Adaptive Neural Network-Based Clustering of *Yeast* Protein–Protein Interactions

Jae-Hong Eom and Byoung-Tak Zhang

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

Abstract. In this paper, we presents an adaptive neural network based clustering method to group protein–protein interaction data according to their functional categories for new protein interaction prediction in conjunction with information theory based feature selection. Our technique for grouping protein interaction is based on ART-1 neural network. The cluster prototype constructed with existing protein interaction data is used to predict the class of new protein interactions. The protein interaction data of *S.cerevisiae* (bakers yeast) from MIPS and SGD are used. The clustering performance was compared with traditional *k*-means clustering method in terms of cluster distance. According to the experimental results, the proposed method shows about 89.7% clustering accuracy and the feature selection filter boosted overall performances about 14.8%. Also, inter-cluster distances of cluster constructed with ART-1 based clustering method have shown high cluster quality.

1 Introduction

These days, with the advancement of genomic technology and genome-wide analysis of organisms, one of the great challenges of the post–genomic era of today is to understand how genetic information of proteins results in the predetermined action of gene products both temporally and spatially to accomplish biological functions, and how they act together to build an organism. 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. Therefore the 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, many researchers have undertaken the functional analysis of the yeast genome comprising more than 6,300 proteins (YPD) [1] and abundant interaction data has been produced by many research groups. Thus, the demands for the effective methods to discover novel knowledge from the interaction data through the analysis of these data are more increasing than ever before.

Many attempts have been made to predict protein functions (interactions) using such data as gene expressions and protein–protein interactions. Clustering analysis of gene expression data has been used to predict functions of un-annotated proteins based on the idea that genes with similar functions are likely to be co-expressed [2, 3]. Park *et al*. [4] analyzed interactions between protein domains in terms of the interactions between structural families of evolutionarily related domains. Iossifov *et al*. [5] and Ng *et al*. [6] inferred new interaction from the existing interaction data. And there are many other approaches for analyzing and predicting protein interactions.

G. Das and V.P. Gulati (Eds.): CIT 2004, LNCS 3356, pp. 49–57, 2004.

However, many approaches to protein interaction analysis suffer from high dimensional property of the data which have more than thousand features [7].

In this paper, we propose an adaptive neural network based clustering method for clustering protein–protein interactions in the context of their feature association. We use ART-1 version of adaptive resonance theory [8] as an adaptive neural network clustering model. The ART-1 [9] is a modified version of ART [10] for clustering binary vectors. The advantage of using ART-1 algorithms to group feature abundant interaction data is that it adapts to the change in new protein interactions over various experiment data without losing information about other protein interaction data trained previously. Here, we assume protein–protein interaction of yeast as feature-to-feature association of each interacting proteins. To analyze protein–protein interactions with respect to their interaction class with their feature association, we use as many features as possible from several major databases such as MIPS and SGD [11, 12] to construct a rich feature vector for each protein interaction which is provided to the proposed clustering model. Here, we use the same approach of Rangarajan *et al*. [13] for the design of clustering model and employ the feature selection filter of Yu *et al*. [14] to reduce computational complexity and improve the overall clustering performance by eliminating non-informative features.

The remainder of this paper is organized as follows. In Section 2, we introduce the concept of the feature selection filter and the overall architecture of the clustering model with its learning algorithm. Section 3 describes the representation scheme of the protein–protein interaction for the adaptive neural network based clustering and also presents the experimental results in comparison with the conventional *k*-means clustering. Finally, concluding remarks and future works are given in Section 4.

2 Feature Dimension Reduction and Clustering

Feature Dimension Reduction by Feature Selection

Here, we consider each protein–protein interaction as the feature to feature associations. We constructed massive feature sets for each protein and interacting protein pairs from public protein description databases [11, 12, 15, 16, 17]. However, there exist also many features which have no information of their association with other proteins. Therefore, feature selection may be needed in advance of the clustering of protein–protein interactions. Especially, this feature selection is indispensable when dealing with such high dimensional (feature dimension) data.

Feature selection is the process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion. Generally, a feature is regarded as a good feature when it is relevant to the class concept but is not redundant with 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, one is based on classical linear correlation and the other is based on information theory. Linear 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 with the class variable 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 computation requires all features are numeric-valued [14]. In this paper, we use an information theory-based correlation measure to overcome these drawbacks.

