A Semi-Supervised Learning Approach to Disease Gene Prediction

Thanh Phuong Nguyen and Tu Bao Ho
Japan Advanced Institute of Science and Technology
{phuong,bao}@jaist.ac.jp

Abstract

Discovering human disease-causing genes (disease genes in short) is one of the most challenging problems in bioinformatics and biomedicine, as most diseases are related in some way to our genes. Various methods have been proposed to exploit existing data sources for solving the problem. We aim to develop a novel method to predict disease genes that takes into account the imbalance between known disease genes and unknown disease genes. To this end, our method makes the best of semi-supervised learning, integrating data of human protein-protein interactions and various biological data extracted from multiple proteomic/genomic databases. Experimental evaluation shows high performance of our proposed method. Also, a considerable number of potential disease genes were discovered.

Supplementary materials are now available from http://www.jaist.ac.jp/∼s0560205/DiseaseGenes/.

1. Introduction

One of the ultimate goals of biological sciences, and certainly one with a high impact on society, is to improve our understanding of the processes and events related to diseases. The information contained in our genes is so critical that simple changes can lead to a severe inheritable disease, make us more inclined to develop a chronic disease, or make us more vulnerable to an infectious disease. Genes related to causing some diseases are called disease-causing genes or disease genes [11]. Intuitively, proteins corresponding to disease genes are disease proteins. Previous biological and medical methods in this area were expensive and laborious. There is a great need to develop computational methods to effectively discover disease genes, to support biologists and pharmacists in their work.

Many studies have tried to discover disease genes with various methods and data sources. Some work related to disease gene prediction was based on annotations [15], or based on sequences [1]. They often treated disease genes as separate and independent genes. However, it is well known that biological processes are not realized by the single molecule, but rather by the complex interactions of proteins, and the breakdown in protein interaction networks could result in diseases [13]. From the interactomics point of view, disease genes could then be investigated through the interaction networks of disease proteins.

Research on protein-protein interactions (PPI) and diseases has been rapidly increasing in recent years. Disease genes were discovered by topological features in human PPI networks using the k-nearest neighbor algorithm [17]. Using a phenomic ranking of protein complexes linked to human diseases, a Bayesian model was proposed to predict new candidates for disorders [8]. In [2], the authors integrated graph kernels for gene expression and human PPI to predict disease genes. Also, some work concentrated on using PPI to discover disease genes for specific diseases, i.e. Alzheimer disease, using heuristic score functions [4], [7].

These previous works tried to apply supervised learning based on known disease genes (labeled data) to predict new gene candidates causing diseases. However, nowadays the ratio of known disease genes to the total number of human genes is very small. It is shortcoming if biological information of genes closed to diseases genes is omitted. We propose a novel method for disease gene prediction that makes the best of semi-supervised learning, integrating data of human protein-protein interactions and various biological data extracted from multiple proteomic/genomic databases. The key idea is based on the assumption that disease genes have close biological associations with other genes whose proteins interact with respective proteins of disease genes.

We employ semi-supervised learning methods to determine the extended set of candidate proteins from human protein interaction networks, and to predict putative disease genes from the extended set. We extract various proteomic/gemomic features such as protein domains, GO terms, protein keywords, and coded enzymes of protein candidates, to comprehensively infer disease genes.

We carefully carried out various experiments with disease genes extracted from OMIM database – Online Mendelian Inheritance in Man database (version 2007) [5]. We did five experiments with different sizes of labeled data,

0-7695-3031-1/07 $25.00 © 2007 IEEE
DOI 10.1109/BIBM.2007.30

and twenty trials for each experiment to evaluate accuracy of the method. Accuracy of the prediction was 82%, which showed that the proposed method is useful for the disease gene prediction problem. About fifty potential disease proteins were predicted and some of them have been validated in the scientific literature.

2. Semi-Supervised Learning

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning. SSL considers both labeled data (supervised learning) and unlabeled data (unsupervised data). A given data set $\mathcal{X} = \{x_1, ..., x_l, x_{l+1}, ..., x_n\}$ can always be divided into two parts. The first one is the set of l data points $\mathcal{X}_l = \{x_1, ..., x_l\}$ which are labeled by the label set $\mathcal{Y}_l = \{y_1, ..., y_l\}$, and the other one is the data set of u data points $\mathcal{X}_u = \{x_{l+1}, ..., x_n\}$, the labels of which are not known. The goal is to predict labels of unlabeled data. Some often-used semi-supervised learning methods include EM with generative mixture models, self-training, co-training, transductive support vector machines, and graph-based methods [3].

