7

Disease Gene Prioritization

Carlos Roberto Arias[1], Hsiang-Yuan Yeh[2] and Von-Wun Soo[1,2]
[1]*Institute of Information Systems and Applications, National Tsing Hua University*
[2]*Computer Science Department, National Tsing Hua University*
Taiwan, ROC

1. Introduction

The identification of genes is an ongoing research issue in the biomedical and bioinformatics community. The Human Genome Project which was completed in 2003, identified approximately 20,000+ genes in the human DNA, but there are still many of these genes for which their function or role is unknown, and this accounts only for healthy DNA. Genetic diseases like Cancer, Alzheimer, Hemophilia and others, have mechanisms that we currently just started to understand. For instance, genes BRCA1 and BRCA2, famous for their role in breast cancer (Friedman et al., 1994), only account for 5% of the incidence of the aforementioned cancer (Oldenburg et al., 2007). Many questions rise: What are the rest of the mechanisms involved in this cancer type? Are there other genes involved? How? This only accounts for one type of cancer, and there are at least 177 different types according to the National Cancer Institute [1]. The straightforward method to deal with this problem is to do wet lab experiments with large samples of normal and disease tissue, to test under different conditions the reactions, and check the expression or lack of it in different genes. The complication with this method is the cost, it takes time, it requires specialized equipment, and thus the economic price tag is high. Fortunately the bioinformatics area has acquired maturity during the recent years, biological data is becoming available in different formats throughout different databases and publications are providing new insights. Thanks to these, computational methods can be developed, methods that would save time, effort and money, methods that could help biomedical researchers get clues on which genes to explore on the wet laboratory, so that time is not wasted on genes that are unlikely to contribute in a given disease.

Gene Prioritization methods can be used to find genes that were previously unknown to be related to a given disease. The general definition of gene prioritization is: Given a disease D, a candidate gene set C, and the training data T, then input all these data to the method and it will compute a score for each of the candidate genes, higher scoring genes are supposed to be the genes that are most likely related to disease D, see fig. 1. Methods can be classified according to the type of input data that the method uses, as Text and Data Mining Methods and Network Based Methods. Text and Data Mining methods use training data like genetic localisation, gene expression, phenotypic data (van Driel et al., 2003), PubMeb abstracts (Tiffin et al., 2005), spatial gene expression profiles, linkage analysis (Piro et al., 2010), gene ontology and others (Adie et al., 2005; Ashburner et al., 2000; Schlicker et al., 2010); as the name suggests this

[1] http://www.cancer.gov/cancertopics/alphalist

methods mine the genome or mine the available biomedical literature to produce the scores of the candidate genes. Network Based Methods, use biological networks (Chen et al., 2009; Morrison et al., 2005) as the back bone of the prioritization method, however, some network based methods also combine some data and text mining techniques to improve their results (Aerts et al., 2006; Hutz et al., 2008).

The purpose of this chapter is to give an introduction into the Gene Prioritization Problem. Following the introduction a section explaining Biological Networks is presented as this is necessary background to understand the network based prioritization methods. After this section, we discuss about current state of the art prioritization methods with emphasis in network based methods. Next sections discuss our own prioritization method that is a network based method with a novel microarray data integration. A discussion on Challenges and Future Research opportunities follows and finally the conclusions of this chapter along with a list of available resources for Gene Prioritization.

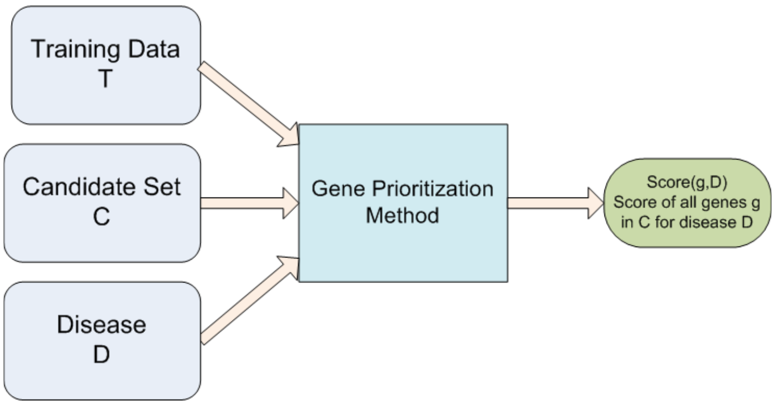


Fig. 1. General Gene Prioritization Overview

2. Biological networks

2.1 Graph theory background

A graph is a data structure that represents a set of relationships between elements or objects. Formally a graph G is a pair defined by $G = (V, E)$, where V is a set of elements that represent the nodes or vertices of the graph, the vertices in most applications hold the name of the attribute being represented. E is the set of edges, where each edge represent a relation between two vertices, an edge is defined by $E = \{(u, v)|u, v \in V\}$, this edges may hold additional information as weight, confidence or distance between nodes, therefore having $E = \{(u, v, w)|u, v \in V \text{ and } w \in Re\}$. The edges may represent direction, where $(u, v) \neq (v, u)$, in which case the graph is called directed graph, and when direction is not important, the graph is called undirected graph.

Graph Properties

Among the intrinsic properties of a graph we have: **Nodes**, the number of nodes in the network, formally $n = |V|$. **Edges**, the number of edges in the graph, formally $e = |E|$. **Connectivity** is a property of the graph, it is defined to be $connectivity(G) = \frac{e}{N}$ where $N = \binom{n}{2}$ is the maximum number of possible edges the graph can have. A graph with connectivity values closer to one would be called dense, and if the connectivity value is close to zero the graph would be called sparse, it is worth mentioning that there is no agreed exact value to consider a graph sparse or dense among the graph theory community. The

diameter of a graph is the distance of the longest shortest path on the graph. **Graph Path**, is a sequence of vertices of the form $\{v_1, v_2, v_3, ..., v_k\}$ where v_1 is the starting node and v_k is the destination node, and $(v_i, v_{i+1}) \in E$; the length of the path is defined by $l = \sum_{i=1}^{k-1} w_i$ where $w_i \in (v_i, v_{i+1}, w_i)$, when all weights are equal to 1 then the length of the path is $k-1$. A shortest path from vertex v to u is one of the paths that has the least accumulated weight from u to v, note that there can be multiple shortest paths from one node to another.

Nodes Properties

The most basic node property is the **degree** that denotes the number of connections a node has; in directed graphs there can be a distinction between incoming and outgoing connections, called **in-degree** and **out-degree** respectively. Several measures of centrality have been created to represent how "central" a node with respect to the other nodes in the graph, this measures are: **Closeness Centrality**, based in the average shortest path to the other vertices in the network; **Betweenness Centrality**, based on the occurrence of the vertex in the shortest paths of the network, **Eigenvalue Centrality**, based in the eigenvector of the adjacency matrix that represents the graph (Freeman, 1979).

