ppiPre - an R package for predicting protein-protein interactions

Yue Deng
School of Computer Science and Technology,
School of Software Engineering
Xidian University
Xi'an, China
anfdeng@163.com

Lin Gao*
School of Computer Science and Technology
Xidian University
Xi'an, China
lgao@mail.xidian.edu.cn

***Abstract*—Since the existing experimental techniques for identifying protein-protein interactions (PPIs) are expensive and time-consuming, and the results are incomplete and/or noisy, developing computational methods for effectively predicting PPIs is of great importance. Therefore, we develop the R package *ppiPre*, which predicts PPIs using heterogeneous data sources, including Gene Ontology (GO) annotations, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations and topological properties of the PPI network. *ppiPre* supports up to 20 species and provides useful functions for predicting PPIs and calculating semantic and topological similarities between proteins. *ppiPre* is open source and freely available from http://cran.r-project.org/web/packages/ppiPre).**

Keywords—protein-protein interactions; prediction; semantic similarity; network topology; R

I. INTRODUCTION

Protein-protein interactions (PPIs) are critical for most cellular processes. High-throughput methods such as Y2H [1][2] and TAP-MS [3] have produced enormous PPI data for several organisms [4] in recent years. However, data generated from these experiments are often erroneous. Thus, computational methods can be very useful for validating experimental data or for choosing potential targets for further small-scale experimental screening. Researchers have suggested that direct data on protein interactions can be combined with indirect data in a supervised learning framework such as support vector machine (SVM), random forest and other classifiers [5][6][7][8][9], and that integrating heterogeneous data sources can improve the result of PPIs prediction [10][11] .

Supervised learning aims at training a classifier using positive and negative examples (truly interacting and non-interacting protein pairs) to filter false positive interactions and to discover false negative interactions in the PPI data. Features used in the training process may be extracted from different kinds of biological evidences, including protein sequences [10][12], GO [13][14], co-expressed pairs [10], domain compositions [15], motif pairs and related mRNA expression [16]. These approaches use similar classification framework to integrate heterogeneous data sources, while they mainly differed in two issues: the set of features used for prediction, and the learning method employed.

Since biological similarities mentioned above don't work well for the poorly studied organisms or proteins, topological similarities based solely on PPI network structure should also be integrated into the prediction framework [17].

Several software tools have been developed for the prediction of PPIs[18][19][20][21][22][23]. These tools use different kinds of features including literature, protein sequences, interaction domain, functional annotation, gene expression, and genome context. Generally, these existing tools have two major disadvantages. First, the species supported are limited. Well studied model organisms such as yeast and human are often supported, while some organisms which are lack of research are not. Second, additional data are often required while using these tools, such as homologous interactions, protein sequence, expression profiles and protein domains.

In this paper we present an R package named *ppiPre* to predict PPIs from the PPI networks given by users and calculate similarity between two proteins. We chose R because it is open source and there already exist packages to handle biological data and graphs [24]. *ppiPre* uses a combination of data sources, including Gene Ontology annotations, KEGG pathway annotations and topological properties of the network. Twenty species are supported by the current version of *ppiPre*, and the package only need original PPI network as input.

II. METHODS

In *ppiPre*, three types of features are integrated, which are semantic similarities based on GO, similarity based on KEGG co-pathway membership and similarities based solely on PPI network topology.

A. Semantic similarities based on GO

Semantic similarities are useful to assess the functional relevance of proteins. The GO is one of the most widely used knowledge source in bioinformatics, and has become the *de facto* standard for the annotation of proteins. The GO annotates proteins with terms from three ontologies: Molecular function (MF), biological process (BP), and cellular component (CC). Ontologies are organized as directed acyclic graphs (DAGs). Proteins that interact in the cell are likely to be in similar locations or involved in similar

biological processes compared to proteins that do not interact. Thus the more semantically similar the gene function annotations are among the interacting proteins, more likely the interaction is reliable. Several metrics have been proposed to measure the semantic similarity between GO annotations, and have been verified in terms of the correlations with other biological evidences such as sequence similarity and protein structure [27][28][29][30]. These measures often involve the information content (IC) of GO aspects or the GO graph structure.