Since labeling often requires much human labor, whereas unlabeled data is far easier to obtain, semi-supervised learning is very useful in many real-word problems, and has recently attracted an increasing number of researchers [3]. In bioinformatics, SSL is also applied to solve many problems and has achieved considerable results, for example, in the study of protein classification [16] and in the functional genomics [9], etc.

In this paper, we employed Harmonic Gaussian method [18] - the graph-based semi-supervised learning algorithm - in the proposed framework. Because we integrated human protein-protein interaction networks, semi-supervised learning based on the graph was considered to be suitable for predicting disease genes. The details of our proposed method are presented in Section 3.

3. Materials and Methods

In this section, we describe our method to predict disease genes using semi-supervised learning. Subsection 3.1 describes the semi-supervised learning framework for disease gene prediction. In Subsection 3.2, the score functions are presented, to estimate the biological significance of extracted features for disease gene prediction.

3.1. Semi-supervised Learning Framework for Predicting Disease genes

Figure 1 briefly describes our semi-supervised learning framework for disease gene prediction which uses integrated human PPI and proteomic/genomic data.

Corresponding to Figure 1, the proposed framework consists of four main tasks, as follows:

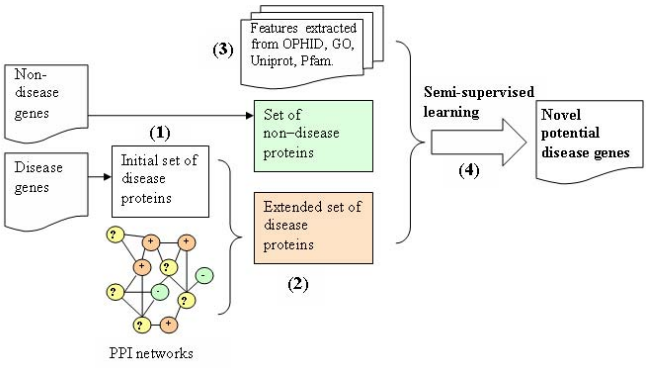

Figure 1. Semi-supervised learning framework for disease gene prediction.

1. Identify disease genes as positives, and non-disease genes as negatives, and map them to the corresponding proteins, called disease proteins and non-disease proteins, respectively.
2. Extend the initial set of positives by extracting their interacting proteins as positive candidates from a human PPI database.
3. Extract and represent human PPI and proteomic/genomic data as feature vectors.
4. Apply a semi-supervised learning algorithm to predict disease genes.

Algorithm 1 presents in detail the algorithm for disease gene prediction using semi-supervised learning. The input of the algorithm are positive examples (known disease genes), negative examples (non-disease genes), the set of human protein-protein interactions, and the set of proteomic/genomic feature data. The training data sets are described in Subsection 4.1. The output of the algorithm is the set of new disease genes.

In Algorithm 1, there are 12 steps corresponding to the four main tasks. Steps 1 to 3 are for the first task. Until Step 3, all disease proteins p_i and non-disease proteins p_i^- are identified by the Uniprot names, and we have the initial set of disease proteins $\mathcal{P}$. From the human PPI network Ω, Step 4 does the second task to generate the extended set of disease proteins $\mathcal{P}^+$ including the interacting proteins p_i^+ (positive candidates) of disease proteins in the initial set. In Step 5, the union set $\mathcal{P}^*$ is formed consisting of positives, positive candidates and negatives. For the third task, Steps 6 to 9 extract various features f^k from databases OPHID, Uniprot, GO, and Pfam, and estimate the scores $score_k$ of features for each protein. The *k*-dimension feature vectors v_i are determined to integrate all feature scores as the input of a semi-supervised learning algorithm. Steps 10 to 12 correspond to the fourth task, where we apply a semi-supervised learning algorithm to predict new disease genes.

In Step 10, we used the SemiL software developed by Huang *et al.* [6] which implements Harmonic Gaussian

Algorithm 1 Semi-supervised learning framework to predict disease genes.

Input:

The set $\mathcal{G}$ of disease genes g_i and the set $\mathcal{G}^-$ of non disease genes g_i^-.
The protein-protein interaction network Ω.
PPI feature and proteomic/genomic features f^k extracted from databases (OPHID, Uniprot, GO, and Pfam).