2.2 Biological networks overview

In this section a brief background on biological networks is presented. As it was explained in the previous section, a graph, or network, is a set of relationship between objects, in the specific case of biological networks those objects are related to biological processes. Typical biological networks include: gene regulation networks, signal transduction networks, metabolic networks and protein interaction networks (PIN) (Junker & Schreiber, 2008). Gene regulation networks, also known as signal transcriptional regulation networks, represent how genes control the expression of other genes; these networks are often represented by directed graphs. Signal transduction networks are an extension of gene regulation networks that represents the links between intracellular processes to extracellular functions in response to diverse external events and stimuli. The final target in a signal transduction pathway is either a transcription factor or a metabolic enzyme. Metabolic networks are determined through biochemical experiments, and consist in metabolites converting into each other with the interaction of enzymes. The last of the typical biological networks are the PINs, they represent the interaction between different gene products, they are usually modeled with undirected networks, indicating only that there is a probable functional relation between the two related proteins without indicating direction. Some other networks exist that represent specific problem oriented networks, like (Yeh et al., 2009) that identifies genetic regulatory network in prostate cancer using microarray data. Fig. 2 and Fig. 3 show the general structure of a biological network and a sample of a PIN.

2.3 Protein interaction networks

These networks are the central focus of attention in the network based disease gene prioritization, so they deserve special attention. There are four main approaches to create PIN: high throughput technology, manual curation from published experiments results, automatic text mining from published literature and computational prediction from diverse genomic data (Wu et al., 2008). Some publicly available databases hold high quality, manually curated PIN, such as HPRD (Prasad et al., 2009), BIND (Bader et al., 2003) and BioGRID (Breitkreutz et al., 2008), in our work we have used BioIR (Liu et al., 2009) which integrates the previously mentioned databases along with DIP (Salwinski et al., 2004), IntAct (Aranda et al., 2010), MIPS (Pagel et al., 2005) and MINT (Ceol et al., 2010).

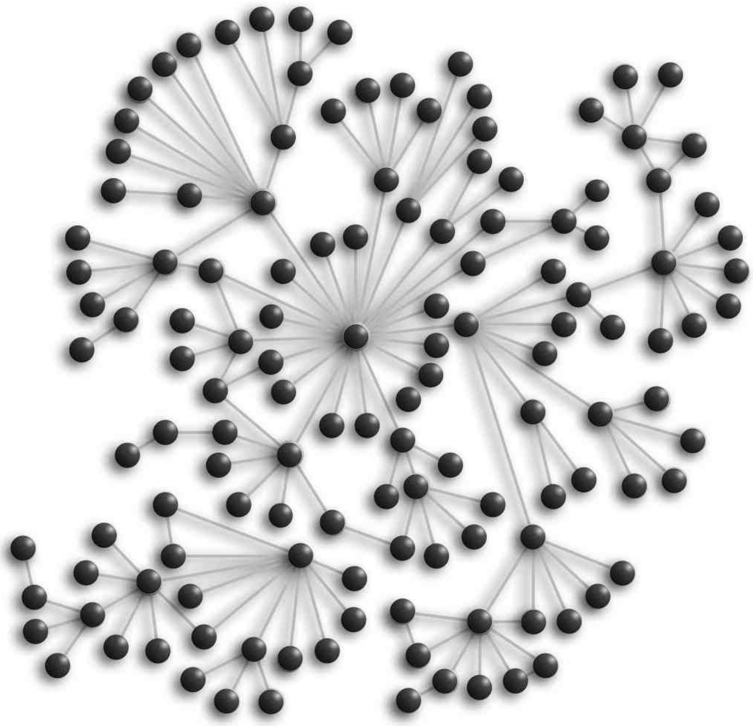


Fig. 2. Structure of a Biological Network

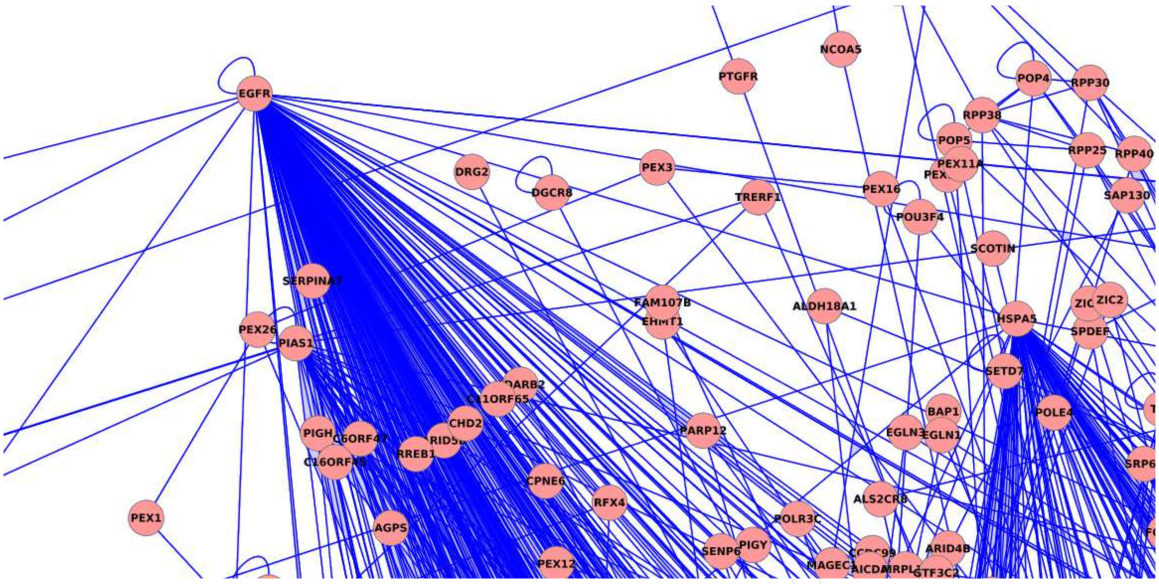


Fig. 3. Sample of Human PIN

2.3.1 Creation and curation of PIN

Current PIN databases are a rich resource of protein interactions, they mostly differ on the way they acquire their data, or on the way they validate it. For instance HPRD, BIND, BioGRID, MINT and MIPS are manually curated, this means a team of biologists check the literature to find new interactions, and once an interaction is confirmed it is added to the database. On the other hand DIP and IntAct are based on literature mining, they achieve this using computational methods that retrieve the interaction knowledge automatically from published papers. Another method to create PIN is using microarray data samples, these methods rely on the principle that co-expressing genes must be related, so by using statistical methods they can produce the list of likely relationships from the list of gene products in the microarray data.

2.3.2 Properties of PIN

PIN are known to have the following properties:

- Sparseness, although there is no one preset value of connectivity, it has been showed that biological networks are sparse containing much less than $O(n^2)$ edges in the network. Due to this property biological networks can be stored more efficiently in memory, and some algorithms exploit it to improve significantly their time performance.
- Small World, the concept was originated in the social sciences to explain how inside social networks the path length to go from one node to another is very small. However this is a subjective measure that lacks statistical or objective measure for actual networks. A more precise property that is seen in most empirical networks is Power law degree distribution, where the networks show that some few vertices have high degree and much more vertices have very low degree (Barabasi & Albert, 1999). Recent research has shown that biological networks do not necessarily follow power law degree distribution, but confirm that the distribution of degrees is heavy tailed (Garcia De Lomana et al., 2010).
- One of the disadvantageous properties of PIN is that they have a noisy nature, there is a enormous amount of missing information and false positives in the data (Edwards et al., 2002), therefore this fact must be taken into consideration when dealing with this kind of networks.
- As a consequence of the small world property, few nodes have high degree value, and it has been discovered that these nodes play an important role in the network, as opposed to other nonessential genes.
- Motifs, deep analysis in PIN has shown that there are recurrent subnetworks appearing in the full network, these subnetworks are called motifs. They have been discovered using statistical tools and showing that they occur more in the network than just by random coincidence (Junker & Schreiber, 2008).