The IC-based similarity measures depend on the frequencies of two GO terms involved. The IC of a term can be quantified in terms of the probability of its occurrence and gives a measure of how specific and informative a term is. It is defined as follows:

$$IC(t) = -\log(p(t)) \tag{1}$$

where $p(t)$ is the number of proteins annotated to term t and its descendants divided by the total number of proteins annotated to GO. Two newly published IC-based semantic similarity measures, namely IntelliGO [30] and Topological Clustering Semantic Similarity (TCSS) [31], are integrated in *ppiPre*.

The IntelliGO similarity measure integrates complementary properties in a novel annotation vector space model representation of protein annotations with coefficients based on both IC and annotation origin through evidence codes which trace the procedure that was used to assign specific GO terms to given proteins [32]. The coefficients assigned to each GO term are composed of two measures. A weight $w(g, t)$ is assigned to the evidence code that qualifies the importance of the association between a GO term t and a protein g. The *Inverse Annotation Frequency* (*IAF*) measure is defined as the ratio between the total number of proteins and the number of proteins annotated by the term t. The coefficient α_t is defined as

$$\alpha_t = w(g, t) * IAF(t) \tag{2}$$

The IntelliGO semantic similarity measure between two proteins g and h represented by their vectors $\vec{g}$ and $\vec{h}$ is given by the following formula:

$$SIM_{IntelliGO}(g,h) = \frac{\vec{g} * \vec{h}}{\sqrt{\vec{g} * \vec{g}} * \sqrt{\vec{h} * \vec{h}}} \tag{3}$$

where

$\vec{g} = \sum_i \alpha_i * \vec{e}_i$ is the vectorial representation of the protein g.

$\vec{h} = \sum_j \beta_j * \vec{e}_j$ is the vectorial representation of the protein h.

$\alpha_i = w(g, t_i) * IAF(t_i)$ is the coefficient of term t_i for protein g.

$\beta_j = w(h, t_j) * IAF(t_j)$ is the coefficient of term t_j for protein h.

$\vec{g} * \vec{h} = \sum_{i,j} \alpha_i * \beta_j * \vec{e}_i * \vec{e}_j$ is the dot product between the two protein vectors.

$\vec{e}_i * \vec{e}_j = \frac{2 * Depth(LCA)}{MinSPL(t_i, t_j) + 2 * Depth(LCA)}$ is the dot product between $\vec{e}_i$ and $\vec{e}_j$. LCA is the lowest common ancestor of the two terms. MinSPL is the minimal shortest path length between the two terms passing through this LCA.

The TCSS algorithm considers unequal depth of biological knowledge representation in different branches of the GO DAG. The main idea of TCSS is to divide the GO DAG into sub-graphs defining similar concept and score a PPI higher if participating proteins belong to the same sub-graph.

In the first step, sub-graphs are defined based on a threshold on the IC of all terms. Terms below a previously defined cutoff of IC are selected as sub-graph roots. And two sub-graphs are merged to increase the dissimilarity between sub-graphs if their roots have similar IC values. GO terms often have multiple parents, which could result in overlapping sub-graphs. Sub-graph overlap is then removed in two ways. Edges are removed by transitive reduction of GO graph G, which results in the smallest graph $R(G)$ such that the transitive closure of G is same as the transitive closure of $R(G)$. Terms that still belong to more than one sub-graph after edge reduction are replicated in each sub-graph, as well as the descendants of the terms. Semantic similarity between two terms is calculated within a sub-graph instead of the complete GO DAG. After the first step, all sub-graphs are connected to construct a meta-graph.

The second step is normalized scoring. Semantic similarity is scored on the meta-graph to get more balanced semantic similarity scores compared to on the complete GO DAG.

The annotation information content (ICA) of term t is calculated based on the frequency of gene products annotated to t and its children and is defined as follows:

$$ICA(t) = -\ln\left(\frac{annot(t)}{\sum_{t \in O} annot(t)}\right) \tag{4}$$

$$annot(t) = \left| P_t \bigcup_{c \in N(t)} P_c \right| \tag{5}$$

where t is a term in the ontology O and P_t is the set of gene products annotated to t. $N(t)$ is the set of child terms of t.

