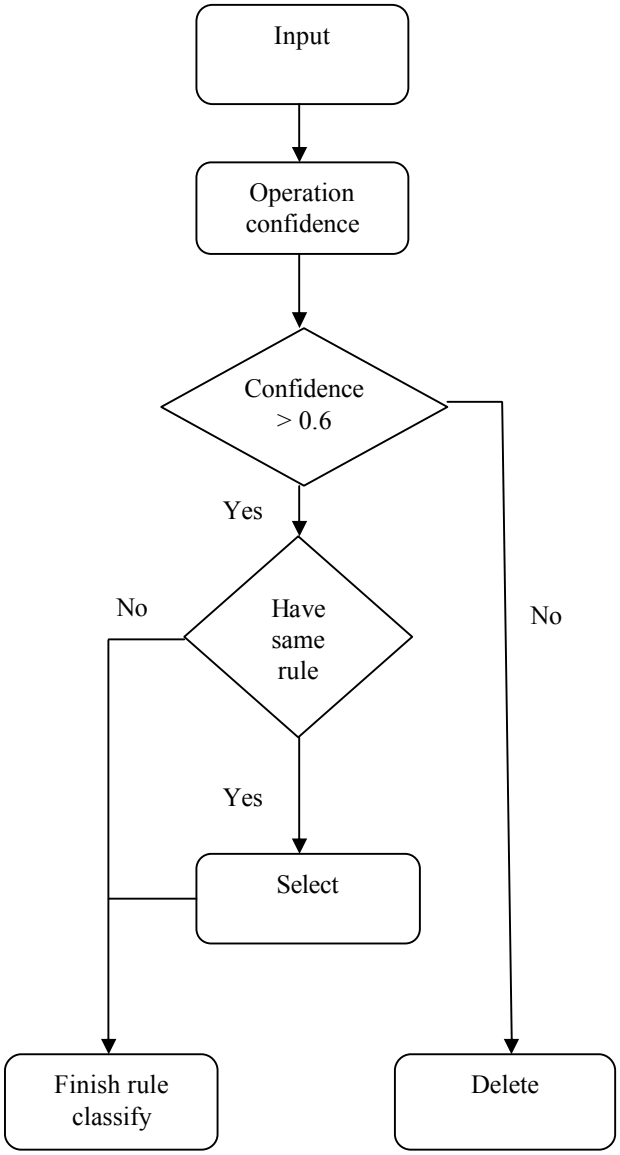

Fig 2. Applying the rules to classify

Taking association rule operation value for PDB file[14] called confidence. If the value of confidence is less than 0.6, delete this rule. Next check have same rule with different type, if have same rule, delete lower confidence rule.

3.2 Obtaining Data

Frequent pattern mining results can be obtained such as: identification of protein complexes in the same body, which is electrically common to the residues in the concave joint surface, and hydrophilic (or hydrophobic) residue at the convex joint surface. Association rule mining results can be obtained, such as: polar residues

on the concave joint surface and hydrophilic (or hydrophobic) residue at the convex joint surface. The identification of protein complexes mostly is interaction. Combination of surface materials can have physical and chemical properties of amino acids. As well as numerical features such as accessible surface area (ASA). The appropriate numeric features discrete (discretize) for the interval, as a project (item) mining of association rules. At this stage, we just take the residual basic body styles and frequently used as input data mining of association rules.

We combine two protein complexes uneven surface residues projected on the surface of the joint surface of the cross to a radius of 10 Å circular motion in the transverse plane. The radius of 10 Å successive increases in a circular motion to form concentric rings[21]. Each ring will cut the number of between zones of equal area (sector), and different ring every area roughly equal to the formation of residues in each district a deal. This method will be divided residues from the past in the same transaction. But the disadvantage is that the district boundary residues are rigidly assigned to a transaction that is on the boundary of two similar residues are divided into different deal.

Another way for the direct use of the tertiary structure coordinates to each binding surface residues as a benchmark. In concave surface of combination site. For example, take one of the residues r, the same as in the concave and r similar residues put together, then convex. The similar residues with r is added and assigned to the same transaction. The advantage is that residues close to each other are put into a transaction. But the drawback is that a residue may repeatedly appears in some transactions.

Then we find on the amount of data that is significantly the number of type biological transient less than obligate, which causes production rules and calculations during supper. The value of biological transient is underrated, so we have to do to adjust the value of biological transient. So biological obligate and biological transient are at the fair situation.

3.3 Associative Classification Rule

Various protein complexes bind frequency distribution of different surface amino acids, which the reason we believe that the association rules can be used as the basis to predict classification. In addition, the combination of surface irregularities are made of a complex combination of the two surfaces.