3. Previous and on-going research

As was discussed in the introduction of the chapter prioritization methods can be classified as text and data mining based and network based methods. The main difference between the different approaches is the kind of data they use to do the prioritization of the candidate set.

3.1 Text and data mining methods

These methods usually rank candidate genes by matching their information and profile across multiple biological data sources. GeneSeeker is a web tool that selects candidate genes of the interest disease based on gene expression and phenotypic data from human and mouse (van Driel et al., 2003). eVOC system performs candidate gene selection based on the co-occurrence of disease name in PubMed abstracts through data-mining methods (Tiffin et al., 2005). DGP (Disease Gene Prediction) (López-Bigas & Ouzounis, 2004) and PROSPECTS (Adie et al., 2005) use basic sequence information to classify genes as likely or unlikely to be involved with the disease under study. The extended version of PROSPECTS, SUSPECTS (Adie et al., 2006), is developed by integrating annotation data from Gene

Ontology (GO) (Ashburner et al., 2000), InterPro and expression data. However, many of the methods suffer from limitations imposed by the data source which has little knowledge about the disease. GO terms include a brief description of the corresponding biological function of the genes but only 60% of all human genes have associated GO terms and these terms may be inconsistent due to differences in curators' judgment (Dolan et al., 2005). Due to the incomplete data, the approaches reduce the probability to rank the candidate genes of a specific disease.

Most recent methods of this kind of prioritization include MedSim (Schlicker et al., 2010) and a method based on spatially mapped gene expression (Piro et al., 2010). MedSim uses GO enrichment and applies their own similarities measures (Schlicker et al., 2007), by doing so they manage to extend the existing annotations to achieve the assignment of known disease genes to the correct phenotypes. The spatially mapped gene expression method uses a combination of data including linkage analysis, differential expression to acquire the list of candidates genes, then by using the phenotypes and associated phenotypes they find reference genes which in turn are filtered with the spatial gene-expression data; then by using both the candidate genes and the reference genes they apply their method to do the gene ranking.

3.2 Network based methods

As the name suggests these methods primarily use biological networks to do the prioritization process, this is mainly due to the increasing availability of human protein interaction data, and the emergence of network analysis. These methods usually rely on the the assumption that genes that are associated with diseases have a heavy interaction with each other (Erten & Koyutürk, 2010). Fig. 4 shows a general overview of these type of methods. The input that they commonly receive is the set of seed genes S that represent the previous knowledge to the method, genes that are known to be related to some disease D, along with these genes a score of how much they are related to the disease is given, denoted by $\sigma(v, D)$. The other part of the input is the genome of the organism represented by its PIN, denoted in the picture as the candidate set C. After the method calculates the score, just like any other method, it outputs the set of candidate genes with their score, where higher scores have higher probability of being related to disease D.

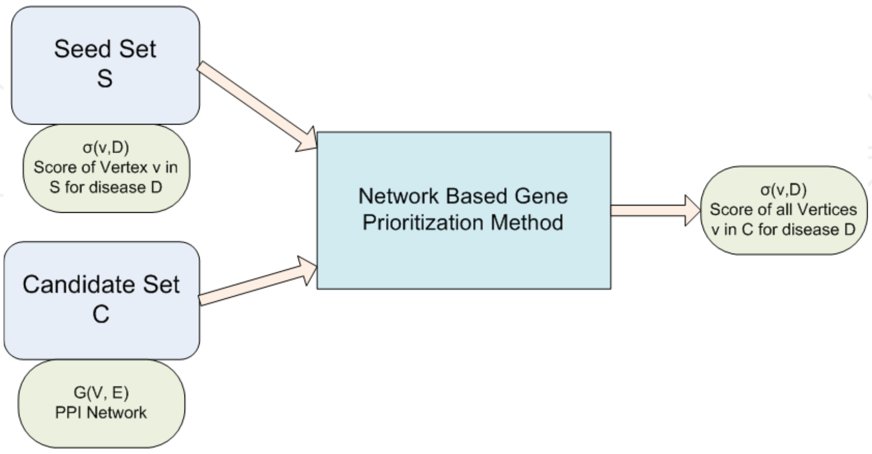

Seed Set
S
σ(v,D)
Score of Vertex v in S for disease D
Candidate Set
C
G(V, E)
PPI Network
Network Based Gene Prioritization Method
σ(v,D)
Score of all Vertices v in C for disease D

Fig. 4. Network Based Gene Prioritization Methods

Furthermore, network based methods can be classified in local and global methods. Local methods use local information to the seed genes, basically classifying by network proximity through the inspection of the direct neighbors of the seed genes or higher order neighbors in other words nodes in the network that are not directly adjacent to the seed nodes but are easily reached by them. Global methods model the flow through the whole network to provide a score of the connectivity and impact of the seed genes or previous knowledge on the rest of the nodes.

Network based gene prioritization is performed by assessing how much genes interact together and are close to known disease genes in protein networks. Endeavour takes a machine learning approach by building a model with initial known disease-related genes as training set, then that model is used to rank the test set of candidate genes according to the similarity score using multiple genomic data sources(Aerts et al., 2006). Chen et al. applied link based strategies widely used in social and web network analyses such as HITS with priors, PageRank, and K-step Markov to prioritize disease candidate genes based on protein networks (Chen et al., 2009). Ma et al. developed a system for gene prioritization by Combining Gene expression and protein-protein Interaction network (CGI) using Markov random field theory (Ma et al., 2007). CANDID used information from publications, protein domain descriptions, cross-species conservation measures, gene expression profiles and protein-protein interactions to do a prioritization algorithm on candidate genes that influence complex human traits (Hutz et al., 2008). GeneRank ranks genes based on Google's PageRank algorithm and expression data to do gene prioritization(Morrison et al., 2005). Ozgur et al. explored the connectivity properties of biological networks to compute an association score between candidate and disease-related genes (Özgür et al., 2008). Mani et al. proposed a method called Interactome Dysregulation Enrichment Analysis (IDEA) to predict cancer related genes using interactome and microarray data (Mani et al., 2008). Karni, Soreq, and Sharan attempted to predict the causal gene from expression profile data and they identified a set of disease-related genes that could best explain the expression changes of the disease-related genes in terms of probable pathways leading from the causal to the affected genes in the network (Karni et al., 2009).Tables 1 and 2 show a summary of the aforementioned methods.

