Parzen Windows Based Protein Function Prediction Using Protein-Protein Interaction Data

A.M. Koura †, A. H. Kamal †, and I. F. Abdul-Rahman †

†Faculty of Computers and Information, Cairo University

Summary
Determining protein function on a proteomic scale is a major challenge in the post-genomic era. Right now only less than half of the actual functional annotations are available for a typical proteome. The recent high-throughput bio-techniques have provided us large-scale protein– protein interaction (PPI) data, and many studies have shown that function prediction from PPI data is a promising way as proteins are likely to collaborate for a common purpose. However, the protein interaction data is very noisy, which makes the task very challenging.
In this paper, a Parzen Window classifier is proposed to predict protein functions using IntAct protein interaction dataset. We present a probabilistic framework for predicting functions of unknown proteins based on incorporating Parzen Windows in the Bayesian formula. We use the leave-one-out cross validation to compare the performance. The experimental results demonstrate that our algorithm performs better than other competing methods in terms of prediction accuracy

Keywords:
Parzen Windows, protein function, protein-protein interactions, Bayesian classifier

1. Introduction

Since the completion of sequencing the human genome [1], discovering the underlying principles of interactions and the functional roles of proteins has been in the spotlight in the post-genomic era. The functional characterization of newly determined proteins has become one of the most crucial challenges. The classical way to predict protein functions is to find homologies between a non-annotated protein and other proteins using sequence similarity algorithms, such as FASTA [2] and PSI-BLAST [3]. The function of the non-annotated protein can then be assigned according to the annotated proteins with similar sequences. In addition, several computational approaches are proposed based on correlated evolution mechanisms of genes. For example, the domain fusion analysis infers that a pair of proteins interacts with each other and thus performs related functions [4]. In recent years, the data generated by high-throughput techniques have facilitated the functional classification. For example, microarrays monitor the expression levels of thousands of genes, and the correlated expression profiles of the genes can be interpreted as their functional relatedness [5]. Protein-protein interaction data, enriched by high-throughput experiments including yeast two-hybrid systems [6] and mass spectrometry [7], have provided the important clues of functional associations between proteins. The integrated protein interaction networks have been built from the heterogeneous interaction data sources. Accordingly, numerous computational methods have been supplemented for uncovering the functional information of uncharacterized proteins in the networks.

There are several approaches proposed to predict protein functions with protein interaction networks. The neighbor counting method [8] uses the majority-rule to label a protein with the functions that occur most frequently in its interaction partners. Some caveats of this approach are that it can only predict up to three functions and it doesn't take into account any significance value and the full topology of the network. To solve the above problem, Hishigaki et al. [9] use a chi-square statistics to calculate the significance of the functions of neighbor proteins. In detail, they examine the n-neighborhood of a protein. For a protein p, each function f is assigned a score. Those functions with higher score than a threshold will be kept as predicted functions for protein p. A shortcoming of this approach is that within the n-neighborhood, proteins at different distances from p are treated in the same way. Chua et al. [10] try to tackle the problem by investigating the relation between network distance and functional similarity. They focus on the 1- and 2-neighbourhoods of a protein, and devise a functional similarity score that gives different weights to proteins according to their distances from the target protein. In addition, these methods can only predict the proteins which have at least one interaction partner. This means lots of unknown proteins cannot be predicted by these methods. Moreover, the predicted annotations for an unknown protein are limited by the annotations of its interacting partners.

To avoid those limitations, several other approaches are proposed to use the global topology of protein interaction networks. Vazquez et al. [11] assign a function f to each non-annotated protein p so as to maximize the number of edges that connect proteins assigned with the same function. This optimization problem, which generalizes the computationally hard problem of minimum multi-way cut, is heuristically solved using simulated annealing. Karaoz

