Proceedings of IEEE International Conference on Data Mining (ICDM) 2006, Workshop on Data Mining in Bioinformatics: IEEE Computer Society Press; 2006. p. 114 - 119.

Predictive Integration of Gene Ontology-Driven Similarity and Functional Interactions

Francisco Azuaje[1,*], Haiying Wang[1], Huiru Zheng[1],Olivier Bodenreider[2] and Alban Chesneau[3]
[1]School of Computing and Mathematics, University of Ulster, UK., [2]National Library of Medicine, National Institutes of Health., USA, [3] High-Throughput Protein Technologies Group, EMBL-Grenoble, France.
E-mail: fj.azuaje@ulster.ac.uk, hy.wang@ulster.ac.uk, h.zheng@ulster.ac.uk, olivier@nlm.nih.gov, chesneau@embl-grenoble.fr

Abstract

There is a need to develop methods to automatically incorporate prior knowledge to support the prediction and validation of novel functional associations. One such important source is represented by the Gene Ontology (GO)™ and the many model organism databases of gene products annotated to the GO. We investigated quantitative relationships between the GO-driven similarity of genes and their functional interactions by analyzing different types of associations in Saccharomyces cerevisiae and Caenorhabditis elegans. Interacting genes exhibited significantly higher levels of GO-driven similarity (GOS) in comparison to random pairs of genes used as a surrogate for negative interactions. The Biological Process hierarchy provides more reliable results for co-regulatory and protein-protein interactions. GOS represent a relevant resource to support prediction of functional networks in combination with other resources.

1. Introduction

The reliable prediction of functional networks of genes may be achieved by integrating multiple data sources, such as gene expression and high-throughput protein-protein interaction experiments. This is necessary because such individual sources may be considered as *weak prediction models*. Several studies have reported significant links between different types of genomic data sets, as well as techniques to combine them and improve prediction quality for relatively simple model organisms [1], [2]. Furthermore, it is crucial to integrate prior knowledge resources, such as annotation databases and the literature, not only for building advanced functional classifiers, but also to assist in the validation of technique-independent predictions (e.g., to detect potential spurious associations). The *Gene Ontology ™* (GO) is one such source of prior knowledge, which has become the *de facto* standard for annotating gene products [3]. Information extracted from model organism databases annotated to the GO has been applied to gene expression analysis and for making *de novo* functional predictions [4]. Methods based on the GO have been proposed for measuring similarity between genes. Previous research showed significant relationships between GO-driven similarity of pairs of genes and their sequence-based similarity [5]. We have also evaluated relevant relationships between GO-driven similarity and gene expression correlation [6].

Prior to integrating a predictive resource, *Res*, it is first necessary to assess its predictive relevance in relation to data sets of known positive and negative interactions. In this case the hypothesis to prove is: Can information extracted from *Res* be in principle applied to distinguish pairs of interacting genes (positives) from those that have not shown evidence to be interacting (negatives)? Are there significant quantitative relationships to indicate that *Res* may be used as an input to different prediction models?

The application of information from model organism databases annotated to the GO to support the prediction of functional networks has not been rigorously investigated. Jansen *et al.* [1] integrated different data sets including annotations derived only from the GO *Biological Process* hierarchy to predict protein-protein (PP) interactions. The GO-driven similarity of a pair of genes was used as an indicator of PP interactions in yeast. Between-gene similarity was calculated by identifying the set of GO terms shared by

the two sets of annotations. For a given database of protein pairs, the total number of protein pairs sharing the same set of annotations was used as an estimator of similarity. Thus, the lower this frequency, the more similar the gene pair under consideration. They found that lower term frequencies were correlated with a higher likelihood of finding two proteins in the same complex. Nevertheless, such a similarity assessment approach does not fully exploit relevant topological and information content features that may be useful for estimating between-gene similarity. In some cases genes annotated to closely related but distinct GO terms may actually exhibit no similarity according to this method.