Each feature of a 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 value of $Y=y_j$. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called *information gain* [18], 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 correlated more to the feature X than to the feature Z, if $IG(X \mid Y) > IG(Z \mid Y)$. It is said that symmetry is a desired property for a measure of correlation between features and information gain [14]. 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 effect. Therefore, here we use the *symmetrical uncertainty* as the measure of feature correlation [14, 19], which is 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*. [14], named the fast correlation-based filter (FCBF). In this paper, we call this FCBF procedure as the 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 main parts. In the first part (Step 1 and Step 2), it calculates the symmetrical uncertainty (SU) value for each feature, selects the relevant features into S'_{list} on the basis of 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, the conditional protein class (C_C) and the resulting 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. The two correlations are defined as follows [14].

Definition 1 (*c*–correlation $SU_{i,c}$, *f*–correlation $SU_{j,i}$)**.** Assume that a 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 $S' \subseteq S$ of relevant features can be decided by a threshold SU value δ,

such that $\forall f_i \in S'$, $1 \le i \le N$, $SU_{i,c} \ge \delta$. And the pairwise correlation between all features (called *f*–correlation) can be defined in the same way as the *c*–correlation with a threshold value δ. The value of *f*–correlation is used to decide whether a relevant feature is redundant or not when considered with other relevant features.

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 $SU_{i,c} \ge \delta$.
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 $SU_{p,q} \ge SU_{q,c}$,
            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 the next 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 procedure of the feature dimension reduction filter (FDRF)

Clustering Protein Interactions

We use ART-1 neural network for grouping the class of protein–protein interactions. In our ART-1 based clustering, each protein interaction is represented by a prototype vector that is a generalized representation of the set of features of each interacting proteins. Figure 2 presents the architecture of ART-1 based clustering model. The PPI_i stands for each protein interaction and it includes the set of features of two interacting proteins. The degree of similarity between the members of each cluster can be controlled by changing the value of the vigilance parameter. The overall procedure for clustering protein–protein interactions with the ART-1 based clustering model described in Appendix A. The basic layout of this procedure is identical with the work of Rangarajan *et al*. [13].

3 Experimental Results

Protein Interaction as Binary Feature Vector

An interaction is represented as a pair of two proteins that directly binds to each other. This protein interaction is represented by a binary feature vector of interacting proteins and their associations. Figure 3 describe this interaction representation processes. Interaction data prepared in this manner are provided to the ART-1 based clus-

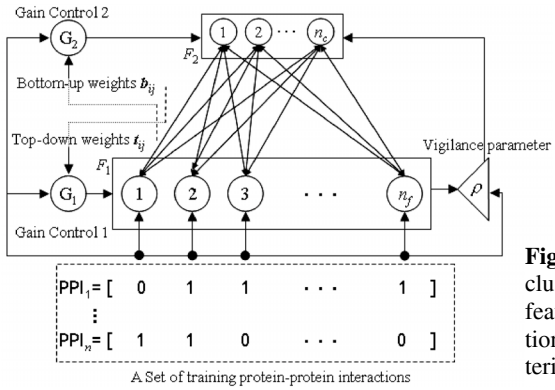

Fig. 2. The architecture of ART-1 based clustering model. PPI_i stands for the feature vector of protein–protein interaction which represents the feature characteristic of interacting proteins.

ter model to group each protein interaction class and learn the association of features which generalize the interactions. Then, the constructed cluster prototype is used to predict the classes of protein interactions presented in the test step. The class of interacting protein from MIPS [11] which is known for the most reliable curated protein interaction database in current literature is used to evaluate the clustering accuracy.