Using arg amino acids for example, in the identification of the complex, if there is a combination of concave and phe ser, then there will be 90% across arg, there is the association rule mining we can get rule: {phe, ser} → {arg} , support for the rule of 1.9%, confidence is 90%. After exploration, it identify with the non-binding protein complex identification of amino acid side connected to the rules. It can assist in predicting the style to get a likelihood classification combined with another side effect of the combination of face recognition occurs. If one party has a combination of surface phe and ser. The other party has arg, and then the binding surface increases the likelihood of identification. But if the two sides should also consider combining amino acid side of the amino acid pattern of non-compliance with the identification of the joint surface. The likelihood of this occurrence to identify the combination of surface and reduced. In addition, detailed rules amino acid position for the application of the rules is likely to affect, such as: if ser phe and the distance to the amino acid level, it is very far away. Even if there is another combination of surface arg. The applicability of this rule may play on a discount of, on the contrary. If it is very close distance, the influence of this rule should be raised. We will consider the overall impact is to identify a set of joint surface when the amino acid pattern recognition and non-recognition of the joint surface binding amino acid pattern in the surface of the judge[22,23]. Obtain association rules method, divided into two steps:

1. Delete unimportant or conflict association rules.
2. Select the association rules, set for unknown objects.

And related forms of association rule is X => C, X is the project set (also called itemset), C is a category. Association rules in the form of <X, Y> => C, X and Y are itemsets, representing convex and convex surface binding residues; C is a correlation between categories[24].

4. Association Rules Deletion

Data for the transaction or relation, in the training data set (hereinafter referred to as DB) data attach each category. The following instructions to the P and N two categories. For example, Arising class association rule (hereinafter referred to as CAR) format X => Y, X is an item set, Y is a type.

In [25] algorithm, which depend on the sort of confidence class association rule. The pattern with the same confidence are sorted according their supports. Then the class association rule is according to the sort order of selection. The selection process is that the data base in line with the current class association rule (called r) case deletion of the conditions of. Data base case after delete the called DB '. Suppose r of category P,

in line for the case N R condition but the number of classes, for r the error; case DB 'is determined by a majority judgment error, assume that the majority of Class P case, the DB' in the number N class called DB 'of the error. So every pick a class association rule, there will be one pair of class association rule of error, and choose class association rule to the whole process lasted until the lowest error

Its algorithm is as follows:

1 *R* = sort(*R*);

2 **for** each rule *r* in *R* in sequence **do**

3 *temp* = empty;

4 **for** each case *d* in *D* **do**

5 **if** *d* satisfies the conditions of *r* **then**

6 store *d*.id in *temp* and mark *r* if it correctly classifies *d*;

end

7 **if** *r* is marked **then**

8 insert *r* at the end of *C*;

9 delete all the cases with the ids in *temp* from *D*;

10 select a default class for the current *C*;

11 compute the total number of errors of *C*;

12 **end**

13 **end**

14 Find the first rule *p* in *C* with the lowest total number of errors and drop all the rules after *p* in *C*;

15 Add the default class associated with *p* to end of *C*, and return *C* (our classifier).

5. Applying Rules for Classification

When the classification of an unknown category of object, there are some methods of selecting rules:

1. Confidence sum
2. Higher confidence
3. The number of qualified rule

The methods are as follows:

1. Confidence sum: If this object contains a number of rules Ri and Rj, Ri is obligate type rule set and Rj is non- obligate type rule set. Let X is sum of confidence for Ri, Y is sum of confidence for Rj. If X > Y, we surmise this object type is obligate, else we surmise this object type is transient.
2. Higher confidence: If this object contains a number of rules R, R is rule set contain obligate and non-obligate two type. Let Ri is descending order of confidence value rule set. Determine which type large quantity for the top few rule in set. If object type large, we surmise this object type is obligate, else we surmise this object type is transient.
3. The number of qualified rule: If this object contains a number of rules Ri and Rj, Ri is obligate type rule set and Rj is non- obligate type rule set. If the number of rule in Ri is large than Rj, then surmise this object is obligate type, if not surmise this object is transient type.

6. Support Adjustment

When predicting PPI type, we find that no matter which method of calculation the prediction results are almost always obligate type. There are almost all obligate data rule, transient data rule almost nothing. We judge because the gap between the obligate data and transient quantity data, resulting in a lower number of transient rule, more likely to be filtered out, so we focused on a number of imbalances do numerically adjustment.

The following formula:

$$C(x) = P(x \cap \text{obligate}) * R / (P(x \cap \text{obligate}) * R + P(x \cap \text{non-obligate})) \quad (1)$$

Let's R is assumed that the ratio of transient to obligate. X is a rule, $P(x \cap \text{obligate})$ denote this PPI is obligate and this PPI contain rule x. $P(x \cap \text{non-obligate})$ denote this PPI is transient.

Table1. The number of non-obligate rules.

Factor	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
Before support adjustment	2	1	1	1	1	0	0	1	1	2
After support adjustment	11	14	14	18	34	64	95	173	291	449

This table shows the number two transient rules in range factor from 1.0 to 2.0. Before the change of the number of transient rules are rare and some cases do not even have. After the change, see table 1 the number of transient rules has increased after support adjustment.

7. Experiment and Discussion

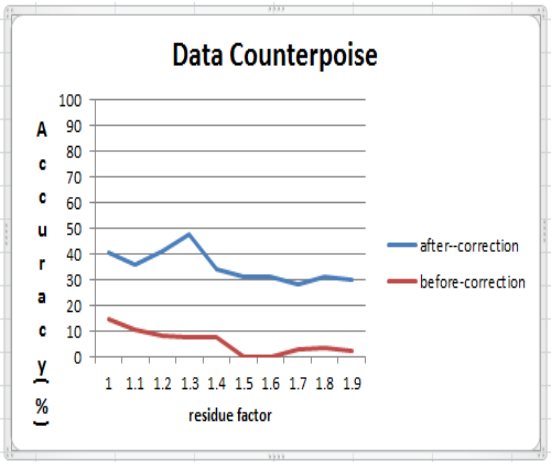

Fig 3 The correct rate of non-obligate data prediction

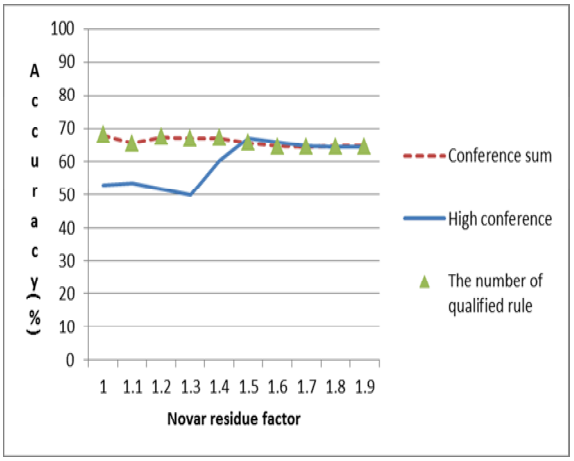

Fig. 4 The result of proposed method.

In figure 3, we find that is low correct rate before data counterpoise. Because the number of non-obligate rule is more less than obligate rule number. From the figure 4, the results of the number of qualified rule and confidence sum rule number the difference of two methods is not large. Because the confidence will be greater the more acceptable when the number of its rule, cause results not dissimilar large. High confidence at lower accuracy rate method beginning, because at factor size is relatively small number of transient rule is not much. It is easy to determine when not to judge him, because of high confidence. The number of the back of transient rule is changed for a long time. It increases the accuracy by High confidences before making a judgment.

8. Conclusions

The amount of protein data is enormous, coupled with environmental variation factors of uncertainty. It takes a lot of time and money to determine protein-protein interaction in wet lab. So there are many experts and scholars toward using known information to predict protein interactions situation, in order to reduce the amount of protein test objectives.

We use a class association rule method for classifying protein-protein interaction type. And we compared several screening methods about screening associated rules. Due to type imbalance, where there are much more obligate protein complexes than transient protein complexes, the interesting measures of the mined rules are tortured. We have designed a method to adjust this effect.

The proposed method can further be used to screen proteins that might have a certain type of protein-protein interaction with a query protein. For biologists, it may take much less time to explore; also it saves time to experiment. Even for pharmaceutical research and development, it has brought many benefits. Mainly proteins and protein interactions experimentally really have to spend a lot of time and money. If a system can quickly provide a list of the list of subjects, there will be a great help.

References

[1] SE Ozbabacan, HB Engin, A Gursoy, O Keskin, "Transient Protein-Protein Interactions," Protein Engineering, Design & Selection, Vol. 24, No. 9, pp. 635-48, 2011.

[2] Ravi Gupta, Ankush Mittal and Kuldip Singh, "A Time-Series-Based Feature Extraction Approach for Prediction

of Protein Structural Class," EURASIP Journal on Bioinformatics and Systems Biology, pp. 1-7, 2008.

[3] Julian Mintseris and Zhiping Weng, "Atomic Contact Vectors in Protein-Protein Recognition," PROTEINS: Structure, Function, and Genetics, Vol. 53, pp. 629–639, 2003.

[4] S. Grosdidier, J. Fernández-Recio, "Identification of Hot-spot Residues in Protein-protein Interactions by Computational Docking," BMC Bioinformatics, 9:447, 2008.

