Artificial Intelligence in Medicine xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Detecting disease genes based on semi-supervised learning and protein–protein interaction networks

Thanh-Phuong Nguyen [a,*], Tu-Bao Ho [b,c]

[a] Microsoft Research – University of Trento Centre for Computational and Systems Biology Piazza Manci 17, Trento 38123, Italy
[b] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan
[c] Vietnam Academy of Science and Technology, Caugiay, Hanoi, Viet Nam

ARTICLE INFO

Article history:
Received 23 September 2009
Received in revised form 24 May 2011
Accepted 1 September 2011

Keywords:
Semi-supervised learning
Protein–protein interaction network
Multiple data resources integration
Disease gene neighbours
Disease-causing gene prediction

ABSTRACT

Objective: Predicting or prioritizing the human genes that cause disease, or "disease genes", is one of the emerging tasks in biomedicine informatics. Research on network-based approach to this problem is carried out upon the key assumption of "the network-neighbour of a disease gene is likely to cause the same or a similar disease", and mostly employs data regarding well-known disease genes, using supervised learning methods. This work aims to find an effective method to exploit the disease gene neighbourhood and the integration of several useful omics data sources, which potentially enhance disease gene predictions.
Methods: We have presented a novel method to effectively predict disease genes by exploiting, in the semi-supervised learning (SSL) scheme, data regarding both disease genes and disease gene neighbours via protein–protein interaction network. Multiple proteomic and genomic data were integrated from six biological databases, including Universal Protein Resource, Interologous Interaction Database, Reactome, Gene Ontology, Pfam, and InterDom, and a gene expression dataset.
Results: By employing a 10 times stratified 10-fold cross validation, the SSL method performs better than the *k*-nearest neighbour method and the support vector machines method in terms of sensitivity of 85%, specificity of 79%, precision of 81%, accuracy of 82%, and a balanced F-function of 83%. The other comparative experimental evaluations demonstrate advantages of the proposed method given a small amount of labeled data with accuracy of 78%. We have applied the proposed method to detect 572 putative disease genes, which are biologically validated by some indirect ways.
Conclusion: Semi-supervised learning improved ability to study disease genes, especially a specific disease when the known disease genes (as labeled data) are very often limited. In addition to the computational improvement, the analysis of predicted disease proteins indicates that the findings are beneficial in deciphering the pathogenic mechanisms.

1. Introduction

One of the ultimate goals of life science is to improve our understanding of the processes related to disease. On the way to this end, much work has been focusing on monogenic diseases caused by the disorder of single genes, and recently on polygenic diseases caused by disorder of multiple genes in combination with lifestyle and environmental factors. We are currently still far from unraveling the molecular mechanisms of most diseases, and thus developing effective methods to uncover disease genes remains a great challenge. In [1], the authors reviewed different work on predicting or prioritizing potential disease genes, varying from distinguishing between disease genes and non-disease genes to finding groups of genes associated to each disease, and classified their approaches into three broad categories which are not mutually exclusive.

The first category related to research basing on intrinsic disease gene properties to systematically study differences between disease genes and non-disease genes, such as higher conservation of disease genes with a broader phylogenetic extent [2], separation of non-disease genes into two groups of housekeeping and non-housekeeping genes [3], extensive correlations between various gene properties and disease characteristics [4]. Until recently, one only knows for sure a relatively limited number of discovered disease genes and the non-disease genes while in between them most human genes are yet-unidentified genes, and thus to detect yet-unidentified genes remains as a challenging task.

The second category related to research basing on links between candidate genes and disease phenotypes. They exploited various kinds of phenotypic traits, such as gene expression patterns [5], gene ontology functional annotation [6], expression overlap with

* Corresponding author. Tel.: +39 0461 28 2821; fax: +39 0461 28 2814.
E-mail addresses: nguyen@cosbi.eu (T.-P. Nguyen), bao@jaist.ac.jp (T.-B. Ho).

0933-3657/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2011.09.003

disease-related anatomical regions [7], or tissue mRNA expression patterns [8]. Relied on relatively poor annotation of many human genes, these approaches have not yet identified well candidate genes for a given disease.

The third category related to research basing on functional relatedness of candidate genes. They mostly assumed that genes leading to the same phenotype were functionally related, and identified candidate genes as those that had functional relations to known disease genes [9–14]. Different ways have been considered to exploit the functional relatedness, notably the network relatedness and combination or integration of different types of functional genomic data.

Since the last few years, inspired by the findings for yeast protein–protein interaction (PPI) networks, several research groups have been exploiting the human PPI network to predict human disease genes via their corresponding product proteins, which are intuitively called disease proteins [15]. The key of those PPI-based methods is the exploration of the neighbourhood relatedness based on the assumption that "the network-neighbour of a disease gene is likely to cause the same or a similar disease" [16,9,11,17].

Concerning the neighbourhood relatedness of disease genes, Ideker and Sharan pointed out in their excellent review four major research areas [15]: (i) properties of disease genes; (ii) prediction of disease-causing genes; (iii) identification of disease-related sub-networks; and (iv) network-based classification of case-control studies. In the area (ii), various supervised learning techniques have been used to solve the binary classification of disease and non-disease gene classes, such as decision tree induction in [2], *k*-nearest neighbour (*k*-NN) in [10], or support vector machines (SVMs) in [18]. In particular, the topological similarity was usually used in protein networks to solve the problem. In the area (iii), disease-related sub-networks were identified by heuristic score functions to predict causing genes of Alzheimer's disease [19], or by literature mining and network analysis for inherited cerebellar ataxias [20], or by cluster analysis for heterogeneous diseases [21], among others.

Due to the complex nature of disease genes, almost the state-of-the-art methods in area (ii) focuses on distinction of disease genes and non-disease genes for a overall view of human genome while those in area (iii) focuses on local view of individual diseases. Apparently some recent fundamental work about the modular nature of genetic diseases [1] or modularity in disease-phenotype network [22] can be the basis for further study in both areas (ii) and (iii).

The new trend of combination and integration of omics data at various levels has shown advantages in prediction or prioritization of disease genes. Borgwardt and Kriegel combined graph kernels for gene expression and human PPI to do the prediction [23]. Smalter et al. built a disease gene classification system using the topological features of PPI networks and other features using SVMs [18]. Radivojac et al. combined various data sources of human PPI network, known gene-disease associations, protein sequence, and protein functional information and exploited them by SVMs [24]. Other work were based on integrating PPI network data with gene expression data [25], or with disease phenotype data [26].