Manuscript received July 5, 2010
Manuscript revised July 20, 2010

et al. [12] use a similar approach but handle one function at a time. They apply a local search procedure in which for every vertex in turn (until convergence), the state of the vertex is changed according to the majority of the states of its neighbors. This procedure guarantees a solution with value at least half of the optimum. Nabieva et al. [13] apply the concept of functional flow which is propagated from an annotated protein to non-annotated proteins. After simulating the spread over time of this functional flow through the network, each non-annotated protein is assigned a score for having the function based on the amount of flow it received during the simulation. Relying on a Markovian assumption that the function of a protein is independent of all other proteins given the functions of its immediate neighbors, Deng et al. [14] adopt the Markov random field (MRF) model to simulate the protein interaction network with functional annotations, which fit the network and got good result. Letovsky and Kasif [15] also use an MRF model but with an assumption that the number of neighbors of a protein that are annotated with a given term is binomially distributed, where that distribution's parameter depends on whether the protein has that function or not. Lee et al. [16] develop a kernel logistic regression (KLR) method, which uses diffusion kernels and incorporated all indirect neighbors in the networks. While these approaches demonstrated that using machine learning and statistical methods can improve prediction performance, they bank on the same functional concept that the interaction partners of a protein are likely to share similar functions with it [10].

In our previous study [18] we used a method, which is based on Gaussian Mixture Model to predict protein function from protein-protein interaction data. In the this method a global information are taken into account by representing a protein using all the functional annotations of all proteins assigned with that term and have a shortest path with target protein in the all protein interaction network.

The current work attempts to provide a more robust probabilistic solution utilizing the fact that the form of the PDF of feature vectors is unknown. We will estimate the distributions with a method known as Parzen Windows. The proposed method uses global information on the whole network. For each function we used a Bayesian approach to compute the posterior probability that the protein posses this function.

The remainder of the paper is organized as follows. In Section 2, we present feature selection stage. In Section 3, we present our Parzen window - based prediction model. Extensive experimental results and comparison with other methods are reported in Section 4. Discussion of the proposed work is introduced in section 5. The paper is concluded in Sections 6.

2. Feature Selection

Typically one protein can have multiple functions, so we transfer function prediction problem into a typical multi label problem with functions as labels and proteins as instances or items. Recently, the issue of learning from multi-label data has attracted significant attention from a lot of researchers in the area of machine learning and pattern recognition.

We construct the protein-protein physical interaction network using the protein interaction dataset IntAct[19]. In this method, a network is represented by a undirected graph $G = (V, E, F)$, i.e. vertex set including each protein as a vertex $V = \{p_1, p_2, \ldots, p_n\}$, and the edge set $E = \{(p_i, p_j)\}$ there is an interaction between protein p_i and p_j, F is a finite alphabet of (annotation) terms (from a function vocabulary, e.g., Gene Ontology(GO) www. geneontology.org).

The problem we then want to solve is to derive the marginal probability of a given protein taking a particular functional label given all the putative functional assignments to the other proteins in the graph. It is based on the Bayesian formula and using Parzen window to estimate the likelihood rate.

2.1 Feature Extraction

We computed a shortest-path vector for each protein using Dijkstra's algorithm from protein interaction network. Each node v is then identified by an n-dimensional feature vector where n is the number of terms. The ith component of the vector is a function of the lengths of the shortest paths in the graph between v and all nodes labeled with the ith term. Let $I_A(t)$ denote the indicator function of a set A that determines whether t belongs to A. i.e.

$$I_A(t) = \begin{cases} 1 & t \in A \\ 0 & otherwise \end{cases} \qquad (1)$$

Let T_p and T_q denote the set of terms assigned to proteins p and q respectively. In this research, we adopt a form of feature vector driven from the global information of the underlying network. The form exploits the observation that the degree of similarity in a certain function between any two proteins in the network depends on the distance between them in the network. The feature vector of a protein p is described as:

$$X_p = \{x_{p_1}, x_{p_2}, \ldots, x_{p_m}\}, \qquad m = |F|$$

With

$$x_{p_t} = \sum_{q \in v} \exp[-d_{min}(p, q)] I_{T_q}(t) \qquad (2)$$

where $d_{min}(p, q)$ is the shortest path length between protein p and q.Note that the contribution of a protein q

to feature element x_{p_t} increases with the decrease in the length of the shortest path between q and p provided that q is annotated with t on the other hand, q has no effect on p if it is not annotated with t. This emphasizes the usefulness of using the above equation.