4. Gene prioritization from microarray data based on shortest paths GP-MIDAS

In this section we present our current advances in our own method: **G**ene **P**rioritization from **MI**croarray **DA**ta on **S**hortest Paths (GP-MIDAS). Our approach differs from other network based methods in the way that we assign the weights to the edges of the PIN, by doing so we manage to get considerable performance compared to other state of the art methods.

4.1 Material

We applied GP-MIDAS for the study of prostate cancer, using the following data sources:

- **PIN**: Taking advantage of the availability of public protein interaction databases, and to have a more complete protein-protein interaction network, we integrated PIN data warehouse including HPRD, DIP, BIND, IntAct, MIPS, MINT and BioGrid databases which has successfully gathered 54,283 available and non-redundant PIN pairs among 10,710 proteins into BioIR database (Liu et al., 2009).
- **Microarray Data**: We integrated microarray data taken from (Lapointe et al., 2004) that consists of 72 primary tumors and 41 normal control sample in Stanford Microarray Database (SMD) (Hubble et al., 2009).

Method	**Brief Description**
Gene Seeker	Gene Expression and Phenotypic Data from Human and Mouse (van Driel et al., 2003)
eVOC	Co-Occurrence of disease name on PubMed abstracts (Tiffin et al., 2005)
DGP	Basic Sequence Information (López-Bigas & Ouzounis, 2004)
PROSPECTS	Basic Sequence Information (Adie et al., 2006)
SUSPECTS	Extension of PROSPECTS, incorporates GO (Adie et al., 2005; Ashburner et al., 2000)
MedSim	Gene Ontology enrichment with their functional similarity measures (Schlicker et al., 2010)
Spatially Mapped Expression	3D Gene Expression Data, Expression Profiles, Phenotype data (Piro et al., 2010)
Limitations	Generally imposed by the data source which carries little knowledge about the disease. For instance GO terms include brief description of the corresponding biological function of the genes but only 60% of all human genes have associated GO terms, and they may be inconsistent due to differences in curators' judgement (Dolan et al., 2005).

Table 1. Data and Text Mining Gene Prioritization Methods

- **Seed Genes**: The initial seed genes known to be related to the prostate cancer are extracted from public Online Mendelian Inheritance in Man (OMIM) database which stores gene-disease associations provided by summaries of publications. The list of the seed genes are shown in Table 3.
- **Test Genes**: We took the KEGG pathway database (Kanehisa et al., 2004) and PGDB database (Li et al., 2003) that are manually curated database for prostate cancer and obtained 102 genes as the truly disease-related genes for prostate cancer. We use this set to test the accuracy of our method.

4.2 Input preparation

The collected material needs to be prepared to be useful for our method, the details on this procedure are presented as follows.

4.2.1 Cope with missing values

The microarray dataset consists of N genes and M experiments and can be represented as an $M \times N$ matrix. It presents different gene expression levels $X_{ij} \mid (i \in M, j \in N)$ in this matrix. Gene expressions either over-expressed or under-expressed can be revealed in terms of two colored channel in the microarray data representing the intensity of the cancer and normal samples, with values ranging from 0 to 255. The gene expression ratios were calculated as

Method	**Brief Description**	**Data Sources**
Endeavor	Machine Learning: Using initially known disease genes; then multiple genomic data sources to rank (Aerts et al., 2006)	BIND
HITS with Priors PageRank K-Step Markov	Prioritization Based on Networks using Social and Web Networks Analysis (Chen et al., 2009)	HPRD, BIND, BioGRID
CGI	Combination of Protein Interaction Network and Gene Expression using Markov Random Field Theory (Ma et al., 2007)	MIPS, DIP
CANDID	Uses Publications, Protein domain descriptions, cross species conservation measures, gene expression profiles and Protein Interaction Networks (Hutz et al., 2008)	NCBI Conserved Domain Database
GeneRank	Based on Google's PageRank algorithm, uses expression data (Morrison et al., 2005)	GO and Synthetic Networks
IDEA	Uses the Interactome and Microarray data (Mani et al., 2008)	B Cell Interactome and OMIM
CIPHER	Based on the assembly of a Gene-Phenotype Network (Wu et al., 2008)	HPRD and OMIM
Özgür et al. (2008)	Using connectivity properties of the networks	Literature Mining by GIN
Karni et al. (2009)	Verifies expression changes of downstream genes	HPRD
Limitations	Most of these approaches include additional interactions predicted from co-expression, pathway, functional or literature data, but still fail to incorporate weights expressing the confidence on the evidence of the interactions.	
GP-MIDAS	Our proposed method, integrates Protein Interaction Network with Normal and Disease Microarray Data, using this integration we apply all-pairs shortest paths to find the significant networks and calculate the score for the genes.	

Table 2. Network Based Gene Prioritization Methods

Gene ID	Gene Symbol	Gene name
367	AR	Androgen receptor
675	BRCA2	Breast cancer type 2 susceptibility protein
3732	CD82	CD82 antigen
11200	CHEK2	Serine/threonine-protein kinase Chk2
60528	ELAC2	Zinc phosphodiesterase ELAC protein 2
2048	EPHB2	Ephrin type-B receptor 2 precursor
3092	HIP1	Huntingtin-interacting protein 1
1316	KLF6	Krueppel-like factor 6
8379	MAD1L1	Mitotic spindle assembly checkpoint protein MAD
4481	MSR1	Macrophage scavenger receptor types I and II
4601	MXI1	MAX-interacting protein 1
7834	PCAP	Predisposing for prostate cancer
5728	PTEN/ PTENP1	Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase and dual- specificity protein phosphatase PTEN
6041	RNASEL	2-5A-dependent ribonuclease
5513	HPC1	Hereditary prostate cancer 1

Table 3. Seed Genes of Prostate Cancer from OMIM Database

the median value of the pixels minus background pixel median value for one color channel divided by the same for the other channel because the mean value of the normalized ratio is much easier to be affected by noise than the median value. We applied the base-2 logarithmic transformation of each gene among experimental dataset and this value carried out the normalization of the gene expression value with mean 0 and standard deviation 1 in every experiment. Although microarray can be used to detect thousands of genes under a variety of conditions, there are still many missing values in microarray (Troyanskaya et al., 2001). The reasons for missing values include insufficient resolution, image corruption, and dust or scratches on the slide. If a gene contains many missing values in experiments, it is not easy to determine a precise expression value for each gene that causes a difficulty in the subsequent analysis of the regulation networks. However, we can not simply remove all gene data that contains missing values because the number of remaining genes will become too small to predict the network correctly. In order to get a better result, the genes that contain less than 20% entries missing in all experiment are picked. In order to get as complete data as possible, we use the K-Nearest-Neighbors (KNN) algorithm (Troyanskaya et al., 2001) to estimate the missing values.