It is worth noting that all the above-mentioned supervise learning methods are based on the assumption about the separation of available disease genes and non-disease genes. However, we can only know for sure a relatively limited number of discovered disease genes and the non-disease genes while in between them most human genes are as yet-unidentified genes, which are not known as disease genes or non-disease genes. It is significant to develop computational methods that take into account those human genes. To this end, we develop a novel and effective computational method for predicting disease genes using a systematic semi-supervised learning (SSL) framework with multifarious biological data related to disease genes. The key idea is to combine useful data regarding both known disease genes and neighbours of disease genes.

This work has two main contributions. On the one hand, it is the first to not only predict the yet-unidentified genes but also use them in the prediction process by the SSL method. It is known that genes associated with a particular phenotype or function are not randomly positioned in the PPI network, but tend to exhibit high connectivity; they cluster together and occur in central network locations [9]. This overriding property supports the fundamental assumptions about the consistency of SSL, and thus SSL enables us to systematically integrate genomic and proteomic features related to diseases from various data sources, which further enriches the proposed computational scheme. On the other hand, the method integrates the suitable multiple features needed for characterizing yet-unidentified genes in the SSL scheme. Six biological databases are extracted, preprocessed and integrated, including Universal Protein Resource (UniProt) [27], Gene Ontology (GO) [28], Pfam [29], InterDom [30], Reactome [31], and a gene expression dataset [32]. Different functions to characterize topological features of the human PPI network, genomic and proteomic features are appropriately defined. By exploiting such integrated data of disease genes neighbours, it is expected to better predict the disease genes.

We carefully performed two experiments to evaluate the performance of the proposed SSL method. By employing a 10 times stratified 10-fold cross validation, the first one was to evaluate the SSL method, the *k*-NN method on only PPI data [10] and the SVMs method on multiple data [18]. The results showed that SSL method predicted more effectively disease genes in terms of sensitivity, specificity, precision, accuracy, and a balanced F-function. The second one was to estimate accuracy of the SSL method and the *k*-NN method with different data sizes *l* of the labeled data set and each set was tested with twenty trials. Higher accuracy of the SSL method was achieved for all of the tests, even though given a small amount of labeled data. We also did six experiments with different combinations of data features to show the advantage of the data integration and the integration of all data features produced the best result.

This work not only proposes an effective method for disease gene prediction, but also hypothesizes a number of putative disease genes. We carefully carried out an experiment with disease gene information extracted from the Online Mendelian Inheritance in Man (OMIM) database [33]. Testing with all interacting partners of disease proteins, we predicted 572 putative disease proteins. The analysis of these proteins through several ways indicates that the findings are beneficial in deciphering the pathogenic mechanisms.

2. Method

In this section, we first briefly introduce the semi-supervised learning with its appropriateness in disease gene prediction, and then present the proposed framework including the pre-processing procedure for multiple data features.

2.1. Semi-supervised learning and disease gene prediction

Semi-supervised learning is halfway between supervised and unsupervised learning. Essentially, it additionally exploits large sets of available unlabeled data to do supervised learning tasks when the amount of supervised data is limited and additionally exploits rich information in available labeled data to do unsupervised learning tasks. SSL is very useful in solving many real-world problems, especially in the domains where labeling requires much human labour while voluminous unlabeled data is far easier to obtain. In biomedicine, SSL has been applied to many problems and

Please cite this article in press as: Nguyen T-P, Ho T-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med (2011), doi:10.1016/j.artmed.2011.09.003

has achieved notable results, for example, in the study of protein classification [34] and in functional genomics [35], among others.

Due to its capability of learning from both labeled and unlabeled data, SSL is potentially more effective in predicting disease genes. Note that the disease gene identification has been raised as one general classification problem because of two biological reasons. The first one is that disease genes (which are known) are apparently different from other human genes (yet-unidentified disease genes). The second one is that disease genes have several sharing features [3,36] that be well appropriate in SSL applications.

Analyzing the appropriateness of SSL in the disease gene prediction, it is found that the topology of PPI networks satisfied the fundamental assumptions about the consistency of SSL. These assumptions of consistency are: (i) nearby points are likely to have the same label and (ii) points on the same structure (typically referred to as a cluster or a manifold) are likely to have the same label [37]. Likewise, genes that are associated with a particular phenotype or function including the progression of disease are not randomly positioned in the network. Rather, they tend to exhibit high connectivity, cluster together, and occur in the network hubs [9]. As a result, a graph-based SSL method can be suitable for the task of disease gene prediction as PPI networks, when being considered as graphs with nodes as proteins and edges as protein interactions, satisfy the SSL assumptions.

There are about 25–30,000 genes in the human body. As reported in [33], some of them are known to cause diseases, and for these we use the term "known disease genes" or "disease genes". Genes that are assumed not to cause any disease are referred to as "known non-disease genes" or simply "non-disease genes". Disease genes and non-disease genes are labeled data (positives and negatives respectively) in SSL, and the yet-unidentified genes are unlabeled data in SSL. The details of our proposed method are presented in Sections 2.2 and 2.3.

2.2. The proposed method for predicting disease genes

The key premise is to enrich the disease gene prediction by (1) incorporate both known disease genes and neighbours of disease genes and (2) integrating multiple data sources in the SSL scheme.

Fig. 1 illustrates three main steps: (i) identify disease proteins, non-disease proteins, and candidate proteins, (ii) extract and preprocess heterogeneous data from multiple data sources, and (iii) use SSL to predict disease genes. Detailed procedures for three steps are described as follows.

2.2.1. Identify disease proteins, non-disease proteins and candidate proteins

First, take available disease genes, i.e., from the OMIM database and identified the corresponding disease proteins by the UniProt accessions. The set of disease proteins is the positive example set $\mathcal{P}^+$.

Second, extract interactions p_{ij} of disease proteins p_i ($p_i \in \mathcal{P}^+$) from a protein interaction network Ψ. A set of candidate proteins were obtained as the set of neighbours p_j of disease proteins p_i in the network Ψ. The set of candidate proteins is the unlabeled example set $\mathcal{P}^c$.

Third, randomly choose non-disease proteins which are not candidate proteins in the set $\mathcal{P}^c$ and also not proteins corresponding to genes in the ubiquitously expressed human genes (UEHGs) set. The reason for excluding UEHGs is that UEHGs are essential genes having features that differ significantly, both from disease genes and from other genes [3]. The set of non-disease proteins is the negative example set $\mathcal{P}^-$.

After Step 1, we prepared the two sets of labeled data (the positive dataset $\mathcal{P}^+$ and the negative dataset $\mathcal{P}^-$), and one set of unlabeled data (the candidate proteins set $\mathcal{P}^c$). Denote by $\mathcal{P}^*$ the set obtained by the union of these three sets, i.e., $\mathcal{P}^* = \mathcal{P}^+ \cup \mathcal{P}^- \cup \mathcal{P}^c$.

