Prediction of Implicit Protein–Protein Interaction by Optimal Associative Feature Mining

Jae-Hong Eom, Jeong-Ho Chang, and Byoung-Tak Zhang

Biointelligence Lab., School of Computer Science and Engineering
Seoul National University
Seoul 151-744, South Korea
{jheom,jhchang,btzhang}@bi.snu.ac.kr

Abstract. Proteins are known to perform a biological function by interacting with other proteins or compounds. Since protein–protein interaction is intrinsic to most cellular processes, protein interaction prediction is an important issue in post–genomic biology where abundant interaction data has been produced by many research groups. In this paper, we present an associative feature mining method to predict implicit protein–protein interactions of *S.cerevisiae* from public protein–protein interaction data. To overcome the dimensionality problem of conventional data mining approach, we employ feature dimension reduction filter (FDRF) method based on the information theory to select optimal informative features and to speed up the overall mining procedure. As a mining method to predict interaction, we use association rule discovery algorithm for associative feature and rule mining. Using the discovered associative feature we predict implicit protein interactions which have not been observed in training data. According to the experimental results, the proposed method accomplishes about 94.8% prediction accuracy with reduced computation time which is 32.5% faster than conventional method that has no feature filter.

1 Introduction

With the advancement of genomic technology and genome–wide analysis of organisms, one of the great challenges to post-genomic biology is to understand how genetic information of proteins results in the predetermined action of gene products, both temporally and spatially, to accomplish biological function and how they act together with each other to build an organism. Also, it is known that protein–protein interactions are fundamental biochemical reactions in the organisms and play an important role since they determine the biological processes [1]. Therefore comprehensive description and detailed analysis of protein–protein interactions would significantly contribute to the understanding of biological phenomena and problems.

After the completion of the genome sequence of *S.cerevisiae*, budding yeast (bakers yeast), many researchers have undertaken the task of functionally analyzing the yeast genome comprising more than 6,300 proteins (YPD) [2] and abundant interaction data have been produced by many research groups. Subsequently, several promising methods have been successfully applied to this field.

Z.R. Yang et al. (Eds.): IDEAL 2004, LNCS 3177, pp. 85–91, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The technique to uncover useful information or facts underlying huge data, which called 'data mining', attracts a lot of attention, and much research applying data mining to bioinformatics have already been conducted. One of the most popular data mining methods is 'association rule discovery' developed by Agrawal *et al.* [3]. Satou *et al.* [4] and Oyama *et al.* [5] applied this method to find rules describing the association among heterogeneous genome data. But, nearly all data mining approaches to bioinformatics suffer from high dimensional property of data which have more than thousand features. In data mining, features describing each data represent the dimensionality of each data item. Oyama *et al.* [5], for example, predicted protein–protein interaction using the data which have more than 5,240 feature dimensions.

In this paper, we propose an efficient protein–protein interaction mining technique which performs efficiently with feature dimension reduction. Here we combine the feature selection approach which was originally introduced by Yu *et al.* [6] and the protein interaction mining based on association rule discovery which was originally introduced by Oyama *et al.* [5]. We formulate the problem of protein-protein interaction prediction as the problem of mining feature-to-feature association of each interacting protein. To predict protein–protein interactions with the association information, we use as many features as possible from several major databases such as MIPS, DIP and SGD [7, 8, 9]. And feature dimension reduction filter (FDRF) method based on the information theory is used to select the most informative features and to speed up the overall mining procedures. After the feature filtering, we apply the association rule discovery method to find optimal associative feature sets and predict additional interactions (i.e., implicit interactions) between unknown proteins using discovered associative features.

The paper is organized as follows. In Section 2, the concept of FDRF and its procedure are described. In Section 3, the concept of association mining and the approach for protein–protein interaction analysis with feature association are described. In Section 4, we show experimental results in comparison with conventional mining method. Finally, Section 5 presents concluding remarks and future direction.

2 Feature Dimension Reduction

In many applications such as genome projects, the data size is becoming increasingly large in both rows (i.e., number of instances) and in columns (i.e., number of features). Therefore, feature selection is necessary in machine learning tasks when dealing with such high dimensional data [6].