4.2.2 Update of microarray expression values

Once the necessary microarray data is collected, we need to preprocess it, so that it becomes ready to be used in our methods. The preprocessing procedure consists of two steps:

1. Transform the Microarray Data Expression Values. The purpose of this transformation is to make the expression values ready to be used as weights in the network. This transformation has two steps, initially the values are updated using a sample of normal expression microarray data, the effect of this operation is that values that are very similar between normal and cancer samples should have less impact on our analysis. To accomplish this we subtract the value from the cancer microarray data to the value of

the normal expression data as shown in Equation 1. The next step is to transform the values, the rationale behind this transformation is that expression values may be negative for under expressed genes, and if these values are used as they are, our network may have negative weights, thus making shortest paths analysis more difficult. Equation 2 shows how the expression values are transformed.

$$ExpressionValue_i = ExpressionValue_i - NormalExpresionValue_i \tag{1}$$

$$TransformedExpresionValue_i = -\ln\left(\frac{|\ ExpresionValue_i\ | - min}{max - min}\right) \tag{2}$$

Considering that the sign of of the value in the microarray data represents over or under expression, and the fact that we want to make a representation of distance, for this is what we want in our quantitative analysis, we use the absolute value of the microarray data, then these results are normalized, using the *max* and *min* values found, by doing these two steps we get values in the range $[0, 1]$, where values closer to 1 mean that they are more expressed (either over expressed or under expressed). Finally we compute the negative of the natural logarithm on the previous results, this is to make smaller numbers (less expression level) become large distances, and bigger numbers (higher expression level) become short distances. The result of this step is a transformation of the gene expression, where more expressed genes have smaller value, and less expressed genes have higher values, in the next step we convert this values into distances between genes, thus more expressed genes relationships will become shorter distances than less expressed genes relationships. In the case the $ExpresionValue_i == min$ we just set the whole result to be a big value, since $\ln(0)$ is not defined.

2. Convert to Human Protein Interaction Network to a Weighted Network. Since we need the network to become a weighted one, where these weights are related to the specific interactions in cancer related network, we use the transformed values of the microarray data. However the microarray data provides transformed expression values for the genes, not for the relationship between genes. The Pearson correlation coefficient for analyzing gene-pair relationships could be unsuitable to explore the true gene relationship because it is overly sensitive to the shape of an expression curve (Kim et al., 2007).To overcome this issue, we combine the values of the two interacting genes together. For instance if we have microarray values {(SEPW1, 4.097), (BRCA1, 1.395), (AKT1, 2.006), (BACH1, 2.823), (AHNAK, 3.597)} and we have the following edges in our graph {(AKT1, AHNAK), (BACH1, BRCA1), (BRCA1, AKT1)}, then the first edge weight would be the addition of the transformed expression values of each of the vertices 2.006 + 3.597 = 5.603 providing the weight of the first edge. The resulting weighted edges of this instance would be {(AKT1, AHNAK, 5.603), (BACH1, BRCA1, 4.218), (BRCA1, AKT1, 3.401)}.

4.3 Method description

Our current method is based on the analysis of the shortest paths between all the pairs of the genes on the input network.

4.3.1 Shortest paths analysis

Genes co-occurring in a particular network tend to participate together in related biological processes based on their linkage with the known disease genes (Tin et al., 2009). Our methodology is based in the the computation of all pairs shortest paths (APSP) in the network and the retrieval of these paths for posterior analysis in our experiments. The computation of APSP is carried out using our implementation of the all pairs shortest paths algorithm (Arias & Soo, 2010), which takes advantage of the topology and special characteristics of biological networks, such as sparseness and singles, nodes that have only one connection. We call our implementation KC-APSP. At this step of the process we input our prepared PIN and get as result a list of all the shortest paths between all the pair combination of the genes.

4.3.2 Scoring of genes on shortest paths

Once all the shortest paths are computed, we traverse the list of shortest paths (*PathList*), to verify if any of the seed genes are on the resulting paths, if so, these paths need to be considered for the scoring. Finally a score is computed for each gene. This analysis is done across M microarray data experiments.

4.3.3 Compute the score function

Having all the paths stored in $PathList_m$ for $m \in M$ we can compute the denominator $denom_m$, to be used in the score function using Equation 3, this is done for each microarray experiment m.

$$denom_m = \sum_{i=1}^{n} \frac{1}{l_{im}} \tag{3}$$

Where l_{im} is the length of the i^{th} path in sample $m \in M$ for n generated and filtered by seed set paths. Once the denominator is ready, we proceed to compute the score. For each experiment m of M, and for each gene g on the network we compute the score for each gene according to Equation 4.

$$Score(Gene_{i,m}) = \sum_{Gene_{i,m} \in Path_{j,m}}^{PathList_m} \frac{\frac{1}{l_{j,m}}}{denom_m} \tag{4}$$

The motivation behind Equation 4 is that a gene that appears in more generated paths is going to achieve higher score, even higher for paths with shorter length, the highest being 1 if the gene appears in all the found paths.

5. Current results

In this section we present our current results, first we discuss the leave one out cross validation, and lastly we present the precision and recall of our method compared to other methods that use similar data sources to GP-MIDAS. As it is shown in this section, our method presents promising results that can lay the foundation for more advanced and accurate approaches.

5.1 Leave-one-out cross validation of our method

The performance of our algorithm was evaluated by leave-one-out cross validation method. In each experimental test on a known-disease gene set S, known as the Seed Set, which contains $|S|$ genes; we delete one gene g from the Seed Set thus having $S' = S - g$. We used S' set to train our prioritization model. Then, we prioritized the Candidate Gene Set to determine the rank of that deleted gene g. We got 100% to cover the deleted genes from the Candidate Gene Set and the rankness of those seed genes are listed in Table 4; LOO Score Position refers to the result of GP-MIDAS after deleting the given gene g from the seed set, Closeness Centrality

Gene	Recovered Subnetworks		All Seed Genes Subnet.
	LOO Score Position	Closeness Centrality Position	Closeness Centrality Position
AR	1	2	2
PTEN	6	23	24
BRCA2	10	36	26
EPHB2	13	46	42
HIP1	14	43	43
CHEK2	15	35	34
RNASEL	19	53	53
MXI1	20	58	58
MAD1L1	21	53	47
ELAC2	22	61	63
KLF6	26	42	40
MSR1	27	61	62
CD82	33	54	50

Table 4. Leave One Out Experiment Results

on Recovered Network refers to this centrality measure on the induced subgraph made from the seed set without the given gene *g* and all the shortest paths generated by this set pairs, and the last column refers to the centrality measure position on the full network for the given gene *g*. In order to realize the performance of gene prioritization with the weighted graph based on the gene expression, we compare the closeness centrality position in the entire PIN and sub-networks reconstructed from seed genes. From the entire network and sub-network of the seed genes using closeness properties, only 1 original seed genes rank among its top 20 ranking genes. However, we recover 8 seed genes among top 20 ranking genes. The results confirmed that PIN without any gene expression have more false positive and our method integrated gene expression is potentially able to perform better in the identification of genes associated with a given disease and should be more informative.