For a term t_i^s belonging to the i^{th} sub-graph G_i^s, the sub-graph information content (ICS) of t_i is defined as follows:

$$ICS(t_i^s) = \frac{ICA(t_i^s)}{\max_{t_i^s \in G_i^s} ICA(t_i^s)} \tag{6}$$

For a term t_i^m in meta-graph G^m, the information content (ICM) is defined as follows:

$$ICM(t_i^m) = \frac{ICA(t_i^m)}{\max_{t_i^m \in G^m} ICA(t_i^m)} \tag{7}$$

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

The semantic similarity between proteins A and B is defined by the maximum approach:

$$\max_{s_i,t_j \in S,T} \begin{cases} ICM_{\max}\left(LCA\left(s_i,t_j\right)\right) & if \quad s_i \in G_i^s \quad and \quad t_j \in G_j^s \\ ICS_{\max}\left(LCA\left(s_i,t_j\right)\right) & if \quad s_i,t_j \in G_i^s \end{cases} \tag{8}$$

where S and T are the sets of GO terms annotated to proteins A and B respectively. $LCA(s_i,t_j)$ is the lowest common ancestor of the terms s_i and t_j.

Besides two IC-based measures, one well-known graph-based measure presented by Wang [33] is integrated in the prediction framework.

In Wang's measure, each edge in the GO DAG is given a weight according to its type ("is-a" or "part-of"). For a term t, a sub-DAG comprised of the term t and all its ancestors can be represented as $DAG_t = (t,T_t,E_t)$, where T_t is the ancestors of term t and E_t is the set of edges connecting to the terms in DAG_t. For a term n in DAG_t, the semantic contribution of n to t, $S_t(n)$, is the product of all the edge weights in the path which has the maximum product among all the paths from term n to t.

The semantic similarity between two terms i and j is calculated as follows:

$$Sim_{Wang}\left(i,j\right) = \frac{\sum_{t \in T_i \cap T_j} S_i\left(t\right) + S_j\left(t\right)}{SV\left(i\right) + SV\left(j\right)} \tag{9}$$

where $SV(x)$ is the total semantic contribution of the term x in DAG_x.

The semantic similarity between two proteins A and B is the maximum semantic similarity between any of the terms in GO term sets GO_A and GO_B that annotate A and B.

B. Similarity based on KEGG co-pathway membership

KEGG contains graphical representations of cellular processes. If two proteins have at least one shared KEGG pathway membership, the interaction between them is considered to be reliable. The similarity is defined in the form of Jaccard similarity [34]:

$$Sim_{KEGG}\left(x,y\right) = \frac{\left|P\left(x\right) \cap P\left(y\right)\right|}{\left|P\left(x\right) \cup P\left(y\right)\right|} \tag{10}$$

where $P(x)$ is the pathways that protein x is annotated to in KEGG.

C. Similarities based on network topology

For the prediction framework to work well on the proteins which are not well annotated in GO and/or KEGG, especially the proteins of poorly studied organisms, three similarity measures based solely on network structure are also integrated into the prediction framework of *ppiPre*.

The classical Jaccard similarity is defined as:

$$Sim_{Jac}\left(x,y\right) = \frac{\left|N\left(x\right) \cap N\left(y\right)\right|}{\left|N\left(x\right) \cup N\left(y\right)\right|} \tag{11}$$

where $N(x)$ denotes the set of direct neighbors of node x.

Adamic-Adar similarity [35] assigns the less connected neighbors more weights, and is defined as:

$$Sim_{AA}\left(x,y\right) = \sum_{z \in N(x) \cap N(y)} \frac{1}{\log k_z} \tag{12}$$

where k_z is the degree of node z.

Resource Allocation similarity [36] is motivated by the resource allocation dynamics on complex networks [37]. The common neighbors of two nodes in a network play the role of transmitters, which will equally distribute a unit of resource to all its neighbors. The similarity between node x and y can be defined as the amount of resource y received from x, which is

$$Sim_{RA}\left(x,y\right) = \sum_{z \in N(x) \cap N(y)} \frac{1}{k_z} \tag{13}$$