Feature selection is the process of choosing a subset of original feature so that the feature space is optimally reduced according to a certain evaluation criterion. Generally, a feature is regarded as s good feature if it is relevant to the class concept but is not redundant to any other relevant features and the correlation between two variables can be regarded as a goodness measure.

The correlation between two random variables can be measured by two broad approaches, based on classical linear correlation and based on information theory. Lin-

ear correlation approaches (e.g., *linear correlation coefficient*, *least square regression error*, and *maximal information compression index*) have several benefits. These approaches can remove features with near zero linear correlation to the class and reduce redundancy among selected features. But, linear correlation measures may not be able to capture correlations that are not linear in nature and the calculation requires all features contain numerical values [6]. In this paper, we use an information theory-based correlation measure to overcome these drawbacks.

Each feature of data can be considered as a random variable. And the uncertainty of a random variable can be measured by *entropy*. The entropy of a variable X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)), \tag{1}$$

And the entropy of X after observing values of another variable Y is defined as

$$H(X \mid Y) = -\sum_j P(y_j)\sum_i P(x_i \mid y_j)\log_2(P(x_i \mid y_j)), \tag{2}$$

where $P(y_j)$ is the prior probability of the value y_j of Y, and $P(x_i \mid y_j)$ is the posterior probability of X being x_i given the values of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called *information gain* [10], which is given by

$$IG(X \mid Y) = H(X) - H(X \mid Y) \tag{3}$$

According to this measure, a feature Y is considered to be more correlated to feature X than feature Z, if $IG(X \mid Y) > IG(Z \mid Y)$. Symmetry is a desired property for a measure of correlation between features and information gain. However, information gain is biased in favor of features with more values and the values have to be normalized to ensure they are comparable and have the same affect. Therefore, here we use the *symmetrical uncertainty* as a measure of feature correlation [11], defined as

$$SU(X,Y) = 2\left[\frac{IG(X \mid Y)}{H(X) + H(Y)}\right],\ 0 \le SU(X,Y) \le 1 \tag{4}$$

Figure 1 shows the overall procedure of the correlation-based feature dimension reduction filter which was earlier introduced by Yu *et al*. [6], named fast correlation-based filter (FCBF). In this paper, we call this FCBC procedure as feature dimension reduction filter (FDRF) for our application. The algorithm finds a set of principal features S_{best} for the class concept. The procedures in Figure 1 are divided into two major parts. In the first part (Step 1 and Step 2), it calculates the symmetrical uncertainty (SU) values for each feature, selects relevant feature into S'_{list} based on the predefined threshold δ, and constructs an ordered list of them in descending order according to their SU values. In the second part (Step 3 and Step 4), it further processes the ordered list to remove redundant features and only keeps principal ones among all the selected relevant features.

With symmetrical uncertainty as a feature association measure, we reduce the feature dimension through the feature selection procedure. In Figure 1, the class C is divided into two classes, conditional protein class (C_C) and result protein class (C_R) of interaction. The relevance of a feature to the protein interaction (interaction class) is decided by the value of *c*–correlation and *f*–correlation, where an SU value δ is used as a threshold value.

Given training dataset $S = (f_1,\ldots, f_N, C)$, where $C = C_C \cup C_R$ and
User-decided threshold δ, do following procedure for each class C_C and C_R.

1. **Repeat** Step 1.1 to 1.2, for all i, $i = 1$ to N.
   1.1 **Calculate** $SU_{i,c}$ for f_i.
   1.2. **Append** f_i to S'_{list} when $SU_{i,c} \geq \delta$.
2. **Sort** S'_{list} in descending order with $SU_{i,c}$ value.
3. **Set** f_p with the first element of S'_{list}.
4. **Repeat** Step 4.1 to 4.3, for all $f_p \neq NULL$.
   4.1 **Set** f_q with the next element of f_p in S'_{list}.
   4.2 **Repeat** Step 4.2.1 to 4.2.3, for all $f_q \neq NULL$.
      4.2.1 **Set** $f'_q = f_q$.
      4.2.2 if $SU_{p,q} \geq SU_{q,c}$,
         Remove f_q from S'_{list} and **Set** f_q with the next element of f'_q in S'_{list}.
         else **Set** f_q with the next element of f_q in S'_{list}.
      4.2.3 **Set** f_q with the next element of f_q in S'_{list}
   4.3 **Set** f_p with next the element of f_p in S'_{list}.
5. **Set** $S_{best} = S'_{list}$

Output the most informative optimal feature subset: S_{best}

Fig. 1. The procedures of feature dimension reduction filter (FDRF).