2.2.2. Extract and preprocess heterogeneous data from multiple data sources

For each protein in $\mathcal{P}^*$, extract numerous data from multiple public databases. We investigated several topological, proteomic, and genomic features $t^{feature}$ associated with diseases. Data regarding these features have indeed mixed binary, categorical and numerical types. For this reason, the data were preprocessed, concretely binary and categorical data were transformed to numerical data and then numerical data were normalised. Score functions $f_{feature}$ were proposed to normalised the data features $t^{feature}$.

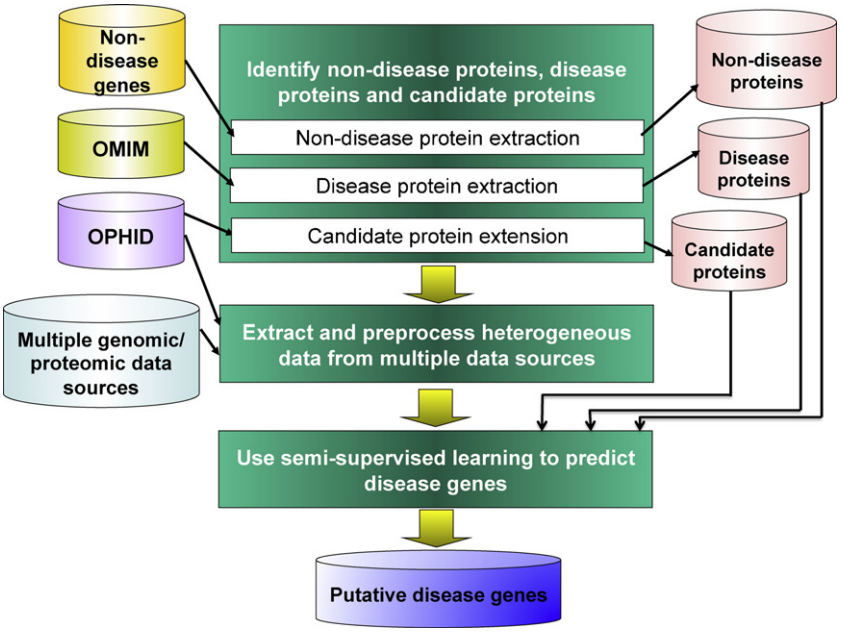

Fig. 1. Semi-supervised learning method for disease gene prediction.

Please cite this article in press as: Nguyen T-P, Ho T-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med (2011), doi:10.1016/j.artmed.2011.09.003

Table 1
Six public databases used for data retrieval (Accessed: January 2009).

Database	Description	URL	Statistics
UniProt [27]	A comprehensive high-quality and freely accessible resource of protein sequence and functional information	http://www.UniProt.org	220,325 entries
i2d [39]	A known experimental and predicted PPIs for five model organisms and human	http://ophid.utoronto.ca/	424,066 entries
Reactome [31]	A curated resource of core pathways and reactions in human biology	http://www.reactome.org	928 pathways for human
GO [28]	A controlled vocabulary to describe gene and gene product attributes in any organism	http://www.geneontology.org/	
Pfam [29]	A large collection of protein families, each represented by multiple sequence alignments and hidden Markov models	http://www.sanger.ac.uk/Software/Pfam/	10,340 families
InterDom [30]	A database of putative interacting protein domains derived from multiple sources	http://interdom.i2r.a-star.edu.sg/	148,938 entries

Table 2
Statistics of two sets $\mathcal{P}^+$ and $\mathcal{P}^*$ with the eight extracted proteomic and genomic features.

Data sources	Feature $t^{feature}$	♯Record		♯Category	
		in $\mathcal{P}^*$	in $\mathcal{P}^+$	in $\mathcal{P}^*$	in $\mathcal{P}^+$
UniProt	t^{length}	4412	1496		
	t^{KW}	31,465	13,597	564	504
	t^{EC}	1123	451	133	106
Gene Ontology	t^{GO}	17,241	6404	2911	1817
Pfam	t^{Pfam}	6817	2426	1796	1413
InterDom	t^{DDI}	3854	1322		
Reactome	$t^{Pathway}$	1167	540	68	62
Gene Expression	$t^{expression}$	696	52		

2.2.3. Use SSL to predict disease genes

A graph-based SSL algorithm was appropriately employed for the prediction task. The labeled and unlabeled datasets (in Step 1) and the multiple data (in Step 2) were learned using Gaussian fields and Harmonic functions [38]. More specifically the learning problem was formulated in terms of a Gaussian random field on the weighted graph representing labeled and unlabeled data, where the mean of the field was characterized in terms of harmonic functions. The output was a set of new putative disease genes.

2.3. Extracting and preprocessing heterogeneous data

Several biological features are known to associate and cause diseases. However, data concerning those features are scattered in a wide range of data sources. A computational scheme for the data integration has been emerged to better study disease genes. Our proposed method employed SSL to combine data features of both labeled examples and unlabeled examples.

Table 1 presents some description about six databases under our investigation. Topological data of the PPI networks were extracted from the i2d database (the formerly known as Online Predicted Human Interaction Database). Other data concerning eight proteomic and genomic features were selected from six data sources (five public databases and one published dataset): UniProt (three features on sequence length, tagged keyword, and coded enzyme), GO (one feature on GO term), Pfam (one feature on protein domain), InterDom (one feature on domain-domain interaction), Reactome (one feature on pathway), and gene expression dataset (one feature on gene expression profile).

Table 2 shows the statistics of the proteomic and genomic features from each data source used. Columns, 3 and 4, present the numbers of records extracted according to their respective features, and the last two columns show the numbers of feature categories. Among the 5557 proteins in $\mathcal{P}^*$, 31,465 data records were found for the keyword features, and 1123 for the enzyme features. These proteins shared the same 564 keywords and 133 enzymes, as shown in Table 2. For example, two records (P05067, Alzheimer disease) and (P01011, disease mutation) where P05067, P01011 are the UniProt accessions; "Alzheimer disease" and "disease mutation" are their keywords, or (O75688, ec3.1.3) where O75688 is the UniProt name and ec3.1.3 is the enzyme commission.

The data types are heterogeneous, since the keyword data, the pathway data and coded enzyme data are in form of a categorical free text, while the sequence length and number of domain–domain interactions are numerical values. We defined score functions to transform and normalise the extracted data. The score functions $f_{feature}$ proposed for topological feature data and genomic, proteomic data features are introduced below and summarized in Table 3. Note that in the experimental evaluation the set P^+ consists of disease genes in the training dataset (not the whole set of disease genes).

Gene expression data are valuable to be incorporated because it is the quantitative trait and highly heritable. A lot of evidence showed the potential causal impact of differential gene expression on complex disease risk [32]. After investigating several gene expression datasets, such as [40–42], we used the gene expression profile that defined colon cell maturation [40]. Mariadason et al. analyzed 17,280 sequences to reveal the maturation of *Caco-2* cells and then studied related biological phenomena, such as patterns of coordinate regulation, cell cycle, xenobiotic and drug detoxification, signal transduction pathways, etc. Because of containing rich information, we chose this set to explore gene expression profile information. It was not necessary to preprocess the gene expression data because they were normalised already.