D. Implementation and Usage

At present, *ppiPre* supports twenty species, which are Human, Yeast, Fly, Worm, Mouse, Arabidopsis, Rat, Zebrafish, Bovine, Canine, Anopheles, E.coli strain Sakai, Chicken, Chimp, Malaria, Rhesus, Pig, Streptomyces coelicolor, Xenopus and E.coli strain K-12. The IC used in *ppiPre* is species specific and calculated from corresponding Bioconductor annotation packages *org.Hs.eg.db*, *org.Sc.sgd.db*, *org.Dm.eg.db*, *org.Ce.eg.db*, *org.Mm.eg.db*, *org.At.tair.db*, *org.Rn.eg.db*, *org.Dr.eg.db*, *org.Bt.eg.db*, *org.Cf.eg.db*, *org.Ag.eg.db*, *org.EcSakai.eg.db*, *org.Gg.eg.db*, *org.Pt.eg.db*, *org.Pf.plasmo.db*, *org.Mmu.eg.db*, *org.Sco.eg.db*, *org.Ss.eg.db*, *org.Xl.eg.db* and *org.EcK12.eg.db*.

Annotation packages *GO.db* and *KEGG.db* are used to obtain the relations of GO terms and the number of shared pathway of two proteins. The *igraph* software package is used to calculate topological similarities.

Besides the features, the classifier is of great significant in a prediction framework. The *ppiPre* package chose the classical SVM [38] to combine heterogeneous features. The function *svm*() provided by the package *e1071* offers an interface to the LIBSVM library [39] and is used to train a SVM. SVM is chosen because it is able to handle small training set. Some other classifiers including random forest and bayesian classifier have also been tested during the development of *ppiPre*, but their performances were inferior to that using the SVM.

The prediction framework is shown in Figure 1. First, SVM is trained using the gold-standard PPI data sets (solid arrows). Then the trained classifier can predict PPIs from the PPI networks given by users (hollow arrows).

Functions for predicting PPIs and calculating similarities are provided within *ppiPre*.

The function *SVMPredict* reads the training set from an input file, computes the features for the training set, trains the SVM classifier, and predicts the false interactions from PPIs given by user.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

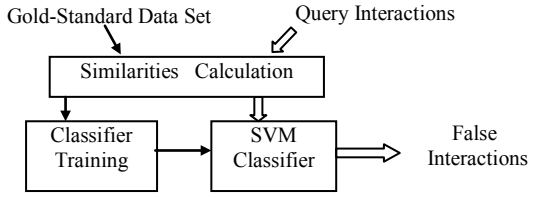

Figure 1. Graphical overview of the prediction framework.

For example:

>SVMPredict(trainingset, predictingset, organism= "human")

The training set is a comma separated values (CSV) file, each line of which is made up of three columns which are names of two proteins and a label. The label is either 1 or 0, indicating that the two proteins are interacting or not. The format of the predicting set is the same as training set. For yeast, ORF IDs from Saccharomyces Genome Database (SGD) are needed as the names of proteins, while Entrez Gene IDs are needed for other species. The result including potential false interactions will be written to a file.

The function *FNPre* predicts the false negative interactions according to three topological similarities as described before. User can predict new PPIs based on one or more topological similarities. A given threshold is the ratio of false negative interactions to positive interactions in the network, which controls the number of false negative interactions to be discovered. The result is also saved in a CSV file.

For example:

>FNPre(file="sample.csv",indicator=c("RA","AA"),threshold=0.1, output="FNPreResul.csv)

The indicator can be any combination of "RA", "AA", and "Jaccard", which indicates the similarities used.

The functions *KEGGSim*, *WangGeneSim*, *TCSSGeneSim and IntelliGOGeneSim* compute the corresponding semantic similarity between two proteins.

The function *GOKEGGSims* and *GOKEGGSimsFromFile* compute the semantic similarities between two proteins or protein pairs stored in a CSV file. The result consists of one KEGG-based and nine GO-based semantic similarities which are calculated by three methods on three GO ontologies.

The functions *JaccardSim*, *AASim*, *RASim* and *TopologicSims* compute the corresponding topological similarity or all of the three similarities between two proteins.

The function *ComputeAllEvidences* reads interactions from a file which contains interactions and compute both biological and topological features of each interaction.

Functional R scripts for all the functions are provided within the package.

III. Results and Discussion

Two commonly used yeast gold-standard data sets, the Munich Information Center for Protein Sequences (MIPS) data set [40] and the binary gold-standard data set [41], have been tested using *ppiPre*. Self-interactions are eliminated since the similarity measures are not appropriate in this case. Table 1 shows the detail of the gold-standard data sets. As negative examples we select random, non-interacting pairs from the interacting proteins, while maintaining the degree of each protein in the PPI network. The number of negative examples was taken as equal to the number of positive examples.