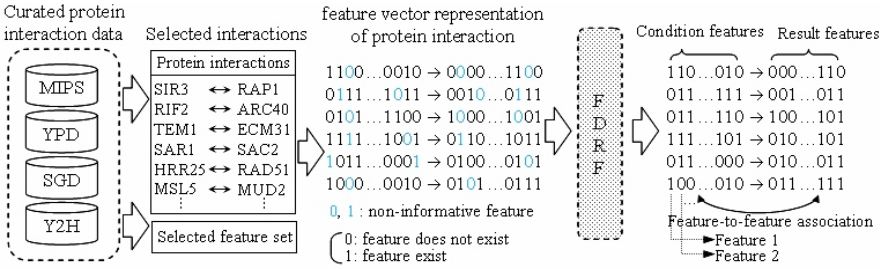

Fig. 3. Representation of protein interaction by feature vectors. Each interaction is represented as a binary feature vector (whether the feature exists or not) and their associations. The FDRF sets those features as 'don't care' which have *SU* value less than an *SU* threshold δ. This is intended to eliminate non-informative features so as to boost up the clustering performance of ART-1 based clustering model. The features marked 'don't care' are also regarded as 'don't care' in ART-1 based clustering model training. These non-informative features are not shown in the vector representation of right side of Figure 3. The resulting binary vector of interaction is provided to the ART-1 based clustering model, described in Figure 2, for model training and testing

Data Sets

Each *yeast* protein has various functions or characteristics which are called 'feature.' In this paper, set of features of each protein are collected from public genome databases [11, 12, 15, 16, 17]. We use similar features of protein interaction of Oyama *et al*. [20] which include YPD categories, EC numbers (from SWISS-PROT), SWISS-PROT/PIR keywords, PROSITE motifs, bias of the amino acids, segment cluster, and amino acid patterns. A major protein pairs of the interactions are also obtained from the same data source of Oyama *et al*. [20]. These interaction dataset include various experimental data such as YPD and Y2H by Ito *et al*. [16] and Uetz *et al*. [17]. Additionally, we used SGD to construct more abundant feature set [12].

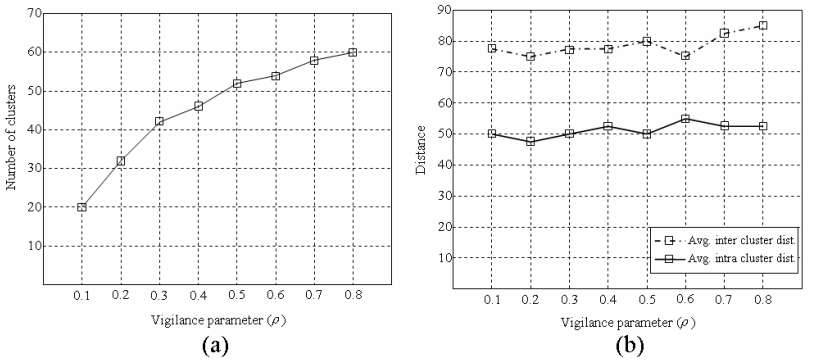

Fig. 4. The effects by the variation of vigilance parameter ρ. Figure 4(a) shows the variation of the number of cluster with the increase of vigilance parameter and Figure 4(b) shows the variation in cluster distances as the vigilance parameter ρ varies

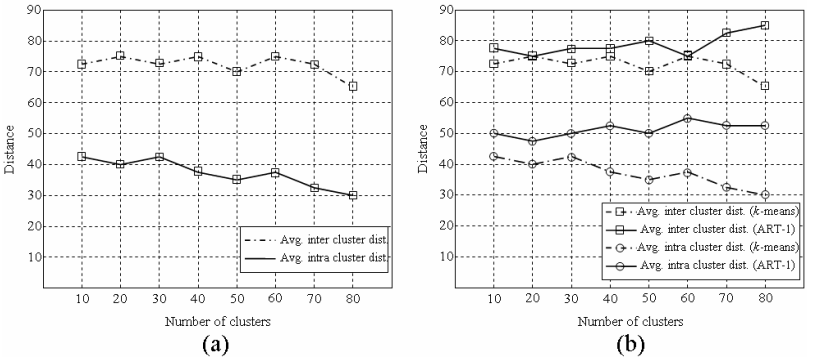

Fig. 5. The effects of the number of clusters on the quality of the clustering result. (a) The cluster distance variation of *k*-means cluster model. (b) The comparison of ART-1 and *k*-means clustering models with respect to average inter–cluster and intra–cluster distances. ART-1 based method shows good clustering qualities (all ART-1 based clustering models were trained with FDRF filtered interaction vectors)

With the constructed interaction class, we predicted the class of new protein–protein interactions. The accuracy of class prediction is measured whether the predicted class of interaction is corretly correspond to the class of MIPS. The results are measured with 10-fold cross-validation. And the quality of the clustering result is evaluated by inter– (average distance between clusters) and intra– (average distance between members of each cluster) cluster distance. The *k*-means method was used as a baseline clustering model. The *k*-means partitions a given dataset into *k* clusters. The same number of clusters corresponding to each variant in the results of ART-1 based model were used for performance comparison.