Output:

The set of new predicted disease genes $\mathcal{G}^+$

1: $\mathcal{G}^+ := \emptyset$;
$\mathcal{P} := \emptyset$; /* $\mathcal{P}$ is the initial set of disease proteins. */
$\mathcal{P}^+ := \emptyset$; /* $\mathcal{P}^+$ is the extended set of disease proteins. */
$\mathcal{P}^- := \emptyset$; /* $\mathcal{P}^-$ is the set of non-disease proteins. */
2: Map all disease genes $g_i \in \mathcal{G}$ to the corresponding proteins p_i; $\mathcal{P} = \mathcal{P} \cup \{p_i\}$.
3: Map all non-disease genes $g_i^- \in \mathcal{G}^-$ to the corresponding proteins p_i^-; $\mathcal{P}^- := \mathcal{P}^- \cup \{p_i^-\}$.
4: **for each** $p_i \in \mathcal{P}$
Extract all interacting proteins p_i^+ of the disease protein p_i from Ω; $\mathcal{P}^+ := \mathcal{P}^+ \cup \{p_i^+\}$.
5: $\mathcal{P}^* := \mathcal{P}^+ \cup \mathcal{P}^- \cup \mathcal{P}$. /* P^* is the union set of all considered proteins. */
6: **for each** $p_i \in \mathcal{P}^*$
7: Extract features f^{ppi}, f^{length}, f^{kw}, f^{ec}, f^{go}, f^{pfam} from databases OPHID, Uniprot, GO, and Pfam.
8: Estimate the score $score_k$ of each feature f^k.
9: Determine *k*-dimension feature vectors $v_i = \{v_i^k\}$ where the dimension v_i^k corresponds to the feature f^k.
10: Run a semi-supervised learning algorithm on the set $\mathcal{P}^*$ and the set of feature vectors v_i to predict disease proteins.
11: Map new predicted disease proteins to their corresponding disease genes g_i^+; $\mathcal{G}^+ := \mathcal{G}^+ \cup \{g_i^+\}$.
12: **return** $\mathcal{G}^+$.

method. SemiL software is an efficient software for solving large-scale semi-supervised learning problems. SemiL provides various options for distance weight, hard or soft label, normalization, etc.

3.2. Scores of proteomic/genomic features

To have a comprehensive view of the relationship between disease genes and other proteomic/proteomic features, in addition to PPI features from OPHID database, we also utilized three other databases, Uniprot, GO, and Pfam, to look for relevant and useful features for disease gene prediction. When integrating proteomic/genomic features, one difficulty that we have to overcome is representing various data types of feature data. Extracted data may be numerical such as sequence length, or categorical such as keywords, and coded enzymes. Thus, by using score functions we can represent the extracted data in the form of feature vectors.

Table 1 shows the statistics of extracted features for each data source. The first two columns are the number of records extracted according to respective features, and the last two columns are the number of feature categories.

We describe here how we define the score functions. First, the protein-protein interaction score was considered, to determine how close one protein is to other disease proteins in protein interaction networks. Human protein-protein interactions are extracted from OPHID database (http://ophid.utoronto.ca/ophid/). One protein may have many interactions, however, among these protein interactions, the more the protein interacts with disease proteins, the more likely it is to be a disease protein. Also, a protein which is a hub of many interactions is often very important. The score $score_{ppi}$ for the PPI feature f^{ppi} is defined as follows:

$$score_{ppi}(p_i) = \frac{\sum_{p_j \in \mathcal{P}} Int(p_i, p_j)}{\sum_{p_j \in \mathcal{P}^*} Int(p_i, p_j)} * \frac{\sum_{p_j \in \mathcal{P}} Int(p_i, p_j)}{Avg_{ppi}} \quad (1)$$

where

$$Int(p_i, p_j) = \begin{cases} 1 & \text{if there is an interaction between} \\ & \text{proteins } p_i \text{ and } p_j, \\ 0 & \text{otherwise.} \end{cases}$$

Avg_{ppi}: the average of number of protein interactions belonging to disease proteins.

UniProt database (http://www.pir.uniprot.org/) is the world's most comprehensive catalog of information on proteins, and it provides functional, structural or other categories, lengths of protein sequences, and describes enzymes coded.

The following equation is the score for the sequence length feature:

$$score_{length}(p_i) = \frac{length(p_i)}{Avg_{length}} \quad (2)$$

where

$length(p_i)$: the sequence length of a protein p_i.
Avg_{length}: the average sequence length of disease proteins.

Disease proteins may share the same keywords, and be coded by the same enzymes. Some keywords and coded enzymes were found to be common in the set of disease proteins. Then, $score_{kw}$ and $score_{ec}$ show how probable it

Table 1. Statistics for the set of all proteins considered, and the set of disease proteins with the extracted features.