2.2 Feature Reduction

In biological data, feature vector is large, so feature reduction is essential. We applied the principle component analysis PCA for the purpose of dimensionality reduction [20].

A principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are: A) data reduction and B) interpretation.

Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$. Geometrically these linear combinations represent the selection of a new coordinate system obtained by rotating the original system, with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.

As we shall see, principal components depend solely on the covariance matrix Σ (or the correlation matrix p) of $X_1, X_2, \ldots, X_p$.

Let the random vector $X^T = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

Consider the linear combinations

$$Y_1 = a_1^T X = a_{11} X_1 + a_{12} X_2 + \cdots + a_{1p} X_p$$
$$Y_1 = a_2^T X = a_{21} X_1 + a_{22} X_2 + \cdots + a_{2p} X_p$$
$$Y_1 = a_p^T X = a_{p1} X_1 + a_{p2} X_2 + \cdots + a_{pp} X_p \quad (3)$$

Then, we obtain

$$var(Y_i) = a_i^T \sum a_i \quad i = 1,2,\ldots,p \quad (4)$$
$$cov(Y_i, Y_k) = a_i^T \sum a_k \quad i,k = 1,2,\ldots,p \quad (5)$$

The principal components are those uncorrelated linear combinations $Y_1, Y_2, \ldots, X_p$, whose variances in (4) are as large as possible.

The first principal component is the linear combination with maximum variance. That is, it maximizes $var(Y_1) = a_1^T \sum a_1$. It is clear that $var(Y_1) = a_1^T \sum a_1$ can be increased by multiplying any a_j by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length. We therefore define

First principal component = linear combination $a_1^T X$ that maximizes $var(a_1^T X)$ subject to $a_1^T a_1 = 1$

3 The Proposed Method

Our approach predicts multiple functions (terms) for each protein, which is functionally uncharacterized.

First, we define a scoring function $f_t(p)$ for every term $t \in \mathbb{F}$. Terms are then sorted in descending order according to $f_t(p)$. The topmost terms are supposed to have high chance of being considered associating p. We define the score function as being the ratio between the posterior probability that $t \in T_p$ given the feature vector of the protein p and posterior probability that $t \notin T_p$ given the feature vector of the protein p. This is mathematically described as follows

$$f_t(p) = log\left(P(t \in T_p | X_p)\right) - log\left(P(t \notin T_p | X_p)\right) \quad (6)$$

We adopted a Bayesian approach to estimate $p(t \in T_p | X_p)$ and $P(t \notin T_p | X_p)$ and utilize the whole structure information of the network for this purpose as follows:

$$P(t \in T_p | X_p) = \frac{p(X_p | t \in T_p) p(t \in T_p)}{p(X_p)} \quad (7)$$

$$P(t \notin T_p | X_p) = \frac{P(X_p | t \notin T_p) P(t \notin T_p)}{P(X_p)} \quad (8)$$

Notice that it is the product of the likelihood and the prior probability that is most important in determining the posterior probability; the evidence factor, $P(X_p)$, can be viewed as merely a scale factor that guarantees that the posterior probabilities sum to one, as all good probabilities must.

The prior probability $P(t \in T_p)$ could be estimated using a given protein interaction network as:

$$P(t \in T_p) = \frac{n_t}{n} \quad (9)$$

Here n_t is the number of proteins that t annotates and n is the number of all proteins in a given protein interaction network. The prior probability $P(t \notin T_p)$, is estimated as:

$$P(t \notin T_p) = 1 - P(t \in T_p) \quad (10)$$

We propose Parzen Windows Model (PW) for the likelihood probabilities $P(t \in T_p | X_p)$ and $P(t \notin T_p | X_p)$. We randomly select a set of i.i.d. samples of features of proteins annotated with term t as a training data for PW model $P(t \in T_p | X_p)$ and another set not annotated with term t as training data for PW model to build models for the term $P(t \notin T_p | X_p)$.