Definition 1 (*c*–correlation $SU_{i,c}$, *f*–correlation $SU_{j,i}$). Assume that dataset S contains N ($f_1,\ldots, f_N$) features and a class C (C_C or C_R). Let $SU_{i,c}$ denote the SU value that measures the correlation between a feature f_i and the class C (called *c*–correlation), then a subset S' of relevant features can be decided by a threshold SU value δ, such that $f_i \in S'$, $1 \leq i \leq N$, $SU_{i,c} \geq \delta$. And the pairwise correlation between all features (called *f*–correlation) can be defined in same manner of *c*–correlation with a threshold value δ. The value of *f*–correlation is used to decide whether relevant feature is redundant or not when considering it with other relevant features.

3 Mining Associative Feature

Association Mining

To predict protein–protein interaction with feature association, we adopt the association rule discovery algorithm (so-called Apriori algorithm) proposed by Agrawal *et*

al. [3]. Generally, an association rule $R\ (A \Rightarrow B)$ has two values, *support* and *confidence*, representing the characteristics of the association rule. Support (*SP*) represents the frequency of co-occurrence of all the items appearing in the rule. And confidence (*CF*) is the accuracy of the rule, which is calculated by dividing the *SP* value by the frequency of the item in conditional part of the rule.

$$SP(A \Rightarrow B) = P(A \cup B),\ CF(A \Rightarrow B) = P(B \mid A) \tag{5}$$

where $A \Rightarrow B$ represents association rule for two items (set of features) *A* and *B* in that order. Association rule can be discovered by detecting all the possible rules whose supports and confidences are larger than the user-defined threshold value called minimal support (SP_{min}) and minimal confidence (CF_{min}) respectively. Rules that satisfy both minimum support and minimum confidence threshold are taken as to be *strong*. Here we consider these strong association rules as interesting ones.

In this work, we use the same association rule mining and the scoring approach of Oyama *et al*. [5] for performance comparison.

Protein Interaction with Feature Association

An interaction is represented as a pair of two proteins that directly bind to each other. To analyze protein–protein interactions with feature association, we consider each interacting protein pair as a transaction of data mining. These transactions with binary vector representation are described in Figure 2. Using association rule mining, then, we extract association of features which generalize the interactions.

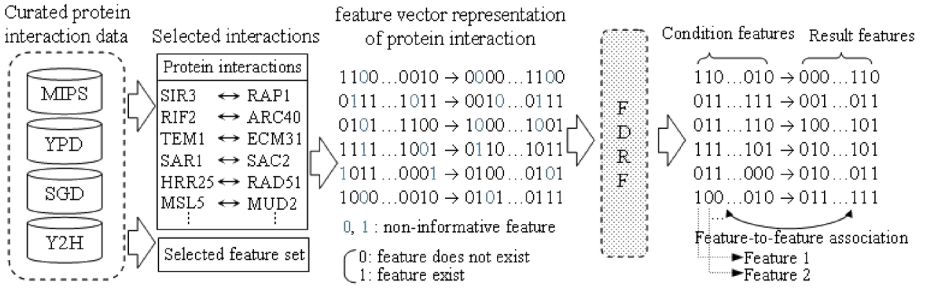

Fig. 2. Representation of protein interaction by feature vectors. Each interaction is represented with binary feature vector (whether the feature exists or not) and their associations. The FDRF sets those features as 'don't care' which have *SU* value less than given *SU* threshold δ. This is intended to consider in association mining only those features that have greater *SU* value than a given threshold. The features marked 'don't care' are regarded as 'don't care' also in association rule mining (i.e., these features are not counted in the calculation of support and confidence). These features are not shown in the vector representation of right side of Figure 2.

4 Experimental Results

Data Sets

Major protein pairs of the interactions are obtained from the same data source of Oyama *et al*. [5]. It includes MIPS [7], YPD [12] and two Y2H data by Ito *et al*. [13]

and Uetz *et al*. [14]. Additionally, we use SGD to collect additional feature set [8]. Table 1 shows the statistics of each interaction data source and the number of features before and after the application of FDRF.