Feature f^k	#Record		#Category	
	The whole data set	Set of disease proteins	The whole data set	Set of disease proteins
f^{GO}	17241	6404	2911	1817
f^{KW}	31465	13597	564	504
f^{EC}	1123	451	133	106
f^{Pfam}	6817	2426	1796	1413

is that a protein is a disease protein, in terms of keywords and enzymes. In Uniprot database, keywords are classified into 10 categories, i.e. biological process, developmental stage, disease, molecular function, etc.

Among 5,557 proteins (details in Section 4.1), there are 31,465 data records extracted for keyword features, and 1,123 enzymes. These proteins share the same 564 keywords and 133 enzymes.

We proposed similar scores for keyword feature f^{kw} and coded enzymes feature f^{ec}. For each protein, we extracted the corresponding keywords kw_i and coded enzymes ec_i. The keyword and enzyme data are categorical, for example, (P05067, alzheimer disease) and (P01011, disease mutation) where P05067, P01011 are the Uniprot names, and "alzheimer disease", "disease mutation" are their keywords; or (O75688, ec3.1.3) where O75688 is the Uniprot names, and ec3.1.3, is enzymes coded.

Since each protein may have many different keywords, each keyword kw_i is assigned to its significant weight, as follows:

$$w_i^{kw} = \overline{freq}(kw_i) * freq(kw_i),$$

where

$\overline{freq}(kw_i)$: the frequency count of kw_i observed in the set of disease proteins $\mathcal{P}$.

$freq(kw_i)$: the frequency count of kw_i observed in the set of proteins $\mathcal{P}^*$.

Equation 3 shows the score for the keyword feature:

$$score_{kw}(p_i) = \frac{1}{\sum\limits_{\forall kw_j \in p_i} w_j^{kw}} \tag{3}$$

Unlike the keyword feature, each protein p_i is coded by only one enzyme ec_i. The score for the coded enzyme feature is defined in Equation 4.

$$score_{ec}(p_i) = \overline{freq}(ec_i) * freq(ec_i) \tag{4}$$

where

$\overline{freq}(ec_i)$: the frequency count of ec_i observed in the set of disease proteins $\mathcal{P}$.

$freq(ec_i)$: the frequency count of ec_i observed in the set of proteins $\mathcal{P}^*$.

It is useful to investigate the relationship between GO terms (http://www.geneontology.org/) and disease proteins. GO terms are divided into three groups: molecular function, biological process and cellular component. GO terms related to the set of proteins considered go_i were extracted, and each of them has its own weight, defined by the following equation:

$$w_i^{go} = \frac{\#\overline{go}_i + 1}{\#go_i + 1},$$

where

$\#\overline{go}_i$: the count of go_i observed in the set of disease proteins $\mathcal{P}$.

$\#go_i$: the frequency counts of go_i observed in the set of proteins $\mathcal{P}^*$.

Then, the score for GO term feature is proposed as follows:

$$score_{go}(p_i) = \frac{1}{\sum\limits_{\forall kw_j \in p_i} w_i^{go}} \tag{5}$$

Protein domains are the building blocks of proteins. Disease proteins may structurally or functionally depend on their domains. Pfam database (http://www.sanger.ac.uk/Software/Pfam/) is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. Pfam domains d_j of all considered proteins are extracted and scored by Equation 6.

$$score_{pfam}(p_i) = \frac{\#\overline{pfam}_i + 1}{\#pfam_i + 1}, \tag{6}$$

where

$\#\overline{pfam}_i$: the number of domains d_j of a protein p_i observed in the set of domains belonging to disease proteins.

$\#pfam_i$: the number of domains d_j of the protein p_i.

4. Experiments

4.1. Experiment design

We first prepared three data sets to carry out the experiments: (i) the set of disease genes, (ii) the set of non-disease genes, (iii) the set of protein-protein interactions. Then, we carried out experiments with various parameters to computationally evaluate accuracy of the proposed method. Finally, we looked up newly-predicted disease genes in the scientific literature to biologically verify the findings of the proposed method.

The database OMIM is a catalog of human genes and genetic disorders. In OMIM, the list of hereditary disease

genes is described in the OMIM morbid map. There are 4,512 records with 3,053 unique OMIM ID in the catalog. As shown in Algorithm 1, the total of 3,053 human disease genes were mapped, to look for their disease proteins identified by Uniprot names. The results showed 3,590 corresponding disease proteins. Some of these proteins have published interactions.