- *The topological function*: This function measures the topological association between a given protein and disease proteins based on the PPI network. We can assume that if one protein has many interactions with disease proteins, this protein is likely to be a disease protein. The function $f_{ppi}(p_i)$ was defined for the feature t^{ppi}.
- *The keyword function*: Disease proteins may have the same keywords, and these common keywords are tagged more frequently in the set of disease proteins than other proteins. Categorical data of the keyword feature were converted and normalised by their frequency and assigned to each protein p_i by the function $f_{kw}(p_i)$.
- *The enzyme function*: Enzymes perform a wide variety of functions inside living organisms. The relationship between enzymes and diseases has been studied in many research. Like the keyword feature, some specific enzymes are often shared among the group of disease proteins. This data feature was scored by function $f_{ec}(p_i)$.
- *The sequence length function*: We investigated the protein sequence length feature to study how the sequence length of a protein relates to disease-causing mechanisms. Function $f_{length}(p_i)$ was the ratio of sequence length of a protein over the average length of disease proteins.
- *The GO term function*: The GO terms are divided into three groups: molecular function, biological process, and cellular component.

Please cite this article in press as: Nguyen T-P, Ho T-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med (2011), doi:10.1016/j.artmed.2011.09.003

Table 3
Topological feature, genomic/proteomic features and their corresponding functions.

Functions	Notations and explanations
$f_{ppi}(p_i) = \frac{\sum_{p_j \in \mathcal{P}^+} Int(p_i,p_j)}{\sum_{p_j \in \mathcal{P}^*} Int(p_i,p_j)} \times \frac{\sum_{p_j \in \mathcal{P}^+} Int(p_i,p_j)}{Avg_{ppi}}$	$Int(p_i, p_j) = \begin{cases} 1 & \text{betweenproteins } p_i \text{ and } p_j, \\ 0 & \text{otherwise.} \end{cases}$ Avg_{ppi}: the average of the number of protein interactions belonging to disease proteins.
$f_{kw}(p_i) = \frac{1}{\sum_{\forall kw_i \in p_i} w_i^{kw}}$	$w_i^{kw} = freq^+(kw_i) \times freq^*(kw_i)$ $freq^+(kw_i)$: the frequency count of kw_i observed in $\mathcal{P}^+$. $freq^*(kw_i)$: the frequency count of kw_i observed in $\mathcal{P}^*$.
$f_{ec}(p_i) = freq^+(ec_i) \times freq^*(ec_i)$	$freq^+(ec_i)$: the frequency count of ec_i observed in $\mathcal{P}^+$. $freq^*(ec_i)$: the frequency count of ec_i observed in $\mathcal{P}^*$.
$f_{go}(p_i) = \frac{1}{\sum_{\forall go_i \in p_i} w_i^{go}}$	$w_i^{go} = (\sharp go_i^+ + 1)/(\sharp go_i^* + 1)$ $\sharp go_i^+$: the frequency count of go_i observed in $\mathcal{P}^+$. $\sharp go_i^*$: the frequency count of go_i observed in $\mathcal{P}^*$.
$f_{pfam}(p_i) = \frac{\sharp pfam_i^+ + 1}{\sharp pfam_i^* + 1}$	$\sharp pfam_i^+$: the number of domains d_j of a protein p_i observed in $\mathcal{P}^+$ $\sharp pfam_i^*$: the number of domains d_j of the protein p_i observed in $\mathcal{P}^*$.
$f_{length}(p_i) = \frac{length(p_i)}{Avg_{length}}$	$length(p_i)$: the sequence length of a protein p_i. Avg_{length}: the average sequence length of disease proteins in $\mathcal{P}^+$.
$f_{pathway}(p_i) = \sum_{\forall pathway_i \in p_i} w_i^{pathway}$	$w_i^{pathway} = freq^+(pathway_i) \times freq^*(pathway_i)$ $freq^+(pathway_i)$: the frequency count of $pathway_i$ observed in $\mathcal{P}^+$. $freq^*(pathway_i)$: the frequency count of $pathway_i$ observed in $\mathcal{P}^*$.
$f_{ddi}(p_i) = \frac{1}{\sum_{\forall ddi_i \in p_i} w_i^{ddi}}$	$w_i^{ddi} = \frac{Avg_{ddi}}{\sharp ddi(p_i)}$ $\sharp ddi(p_i)$: the number of DDI observed in $\mathcal{P}^+$ of protein p_i. Avg_{ddi}: the average of the number of DDI of disease proteins in $\mathcal{P}^+$.

These terms present the typical information about the proteins, and the disease proteins probably centralise on some specific terms that were normalised by the function $f_{go}(p_i)$.

- *The protein domain function*: Protein domains are the building blocks of proteins. Disease proteins may structurally or functionally depend on their domains. If a protein has many domains related to disease, this protein is more likely to be a disease protein. Pfam domains d_j of all considered proteins were nomalised by $f_{pfam}(p_i)$.
- *The DDI function*: DDIs are likely to regulate the interactions of proteins, and themselves perform specific functions in cells, particularly in causing diseases. We obtained the DDI data from the InterDom database and weighted them by $f_{ddi}(p_i)$ based on the number of their DDI shared with disease proteins.
- *The biological pathway function*: Many disease processes arise from defects in biological pathways. Extracting data from Reactome database, there were 68 pathways involved by all proteins in the extended protein set and among these pathways 62 pathways were found to contain at least one disease protein. The $f_{pathway}(p_i)$ of feature $t^{pathway}$ was based on the frequency of the pathways observed in both $\mathcal{P}^*$ and $\mathcal{P}^+$.

3. Experiments

To evaluate the performance of the proposed method, we repeated two previous works based on supervised learning, i.e. the *k*-NN method with single data [10] and the SVMs method with multiple data [18]. The *k*-NN method is a typical classification method that assigns class label to a unknown object based on the majority of its nearest neighbours in the known classes. The SVMs method is a typical kernel method that learns a hyperplane to separate two classes with their maximized margin.

We carried out two comparative evaluations. By employing a 10 times stratified 10-fold cross validation, the first one was to evaluate the SSL method, the *k*-NN method and the SVMs method in terms of sensitivity, specificity, precision, accuracy, and a balanced F-function. The second one was to estimate accuracy of the SSL method and the *k*-NN method with different data sizes *l* of the labeled data set and each set was tested with twenty trials. Additionally, we did six experiments with different combinations of data features to show the advantage of the data integration.