3.1 Parzen Windows Classifier

The Parzen window classifier is a kind of suboptimal Bayes classifier. It classifies an input vector $x \in R^d$ according to the Bayes decision rule. When an input vector x is given, x will be classified to class $w_i (1 \le i \le C)$ by comparing

$$f(x; w_i) = p(x|w_i)p(w_i) \quad (11)$$

where $p(wi)$ and $p(x|wi)$ are the a priori probability and the likelihood of class w_i , respectively. The classification result w_x is obtained from the following decision rule:

$$w_x = arg \max_{1 \le i \le c} f(x; w_i), \quad (12)$$

The $p(x|w_i)$ is estimated by the Parzen window method. When the training pattern set $X = \{x_1, \ldots, x_{n_i}\}$ of class w_i is given, its functional form is represented by the mean of the kernels centered at each training vector.

$$P(x|w_i) = \frac{1}{n_i h^d V_k} \sum_{j=1}^{n_i} K\left(\frac{x - x_j}{h}\right) \quad (13)$$

where $K(x)$ is the Parzen window or kernel function, and h is the window-width or smoothing parameter. To ensure $\int P(x|w_i)\, dx = 1$, the kernel function needs to be non-negative, and $V_k = \int K(x)\, dx$ should be finite. Determining h is important issue in this method because it directly affects the quality of estimation.

If we set $p(w_i)$ as the ratio of the number of samples in class w_i to the whole training set,

$$p(w_i) = \frac{n_i}{N} \quad (14)$$

where N is the total number of samples in the training set. By substituting (13), (14), the decision rule (12) becomes[21]

$$w_x = \arg \max_{1 \le i \le c} \left[\sum_{j=1, x_j \in w_i}^{n_i} K\left(\frac{x - x_j}{h}\right)\right] \quad (15)$$

4. Experimental Results

To build a protein interaction network for our experiments, we have used organism Yeast, Fly, and Human specific interaction datasets from IntAct dataset. Figure 1 show the dataset details

IntAct/Organism	#proteins	#interactions
YEAST	4729	35275
FLY	6666	19565
HUMAN	5074	15537

Fig. 1. IntAct Dataset details

To evaluate the effectiveness of our method, we used the function annotations in the Gene ontology(GO).A GO definition file was obtained from the Gene Ontology consortium web site [24]. It includes 19094 GO terms, including 9856 biological process terms, 7559 molecular function terms, and 1679 cellular component terms. Among the 4729 proteins in Yeast; 3610 are annotated with 1084 biological process terms, 3610 are annotated with 1468 molecular function terms, and 4292 are annotated with 610 cellular component terms. Also Among the 6666 proteins in Fly; 4125 are annotated with 1090 biological process terms, 6069 are annotated with 1139 molecular function terms, and 1805 are annotated with 478cellular component terms. In Human organism, there are 5074 proteins, 4574 annotated with 1659 molecular function, 4500 proteins annotated with 1125, and 3252 proteins annotated with 626 terms.

Since it is hard to evaluate the prediction performance directly on the non annotated proteins, we adopted the leave one-out cross-validation method to estimate the performance. That is, for each protein P in a given set of annotated proteins PA, we assumed the functions of P were unknown and used $PA - \{P\}$ to predict the functions of P. We then compared the predicted functions with the true annotation. Since we make experiments on already annotated proteins, we can measure the precision and recall values of the annotation predictions. Let R be the set of (known) annotations of protein P and Q be the set of annotation predictions. Then, we define precision and recall as:

$$Precision(Q, R) = |Q \cap R| / |Q| \text{ and } \quad (16)$$

$$Recall = |Q \cap R| / |R| \quad (17)$$

To achieve high accuracy in a prediction, the technique should have high precision and recall values. Usually there is a tradeoff between having high precision and high recall. Thus, to evaluate predictions of different techniques, we use the F-value of the prediction instead of its precision and recall. F-value is defined [22] as the harmonic mean of precision and recall of a prediction set:

$$F - value(Q, R) = \frac{2 * precision(Q, R) * Recall(Q, R)}{precision(Q, R) + Recall(Q, R)} \quad (18)$$

After running our technique on a dataset, we obtain scores for all GO terms (or other annotation types). We can then obtain a prediction set by either picking the GO terms with scores above a given threshold or picking top k GO terms (with top scores). We use the following method for selecting the value of k for top k cutoff in an experiment:

For each protein, we find the k value that produces the maximum F-value for the top- k predictions of the protein.