Using annotations from the three GO hierarchies: *Molecular Function* (MF), *Biological Process* (BP) and *Cellular Component* (CC), we sought to assess relationships between the GO-driven similarity of a pair of genes and their functional interactions. This study investigated the feasibility of applying GO-driven similarity to support the prediction of functional interactions of genes, including physical and regulatory interactions, in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. A key question addressed was: Can GO-driven similarity be applied to estimate the functional coupling of genes? Our hypothesis is that the GO-driven similarity among genes is a relevant indicator of functional interaction.

2. Materials

2.1. Data sets

Gene co-regulation in *S. cerevisiae* (CoReg)

This data set originated from a comprehensive collection of annotated regulons compiled by Simonis *et al.* [7]. Their data set comprised more than 1400 pairs of gene-factor associations retrieved from public databases and literature. More than 13000 pairs of co-regulated genes were then extracted from these data. These pairs comprised the CoReg reference data set analyzed in this investigation.

Functional network of yeast genes (FunNet)

This data set was obtained from an extensive, high-quality functional gene network investigated by Lee *et al.* [2]. Unlike the CoReg data set, FunNet comprises different types of functional associations, mediated or not by physical interaction. This network was inferred by integrating diverse, high-quality functional data sets (e.g. mRNA coexpression, gene-fusions). A sub-sample of 19,216 pairs of genes representing the most reliable interaction predictions were analyzed in this study.

PP interactions in *C. elegans* (PPInt)

This data set represents another level of complexity, in which 860 protein-protein (PP) interactions were obtained from the *Worm Interactome* (WI5) map. The selected data set, from now on referred to as PPInt, contains the highest-confidence WI5 interactions [8].

2.2. The GO

The GO hierarchies provide controlled terms for describing the role played by a gene product, the biological goals to which a gene product contributes and the cellular localization of the gene product respectively. Within each hierarchy, GO terms are organized in a *directed acyclic graph*, whose nodes are the terms. There are two types of relationships among GO terms: "is a" and "part of". The first type is used when a child term is more specific than its parent term. The second type is used when a parent has the child as its part. This study takes advantage of both types of links for computing similarity between terms as justified elsewhere [5]. The annotations recorded in the model organism databases consist of associations between gene products and GO terms. The evidence supporting such annotations is captured by evidence codes, including *TAS* (Traceable Author Statement), *ISS* (Inferred from Sequence or structural Similarity) and *IEA* (Inferred from Electronic Annotation). While *TAS* refers to peer-reviewed papers and indicates strong evidence, *IEA* and *ISS* denote automated predictions, i.e., generally less reliable annotations. The reader is also referred to [6] and [9] for an introduction to some of the predictive data analysis applications of the GO.

2.3. GO annotation databases

The pairs of interacting genes in the three data sets presented earlier are annotated to the GO. We performed experiments on data excluding the less reliable annotations (i.e., ignoring annotations whose evidence code is either *ISS* or *IEA*). Moreover, we compared these results against those obtained from excluding only *IEA* annotations. The August 2005 database releases of the *Saccharomyces Genome Database* (SGD) and *WormBase* (WB), all available at www.godatabase.org, provided the GO annotations for these data sets. CoReg has 8,839, 10,874, and 11,309 interacting pairs with both genes linked to at least one GO term under the MF, BP and CC hierarchies respectively. FunNet had 11,767, 15,520 and 16,865 pairs of interacting genes with both genes associated with at least one GO term under the MF, BP and CC hierarchies respectively. In PPInt the numbers of

interacting pairs of genes in which both genes were described by at least one GO term were 152 under the BP hierarchy and 5 under the CC hierarchy. This data set did not contain any valid annotations under MF. The number of annotations reported above refers to non-*ISS*/non-*IEA* annotations.

3. Methods

3.1. GO-driven similarity

To estimate the similarity between two genes g_k and g_p, annotated with sets of GO terms A_k and A_p respectively, one must first understand how to calculate the similarity between two GO terms. Several i*nformation-theoretic approaches* to measuring ontology-driven similarity have been studied previously [5], [9]. Unlike traditional *edge-counting techniques*, these methods are based on the assumption that the more information two terms share in common, the more similar they are. *Lin's similarity model*, for example, has shown to produce both biologically meaningful and consistent similarity predictions [5], [6] in comparison to related approaches. Given terms $c_i \in A_k$ and $c_j \in A_p$, the between-term Lin's similarity is defined as:

$$sim(c_i, c_j) = \frac{2 \times \max_{c \in S(c_i, c_j)} [\log(p(c))]}{\log(p(c_i)) + \log(p(c_j))} \quad (1)$$

where $S(c_i,c_j)$ represents the set of ancestor terms shared by both c_i and c_j, 'max' represents the maximum operator, and $p(c)$ is the probability of finding c or any of its descendants in the database analyzed. It generates normalized values between 0 and 1.