We prepared three data sets to carry out the experiments: a set of disease genes, a set of non-disease genes, and a set of PPIs. We repeated the similar procedure of choosing negatives and positives as it was done in [10,18]. The set of disease genes was extracted from the OMIM database that is a catalogue of human genes and genetic disorders. In the OMIM database the list of hereditary disease genes is described in the OMIM morbid map. As reported in [33], there are 4512 records with 3053 unique OMIM identifiers in the catalogue. A total of 3053 human disease phenotypes were mapped to look for their disease proteins identified by UniProt accessions. The results showed 3590 corresponding disease proteins and this disease protein set consisted of 1502 proteins having interactions published in the i2d database. From the set of human genes, we randomly chose negative examples, which did not belong to both the OMIM morbid map and the UEHG set. As the ratio between human disease genes and non-disease genes is currently not known, we generated the equal number of non-disease proteins to the number of disease proteins. The set of human PPIs were extracted from the i2d database. Among 51,934 human PPIs stored in the i2d database, there were 13,368 interactions, which had at least one interacting partner belonging to the set of disease proteins. Based on 13,368 interactions, the initial set of 1502 disease proteins was extended to 5775 proteins.

In the SSL implementation, we have chosen the SemiL software, developed by Huang and Kecman [43] (Accessed: January 2009), to run the Harmonic Gaussian method as it is efficient for solving large-scale semi-supervised learning problems using graph kernels and thus it is suitable for the topological characteristics of PPI networks. In the algorithm, given l labeled points $(x_1, y_1), \ldots, (x_l, y_l)$ and

Please cite this article in press as: Nguyen T-P, Ho T-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med (2011), doi:10.1016/j.artmed.2011.09.003

Table 4
The 10 time 10-folds cross validation performance of the SSL method (SSL1 with Cosine distance and SSL2 with Euclidean distance) compared to the two methods SVMs and *k*-NN. The performance of the SSL method is highlighted in bold.

Method	Precision	Accuracy	Sensitivity	Specificity	F-measure
SSL1	**0.812 ± 0.042**	**0.823 ± 0.019**	**0.852 ± 0.031**	**0.794 ± 0.041**	**0.829 ± 0.013**
SSL2	**0.806 ± 0.039**	**0.820 ± 0.019**	**0.850 ± 0.026**	**0.789 ± 0.036**	**0.825 ± 0.013**
SVMs	0.713 ± 0.032	0.741 ± 0.023	0.804 ± 0.035	0.677 ± 0.038	0.756 ± 0.019
1-NN	0.779 ± 0.033	0.786 ± 0.032	0.798 ± 0.025	0.774 ± 0.042	0.789 ± 0.032
3-NN	0.768 ± 0.037	0.782 ± 0.020	0.806 ± 0.030	0.757 ± 0.037	0.787 ± 0.027
5-NN	0.771 ± 0.031	0.771 ± 0.017	0.819 ± 0.029	0.744 ± 0.037	0.789 ± 0.027
7-NN	0.761 ± 0.042	0.761 ± 0.024	0.822 ± 0.030	0.720 ± 0.022	0.782 ± 0.019
9-NN	0.776 ± 0.030	0.540 ± 0.025	0.770 ± 0.027	0.752 ± 0.034	0.763 ± 0.026

u unlabeled points $(x_{l+1}, \ldots, x_{l+u})$, the data space is represented as a graph $G = (V, E)$. The set of nodes V corresponds to both nodes $L = \{1, \ldots, l\}$ corresponding to labeled points and nodes $U = \{l+1, \ldots, l+u\}$ corresponding to unlabeled ones. The task is to assign (predict) the labels of nodes in the set of unlabeled data. An $n \times n$ symmetric weight matrix W on the edges of the graph is given. When $x \in \mathcal{R}$, W was defined as

$$W_{ij} = exp\left(-\sum_{d=1}^{m} \frac{(x_{id} - x_{jd})^2}{\sigma_d^2}\right)$$

where x_{id} is the dth component of instance x_i represented as a vector $x_i \in \mathcal{R}$, and $\sigma_1, \ldots, \sigma_m$ are length scale hyperparameters for each dimension. The nearby points in the Euclidean space are assigned large edge weights. Intuitively, unlabeled points that are nearby in the graph have similar labels. In our experiment, the weight matrices W were calculated with two different distance functions: Euclidean and Cosine, and the degree of graph was 20.

We used the popular machine learning workbench package Weka [44] to run *k*-NN and SVMs. For the *k*-NN method test, the different values for the parameter *k* were chosen exactly as in Xu and Li's work [10]. For the SVMs test, the kernels were RBF and linear kernel functions, and other parameters were default values as in [18].

4. Results

4.1. Computational validation

In the first experiment, we evaluated five measures of precision, accuracy, sensitivity, specificity, and F-measure by a 10 × 10-fold stratified cross validation [45] for the three methods, SSL, *k*-NN, and SVMs. In each fold, the training data set was randomly divided into 10 subsets, 9 subsets for training and the rest one for testing and in each subset, the number of negatives and positives were equal. The performance of the SSL method, the *k*-NN method, and the SVMs method then was statistically tested in terms of confidence intervals, to give an estimate of the amount of error involved in the data. To estimate a 95% confidence interval for each calculated precision, accuracy, specificity, sensitivity and F-measure we used *t* distribution. The experimental results with 95% confidence intervals are shown in Table 4. The experimental results demonstrated that the SSL methods performed better the other methods in the disease gene prediction.

In the second experiment, from the training dataset we randomly selected l data points as labeled data, while the rest $(n - l)$ were unlabeled data. The accuracy was estimated by comparing the number of predicted labels and the number of true labels. For each labeled set size l, we did 20 trials and the final result was the average accuracy of the 20 trials. These procedures were carried out for both the SSL method and the *k*-NN method on the same testing datasets.

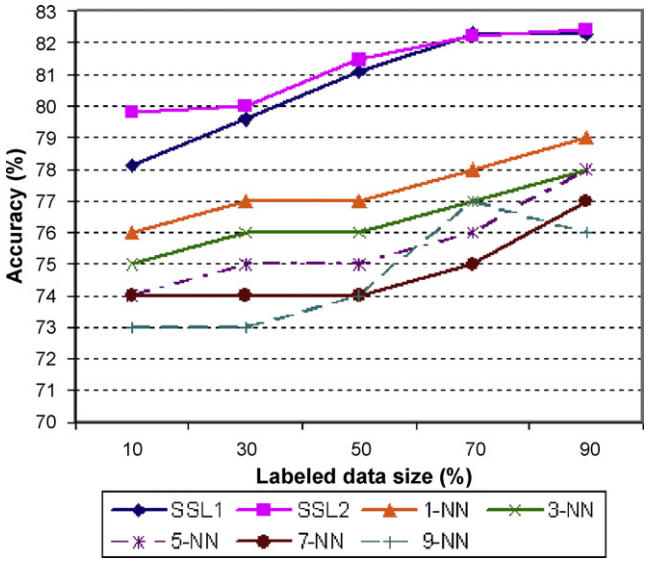

Fig. 2. Accuracy of the proposed method with different labeled set sizes for Cosine distance (SSL1) and the Euclidean (SSL2) compared to the *k*-NN method.