Compiling a list of genes that are known not to be involved in hereditary disease is difficult. A recent study [14] showed that the human genome may contain thousands of essential genes having features that differ significantly, both from disease genes and from other genes. In the absence of a set of well-defined human essential genes, they compiled the list of ubiquitously expressed human genes (UEHG) as an approximation of essential genes. Non-disease genes belong to neither the OMIM morbid map nor the UEHG set. The genes that satisfy this condition are negative examples in our experiments. Mapping to Uniprot names, there are 723 proteins corresponding to UEHG, and 180 proteins overlapping in the set of disease proteins.

We obtained the human protein-protein interactions from OPHID database. Among 51,934 human protein-protein interactions stored in OPHID, there are 13,368 interactions which have at least one interacting partner belonging to the set of disease proteins. We found that there were 1,502 disease proteins having interactions in OPHID. From 13,368 interactions, the initial set of disease proteins extended to 5,775 proteins.

4.2. Experiment Results

As mentioned above, we used SemiL software to implement the Harmonic Gaussian method [19]. The weight matrixes W were calculated with two different distance functions, i.e. Euclidean distance and Cosine distance, and the degree of graph was 20. The kernel was RBF function, and other parameters were default.

From the data set, we randomly selected l data points as labeled data, and the rest *(n-l)* as unlabeled data. Then, accuracy was estimated by comparing the predicted labels and true labels. For each labeled set size l tested, we performed 20 trials. The final result is average accuracy of 20 trials. Accuracy is defined as the ratio of *(true positive/(true positive + false positive))*

We chose similar sets of disease genes, non-disease genes, and protein-protein interactions as those used in the of of Xu and Li [17], but our method provided higher accuracy with similar data. In [17], accuracy ranged from 74% to 76%. Our accuracy ranged from 78% to 82%. In future work, we would like to reproduce the same experiments as in [17] for a comparative evaluation.

Figure 2 shows accuracy of our method. When the size of labeled data is small (10% of the data set), semi-supervised learning obtained non trivial accuracy, 78%. When the number of labeled data is at least half of the total

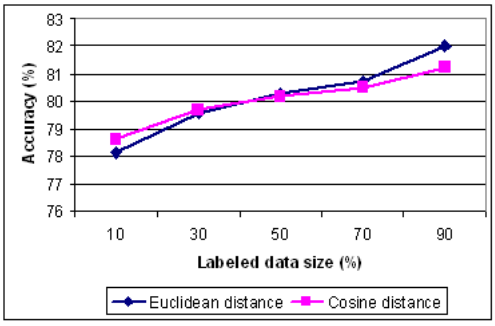

Figure 2. Accuracy of the proposed method with different sizes of labeled data for the Euclidean and Cosine distance.

data set, accuracy is over 80%. This demonstrates that even with a very low percent of labeled data, semi-supervised learning can still predict disease genes with high accuracy.

Table 2 shows some predicted disease proteins and their corresponding disease genes. The list of 50 newly-predicted disease proteins is available as supplementary materials at http://www.jaist.ac.jp/∼s0560205/DiseasePPI/.

Table 2. List of some potential disease proteins and corresponding disease genes.

Disease proteins in Uniprot names	Disease proteins in protein names	Disease genes
O14745	NHERF_HUMAN	SLC9A3R1
P08670	VIME_HUMAN	VIM
P25490	TYY1_HUMAN	YY1
P27348	1433T_HUMAN	YWHAQ
Q13363	CTBP1_HUMAN	CTBP1
Q13813	SPTA2_HUMAN	SPTAN1
O43157	PLXB1_HUMAN	PLXNB1
P02760	AMBP_HUMAN	AMBP
Q8WYH8	ING5_HUMAN	ING5
O43852	CALU_HUMAN	CALU

5. Discussion

In addition to computational evaluation, we endeavored to look for biological evidence to support to our method. And we found some interesting evidence when verifying the novel potential disease proteins. As in [14], ubiquitously expressed human genes, also known as house keeping genes, should be regarded as most severe ”disease” genes. Among 50 new predicted disease proteins, there are 6 proteins which correspond to UEHG genes, i.e., *nherf_human, ddx3x_human, tyy1_human, 1433t_human, ctbp1_human, spta2_human.*

Hepatitis C virus (HCV) core influences the expression of host genes [12]. *Ddx3x_human* (ATP-dependent RNA helicase DDX3X) acts as a cofactor for XPO1-mediated nuclear export of incompletely spliced HIV-1 Rev RNAs, and is also involved in HIV-1 replication. This protein interacts specifically with hepatitis C virus core protein, resulting in a change in intracellular location.