Between-gene similarity results from the aggregation of similarity values between the annotation terms of these genes. In practice, given a pair of gene products, g_k and g_p, with sets of annotations A_k and A_p comprising m and n terms respectively, the between-gene similarity, $SIM(g_k, g_p)$, is defined as the *simple average* (inter-set) *similarity* between terms from A_i and A_j:

$$SIM(g_k, g_p) = \frac{1}{m \times n} \times \sum_{c_i \in A_k, c_j \in A_p} sim(c_i, c_j) \quad (2)$$

where $sim(c_i,c_j)$ may be calculated using (1). Nevertheless, this method might not always produce consistent results. For example, intuitively, the similarity between two genes having the same sets of annotation terms is expected to be equal to 1. However, this is not true when several annotations within a hierarchy are assigned to the genes. It will estimate, for instance, $SIM(g_i, g_j) = 0.5$, for $g_i = g_j$ when A_i and A_j are described by the same set of annotations with more than one GO term within a hierarchy. In order to address this limitation, we have introduced an alternative approach that selectively aggregates *highest average* (inter-set) *similarity* values [9] as follows:

$$SIM(g_i, g_j) = \frac{1}{m+n} \times \left(\sum_k \max_p (sim(c_k, c_p)) + \sum_p \max_k (sim(c_k, c_p)) \right) \quad (3)$$

These approaches and their relationships to sequence-based similarity and co-expression have been investigated in [5] and [6]. From now on we will refer to (2) and (3) as the *simple* and *highest average similarity* methods respectively.

3.2. Linking GO-driven similarity and functional interactions

Comparing GO-driven similarity to other indicators of functional relations. GO-driven similarity values were calculated for all the annotated pairs of genes in the data sets described in Section 2. These data represented our sets of true positive interactions, which were statistically analyzed to show significant relationships with GO-driven similarity. In order to illustrate such links, similarity values from these sets of true positive interactions were compared to similarity values measured in a set of randomly associated genes, used as a surrogate for negative interactions, i.e. pairs of genes not showing evidence of interaction. In practice, a set of "non-interacting genes" was produced as follows. For a given data set, **P**, comprising M true positive interactions, a set **N**, with M negative interactions was built by randomly pairing genes from **P**. Moreover, the resulting sets were verified to ensure that newly formed pairs were not included in **P**. One has to take into account that some of the pairs included in **N** may actually be false negatives (i.e., interacting genes whose interaction has not been not recorded in **P**) and this might influence the comparisons performed. However, at least with regard to the data sets analyzed (evidence available) this could not be demonstrated. The resulting data sets **N** represent a valid approximation of counter-examples, which are essential to explore potential associations between functional interactions and GO-driven similarity. Furthermore, the random effects and variability linked to this data sampling procedure is reduced by generating K independent **N** sets. These K sets are then analyzed as an aggregated set, **N'**, consisting of K x M pairs of (non-interacting) genes.

Fundamental relationships between GO-driven similarity and the existence/absence of interactions were estimated by comparing similarity values exhibited in **P** versus values observed in **N'**. This was done for each of the three problems described in Section 2 and for the three GO hierarchies independently. Differences between **P** and **N'** were summarized by estimating their respective mean similarity values. The significance of their differences was tested by applying the Student's *t*-Test. The relevant null hypothesis tested was that these mean similarity values originated from the same sample, i.e. there are no significant differences between mean values in **P** and **N'**.