We name this value as ‘‘Maximum F-value with Local Cutoff’’ (MLC). Then, we average all the MLCs (i.e.,avgMLC) corresponding to all proteins in order to indicate the accuracy of a technique.
In this experiment, we compare protein annotation prediction performances of three techniques, namely, correlation mining (CM)[23], neighbor counting (NC)[8], and our technique Parzen Windows (PW). For each technique, we compute avgMLCs over all proteins . In tables 2,3 we list the avgMLC values of NC, CM, and PW techniques on three methods employing the molecular functionality GO annotations of proteins. Table 2 shows that the PW technique produces better avgMLC values than CM and NC techniques respectively over all proteins in Yeast . In table 2 our method PW performs better than NC and CM, in precision. On the other hand recall values are very close to each other, best prediction accuracy is obtained by the CM method. Table 3 display the avgMLC in Fly Organism for three techniques. Our method (PW) outperform high percentage than NC and CM respectively. Our approach also has higher precision than NC and CM.In the FLY dataset, although Recall values are very close to each other, best prediction accuracy is obtained by the CM method.

Table 2 Comparison of techniques by avgMLCs over all proteins in Yeast organism

Technique	avgMLC	Prec.	Rec.
PW	84.6%	90.7%	86.6%
CM	68.7%	69.2%	87.4%
NC	54.4%	66.6%	85.5%

Table 3 Comparison of techniques by avgMLCs over all proteins in Fly organism

Technique	avgMLC	Prec.	Rec.
PW	86.2%	94.7%	88.6%
CM	78.7%	69.2%	92.4%
NC	68.4%	70.6%	91.5%

We computed the F-value for each k value in top-k prediction tests. To sum up the prediction results, for each individual protein, we picked the k value that produces the highest F-value for that protein. Therefore F-values of techniques represent the highest possible accuracy of the technique, rather than the accuracy specific to the value of k.Next, we test the accuracy of PW, NC, and CM techniques on FLY,YEAST and HUMAN datasets for Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) sub-ontologies of GO. Table 4 (a-c) displays F-values of this experiment. We find that all three techniques produce best results on the CC ontology. We explain this observation as follows. Physical protein interactions occur in the same cellular location; therefore protein interaction partners are usually annotated by the same CC annotations.

Table 4(a-c) Ontology comparison of F-values on (a) FLY, (b) YEAST, and (c) HUMAN.

(a)

FLY	BP	MF	CC
PW	79.1%	84.6%	68.2%
CM	50.2%	54.2%	58.6%
NC	41.5%	39.6%	46.2%

(b)

Yeast	BP	MF	CC
PW	82.6%	86.2%	84.1%
CM	67.6%	72.3%	81.6%
NC	80.7%	76.0%	82.1%

(c)

Human	BP	MF	CC
PW	80.2%	79.6%	84.0%
CM	60.1%	65.8%	76.1%
NC	44.2%	41.3%	64.7%

5 Discussion

We developed probabilistic model Parzen windows to predict protein functions. We estimated the posterior probability that the protein has the function of interest given all of the available information. The posterior probability indicates how confident we are about assigning the function to the protein. Our method is a global approach taking into consideration the entire interaction network and the functions of known proteins. We applied our approach to predict functions of yeast, Fly, and Human proteins based upon Gene Ontology (GO) classifications and upon the interaction networks based on IntAct dataset. We have been studied the precision, recall, and average Maximum F-value with Local Cutoff (MLC) which called avgMLC by the leave-one-out approach**.**
Our method treats each function independently and separately, generally, that fact that a protein has one function does not prevent it from having other functions. Therefore, our model determines each function for each protein without a bias.Comparing the avgMLCs, the PW

technique gives the best results, and produces better predictions than the CM and NC techniques, respectively. Since we obtained the highest avgMLC values with IntAct dataset. The precision/recall values in Table 2,3 are obtained by using the given k values and picking the top k GO terms with highest scores.

6 Conclusions

In this paper, we derived a model of probability density function that uses the Parzen-window approach combined with Bayesian formula to predict the functions of proteins in a protein-protein interaction networks. One of the attractive advantages of the proposed method was that it considers the effect global information on the protein function prediction. Experimental results showed that the proposed method is highly promising and outperforms other methods.