5.2 The precision and recall comparison with previous methods

We evaluated the performance of our algorithm in terms of overall precision versus recall when varying the rank threshold. Precision is the fraction of true gene-disease associations that ranked within the top k% in the corresponding trial of the cross validation procedure. Recall is the fraction of trials in which the disease-related genes from PGDB was recovered as one of the top k% scoring ones. We compare the performance with the following network based methods: GeneRanker, ENDEAVOUR, HITS with priors, PageRank, K-step Markov and CIPHER. In GeneRanker, we do not use 543 genes reported to be associated to prostate cancer in the literature but applied the seed as presented in the list of genes in Table 3. We set a back probability of 0.3 for PageRank with priors and HITS with priors, this value is selected because (Chen et al., 2009) express that this is the optimal value for back probability, and step size 6 for K-Step Markov method in ToppNet. Further, we reason that the use of literature evidence in this benchmark test would unfairly improve ENDEAVOUR's performance because these literatures may include direct evidence that reports the association between the gene and the disease. Neither one of these methods rank the seed genes therefore we only compare the performance with all the genes in the candidate set except the seed genes. Among the top 10 genes, we got 4 prostate cancer-related genes while applying both normal and cancer

	Precision (%)				
Top K	10	20	30	40	50
Methods					
HITS with Priors	40	25	23.3	17.5	18
K-Step Markov	20	20	20	22.5	18
PageRank	20	15	20	17.5	18
GeneRanker	30	20	20	20	18
Endeavour	10	15	20	20	20
CIPHER	10	10	10	20	20
GP-MIDAS	40	30	23.3	17.5	20

Table 5. Precision for Top K Rank Comparison Across Methods

	Recall (%)				
Top K	10	20	30	40	50
Methods					
HITS with Priors	4.5	5.6	7.9	7.9	10.1
K-Step Markov	2.2	4.5	6.7	10.1	10.1
PageRank	2.2	3.4	6.7	7.9	10.1
GeneRanker	3.4	4.5	6.7	9	10.1
Endeavour	1	2.9	5.9	7.8	14.6
CIPHER	1.1	2.2	3.4	9.0	11.2
GP-MIDAS	4.5	6.7	7.9	7.9	14.6

Table 6. Recall for Top K Rank Comparison Across Methods

samples and the performance is equal to the HITS with priors which is the highest one from the previous methods. We also get the highest precision among the top 50 ranking genes. Tables 5 and 6 denote that our method gets the highest precision and recall. Using the different expression values between cancer and normal samples may help us to extract more significant genes and rank them to be higher. Fig. 5 shows the combined precision and recall value using F-Measure, in the figure can be clearly seen that in most instances GP-MIDAS outperforms other methods.

6. Challenges and future research opportunities

As it was previously discussed gene prioritization methods can be either data and text mining based or network based, however the division line between these two approaches is less clear every day as some methods integrate both approaches and use more information to improve the accuracy of the results. Despite the increase of accuracy, the main challenge is to find novel genes that are actually involved with a given disease, genes that have not been reported before, presenting the problem of proving that the newly found genes are in fact involved with the disease. Therefore it becomes essential to present more and better biological explanations on the definition of newer approaches, by doing so biomedical researchers will have more confidence in trying the novel genes in in-vitro experiments. One clear research opportunity is presented, and it is the combination of different network based approaches, by using local and global information. The work of (Erten & Koyutürk, 2010) shows promising results, aiming at the discovery of loosely connected genes using statistical correction schemes that help overcome the preference of straightforward method for genes with high centrality values; this

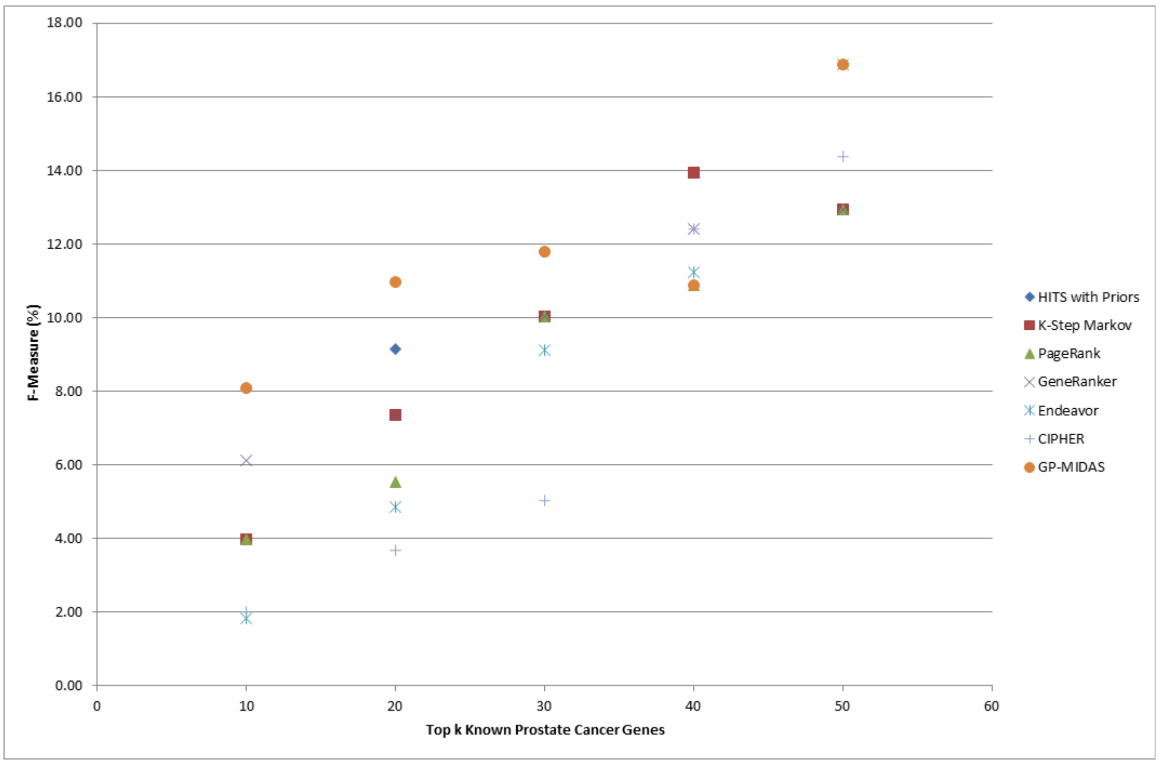


Fig. 5. F-Measure of different methods applied to Prostate Cancer

aproach uses only global methods, (Navlakha & Kingsford, 2010) combines several network based methods to produce a new score which also shows a potential new research line.

7. Conclusions

The past few years have shown an increasing interest in the disease gene prioritization problem and thanks to the availability of more and better data sources there has been a growing number of methods and approaches to this problem. A plethora of methods have become available for the genetics disease research community, and as the methods become more mature the results will become increasingly accurate and more biologically meaningful. Our own approach GP-MIDAS has proven to be promising showing a better performance in most instances to related methods, exposing that by setting the weights of the PIN to have more related meaning to the given disease the results can be better than previous plain shortest paths methodologies.

8. Acknowledgments

This research is partially supported by the Bioresources Collection and Research Center of Linko Chang Gung Memorial Hospital and National Tsing Hua University of Taiwan R. O. C. under the grant number 98N2424E1, and the Universidad Tecnológica Centroamericana in Honduras.