Using GO-driven similarity to predict interactions. After identifying significant differences, the capacity of GO-driven similarity to predict functional interactions was analyzed. Given a similarity value, $SIM(g_k, g_p)$, and a pre-defined *predictive similarity threshold* value, *GOS-Th*, genes g_k and g_p are predicted to be an interacting pair (positive interaction) if $SIM(g_k,g_p) \geq GOS\text{-}Th$. Some of these predictions will obviously be false. Therefore, the next task was to estimate the rate of falsely predicted interactions. More generally, this is related to the problem of estimating the *decisive false discovery rate*, which has shown to be a robust and conservative estimator of the probability, *P*, of detecting spurious associations [10]. To estimate *P*, *AbN'* and *AbP* are calculated. *AbN'* represents the number of interactions that would occur by chance and *AbP* the number of pairs correctly predicted as positive interacting pairs. The ratio *AbN'/AbP* represents the rate of falsely predicted interactions. *AbN'* was estimated using the mean number of interacting pairs obtained from the *K* data sets, **N**, i.e. the total number of interactions observed in **N'**, divided by *K*. A rate of falsely predicted interactions, *P*, close to 1 corresponds to random prediction. In contrast, low *P* values indicate strong evidence to support the validity of the positive interactions detected by the GO-driven similarity method. *P* values were calculated for the data sets described above using different *GOS-Th* values. This analysis allows one to have a better idea about how many false positive predictions may potentially be made when applying the GO-driven similarity method as a single prediction model. The analysis tasks described above were carried out with $K = 10$.

4. Results

4.1. Results from CoReg

Differences between the sets **P** (positives) and **N'** (negatives) with regard to their mean similarity values from the simple and highest average methods respectively were summarized for. Unknown, *IEA* and *ISS* annotations were excluded. Interacting pairs of genes generally exhibit higher similarity values than non-interacting pairs using both methods. Significant differences ($p < 0.001$) for all GO hierarchies were obtained. This suggests the feasibility of applying GO-driven similarity to support the distinction of co-regulated from non-co-regulated pairs of genes. Figure 1 shows the estimated probabilities, *P*, that such predictions are false as a function of the predictive threshold, *GOS-Th.*