Results

Figure 4(a) shows the increase in the number of cluster with the increase of vigilance parameter and Figure 4(b) shows the variation in cluster distances of ART-1 based cluster model. Figure 5(a) illustrates the distance variation of *k*-means clustering.

Figure 5(b) compares the variation in average inter–cluster distance for the two clustering models as the number of cluster increases. Both ART-1 and *k*-means show distance varying at a steady rate with slight fluctuations. However, the results of ART-1 based model show quite uniform fashion compared to the result of the *k*-means. We can consider this uniformity of ART-1 based model indicates clustering stability, which is an important attribute of high-quality clustering models [13].

Table 1 shows the cluster prediction performance of ART-1 based model and traditional *k*-means model. ART-1 based model outperformed on prediction accuracy about 20% (when it trained with the interaction vectors filtered with FDRF). Thus, we can guess that the ART-1 based cluster model is very useful for the clustering of data which have many features and the proposed clustering model could be used for the analysis of protein–protein interactions. The overall prediction acuracy was improved about 14.8% by FDRF feature selection. Thus we can say that the information theory based feature selection procedure contributes to the improvement of predection acuracy and it is useful as a data preprocessing methods, especilly, when we handle the data which have many features (i.e., high dimensional data).

Table 1. Result of protein interaction class prediction by *k*-means and ART-1 model. The ART-1 based model with FDRF filtered interaction vectors shows the best performance

Method for prototype cluster construction	Number of interactions (T)	Number of interactions predicted correctly (P)	Accuracy (|P|/|T|)
k-means	4,628	3,202	69.2 %
ART-1 (without FDRF)	4,628	[a] 3,466	[a] 74.9 %
ART-1 (with FDRF)	4,628	[b] 4,151	[b] 89.7 %
Diffrence	–	[a] 264 [b] 949	[a] 5.7 % [b] 20.5 %

4 Conclusions

In this paper, we presented an adaptive neural network (ART-1) based clustering method for clustering protein–protein interaction. We applied an information theory-based feature selection procedure to improve the performance of trained clustering model. The proposed method achieved the improvement of accuracy about 20%. From the experimental result of the quality of the clustering result and clustering accuracy, it is suggested that the neural network-based clustering model can be used for a more detailed investigation of the protein–protein interactions, since the proposed model can learn effectively the hidden patterns of the data which have many features.

The current public interaction data produced by a high-throughput method (e.g., Y2H) have many false positives and some of these false positives are corrected as true positives by recent researches with new modern experimental approaches. Thus, the study on the new method for adapting these changes in data set, which is related to false positive screening, remains as future works.

Acknowledgements. This research was supported by the Korean Ministry of Science and Technology under the NRL Program and the Systems Biology Program. The RIACT at Seoul National University provided research facilities for this study.

Appendix A

Input: an array of input protein interaction vectors ***PPI*** and vigilance parameter ρ.

1. **Initialize**
   1.1 **Set** the value of gain control G_1 and G_2,
   $$G_1, G_1 = \begin{cases} 1 \; if \; input \; PPI_I \neq 0 \; and \; output \; from \; F_2 \; Layer = 0 \\ 0 \; for \; all \; other \; cases \end{cases}$$
   1.2 **Set** all nodes in F_1 layer and F_2 layer to 0.
   1.3 **Set** all weight of top-down weight matrix, $t_{ji} = 1$.
   1.4 **Set** all weight of bottom-up weight matrix,
   $b_{ij} = \frac{1}{n_f + 1}$ (n_f = the size of the input feature vector, here n_f = 5,240).
   1.5 **Set** the vigilance parameter ρ (0.2 to 0.7).