Protein *tyy1_human* acts as a repressor in absence of adenovirus E1A protein, but as an activator in its presence. A group of viruses that infect the membranes (tissue linings) of the respiratory tract, the eyes, the intestines, and the urinary tract, adenoviruses account for about 10% of acute respiratory infections in children, and are a frequent cause of diarrhea.

Protein *trrap_human* is the isolation of highly conserved 434 kDa protein, and interacts specifically with the c-Myc N terminus, and has homology to the ATM/PI3-kinase family. *Trrap_human* (related to gene *trrap*) also interacts specifically with the E2F-1 transactivation domain. Expression of transdominant mutants of the protein *trrap_human* or antisense RNA blocks c-Myc- and E1A-mediated oncogenic transformation. Then, *trrap* was suggested as an essential cofactor for both the c-Myc and E1A/E2F oncogenic transcription factor pathways [10].

Though accuracy is the most common evaluation measurement in the disease gene prediction problem, other measurements such as sensitivity and specificity or the area under the ROC curve, should also be used for evaluation in future work. Many semi-supervised learning algorithms have been proposed, and each of them is suitable for a particular problem. In this paper, we proposed the general semi-supervised learning framework for disease gene prediction. In addition to the Harmonic Gaussian algorithm already applied, we may also investigate and use other algorithms that may achieve better results.

6. Conclusion

We have presented an approach using semi-supervised learning to predict disease genes. The experimental results demonstrated that our proposed method performed well with high accuracy, and at the same time, predicted some new disease genes. In future work, we would like to investigate further the biological significance of novel disease genes obtained by our method. Integrating more biological features, like signal transduction pathway, gene loci, and gene-expression data, is also a potential method to improve our method results. Other protein-protein interaction databases should be combined to widen the interaction networks of disease genes.

References

[1] E. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, and B. Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6(55), 2005.

[2] K. Borgwardt and H. Kriegel. Graph kernels for disease outcome prediction from protein-protein interaction networks. In *Pacific Symposium on Biocomputing*, volume 12, pages 4–15, 2007.

[3] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2006.

[4] J. Chen, C. Shen, and A. Sivachenko. Mining Alzheimer Disease Relevant Proteins from Integrated Protein Interactome Data. In *Pacific Symposium on Biocomputing*, volume 11, pages 367–378, 2006.

[5] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33 Database Issue, January 2005.

[6] T. M. Huang and V. Kecman. SemiL, Software for solving semi-supervised learning problems, 2004.

[7] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, and A. Rzhetsky. Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *PNAS*, 101(42):15148–15153, 2004.

[8] K. Lage, O. E. Karlberg, Z. M. Strling, Pll, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tmer, F. Pociot, Y. Tommerup, N.and Moreau, and S. Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316, March 2007.

[9] M. Mark-A and T. Scheffer. Multi-relational learning, text mining, and semi-supervised learning for functional genomics: Special issue: Data mining lessons learned. *Machine Learning*, 57(1-2):61+, 2004.

[10] S. McMahon, H. A. Van Buskirk, K. Dugan, T. D. Copeland, and M. D. Cole. The novel atm-related protein trrap is an essential cofactor for the c-myc and e2f oncoproteins. *Cell*, 94:363–374, 1998.

[11] NCBI. *Genes and disease*. National Library of Medicine (US), NCBI., 2007.

[12] A. Owsianka and A. H. Patel. Hepatitis c virus core protein interacts with a human dead box protein ddx3. *Hepatitis C Virus Core Protein Interacts with a Human DEAD Box Protein DDX3*, 257:330 – 340, 1999.

[13] L. Sam, Y. Liu, J. Li, C. Friedman, and Y. A. Lussier. Discovery of protein interaction networks shared by diseases. In *Pacific Symposium on Biocomputing*, volume 12, pages 76–87, 2007.

[14] Z. Tu, L. Wang, M. Xu, X. Zhou, T. Chen, and F. Sun. Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, 7(31), 2006.

[15] F. S. Turner, D. R. Clutterbuck, and C. Semple. Pocus: mining genomic sequence annotation to predict disease genes. *Genome Biology*, 4(R75), 2003.

[16] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247, August 2005.

[17] J. Xu and Y. Li. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22(22):2800–2805, 2006.

[18] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schlkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16:321–328, 2004.

[19] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and Harmonic functions. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 2003.