Fig. 2 shows the accuracy of our method and the *k*-NN method with various parameters *k*. When the labeled set size was small (10% of the dataset), the SSL method obtained non-trivial accuracy of 78%. When the amount of labeled data was at least half of the total dataset, accuracy of the SSL method is over 80%. By comparison with the *k*-NN method, the SSL method obtained higher accuracy with all tested sizes of the labeled data set. The results demonstrated that when there was a limited number of labeled data, SSL could predict disease genes better than supervised learning did.

Additionally, to see the advantages of data integration, we did six experiments with six different groups of the data features. For each experiment, we calculated the error rate [43] of the prediction. Table 5 and Supplementary material 1 present the results of the six experiments. It is shown that the combination of all investigated data features (as presented in Section 2.3) produces the best result.

Table 5
The error rate of the SSL method with different combinations of data features (e1 with Cosine distance and e2 with Euclidean distance).

Experiment	e1	e2
Exp1: All data features excluding the PPI data feature	0.212	0.216
Exp2: All data features excluding two data features PPI and the DDI	0.216	0.221
Exp3: All data features excluding the gene expression data feature	0.204	0.207
Exp4: All data features excluding the sequence length data feature	0.208	0.211
Exp5: All data features excluding two data features domain and DDI	0.206	0.210
Exp6: All of investigated data features	0.185	0.190

Please cite this article in press as: Nguyen T-P, Ho T-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med (2011), doi:10.1016/j.artmed.2011.09.003

Table 6
List of some high-ranked disease genes by Endeavour system.

♯ Rank	Gene symbol	Gene name	Q-int	*p*-value
2	MYH10	Myosin, heavy chain 10, non-muscle	1.16E− 08	0.00026408
4	FYN	FYN oncogene related to SRC, FGR, YES	3.70E− 07	0.001516515
7	ALK	Anaplastic lymphoma receptor tyrosine kinase	2.47E− 06	0.003953699
10	ERBB3 (HER3)	v-erb-a erythroblastic leukemia viral oncogene homolog 3 (avian)	4.90E− 06	0.005591587
11	ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	5.64E− 06	0.006002614
13	LAMA4	Laminin, alpha 4	6.97E− 06	0.006679875
14	JAK1	Janus kinase 1 (a protein tyrosine kinase)	7.68E− 06	0.00701467
16	SPTBN1	Spectrin, beta, non-erythrocytic 1	9.49E− 06	0.007801282
17	LRP1	Low density lipoprotein-related protein 1 (alpha-2-macroglobulin receptor)	9.80E− 06	0.007931698
22	ACTN2	Actinin, alpha 2	1.26E− 05	0.00899866

4.2. Biological validation

We biologically validated the putative disease genes through different techniques. Testing the whole network of protein interactions, we detected 572 putative disease proteins. The list of predicted disease proteins and their corresponding genes is presented in Supplementary material 2. The biological analysis supported our findings that could be the starting point for new studies on pathophysiology of various diseases.

We investigated the findings by: (i) validating the putative disease gene's keywords and pathways shared with known disease genes, (ii) checking their functional category and gene similarity via the DAVID tools [46] (Accessed: January 2009), (iii) testing them with the ranking system Endeavour – Computer Program For Identifying Disease Genes [47] (Accessed: January 2009), and (iv) studying their disease-related information through biomedical literature. This section discusses some interesting findings.

Firstly, we checked whether the putative disease proteins had keywords and pathways of known disease proteins. Among 47 Reactome pathways shared with known disease proteins, we found that the set of putative proteins belonged to many pathways associated with disease traits, such as '*Signalling in Immune system*' (29 putative proteins), e.g. PI3-kinase p85 subunit beta, GTPase HRas, HLA class I histocompatibility antigen, A-3 alpha chain; '*Hemostasis process*' (25 putative proteins), e.g. PI3-kinase p85 subunit beta, Platelet-derived growth factor subunit A; and '*Gene expression process*' (21 putative proteins), e.g. Cyclin-T1. There were 167 over 572 putative disease genes participating in disease-related pathway. Similarly, 270 Uniprot keywords of predicted disease genes were tagged for known disease proteins. Among them, many putative disease proteins shared the same keywords, e.g., '*alternative splicing*' with 212 proteins, '*polymorphism*' with 195 proteins, and '*glycoprotein*' with 187 proteins.

Secondly, we checked the functional category, and the gene similarity of the putative disease genes via the DAVID tools. Interestingly, 29 genes were found in 67 records in the Genetic Association Database (GAO).[1] For example, *IGFBP2* (insulin-like growth factor binding protein 2, 36 kDa), and *TNFSF8* (tumour necrosis factor (ligand) superfamily, member 8) were related to the term '*diabetes, type 1*'; *IFNAR1* (interferon alpha, beta and omega receptor 1) was related to the term '*Hepatitis B, Chronic*', and *ITGA3* (Integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)) is related to the term '*breast cancer*'. Checking the putative disease genes in the OMIM database, 2 genes were related to 8 records found in database OMIM with the term '*Colorectal cancer*', e.g., *BAX* (bcl2-associated x protein) and *HRAS* (v-Ha-ras harvey rat sarcoma viral oncogene homolog). The detailed results with enrichment analysis are presented in Supplementary material 3.

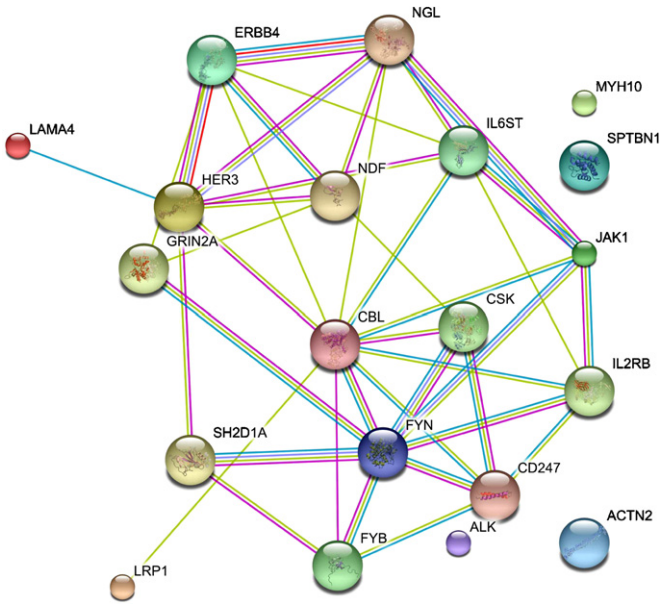

Fig. 3. Functional networks of the high-ranked disease genes.