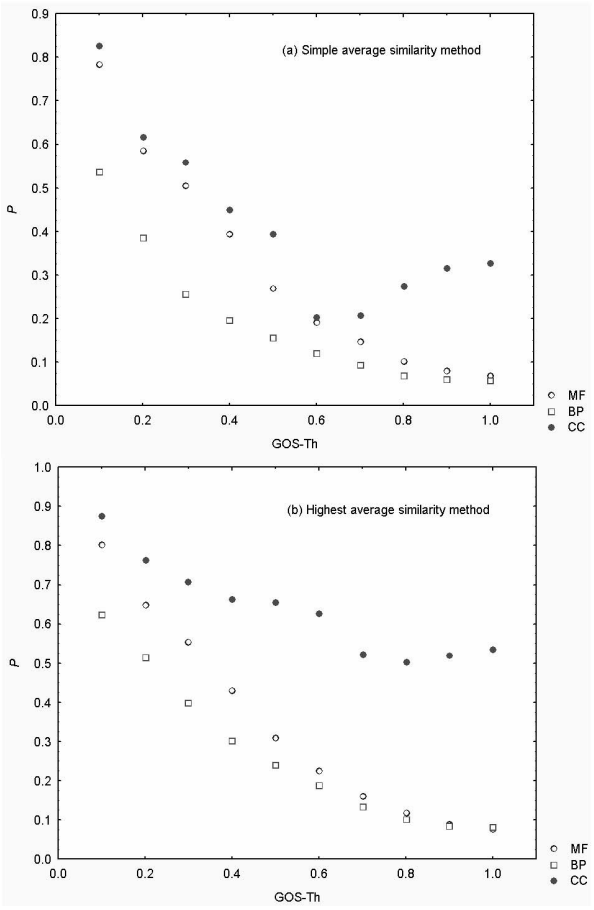

Figure 1 CoReg: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

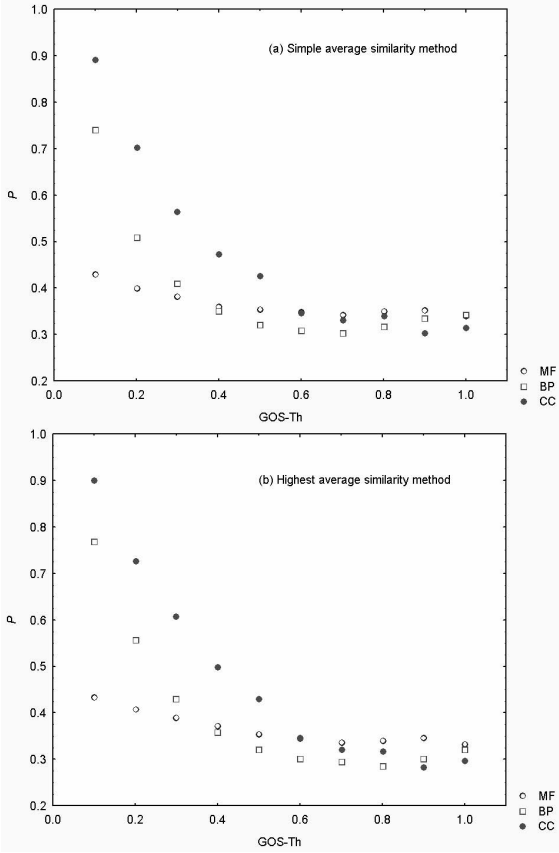

Figure 2 FunNet: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

4.2. Results from FunNet

Significant differences ($p < 0.001$) for all GO hierarchies were obtained. With both methods, pairs of interacting genes tend to exhibit higher similarity values than pairs of non-interacting genes. This suggests the feasibility of using GO-driven similarity to help to distinguish interacting from non-interacting genes (including physical and non-physical interactions). Figure 2 shows the estimated probabilities, *P*, that such predictions are false as a function of *GOS-Th*

4.3. Results from PPInt

Significant differences ($p < 0.05$) were observed only in connection to the BP hierarchy. Figure 3 presents the estimated probabilities, *P*, that such predictions are false as a function of *GOS-Th*. Interpretations should also take into account the very low number of gene pairs with CC annotations.

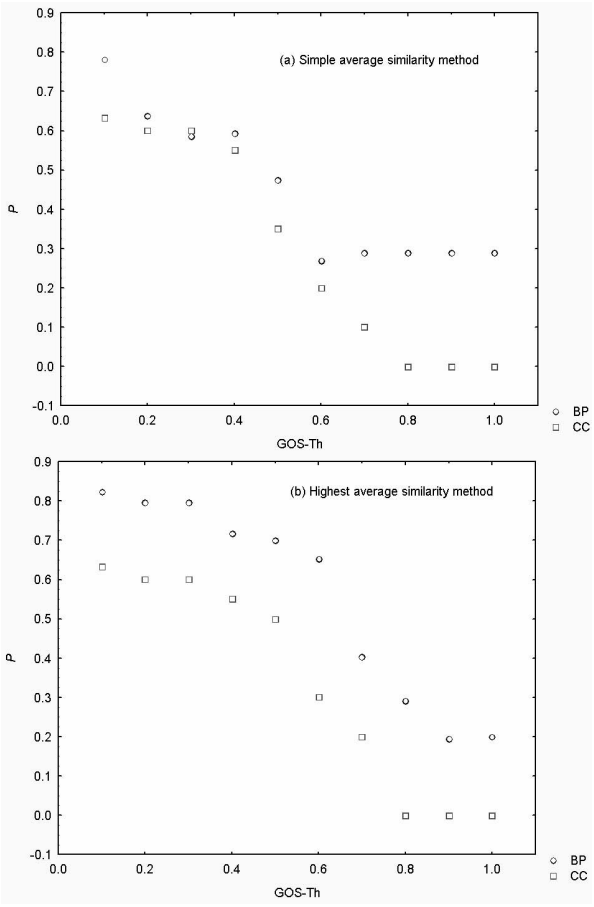

Figure 3. PP-Int: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

5. Discussion and Conclusions

This study demonstrated significant relationships between functional similarity and known interactions. This pattern was remarkably observed under all hierarchies for CoReg and FunNet. GO-driven similarity of pair of genes may be applied to support the prediction of functional interactions (including co-regulatory and PP interactions) in yeast. We also performed a manual verification to assess the potential biological significance of some of the "false positive" (novel) links. This procedure reported nine pairs of proteins (with unknown interactions), which are feasible candidates to be interacting partners in *C. elegans*, such as F28D1.2 and B0547.1, which are involved in DNA repair and ubiquitination, respectively.