2. **Repeat** Step 2.1 to 2.7, for all protein-protein interaction vector PPI_I.
   2.1 **Read** randomly chosen interaction vector
   $PPI_I = (P_1, P_2, \cdots, P_{I=5,240}),$ *where* $P_1 = 0$ *or* 1.
   2.2 **Compute Input** y_j for each node in F_2, $y_j = \sum_{i=1}^{5,240} P_i \times b_{ij}$.
   2.3 **Determine** ***k***, $y_k = \sum_{j=1}^{\#\ of\ nodes\ in\ F_2} \max(y_j)$.
   2.4 **Compute Activation**, $X_k^* = (X_1^*, X_2^*, \cdots, X_{i=5,240}^*)$ for the node k in F_1
   Where, $X_i^* = t_{ki} \times P_i$ $(i = 1, \cdots, 5240)$.
   2.5 **Calculate similarity** δ, between X_i^* and PPI_i:
   $$\delta = \frac{\left\| X_k^* \right\|}{\left\| PPI_I \right\|} = \frac{\sum_{i=1}^{5,240} X_i^*}{\sum_{i=1}^{5,240} P_i}.$$
   2.6 **Update weight** of top-down weight matrix with PPI_I and node k
   If $\delta > \rho$, update top-down weight of node k, $t_{ki}(new) = t_{ki} \times P_i$ where $i = 1, \cdots, 5240$.
   2.7 **Create a new node** in F_2 layer
   2.7.1 **Create** a new node l.
   2.7.2 **Initialize** top-down weight t_{li} to the current input feature pattern.
   2.7.3 **Initialize** button-up weight for the new node l.
   $$b_{il}(new) = \frac{X_i^*}{0.5 + \sum_{i=1}^{5,240} X_i^*} \quad \text{where } i = 1 \ldots 5240.$$

The procedure of ART-1 model based clustering of protein–protein interaction. The basic procedure and the update formula are identical with the work of Rangarajan *et al*. [13].

References

1. Goffeau, A., Barrell, B.G., *et al*.: Life with 6000 genes. *Science* **274** (1996) 546–67.
2. Eisen, M.B., Spellman, P.T., *et al*.: Cluster analysis and display of genomewide expression patterns. *Proc. Natl. Acad. Sci. USA* **95** (1998) 14863-68.
3. Pavlidis, P. andWeston, J.: Gene functional classification from heterogeneous data. In *Proceedings of the 5th International Conference on Computational Molecular Biology* (*RECOMB-2001*) (2001) 249-55.
4. Park, J., Lappe, M., *et al*.: Mapping protein family interactions: intra-molecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* **307** (2001) 929-39.
5. Iossifov, I., Krauthammer, M., *et al*. Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics* **20**(8) (2004) 1205-13.
6. Ng, S.K., Zhang, Z., *et al*. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* **19**(8) (2003) 923-29.
7. Eom, J.-H., Chang, J.-H., *et al*.: Prediction of implicit protein–protein interaction by optimal associative feature mining. In *Proceedings of the 5th International Conference on Intelligent Data Engineering and Automated Learning* (*IDEAL'04*), (2004) 85-91.
8. Carpenter, G.A. and Grossberg, S.: A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision*, *Graphics and Image Processing* **37** (1987) 54-115.
9. Barbara, M.: ART1 and pattern clustering. In proceedings of the 1988 connectionist models summer 1988. Morgan Kaufmann (1988) 174-85.
10. Heins, L.G. and Tauritz, D.R.: Adaptive resonance theory (ART): an introduction. *Internal Report* 95-35, Dept. of Computer Science, Leiden University, Netherlands (1995) 174-85.
11. Mewes, H.W., Amid, C., *et al*.: MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* **32**(1) (2004) D41-44.
12. Christie, K.R., Weng, S., *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**(1) (2004) D311-14.
13. Rangarajan, S.K., Phoha, V.V., *et al*.: Adaptive neural network clustering of web users. *IEEE Computer* **37**(4) (2004) 34-40.
14. Yu, L. and Liu, H.: Feature selection for high dimensional data: a fast correlation-based filter solution. In *Proceedings of the 12th International Conference on Machine Learning (ICML'2003)*, (2003) 856–863.
15. Csank, C., Costanzo, M.C., *et al*.: Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol*. **350** (2002) 347-73.
16. Ito, T., Chiba, T., *et al*.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98** (2001) 4569-4574.
17. Uetz, P., Giot, L., *et al*.: A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature* **403** (2000) 623-627.
18. Quinlan, J.: C4.5: Programs for machine learning. Morgan Kaufmann. (1993)
19. Press, W.H., Teukolsky, S.A, *et al*.: Numerical recipes in C. *Cambridge University Press*. (1988)
20. Oyama, T., Kitano, K., *et al*.: Extraction of knowledge on protein–protein interaction by association rule discovery. *Bioinformatics* **18** (2002) 705-14.