Thirdly, we evaluated the putative disease genes through the Endeavour system. Endeavour is a software application for the computational prioritization of candidates genes, based on a set of training genes. In order to estimate the reliability of the predicted genes, they were ranked by the Endeavour system with all available data sources. There were 42 predicted genes with *p-value* $\leq$0.05 which were found in the set of predicted disease genes. Some of them obtained a very high rank with a statistically significant *p-value*. Table 6 lists the top 10 putative genes ranked by the Endeavour system.

In order to study how the predicted genes related to some specific diseases, we did the second test for 572 candidate proteins with three diseases, cancer, diabetes, and Alzheimer's. The similarity measure between the candidate genes and the known disease genes was calculated by two systems Ouzounis [48] and ProspectR[2] (Accessed: January 2009). It was absorbing that the Endeavour system returned high ranked genes with *p-value* $\leq$0.01, for example, genes *MYH10*, *FYN*, *ALK*, *LAMA4*, *ERBB4*, *LRP1*. We illustrated the functional associations of the top ten proteins ranked by Endeavour using the visualisation tool of the STRING database[3] (Accessed: January 2009). Fig. 3 shows the functional network of these proteins. This network demonstrates how these proteins functionally connect with others; and the thickness of an edge (a connection) displays the strength of that connection. In the network, four

[1] http://geneticassociationdb.nih.gov.

[2] http://www.genetics.med.ed.ac.uk/prospectr/.

[3] http://string.embl.de.

Please cite this article in press as: Nguyen T-P, Ho T-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med (2011), doi:10.1016/j.artmed.2011.09.003

proteins *ERBB3 (HER3)*, *ERBB4*, *NGL (ERBB2)* and *JAK1*, closely associated with each other. It suggested that these proteins might relate to or cause the same diseases. Besides, protein *FYN* was found that it was highly connected with other proteins as a hub of the networks.

Considering four aforementioned proteins *ERBB3 (HER3)*, *ERBB4*, *JAK1*, and *FYN*, there are biological evidence presented their relevance with cancer.

Signalling pathways regulated by *ErbB* receptor family are involved in cancer progression [49]. Studies of *ERBB3* expression in primary cancers and of its mechanistic contributions in cultured cells have implicated it, with varying degrees of certainty, with causation or sustenance of cancers of the breast, ovary, prostate, certain brain cells, retina, melanocytes, colon, pancreas, stomach, oral cavity and lung [50]. *ERBB4* receptor tyrosine kinase in breast cancer is generally regarded as a marker for patient prognosis, controversial exceptions [51].

Studying the family of *Jak kinases*, it is known to be composed from at least four different tyrosine kinases (*Tyk2*, *Jak1*, *Jak2*, *Jak3*) that share significant structural homology with each other. The members of this family of kinases associate constitutively with a variety of cytokine and hormone receptors. In their review [52], Verma et al. affirmed that the *JAK family* played important roles in the generation of responses for interferons, which were cytokines that exhibit important antitumour activities. Recent discoveries have suggested that mutated JAK proteins would be potent targets for anti-cancer therapy [53].

Proto-oncogene tyrosine-protein kinase *Fyn* is implicated in the control of cell growth. And this kinase is required in brain development and mature brain function with crucial roles in the regulation of axon growth, axon guidance, and neurite extension [54,55]. It has been discovered that *FYN* is downregulated in prostate cancer by both chromosomal deletion and promoter hypermethylation, and therefore is a novel prostate tumour suppressor gene candidate [56].

The above analysis showed that the predicted disease proteins were practically useful when studying the genes involved in diseases of interest. Doctors and biologists would employ these results as the hypothetical disease genes for aiding or guiding the wet experiments. This is where the findings of this work become beneficial.

5. Conclusion

In this paper, we have introduced a method based on semi-supervised learning, integrating multiple data features, for the disease gene prediction. The method proposed here is a systematic framework that can be applied to not only a general disease study, but also to a particular disease. Several biological features associating with diseases were examined and extracted and they were effectively combined in the proposed method. The experimental results demonstrated that our method performed well with high accuracy, and at the same time, predicted some new putative disease genes. Performing the experimental with small amounts of labeled data, the results demonstrated the advance of the method in studying specific diseases in case that the known disease genes (as labeled data) are often very limited.

In future work, we would like to validate the predicted disease genes by experiments. It is interesting to extend the proposed method by firstly clustering disease phenotypes and then identifying disease genes for each disease cluster. Other work will involve applying and comparing the performance of the Harmonic Gaussian algorithm with other semi-supervised learning algorithms for disease genes prediction. Feature selection techniques will be developed to additionally preprocess data.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.artmed.2011.09.003.