Our research does not of course suggest that this approach is sufficient or even necessary to detect relevant interactions. It motivates the application of this functional similarity measure as a *complementary predictive resource*. This, in combination with other sources, such as gene co-expression, may support more accurate and biologically-meaningful predictions.

P.H Lee and D. Lee [11] recently integrated ontology-driven similarity information as part of their *modularized network learning method* (MONET). They first identified modules of interrelated genes using gene expression correlation and MIPS (Munich Information center for Protein Sequences database) annotations. *Bayesian networks* were then inferred from the detected modules that successfully predicted relevant gene regulation networks in yeast. Ontology-driven similarity was used to aid in the identification of clusters of genes on the basis of their MIPS annotations. Between-gene similarity was estimated using the between-term Resnik's method [12]. We showed that these relationships go beyond the regulatory level and can support applications involving uni- and multi-cellular organisms. Previous research has shown that Lin's technique may outperform Renisk's and other information-theoretic approaches [6], [12].

The results suggest that in general the larger the *GOS-Th*, the lower the probability of making false positive predictions. But it also highlights the fact that many of the false positive interaction predictions might show relatively high similarities. This may be explained by the difficulties in creating exact true negative data sets. Nevertheless, the results strongly suggest that there is a tendency to reduce the number of false positive interactions by applying more rigorous thresholds.

Alternative assessments may incorporate other estimators of *P*, including less conservative methods. We are applying the GOS assessment approach to support the prediction of integrated, large scale functional networks [13].

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). This paper was revised while FA was a visiting scholar at the Lister Hill National Center for Biomedical Communications, NLM, NIH.

REFERENCES

[1] R. Jansen, H. Yu, D. Greenbaum, and *et al.* "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol.302, 2003, pp.449-453.

[2] I. Lee, S. V. Date, A. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol.306, 2004, pp.1555-1558.

[3] The Gene Ontology Consortium, "Creating the gene ontology resource: Design and implementation," *Genome Research*," vol.11, 2001, pp.1425-1433.

[4] P. Khatri and S. Drăghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol.21, 2005, pp.3587-3595.

[5] P. Lord, R. Stevens, A. Brass, and C. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, 2003, pp.1275--1283.

[6] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, "Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships," In *Proc. of IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, La Jolla, CA, USA, 2004, pp.25-31.

[7] N. Simonis, S. J. Wodak, G.N. Cohen, and J. van Helden, "Combining pattern discovery and discriminant analysis to predict gene co-regulation," *Bioinformatics*, vol.15, 2004, pp.2370-2379.

[8] S. Li, C. M. Armstrong, N. Bertin, and *et al.* "A map of the interactome network of the metazoan *C. elegans*," *Science*, vol.303, 2004, pp.540-543.

[9] F. Azuaje, H. Wang, and O. Bodenreider, "Ontology-driven similarity approaches to supporting gene functional assessment," In *Proc. Of The Eighth Annual Bio-Ontologies Meeting*, Michigan, 25 June, http://bio-ontologies.man.ac.uk/.

[10] D. R. Bickel, "Probabilities of spurious connections in gene networks: application to expression time series," *Bioinformatics*, vol. 21, 2005, pp.1121-1128.

[11] P. H. Lee and D. Lee, "Modularized learning of genetic interaction networks from biological annotations and mRNA expression data," *Bioinformatics*, vol. 21, 2005, pp. 2739-2747.

[12] D. Lin, "An information-theoretic definition of similarity," in *Proc. of 15th International Conference on Machine Learning*, San Francisco, 1998, pp.296-304.

[13] F. Browne, H. Wang, H. Zheng, F. Azuaje, "An assessment of machine and statistical learning approaches to inferring networks of protein-protein interactions", *Journal of Integrative Bioinformatics,* 2006, vol. 3 (2).