Resources

In this last section we present online resources, please note that since this is an evolving field, some of these resources can change with time. For a list of projects hosting biological networks see Table 7; these sites have the capability of being queried for specific proteins, or the user can also download the interaction network that is needed for his particular research. For a list of sites that offer online diseases information or software tools for disease information see Table 8. For a list of sites that offer online ontologies or software tools for ontologies see Table 9. And for a list of sites that offer online prioritization or software tools to do prioritization see Table 10.

Human Protein Reference Database	HPRD	http://www.hprd.org
Biomolecular Interaction Network Database	BIND	http://bond.unleashedinformatics.com
Biological General Repository for Interaction Datasets	BioGRID	http://thebiogrid.org/
Database of Interacting Proteins	DIP	http://dip.doe-mbi.ucla.edu/
IntAct Molecular Interaction Database	IntAct	http://www.ebi.ac.uk/intact/
The MIPS Mammalian Protein-Protein Interaction Database	MIPS	http://mips.helmholtz−muenchen.de/proj/ppi/
Molecular Interaction Database	MINT	http://mint.bio.uniroma2.it/mint/
Kyoto Encyclopedia of Genes and Genomes	KEGG	http://www.genome.jp/kegg/
National Center for Biotechnology Information	NCBI	http://www.ncbi.nlm.nih.gov/

Table 7. Available Biological Networks Sites

Online Mendelian Inheritance in Man	OMIM	http://www.ncbi.nlm.nih.gov/omim
The Human Gene Compendium	GeneCards	http://www.genecards.org/
Genetic Association Database	GAD	http://geneticassociationdb.nih.gov/
Catalog of Published Genome Wide Association Studies	GWAS	http://www.genome.gov/gwastudies/

Table 8. Available Disease Information Sites

Gene Ontology (The Gene Ontology Consortium, 2008)	http://www.geneontology.org/
eVOC Ontology (Kelso et al., 2003)	http://www.evocontology.org/
InterPro (Hunter et al., 2009)	http://www.ebi.ac.uk/interpro/

Table 9. Available Biological Ontology Sites

MedSim	http://www.funsimmat.de/
Endeavor	http://homes.esat.kuleuven.be/ bioiuser/endeavour/index.php
ToppGene	http://toppgene.cchmc.org/
Cypher	http://rulai.cshl.edu/tools/cipher/
CANDID	https://dsgweb.wustl.edu/hutz/candid.html
SUSPECTS	http://www.genetics.med.ed.ac.uk/suspects/
GP-MIDAS	http://bioir.cs.nthu.edu.tw/bne

Table 10. Available Gene Prioritization Sites

9. References

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J. & Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization., *BMC Bioinformatics* 6(1). URL: *http://dx.doi.org/10.1186/1471-2105-6-55*

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J. & Pickard, B. S. (2006). Suspects: enabling fast and effective prioritization of positional candidates., *Bioinformatics* 22(6): 773–4. URL: *http://www.biomedsearch.com/nih/SUSPECTS-enabling-fast-effective -prioritization/16423925.html*

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P. & Moreau, Y. (2006). Gene prioritization through genomic data fusion., *Nature biotechnology* 24(5): 537–544. URL: *http://dx.doi.org/10.1038/nbt1203*

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K. & Hermjakob, H. (2010). The IntAct molecular interaction database in 2010, *NUCLEIC ACIDS RESEARCH* 38(Suppl. 1): D525–D531.

Arias, C. R. & Soo, V.-W. (2010). Computing all pairs shortest paths on graphs with articulation points, *Under Review* .

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium., *Nature genetics* 25(1): 25–29.
URL: *http://dx.doi.org/10.1038/75556*

Bader, G., Betel, D. & Hogue, C. (2003). BIND: the Biomolecular Interaction Network Database, *NUCLEIC ACIDS RESEARCH* 31(1): 248–250.

Barabasi, A. & Albert, R. (1999). Emergence of scaling in random networks, *SCIENCE* 286(5439): 509–512.

Breitkreutz, B.-J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V., Dolinski, K. & Tyers, M. (2008). The BioGRID interaction database: 2008 update, *NUCLEIC ACIDS RESEARCH* 36(Sp. Iss. SI): D637–D640.

Ceol, A., Aryamontri, A. C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L. & Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update, *NUCLEIC ACIDS RESEARCH* 38(Suppl. 1): D532–D539.

Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. (2009). Toppgene suite for gene list enrichment analysis and candidate gene prioritization., *Nucleic acids research* 37(Web Server issue): gkp427+.
URL: *http://dx.doi.org/10.1093/nar/gkp427*

Dolan, M. E., Ni, L., Camon, E. & Blake, J. A. (2005). A procedure for assessing go annotation consistency, *Bioinformatics* 21(suppl_1): i136–143.
URL: *http://dx.doi.org/10.1093/bioinformatics/bti1019*

Edwards, A., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J. & Gerstein, M. (2002). Bridging structural biology and genomics: assessing protein interaction data with known complexes, *TRENDS IN GENETICS* 18(10): 529–536.

Erten, S. & Koyutürk, M. (2010). Role of centrality in network-based prioritization of disease genes, *in* C. Pizzuti, M. Ritchie & M. Giacobini (eds), *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, Vol. 6023 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 13–25.

Freeman, L. (1979). Centrality in social networks: Conceptual clarification, *Social Networks* 1(3): 215–239.
URL: *http://dx.doi.org/10.1016/0378-8733(78)90021-7*

Friedman, L., Ostermeyer, E., Szabo, C., Dowd, P., Lynch, E., Rowell, S. & King, M. (1994). Confirmation of brca1 lay analysis of germline mutations linked to breast and ovarian-cancer in 10 families, *Nature genetics* 8(4): 399–404.

Garcia De Lomana, A. L., Beg, Q. K., De Fabritiis, G. & Villa-Freixa, J. (2010). Statistical Analysis of Global Connectivity and Activity Distributions in Cellular Networks, *Journal of Computational Biology* 17(7): 869–878.

Hubble, J., Demeter, J., Jin, H., Mao, M., Nitzberg, M., Reddy, T. B., Wymore, F., Zachariah, Z. K., Sherlock, G. & Ball, C. A. (2009). Implementation of genepattern within the stanford microarray database., *Nucleic acids research* 37(Database issue).
URL: *http://dx.doi.org/10.1093/nar/gkn786*

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H. & Yeats, C. (2009). InterPro: the integrative protein signature database, *NUCLEIC ACIDS RESEARCH* 37(Sp. Iss. SI): D211–D215.

Hutz, J., Kraja, A., McLeod, H. & Province, M. (2008). Candid: a flexible method for prioritizing candidate genes for complex human traits., *Genetic Epidemiology* 32(8): 779–790.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/18613097*

Junker, B. H. & Schreiber, F. (eds) (2008). *Analysis of Biological Networks*, Wiley Series on Bioinformatics: Computational Techniques and Engineering, John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. (2004). The kegg resource for deciphering the genome., *Nucleic acids research* 32(Database issue): D277–280.
URL: *http://dx.doi.org/10.1093/nar/gkh063*

Karni, S., Soreq, H. & Sharan, R. (2009). A network-based method for predicting disease-causing genes, *Journal of Computational Biology* 16(2): 181–189.
URL: *http://dx.doi.org/10.1089/cmb.2008.05TT*

Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., Jongeneel, C., McCarthy, M., Hide, T. & Hide, W. (2003). eVOC: A controlled vocabulary for unifying gene expression data, *GENOME RESEARCH* 13(6): 1222–1230.