Table 1. Gold-standard positive yeast protein interaction data sets

Data set	No. of Interactions	No. of Proteins	Interaction Type
MIPS	8250	871	co-complex
Yu	1263	1078	binary

Although the similarity measures that depend on GO or KEGG cannot work well with proteins with unknown annotations, the effect on the two data sets above can be ignored because interactions which are lack of annotations account for only 0.2% (16 in MIPS data set and 2 in binary data set). However, when studying PPIs of poorly annotated species, the effect of lacking of annotations must be taken into account.

The performance of *ppiPre* is studied using 10-fold cross validation. Of the MIPS data set, over 98% of the true positive interactions can be classified correctly. Of the binary data set, since the network is very sparse, about 81% of the true positive interactions can be classified correctly. The result shows that *ppiPre* is capable of handling both large and small PPI data.

IV. Conclusions

In this paper, an R package *ppiPre* for predicting PPIs is introduced. The PPIs prediction problem is formalized as a binary classification problem, and seven similarities based on heterogeneous sources are integrated in the classification framework, including one similarity based on KEGG co-pathway membership, three similarities based on GO annotation and three similarities based solely on topology of PPI network. The package works well on predicting PPIs from both large and small PPI networks.

At present, *ppiPre* supports twenty species. In future work, we plan to integrate new effective features and improve efficiency of the algorithms.

Acknowledgment

We want to thank the helpful comments and suggestions from the anonymous reviewers. This work was supported by the NSFC [grant numbers 91130006, 60933009, 61072103, 61100157]; and the Fundamental Research Funds for the Central Universities.

References

[1] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae", Nature, vol. 403, no. 6770, pp. 623–627, 2000.

[2] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein

interactome", Proceedings of the National Academy of Sciences, vol. 98, no. 8, pp. 4569 –4574, 2001.

[3] A.-C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes", Nature, vol. 415, no. 6868, pp. 141–147, 2002.

[4] J. De Las Rivas and C. Fontanillo, "Protein–Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks", PLoS Comput Biol, vol. 6, no. 6, p. e1000807, Jun 2010.

[5] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions", Bioinformatics, vol. 21, no. suppl_1, pp. i38–46, Jun 2005.

[6] X.-W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework", Bioinformatics, vol. 21, no. 24, pp. 4394–4400, Dec 2005.

[7] A. Patil and H. Nakamura, "Filtering high-throughput protein-protein interaction data using a combination of genomic features", BMC Bioinformatics, vol. 6, no. 1, p. 100, 2005.

[8] X. Lin, M. Liu, and X. Chen, "Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms", BMC Bioinformatics, vol. 10, no. Suppl 4, p. S5, 2009.

[9] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph, "Random forest similarity for protein-protein interaction prediction from multiple sources", Pac Symp Biocomput, pp. 531–542, 2005.

[10] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph, "A mixture of feature experts approach for protein-protein interaction prediction", BMC Bioinformatics, vol. 8, no. Suppl 10, p. S6, 2007.

[11] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction", Proteins, vol. 63, no. 3, pp. 490–500, May 2006.

[12] C. Wang, J. Cheng, and S. Su, "Prediction of Interacting Protein Pairs from Sequence Using a Bayesian Method", The Protein Journal, vol. 28, no. 2, pp. 111–115, Feb 2009.

[13] M. Mahdavi and Y.-H. Lin, "False positive reduction in protein-protein interaction predictions using gene ontology annotations", BMC Bioinformatics, vol. 8, no. 1, p. 262, 2007.

[14] O. Kuchaiev, M. Rašajski, D. J. Higham, and N. Pržulj, "Geometric De-noising of Protein-Protein Interaction Networks", PLoS Comput Biol, vol. 5, no. 8, Aug 2009.

[15] T.-P. Nguyen and T.-B. Ho, "An integrative domain-based approach to predicting protein-protein interactions", J Bioinform Comput Biol, vol. 6, no. 6, pp. 1115–1132, Dec 2008.