Table 1. The statistics for the dataset.

Data Source	# of interactions	# of initial features	# of filtered features
MIPS	10,641	6,232 (total)	1,293 (total)
YPD	2,952		
SGD	1,482		
Y2H (Ito *et al*.)	957		
Y2H (Uetz *et al*.)	5,086		

Results

First, we selected more informative features using the filtering procedure of Figure 1 (δ=0.73). Then, we performed association rule mining under the condition of minimal support 9 and minimal confidence 75% on the protein interaction data, resulting in 1,293 features from 6,232 features in total. With the mined feature association, we predicted new protein–protein interaction which were not used in association training step. We run 10-fold cross-validation and predictions were validated with comparison to the collected dataset.

Table 2 shows advantageous effects obtained by the filtering of non-informative (redundant) features. By the application of FDRF, the accuracy increased about 3.4% and the computation time decreased by the fraction of 32.5%, even including the FDRF processing time, compared with the conventional method which perform association rule mining without feature filtering. To summarize, we can see that our proposed method which combines feature filtering and association mining not only prevents wrong associations by eliminating a set of misleding or redundnt features of interaction data but also contribute to the reduction of computational complexity accompanying rule mining procedure.

Table 2. Accuracy of the proposed method and the effect of the FDRF-based feature selection in terms of computation time. The elapsed time was measured on Pentium IV 2.4GHz and 1GB RAM system running Windows.

Prediction method	# of interactions			Accuracy (|P|/|T|)	Elapsed Time
	Training set	Test set (T)	Correctly predicted (P)		
Without FDRF	4,628	463	423	91.4 %	212.34 sec
With FDRF	4,628	463	439	94.8 %	143.27 sec
Improvement	–	–	–	3.4 %	32.5 %

5 Conclusions

In this paper, we presented a novel method for predicting protein–protein interaction by combining information theory based feature dimension reduction filter with fea-

ture association mining. The proposed method achieved the improvement in both the prediction accuracy and the processing time. In Table 2, it is also suggested that further detailed investigation of the protein–protein interaction can be made with smaller granularity of interaction (i.e., not protein, but a set of features of proteins). As a result, we can conclude that the proposed method is suitable for efficient prediction of the interactive protein pair with many features from the experimentally produced protein–protein interaction data which have moderate amount of false positive ratios. But, the current public interaction data produced by such as high-throughput methods (e.g., Y2H) have many false positives. And several interactions of these false positives are corrected by recent researches through the reinvestigation with new experimental approaches. Thus, studies on new methods for resolving these problems remain as a future work.

Acknowledgements

This research was supported by the Korean Ministry of Science and Technology under the NRL Program and the Systems Biology Program.

References

1. Deng, M., *et al*.: Inferring domain–domain interactions from protein–protein interactions. *Genome Res.* 12, 1540–1548, 2002.
2. Goffeau, A., *et al*.: Life with 6000 genes. *Science* 274, 546–567, 1996.
3. Agrawal, R., *et al*.: Mining association rules between sets of items in large databases. In *Proc. of ACM SIGMOD-93* 207–216, 1993.
4. Satou, K., *et al*.: Extraction of substructures of proteins essential to their biological functions by a data mining technique. In *Proc. of ISMB-97* 5, 254–257, 1997.
5. Oyama, T., *et al*.: Extraction of knowledge on protein–protein interaction by association rule discovery. *Bioinformatics* 18, 705–714, 2002.
6. Yu, L. and Liu, H.: Feature selection for high dimensional data: a fast correlation-based filter solution. In *Proc. of ICML-03* 856–863, 2003.
7. Mewes, H.W., *et al*.: MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34, 2002.
8. Xenarios, I., *et al*.: DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305, 2002.
9. Christie, K.R., *et al*.: Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res*. 32, D311–D314, 2004.
10. Quinlan, J.: C4.5: Programs for machine learning. Morgan Kaufmann, 1993.
11. Press, W.H., *et al*.: Numerical recipes in C. *Cambridge University Press*., 1988.
12. Csank, C., *et al*.: Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol*. 350, 347–373, 2002.
13. Ito, T., *et al*.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* 98, 4569–4574, 2001.
14. Uetz, P., *et al*.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403, 623–627, 2000.