References

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome**.** *Nature* 2001; 409**:**860-921.

[2] Pearson WR, and Lipman, DJ, "Improved tools for biological sequence comparison." In *Proceedings of National Academy of Sciences USA 85*, pages 2444 - 2448, 1988.

[3] Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ, " Basic local alignment search tool." *J. Mol. Biol.*, 215:403–410, 1990

[4] Marcotte EM., Pellegrini M., Ng HL., Rice DW, Yeates TO., & Eisenberg D., "Detecting protein function and protein–protein interactions from genome sequences," *Science*, 285 (5428), 1999, 751–753.

[5] Eisen MB, Spellman PT, Brown PO, and Botstein D, "Clustering analysis and display of genome-wide expression patterns." *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.

[6] Parrish JR, and Gulyas KD, Finley RL., "Yeast two-hybrid contributions to interactome mapping**."** *Current Opinion in Biotechnology* 2006,17**:**387-393.

[7] Aebersold R, and Mann M.," Mass spectrometry-based proteomics**."** *Nature* 2003, 422**:**198-207.

[8] Schwikowski B, Uetz P, & Fields S, "A network of protein-protein interactions in yeast", *Nature Biotechnology*, 2000, 18 (12): 1257–1261.

[9] Hishigaki H, Nakai K, Ono T, Tanigami A, and Takagi T, "Assessment of prediction accuracy of protein function from protein–protein interaction data", *Yeast*, 18 (6), 2001, 523–531.

[10] Chua HN, Sung WK, and Wong L, "Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions", *Bioinformatics*, 22 (13), 2006,1623–1630.

[11] Vazquez A, Flammini A, Maritan A, and Vespignani A, "Global protein function prediction from protein–protein interaction networks", *Nature Biotechnology*, 21 (6), 2003, 697–700.

[12] Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, and Kasif S, "Whole-genome annotation by using evidence integration in functional-linkage networks", *Proceedings of the National Academy Sciences of the United States of America*, 101 (9), 2004, 2888–2893.

[13] Nabieva E, Jim K, Agarwal A, Chazelle B, and Singh M," Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps", *Bioinformatics*,21 (Suppl 1) 2005, i302–i310.

[14] Deng M, Zhang K, Mehta S., Chen T., and Sun F., "Prediction of protein function using protein–protein interaction data", *Journal of Computational Biology*, 10 (6), 2003, 947–960.

[15] Letovsky S. & Kasif S.," Predicting protein function from protein/protein interaction data: A probabilistic approach," *Bioinformatics*, 19 (Suppl 1) 2003, i197–i204.

[16] Lee H., Tu Z., Deng M., Sun F., and Chen T., "Diffusion kernel based logistic regression models for protein function prediction", *OMICS*, 10 (1), 2006, 40–55.

[17] Sprinzak E., Sattath S., and Margalit H., "How reliable are experimental protein–protein interaction data? "*,Journal of Molecular Biology*, 327 (5), 2003, 919–923.

[18] Koura A, Kamal A, and Abdul-Rahman I,"Prediction Protein Function using Gaussian Mixture Model in Protein-Protein Interaction Networks,"*IJCSNS April* 2010,114-119.

[19] Hermjakob H et al., "IntAct, an open source molecular interaction database," *Nucleic Acids Research*, 2004, Vol. 32.

[20] Hotelling H.," Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, 24:417-441,498-520,1933

[21] Kang K and Shibata T "A Parzen-Window Classifier Architecture for Massively-Integrated Nanoscale Resonant Devices." *Proceedings of the 10th International conference on ULtimate Integration of Silicon*, 217 – 220,2009

[22] Shaw W, M.,Jr et al. "Performance standards and evaluations in IR test collections:Vector-space and other retrieval models.," *Info. Proc. Manag.*, 33 (1), 15–36.

[23] Kirac M, Ozsoyoglu G, and Yang J "Annotating proteins by mining protein interaction networks." *Bioinformatics* 2006, 22:e260–e270.

[24] Gene Ontology Annotations Database, available at http://www.geneontology.org/GO.current.annotations.shtml