References

[1] Oti M, Brunner HG. The modular nature of genetic diseases. Clinical Genetics 2006;71:1–11.
[2] Adie EJ, Adams RR, Evans KL, Porteous DJ, Pickard BS. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics 2005;6(55).
[3] Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics 2006;7(31).
[4] Lopez-Bigas N, Blencowe BJ, Ouzounis CA. Highly consistent patterns for inherited human diseases at the molecular level. Bioinformatics 2006;22(3):269–77.
[5] Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nature Genetics 2002;31(3):316–9.
[6] Turner FS, Clutterbuck DR, Semple CAM. Pocus: mining genomic sequence annotation to predict disease genes. Genome Biology 2003;4(R75).
[7] Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Research 2005;33(5):1544–52. ISSN: 1362-4962.
[8] Masseroli M, Galati O, Pinciroli F. Gfinder: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Research 2005;33:W717–23.
[9] Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein–protein interactions. Journal of Medical Genetics 2006;43:691–8.
[10] Xu J, Li Y. Discovering disease-genes by topological features in human protein–protein interaction network. Bioinformatics 2006;22(22):2800–5.
[11] Benjamin SB, Alex B. Protein interactions in human genetic diseases. Genome Biology 2008;9(1):R9.1–12.
[12] Erten S, Bebek G, Koyuturk M. Disease gene prioritization based on topological similarity in protein–protein interaction networks. In: Bafna V, Sihinalp SC, editors. RECOMB 2011, vol. LNBI 6577. Vancouver: Springer; 2011. p. 54–68.
[13] Nguyen TP, Ho TB. A semi-supervised learning approach to disease gene prediction. In: IEEE international conference on bioinformation and biomedicine (BIBM 2007). Silicon Valley: IEEE; 2007. p. 207–6214.
[14] Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics 2010;26(8):1057–63.
[15] Ideker T, Sharan R. Protein networks in disease. Genome Research 2008;18(4):644–52.
[16] Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. Briefings in Bioinformatics 2007;8(5):333–46.
[17] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proceedings of the National Academy of Sciences of the United States of America 2007;104(21):8685–90.
[18] Smalter A, Lei SF, Chen XW. Human disease-gene classification with integrative sequence-based and topological features of protein–protein interaction networks. In: Hu X, Mandoiu I, Obradovic Z, Xia J, editors. IEEE international conference on bioinformatics and biomedicine (BIBM 2007). California: CPS; 2007. p. 209–16.
[19] Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. Proceedings of the National Academy of Sciences of the United States of America 2004;101(42):15148–53.
[20] Lim J, Hao T, Shaw C, Patel AJ, Szab G, Rual JF, et al. A proteinprotein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. Cell 2006;125(4):801–14.
[21] Sun PG, Gao L, Han S. Prediction of human disease-related gene clusters by clustering analysis. International Journal of Biological Sciences 2011;7(1):61–73.
[22] Jiang X, Liu B, Jiang J, Zhao H, Fan M, Zhang J, et al. Modularity in the genetic disease-phenotype network. FEBS Letters 2008;582:2549–54.
[23] Borgwardt KM, Kriegel H. Graph kernels for disease outcome prediction from protein–protein interaction networks. In: Pacific symposium on biocomputing, vol. 12. Singapore: World Scientific Publishing Company; 2007. p. 4–15.
[24] Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, et al. An integrated approach to inferring gene-disease associations in humans. Proteins: Structure, Function, and Bioinformatics 2008;72(3):1030–7.
[25] Karni S, Soreq H, Sharan R. A network-based method for predicting disease-causing genes. Journal of Computational Biology 2009;16(2):181–9.
[26] Wu X, Jiang R, Zhang MQ, Li Sh. Network-based global inference of human disease genes. Molecular Systems Biology 2008;4(May).
[27] Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (uniprot). Nucleic Acids Research 2005;33:D154–9.
[28] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 2004;32(Suppl. 1):D258–61.
[29] Finn RD, Tate J, Mistry PC, Coggill J, John Sammut S, Hotz HR, et al. The Pfam protein families database. Nucleic Acids Research 2008;36(Suppl. 1):D281–8.
[30] Ng SK, Zhang Z, Tan SH, Lin K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Research 2003;31(1):251–4.

Please cite this article in press as: Nguyen T-P, Ho T-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med (2011), doi:10.1016/j.artmed.2011.09.003

[31] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Research 2005;33(Suppl. 1):D428–32.
[32] Dermitzakis ET. From gene expression to disease risk. Nature Genetics 2008;40(5):492–3.
[33] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 2005;33(Database Issue (January)). ISSN: 1362-4962.
[34] Weston J, Leslie C, Ie E, Zhou D, Elisseeff A, Noble WS. Semi-supervised protein classification using cluster kernels. Bioinformatics 2005;21(August (15)):3241–7. ISSN: 1367-4803.
[35] Mark-A M, Scheffer T. Multi-relational learning, text mining, and semi-supervised learning for functional genomics: special issue: data mining lessons learned. Machine Learning 2004;57(1–2):61+.
[36] Smith NG, Eyre-Walker A. Human disease genes: patterns and predictions. Gene 2003;318:169–75. ISSN: 0378-1119.
[37] Chapelle O, Scholkopf B, Zien A. Semi-supervised learning. Massachusetts: The MIT Press; 2006.
[38] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Fawcett T, Mishra N, editors. Proceedings of the twentieth international conference on machine learning (ICML-2003). Menlo Park, California: The AAAI Press; 2003. p. 912–9.
[39] Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics 2005;21(May (9)):2076–82.
[40] Mariadason JM, Arango D, Corner GA, Aranes MJ, Hotchkiss KA, Yang LH, et al. A gene expression profile that defines colon cell maturation in vitro. Cancer Research 2002;62(16):4791–804.
[41] Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y, Ihara S, Wang SM, et al. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics 2005;86(2):127–41. ISSN: 0888-7543.
[42] Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the United States of America 2004;101(16):6062–7.
[43] Huang TM, Kecman V. SemiL – software for solving semi-supervised learning problems. <http://www.support-vector.ws/html/semil.html>; 2004 [accessed January 2009].
[44] Witten IH, Eibe F. Data mining: practical machine learning tools and techniques. San Fransisco: Morgan Kaufmann Publishers Inc.; 2005.
[45] Han J, Kamber M. Data mining: concepts and techniques (the Morgan Kaufmann series in data management systems). 1st ed. Morgan Kaufmann; 2000, September.
[46] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. David: database for annotation visualization and integrated discovery. Genome Biology 2003;4(5). ISSN: 1465-6914.
[47] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nature Biotechnology 2006;24(May (5)):537–44.
[48] Lopez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Research 2004;32(10):3108–14.
[49] Holbro T, Civenni G, Hynes NE. The erbb receptors and their role in cancer progression. Experimental Cell Research 2003;284(1):99–110. ISSN: 0014-4827.
[50] Sithanandam G, Anderson LM. The ERBB3 receptor in cancer and cancer gene therapy. Cancer Gene Therapy 2008;15(7):413–48.
[51] Chuu CP, Chen RY, Barkinge JL, Ciaccio M, Jones RB. Systems-level analysis of ErbB4 signaling in breast cancer: a laboratory to clinical perspective. Molecular Cancer Research 2008;6(6):885–91.
[52] Verma A, Kambhampati S, Parmar S, Platanias LC. Jak family of kinases in cancer. Cancer and Metastasis Reviews 2003;22(4):423–34.
[53] Constantinescu SN, Girardot M, Pecquet C. Mining for jak-stat mutations in cancer. Trends in Biochemical Sciences 2008;33(3):122–31.
[54] Rigley K, Slocombe P, Proudfoot K, Wahid S, Mandair K, Bebbington C. Human p59fyn(T) regulates OKT3-induced calcium influx by a mechanism distinct from PIP2 hydrolysis in Jurkat T cells. Journal of Immunology 1995;154(8):1136–45.
[55] Meriane M, Tcherkezian J, Webber CA, Danek EI, Triki I, McFarlane S, et al. Phosphorylation of DCC by Fyn mediates Netrin-1 signaling in growth cone guidance. Journal of Cell Biology 2004;167(4):687–98.
[56] Sørensen KD, Borre M, Ørntoft TF, Dyrskjøt L, Tørring N. Chromosomal deletion, promoter hypermethylation and downregulation of fyn in prostate cancer. International Journal of Cancer 2008;122(3):509–19.

Please cite this article in press as: Nguyen T-P, Ho T-B. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif Intell Med (2011), doi:10.1016/j.artmed.2011.09.003