Kim, K., Jiang, K., Zhang, S., Cai, L., beum Lee, I., Feldman, L. & Huang, H. (2007). Measuring similarities between gene expression profiles through new data transformations, *BMC Bioinformatics* 8: 29.

Lapointe, J., Li, C., Higgins, J., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A., Tibshirani, R., Botstein, D., Brown, P., Brooks, J. & Pollack, J. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer, *Proceedings of the National Academy of Sciences of the United States of America* 101(3): 811–816.

Li, L.-C., Zhao, H., Shiina, H., Kane, C. J. & Dahiya, R. (2003). Pgdb: a curated and integrated database of genes related to the prostate, *Nucleic Acids Research* 31(1): 291–293.

Liu, H.-C., Arias, C. R. & Soo, V.-W. (2009). Bioir: An approach to public domain resource integration of human protein-protein interaction, *The proceeding of the Seventh Asia Pacific Bioinformatics Conference*.

López-Bigas, N. & Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease., *Nucleic Acids Res* 32(10): 3108–3114.
URL: *http://dx.doi.org/10.1093/nar/gkh605*

Ma, X., Lee, H., Wang, L. & Sun, F. (2007). Cgi: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data, *Bioinformatics* 23(2): 215–221.
URL: *http://dx.doi.org/10.1093/bioinformatics/btl569*

Mani, K. M., Lefebvre, C., Wang, K., Lim, W. K. K., Basso, K., Dalla-Favera, R. & Califano, A. (2008). A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas., *Molecular systems biology* 4.
URL: *http://dx.doi.org/10.1038/msb.2008.2*

Morrison, J. L., Breitling, R., Higham, D. J. & Gilbert, D. R. (2005). Generank: using search engine technology for the analysis of microarray experiments., *BMC Bioinformatics* 6: 233. URL: *http://www.biomedsearch.com/nih/GeneRank-using-search-engine-technology/16176585.html*

Navlakha, S. & Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases, *BIOINFORMATICS* 26(8): 1057–1063.

Oldenburg, R. A., Meijers-Heijboer, H., Cornelisse, C. J. & Devilee, P. (2007). Genetic susceptibility for breast cancer: How many more genes to be found?, *CRITICAL REVIEWS IN ONCOLOGY HEMATOLOGY* 63(2): 125–149.

Özgür, A., Vu, T., Erkan, G. & Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network, *ISMB*, pp. 277–285.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H., Ruepp, A. & Frishman, D. (2005). The MIPS mammalian protein-protein interaction database, *BIOINFORMATICS* 21(6): 832–834.

Piro, R. M., Molineris, I., Ala, U., Provero, P. & Di Cunto, F. (2010). Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR, *BIOINFORMATICS* 26(18): I618–I624. 9th European Conference on Computational Biology, Ghent, BELGIUM, SEP 26-29, 2010.

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. & Pandey, A. (2009). Human Protein Reference Database-2009 update, *NUCLEIC ACIDS RESEARCH* 37(Sp. Iss. SI): D767–D772.

Salwinski, L., Miller, C., Smith, A., Pettit, F., Bowie, J. & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update, *NUCLEIC ACIDS RESEARCH* 32(Sp. Iss. SI): D449–D451.

Schlicker, A., Lengauer, T. & Albrecht, M. (2010). Improving disease gene prioritization using the semantic similarity of Gene Ontology terms, *BIOINFORMATICS* 26(18): i561–i567. 9th European Conference on Computational Biology, Ghent, BELGIUM, SEP 26-29, 2010.

Schlicker, A., Rahnenfuehrer, J., Albrecht, M., Lengauer, T. & Domingues, F. S. (2007). GOTax: investigating biological processes and biochemical activities along the taxonomic tree, *GENOME BIOLOGY* 8(3).

The Gene Ontology Consortium (2008). The gene ontology project in 2008, *Nucl. Acids Res.* 36(36 Database issue): 440–444.
URL: *http://dx.doi.org/10.1093/nar/gkm883*

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B. & Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates., *Nucleic acids research* 33(5): 1544–1552.
URL: *http://dx.doi.org/10.1093/nar/gki296*

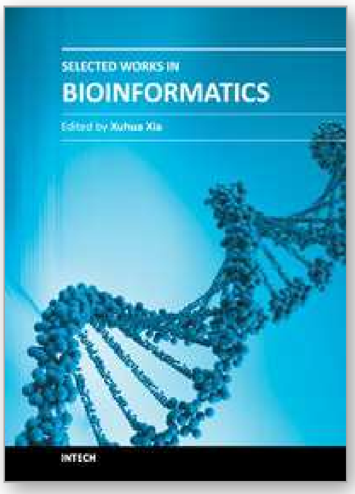
Tin, N., Andrade, M. A. & Perez-Iratxeta, C. (2009). Linking genes to diseases: it's all in the data, *Genome Medicine* 1(8): 77.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. B. (2001). Missing value estimation methods for dna microarrays., *Bioinformatics (Oxford, England)* 17(6): 520–525.
URL: *http://dx.doi.org/10.1093/bioinformatics/17.6.520*

van Driel, M., Cuelenaere, K., Kemmeren, P., Leunissen, J. & Brunner, H. (2003). A new web-based data mining tool for the identification of candidate genes for human genetic disorders, *EUROPEAN JOURNAL OF HUMAN GENETICS* 11(1): 57–63.

Wu, X., Jiang, R., Zhang, M. Q. & Li, S. (2008). Network-based global inference of human disease genes, *MOLECULAR SYSTEMS BIOLOGY* 4.

Yeh, H.-Y., Cheng, S.-W., Lin, Y.-C., Yeh, C.-Y., Lin, S.-F. & Soo, V.-W. (2009). Identifying significant genetic regulatory networks in the prostate cancer from microarray data based on transcription factor analysis and conditional independency, *BMC MEDICAL GENOMICS* 2.



Selected Works in Bioinformatics
Edited by Dr. Xuhua Xia

ISBN 978-953-307-281-4
Hard cover, 176 pages
Publisher InTech
Published online 19, October, 2011
Published in print edition October, 2011

This book consists of nine chapters covering a variety of bioinformatics subjects, ranging from database resources for protein allergens, unravelling genetic determinants of complex disorders, characterization and prediction of regulatory motifs, computational methods for identifying the best classifiers and key disease genes in large-scale transcriptomic and proteomic experiments, functional characterization of inherently unfolded proteins/regions, protein interaction networks and flexible protein-protein docking. The computational algorithms are in general presented in a way that is accessible to advanced undergraduate students, graduate students and researchers in molecular biology and genetics. The book should also serve as stepping stones for mathematicians, biostatisticians, and computational scientists to cross their academic boundaries into the dynamic and ever-expanding field of bioinformatics.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821