[5] JR Perkins, I Diboun, BH Dessailly, JG Lees, C Orengo, "Transient Protein-Protein Interactions: Structural, Functional, and Network Properties," Structure, Vol. 18, No. 10, pp. 1233-43, 2010.

[6] Florian Goebels and Dmitrij Frishman, "Prediction of Protein Interaction Types based on Sequence and Network Features,"BMC Systems Biology, 7(Suppl 6):S5, 2013.

[7] Huang-Cheng Kuo and Ping-Lin Ong, "Classifying Protein Interaction Type with Associative Patterns," IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 143-147, 2013.

[8] Nurcan Tuncbag,Gozde Kar, Ozlem Keskin, Attila Gursoy and Ruth Nussinov, "A Survey of Available Tools and Web Servers for Analysis of Protein-Protein Interactions and Interfaces," Briefings in Bioinformatics, Vol. 10, No. 3, pp. 217-232, 2009.

[9] R. P. Bahadur and M. Zacharias, "The Interface of Protein-protein Complexes: Analysis of Contacts and Prediction of Interactions," Cellular and Molecular Life Sciences, Vol.65, pp. 7-8, 2008.

[10] Huang-Cheng Kuo, Ping-Lin Ong, Jung-Chang Lin, Jen-Peng Huang, "Prediction of Protein-Protein Recognition Using Support Vector Machine Based on Feature Vectors," IEEE International Conference on Bioinformatics and Biomedicine, pp. 200-206, 2008.

[11] Huang-Cheng Kuo, Ping-Lin Ong, Jia-Jie Li, Jen-Peng Huang, "Predicting Protein-Protein Recognition Using Feature Vector," International Conference on Intelligent Systems Design and Applications, pp. 45-50, 2008.

[12] M. Altaf-Ul-Amin, H. Tsuji, K. Kurokawa, H. Ashahi, Y. Shinbo, and S. Kanaya, "A Density-periphery Based Graph Clustering Software Developed for Detection of Protein Complexes in Interaction Networks," International Conference on Information and Communication Technology, pp. 37-42, 2007.

[13] Sung Hee Park, José A Reyes, David R Gilbert, Ji Woong Kim and Sangsoo Kim, "Prediction of Protein-protein Interaction Types Using Association Rule based Classification," BMC Bioinformatics, Vol. 10, January 2009.

[14] Protein Data Bank [http://www.rcsb.org/pdb/home/home.do]

[15] NOXclass [http://noxclass.bioinf.mpi-inf.mpg.de/]

[16] Biomolecular Object Network Databank [http://bond.unleashedinformatics.com/]

[17] Chengbang Huang, Faruck Morcos, Simon P. Kanaan, Stefan Wuchty, Danny Z. Chen, and Jesus A. Izaguirre, "Predicting Protein-protein Interactions from Protein Domains Using a Set Cover Approach," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No. 1, pp. 78-87, 2007.

[18] Frans Coenen and Paul Leng, "The Effect of Threshold Values on Association Rule based Classification Accuracy," Data & Knowledge Engineering, Vol. 60, No. 2, pp. 345-360, 2007.

[19] John Richard Davies, "Statistical Methods for Matching Protein-ligand Binding Sites", Ph.D. Dissertation, School of Mathematics, University of Leeds, 2009.

[20] S. Lukman, K. Sim, J. Li, Y.-P. P. Chen, "Interacting Amino Acid Preferences of 3D Pattern Pairs at the Binding Sites of Transient and Obligate Protein Complexes," Asia-Pacific Bioinformatics Conference, pp. 14-17, 2008.

[21] Huang-Cheng Kuo, Jung-Chang Lin, Ping-Lin Ong, Jen-Peng Huang, "Discovering Amino Acid Patterns on Binding Sites in Protein Complexes," Bioinformation, Vol. 6, No. 1, pp. 10-14, 2011.

[22] Aaron P. Gabow, Sonia M. Leach, William A. Baumgartner, Lawrence E. Hunter and Debra S. Goldberg, "Improving Protein Function Prediction Methods with Integrated Literature Data," BMC Bioinformatics, 9:198, 2008.

[23] Mojdeh Jalali-Heravi, Osmar R. Zaïane, "A Study on Interestingness Measures for Associative Classifiers," ACM Symposium on Applied Computing, pp. 1039-1046, 2010.

[24] Xiaoxin Yin and Jiawei Han, "CPAR: Classification based on Predictive Association Rules," SIAM International Conference on Data Mining, pp. 331–335, 2003.

[25] B. Liu, W. Hsu, Y. Ma, "Integrating Classification and Association Rule Mining," KDD Conference, pp. 80-86, 1998.