[16] A. K. Ramani, Z. Li, G. T. Hart, M. W. Carlson, D. R. Boutz, and E. M. Marcotte, "A map of human protein interactions derived from co-expression of human mRNAs and their orthologs", Mol Syst Biol, vol. 4, Apr 2008.

[17] L. Lü and T. Zhou, "Link prediction in complex networks: A survey", Physica A: Statistical Mechanics and its Applications, vol. 390, no. 6, pp. 1150–1170, Mar 2011.

[18] S. Kim, S.-Y. Shin, I.-H. Lee, S.-J. Kim, R. Sriram, et al., "PIE: an online prediction system for protein-protein interactions from text", Nucleic Acids Research, vol. 36, no. Web Server, pp. W411–W415, May 2008.

[19] Y. Guo, M. Li, X. Pu, G. Li, X. Guang, et al., "PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment", BMC Research Notes, vol. 3, no. 1, p. 145, 2010.

[20] D. Li, W. Liu, Z. Liu, J. Wang, Q. Liu, et al., "PRINCESS, a Protein Interaction Confidence Evaluation System with Multiple Data Sources", Mol Cell Proteomics, vol. 7, no. 6, pp. 1043–1052, Jun 2008.

[21] M. Michaut, S. Kerrien, L. Montecchi-Palazzi, F. Chauvat, C. Cassier-Chauvat, et al, "InteroPORC: Automated Inference of Highly Conserved Protein Interaction Networks", Bioinformatics, vol. 24, no. 14, pp. 1625–1631, Jul 2008.

[22] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, et al., "PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs", BMC Bioinformatics, vol. 7, no. 1, p. 365, 2006.

[23] M. D. McDowall, M. S. Scott, and G. J. Barton, "PIPs: human protein-protein interaction prediction database", Nucleic Acids Research, vol. 37, no. Database, pp. D651–D656, Jan 2009.

[24] G. Csárdi and T. Nepusz, "The igraph software package for complex network research", InterJournal Complex Systems, vol. 1695, no. 1695, 2006.

[25] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et.al., "Gene Ontology: tool for the unification of biology", Nat Genet, vol. 25, no. 1, pp. 25–29, May 2000.

[26] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes", Nucleic Acids Research, vol. 28, no. 1, pp. 27 –30, 2000.

[27] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", in IJCAI, 1995, pp. 448–453.

[28] J. Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", in International Conference Research on Computational Linguistics (ROCLING X), 1997, p. 9008.

[29] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Semantic similarity measures as tools for exploring the gene ontology", Pac Symp Biocomput, pp. 601–612, 2003.

[30] S. Benabderrahmane, M. Smail-Tabbone, O. Poch, A. Napoli, and M.-D. Devignes, "IntelliGO: a new vector-based semantic similarity measure including annotation origin", BMC Bioinformatics, vol. 11, no. 1, p. 588, 2010.

[31] S. Jain and G. Bader, "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology", BMC Bioinformatics, vol. 11, no. 1, p. 562, 2010.

[32] M. F. Rogers and A. Ben-Hur, "The use of gene ontology evidence codes in preventing classifier assessment bias", Bioinformatics, vol. 25, no. 9, pp. 1173 –1177, 2009.

[33] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms", Bioinformatics, vol. 23, no. 10, pp. 1274–1281, May 2007.

[34] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", Bull. Soc. Vaud. Sci. Nat, vol. 37, p. 541, 1901.

[35] L. A. Adamic and E. Adar, "Friends and neighbors on the Web", Social Networks, vol. 25, no. 3, pp. 211–230, 2003.

[36] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information", The European Physical Journal B - Condensed Matter and Complex Systems, vol. 71, no. 4, pp. 623–630, Oct 2009.

[37] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, and B.-Q. Yin, "Power-law strength-degree correlation from resource-allocation dynamics on weighted networks", Phys. Rev. E, vol. 75, no. 2, p. 021102, 2007.

[38] V. N. Vapnik, The nature of statistical learning theory. Springer, 2000.

[39] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines", ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1–27:27, 2011.

[40] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, et al., "Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs", Genome Research, vol. 14, no. 6, pp. 1107 –1118, 2004.

[41] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, et al., "High-Quality Binary Protein Interaction Map of the Yeast Interactome Network", Science, vol. 322, no. 5898, pp. 104–110, Oct 2008.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE