Tel-Aviv University
The Raymond and Beverly Sackler Faculty of Exact Sciences
The Blavatnik School of Computer Science

A Network Based Method for Predicting Disease Causing Genes

This thesis is submitted in partial fulfillment
of the requirements towards the M.Sc. degree
Tel-Aviv University
The Blavatnik School of Computer Science

by
Shaul Karni

The research work in this thesis has been carried out
under the supervision of Prof. Roded Sharan.

January, 2009

Acknowledgments:

I would like to thank my thesis advisor, Roded Sharan, for the initial idea, the close guidance and, and the fruitful ideas. I learned from him a lot while working on this thesis.
I would also like to thank Hermona Soreq for her comments and ideas.

Abstract

A fundamental problem in human health is the inference of disease-causing genes, with important applications to diagnosis and treatment. Previous work in this direction relied on knowledge of multiple loci associated with the disease, or causal genes for similar diseases, which limited its applicability. Here we present a new approach to causal gene prediction that is based on integrating protein-protein interaction network data with gene expression data under a condition of interest. The latter are used to derive a set of disease-related genes which is assumed to be in close proximity in the network to the causal genes. Our method applies a set-cover-like heuristic to identify a small set of genes that best "cover" the disease-related genes. We perform comprehensive simulations to validate our method and test its robustness to noise. In addition, we validate our method on real gene expression data and on gene specific knockouts. We apply it to suggest possible genes that are involved in Myasthenia Gravis. Finally, we use a large collection of gene expression data sets for different diseases to study the topological characteristics of the causal genes as a function of disease type and stage.

Contents

Chapter 1

Introduction

High-throughput technologies such as yeast two-hybrid screens [12] and co-immunoprecipitation [22] are routinely used nowadays to map molecular interactions within the cell. Applications of these maps include the prediction of protein function [41] and orthology [4], the inference of protein modules [40] and more.

In the last two years, global maps of protein-protein interactions (PPIs) have become available for human [38, 44], leading to an array of works aiming at harnessing PPI data to improve the understanding of human disease. In particular, many authors have shown the utility of these networks in inferring disease-causing genes. Franke et al. [13] considered diseases with several associated loci. For such diseases they aimed at identifying a set of genes, spanning the associated loci, whose protein products are connected in a functional network, comprised of PPIs, co-expression relations and gene-ontology similarities. Lage et al. [21] integrated PPI data with information on the phenotype similarity of different diseases. They developed an algorithm for predicting causal genes that relies on the observation that genes causing similar diseases tend to be connected in a PPI network. Kohler el al. [20] grouped diseases into families where using a random walk method from known genes in its family to prioritize candidate genes. Wu et al. [49] scores a candidate gene based on the correlation between the vector of similarities to diseases with known causal genes using a propagation method. Mani et al. [27] used gene expression data in combination with molecular interaction data to identify interactions that exhibit a gain or a loss of expression correlation in a given phenotypic class. They then ranked genes according to the enrichment of their direct neighborhood with such interactions.

Here we present a new algorithm for predicting disease-causing genes. Rather than assuming information on disease loci, or on gene-disease associations, we make use of the abundant information on genes that change their expression levels within the affected tissue under the disease state. We call the latter *disease-related genes.* Our algorithm relies on the assumption that in the disease state, one or more causal genes are disrupted, leading to the expression changes of downstream (disease-related) genes through signaling-regulatory pathways in the network. To uncover the causal genes, we make a parsimonious assumption, seeking the smallest set of 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 a network of physical interactions. Ideally, this network should contain protein-protein and protein-DNA interactions. However, the latter are not available at large scale for human. Hence, in practice, we use PPI data only (see Chapter 6 for further discussion).

In simulations, our algorithm attains very high accuracy on a wide range of parameters, including the size of the input affected set, the noise level within the set, the size of the search space, and the number of causal genes simulated. In validation on real expression data from knockout experiments, our algorithm manages to pinpoint the disrupted gene with high accuracy. Further validations on expression data from different types of cancer show high accuracy in pinpointing known oncogenes. Importantly, we show that our method outperforms a naïve algorithm that ranks disease-associated genes according to their distances in the network to the directly affected genes.

Next, we apply our method to suggest possible genes that are involved in Myasthenia Gravis, a neuromuscular autoimmune disease, for which the causal genes are yet unknown. We also use a large collection of gene expression data sets for different diseases to study the topological characteristics of the causal genes as a function of disease type and stage. We find that monogenic diseases and diseases at early stages exhibit local gene expression patterns, whereas multifactorial and late-stage diseases implement wide changes in gene expression that are not limited to a certain area of the network.

The thesis is organized as follows: Chapter 2 gives a general biological background relevant to the issues addressed in this work. Chapter 3 defines the problem at hand and the algorithm used to tackle it. Chapter 4 describes the results on simulated real data. Chapter 5 provides the results of a large scale analysis of gene expression data. Finally chapter 6 concludes this work.

Part of the results in this thesis have been accepted for publication in the Journal of Computational Biology [19].

Chapter 2

Biological Background

This work is based on integrating protein-protein interactions (PPIs) and gene expression information. Protein-protein interactions refer to the physical associations between protein molecules. These interactions are critical to all cellular processes and along with protein-DNA interactions account for a large part of the dynamics of a cell. For example, signals from outside the cell are propagated inside the cell and to their end targets by interactions along protein signalling pathways. Proteins also interact with one another to form protein complexes, the backbone of most cellular machinery.

Gene expression is a measurement of a gene's transcription level. This is used as a measure of the amount of proteins translated and in turn, as a measure of the activity level of a protein.

2.1 Protein-Protein Interaction Networks

Protein-protein interaction data have increased dramatically throughout the last few years. Along with existing interaction information achieved through decades of molecular biology (e.g. [34, 32]), there are two main high throughput methods for detecting protein interactions: *yeast two-hybrid* (*Y2H*) and *co-immunoprecipitation* (*CoIP*).

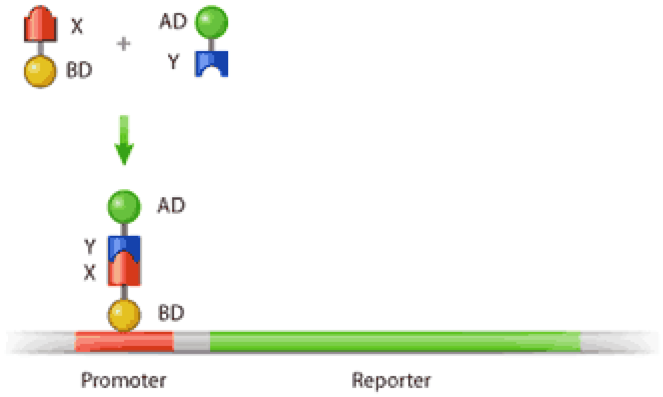

Figure 2.1: The yeast two hybrid process. The expression of the reporter gene is measured in order to detect protein-protein interactions. If proteins X and Y interact, their DNA-binding domain and activation domain will combine to form a functional transcriptional activator, which will induce the transcription of the reporter gene (figure taken from [42]).

2.1.1 Yeast Two-Hybrid

The yeast two-hybrid technique [12] allows the detection of binary protein protein interactions. This is done by causing a reporter gene to be expressed only if two examined proteins physically interact. Y2H uses two protein domains of the yeast GAL4 protein that have specific functions: a DNA-binding domain, capable of binding to a DNA sequence, and an activation domain, capable of activating transcription of the monitored gene. The gene transcription process can only occur when both domains are present. Thus, the two proteins of interest are attached to binding and activation domains of the reporter gene. The protein that is attached to the DNA-binding domain is called the *bait*, and the protein that is attached to the activation domain is call the *prey*. If they interact, an active transcription unit will be formed and the reporter gene will be expressed, forming a protein product that can be detected (see Figure 2.1).

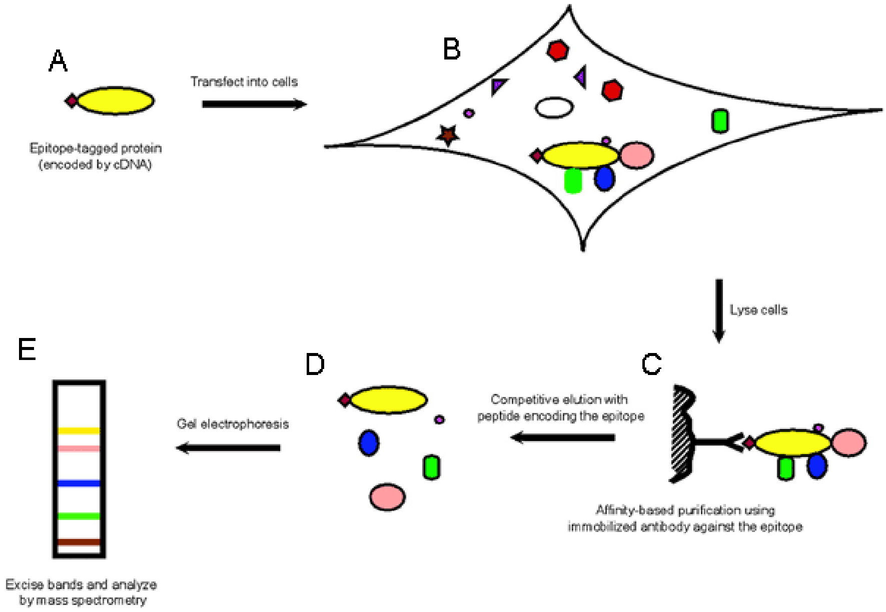

Figure 2.2: The co-IP assay. A) A specific protein bait is prepared and is attached to an affinity tag in order to allow the purification of the bait protein and the associated proteins. B) The bait protein is transfected into the cells. C) The cells are then lysed. D) The lysed protein soup is purified using a pull down assay with an antibody to the tag. E) The proteins are then identified using gel electrophoresis (western blot). (Figure taken from [31]).

2.1.2 Co-Immunoprecipitation

In this method [22], the bait protein is marked by a tag. An antibody which recognizes the tag is used to trap the bait protein and precipitate it. In the precipitation process, any protein physically associated with the bait is precipitated as well. Following this, mass spectrometry is used to determine the identity of the precipitated proteins (see Figure 2.2). The advantage of this method is in its ability to discover interactions between multiple proteins. However, within such a discovered complex, direct interactions cannot be distinguished from indirect interactions mediated by other proteins in the complex.

2.1.3 Interaction Reliability

Employing procedures such as Y2H and coIP, described above, has enabled the discovery of thousands of protein-protein interactions. However, those technologies still suffer from high rates of false positives. Hence, several approaches to assess the reliability of PPIs have been developed. One approach, by [8], estimates the reliabilities of different interaction data sources using the distribution of gene expression correlation coefficients. They separately considered experiments that report pairwise interactions like Y2H and those that report complex membership like mass spectrometry. Pairwise interactions from the literature and membership in protein complexes from Munich Information center for Protein Sequences (MIPS) [30] were used as the gold standard positive set in each case. Randomly chosen protein pairs formed the gold standard negative data set. Reliabilities for each data source were computed using a maximum likelihood scheme based on the expression profiles of each data source, based on the assumption that the distribution of gene expression correlations differs between pairs whose corresponding proteins interact and pairs whose corresponding proteins do not interact. In addition to assigning reliabilities to each such source, the authors also provided a conditional probability scheme to compute probabilities for groups of interactions observed in two or more data sets.

Other methods, suggested by [3] and [45], assign confidence values to protein interactions using a logistic regression model. Briefly, in [45] true positive and true negative interactions were used to train a logistic regression model, which assigns each interaction a reliability score based on the experimental evidence for this interaction (including the type of experiments in which the interaction was observed, and the number of observations in each experimental type). The experiments were partitioned into four categories: co-immunoprecipitation screens, yeast two-hybrid assays, large scale experiments and small scale experiments. Throughout this thesis we used the logistic regression model of [45] to score our PPI network.

2.2 DNA microarrays

DNA microarray is a high throughput technology used to measure gene expression levels. A square area (*chip*) is divided into hundreds of thousands of small squares termed *features*.

Each feature contains short DNA oligonucleotides of 15-50 nucleotides in length, attached to the array at one side, which correspond to, a specific mRNA. Two main assays exist, the single-channel microarray and the dual-channel microarray.

Single-channel microarrays, which are used in this study, are chips designed for a single experiment. Here, along with the features, there are normalization probes that allow the calibration of the expression levels to absolute values (see Figure 2.3).

In dual-channel microarrays, each of two sample is tagged with a different fluorescent dye. The two samples are then added to the same chip at equal amounts. The excess is washed off and the chip is bombarded by laser. This laser excites the fluorescent dye and it emits light. Gene expression is measured as the ratio of the two fluorescent colors.

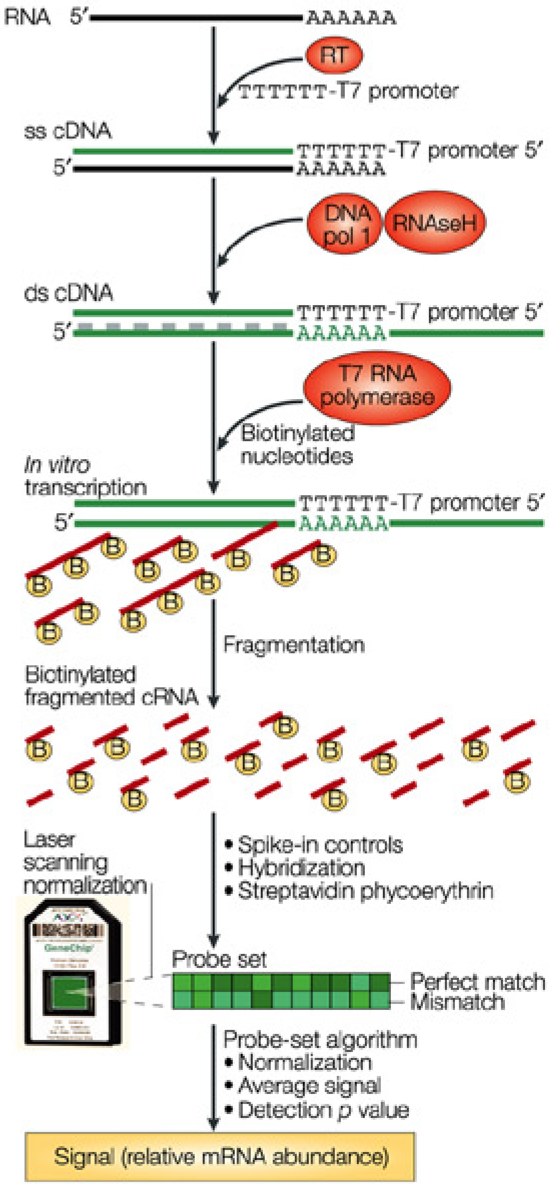

Figure 2.3: A single channel microarray assay. First the mRNA is purified from the sample, and amplified using RT-PCR. The next phase is labeling the fragments with Biotin which are then scanned by a laser scanner to determine the level of Biotin which corresponds to mRNA abundance. (Figure taken from [16]).

Chapter 3

Problem Definition and Algorithmic Approach

We study the problem of predicting one or more disease-causing genes given a set of genes that are implicated in a disease. We propose a network-based framework for it, which relies on the assumption that the protein product of a disease-causing gene should be highly connected in a network of physical interactions to the protein products of genes affected by it. Formally, the basic problem we consider is defined as follows:

Definition 1 (Gene Cover (GC)) *Given a graph $G = (V, E)$, a subset $U \subseteq V$ and a distance threshold l, find a subset of vertices D of minimum size such that for each $u \in U$ there exists a vertex in D of distance at most l from u.*

As we show below, GC is NP-complete as it is polynomially equivalent to Set Cover.

Theorem 1 *The decision versions of GC and Set Cover are polynomially equivalent.*

Proof 1 *Let (S, C, k) be an instance of Set Cover where S is the set of elements, C is a collection of subsets of S and k is a parameter. W.l.o.g., we assume that C covers S. We can easily transform this instance into an instance $(G, S, 1, k)$ of (the decision version of) GC as follows: we construct a bipartite graph $G = (S, C, E)$ with vertices on one side representing elements and vertices on the other side representing subsets. For every $T \in C$ and $s \in T$*

we add an edge (s, T) to E. It is trivial to observe that the Set Cover instance admits a solution if and only if the GC instance admits a solution (the only problematic case is when an GC solution contains an element from S, but such an element can always be substituted by a subset containing it).

In the other direction, suppose we are given an instance (G, U, l, k) of GC. We transform it into an instance (U, C, k) of Set Cover, where C is defined as follows: for each vertex v in G we create a subset $T \subseteq U$ composed of all vertices that are of distance at most l from v (including v if it is part of U). If $T \neq \emptyset$ we add it to C. Again there is a solution to the GC instance if and only if there is a solution to the Set Cover instance. ■

On the positive side, Set Cover can be efficiently approximated to within a logarithmic factor (cf. [7]); as the reduction from Gene Cover to Set Cover is approximation preserving, it implies an $O(\log |U|)$ approximation algorithm for GC as well.

3.0.1 A biologically-motivated formulation

The combinatorial formulation presented above treats all edges of the protein network being analyzed in a uniform manner. Since protein-protein interactions vary greatly in their associated confidence scores (see Section 4.1 below), it is desirable to take edge reliabilities into account. A natural extension to the distance-based formulation above is to quantify the relatedness of a protein v to a set U by the *expected* number of proteins in U that can be reached from it by paths of length at most l. Denote this expectation by $E_l(v, U)$ (we defer the details of its computation to the next section), and consider the following formulation of the gene coverage problem:

Definition 2 (Maximum-expectation Gene Cover (MGC)) *Given a graph $G = (V, E)$, a subset $U \subseteq V$, a distance threshold l, and a parameter k, find a subset of vertices D of size k such that $\sum_{v \in D} E_l(v, U)$ is maximal.*

It is possible to approximate MGC to within a factor of $O(\log |U|)$ by adapting the greedy-based approximation algorithm for Weighted Set Cover. Below we provide a practical heuristic to MGC which is based on this approximation strategy.

In many cases, additional information is available which can help us to limit the search space [49]. Specifically, association studies may provide information on genomic regions which are associated with the investigated disease, reducing the initial search space from thousands of proteins to a few hundreds (see, e.g., [29]). Similarly, copy number variation data can pinpoint areas of the genome whose copy number is modified in the disease state [28]; these areas are then good candidates for causal gene searches.

3.1 Expectation computation

Let $U = \{u_1, \ldots, u_n\}$. Recall that $E_l(v, U)$ denotes the expected number of vertices in U that are reachable from v by paths of length at most l. From the linearity of expectation,

$$E_l(v, U) = \sum_{i=1}^{n} E_l(v, \{u_i\}) = \sum_{i=1}^{n} P_l(v, u_i) \tag{3.1}$$

where $P_l(a, b)$ is the probability of having a path of length at most l between a and b.

For two vertices a and b, let $\Pi_l(a, b) = \{\Pi_1, \ldots, \Pi_m\}$ denote the set of paths of length at most l between a and b. Let π_i be a random variable indicating whether the path Π_i exists. Then $P_l(a, b) = Prob(\cup_{i=1}^{m} \pi_i)$. This probability can be computed using the inclusion-exclusion formula in time that is exponential in m.

To save on running time, one can partition the set of paths into subsets that are edge-disjoint. This is done by constructing a new graph whose vertices represent paths and whose edges connect edge-intersecting paths. The connected components of this graph yield the desired partition. Let $\Delta_1, \ldots, \Delta_t$ denote the resulting subsets of paths, and consider a pair of vertices a, b. Then $P_l(a, b) = 1 - \prod_{i=1}^{t}(1 - Prob(\bigcup_{\pi \in \Delta_i} \pi))$. Each term can be computed by an inclusion-exclusion formula:

$$Prob(\bigcup_{\pi \in \Delta_i} \pi) = \sum_{k=1}^{|\Delta_i|} (-1)^{k-1} \sum_{\substack{\Delta \subseteq \Delta_i \\ |\Delta| = k}} Prob(\bigcap_{\pi \in \Delta} \pi) \tag{3.2}$$

where the probability of an intersection of paths is simply the product of the probabilities of the edges in the intersection.

3.2 The MGC algorithm

We focus on the biologically-motivated MGC. Our algorithmic approach is motivated by the greedy approximation algorithm to Weighted Set Cover. Given a protein network G and a subset of disease-related proteins U, we apply an iterative algorithm to infer the disease-causing genes.

Intuitively, at each iteration the protein, whose "coverage" expectation with respect to the current subset U is maximal, is chosen and the diseased proteins that it "covers" are removed from U. However, the expectation computation gives an advantage to high degree proteins. To circumvent this problem, we compare the original expectation to that obtained with respect to 100 random disease-related subsets of the same size as U. The results of the random runs are used to derive a z-score for each vertex, and the highest-scoring vertex is chosen at each iteration. The algorithm terminates when the highest score attained is below a predefined threshold (1.65, corresponding to a p-value of 0.05), or when all the disease-related genes have been "covered". The pseudo-code of the algorithm is given in Figure 3.1. Due to the randomized nature of the algorithm (in computing the z-score), the results may change slightly between runs. Hence, each experiment is repeated 50 times, and the genes are ranked based on their average ranks in these 50 runs.

```
$\underline{MGC(G, U, l)}$
$D \leftarrow \emptyset$, $U' \leftarrow U$
**While** $U' \neq \emptyset$ **loop**
        **Foreach** $v \in V$ **loop**
                $E_v \leftarrow 0$
                **Foreach** $u \in U'$ **loop**
                        $E_v = E_v + P_l(u, v)$
                **End loop**
                $Z_v = ZScore(v, E_v, |U'|)$
        **End loop**
        $d = argmax_v\{Z_v\}$
        **If** $Z_d < 1.65$ **then**
           **Return** $D$
        **Else**
             $D = D \cup \{d\}$
             $U' = U' \backslash \{u | P_l(d, u) > 0\}$
        **End if**
**End loop**
**Return** $D$
```

Figure 3.1: The MGC algorithm. $P_l(u, v)$ is the probability of a path of length $\leq l$ between u and v (see 'Expectation computation' section). $ZScore(v, E_v, |U|)$ calculates the z-score of v (see 'The MGC algorithm' section).

Chapter 4

Experimental Results

4.1 Network construction

PPI data were collected from the HPRD database [34, 32] and two large scale yeast-two-hybrid experiments [38, 44]. The constructed network consists of 28,972 interactions among 7,915 proteins. The interactions were assigned confidence scores based on the experimental evidence available for each interaction using a logistic regression model adapted from [40].We removed from this network 8 hubs with more then 150 interaction partners, resulting in a final network of 27,707 interactions among 7,870 proteins.

In order to choose an optimal maximal depth with which to run our algorithm, we applied it with maximal length of 1-5 to three of the test cases discribed below (ATM knockout, NFκB knockout and MLL gene expression data), for which a causal gene is known. For length 1 and 2, the correct results were not significant when segments were larger then a few tens of genes, thus are not displayed. The results for length 3-5 are shown in Figure 4.1. As evident from the figure, the best results (except for the ATM knockout data) were attained at $l = 3$, which was subsequently used in all our runs.

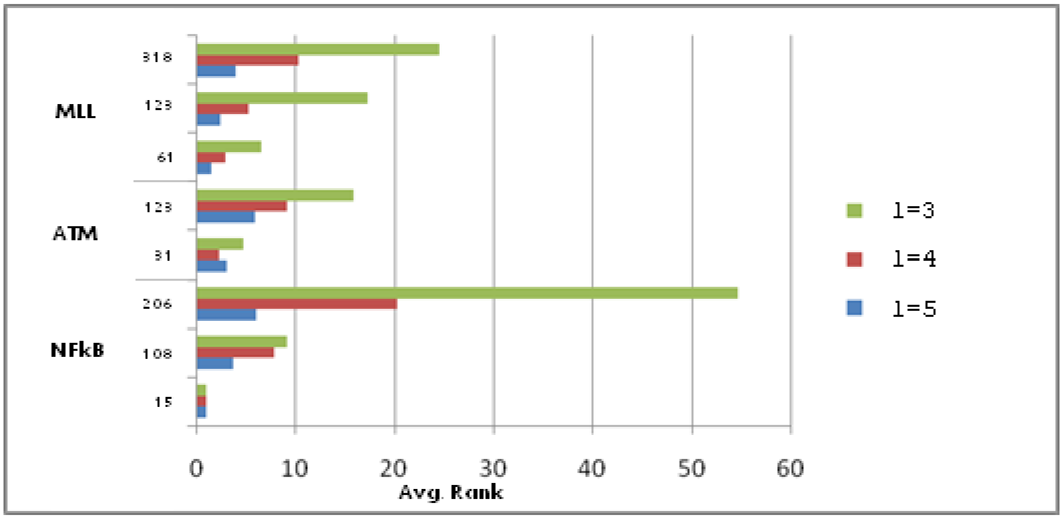

Figure 4.1: Performance with different maximal search length. Simulations were run for ATM and NFκB knockouts and MLL gene expression data. Different segment sizes were compared. See 'Validation on real data' section for a detailed description of these data sets.

4.2 Performance on simulated data

To evaluate the performance of the algorithm, we first applied it to simulated data. In each simulation, one or more "disease-causing" proteins were taken at random from the network, and artificial loci, consisting of 50 to 200 genes each, were constructed around the genes they encode. To construct a "disease-related" subset of a certain size (between 30-180), proteins were chosen at random from the set of proteins of distance at most 3 from the "disease-causing" ones. The results obtained in simulations of a single causal gene are summarized in Figure 4.2. Notably, when limiting the search to a certain locus, the algorithm almost always ranks the simulated causal gene first. The accuracy is lower when searching the entire network: the average rank of the causal gene ranges between 3.8-5.8 and it is ranked first 22-52% of the time, depending on the locus size.

To test the robustness of the algorithm to noise in the input list of disease-related genes, simulations were carried out in which the size of the disease-related set was fixed at 100, and up to 25% of the proteins in the set were replaced by random proteins. Genes whose

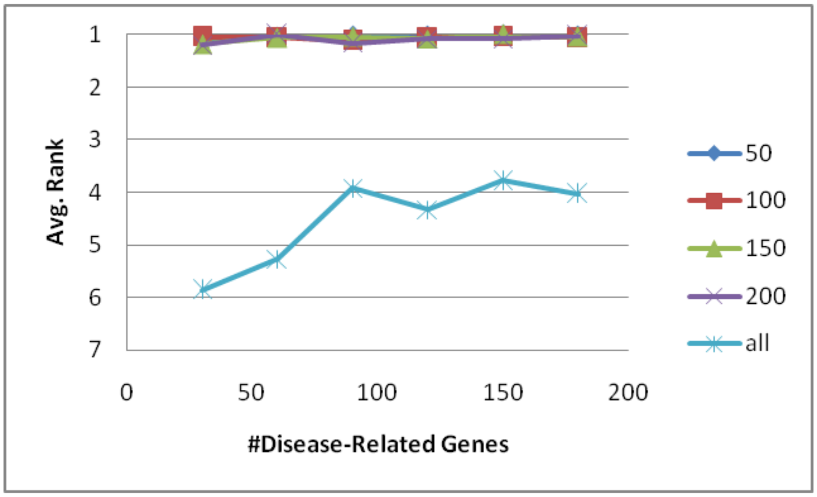

Figure 4.2: Performance on simulated data. The average rank of the correct simulated gene is depicted as a function of the size of the disease-related set. The different plots correspond to loci of different sizes. The all plot depicts the results obtained when searching the entire network.

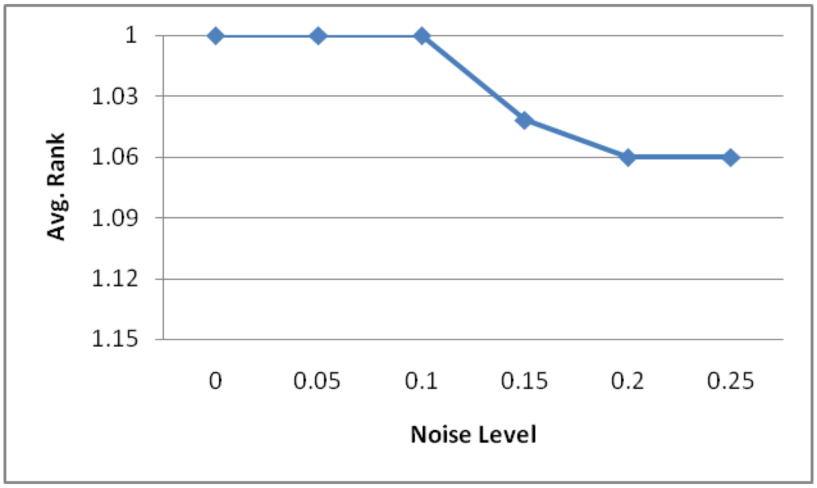

Figure 4.3: Performance in noisy simulations.

"coverage" expectation was equal to that of the simulated causal gene were removed (as they are indistinguishable from it based on our measure). The results are depicted in Figure 4.3 and show that the algorithm's results are robust even at high levels of noise.

We then tested the utility of the algorithm in recovering more than one causal gene. To this end, we conducted experiments in which up to four disease-causing genes were simulated (with a 100-genes sized locus around each one). The results are depicted in Figure 4.4. Evidently, the performance is very good for 1-2 genes, but worsens with three or more genes. E.g., with four disease-causing genes, the algorithm detects at least one (as the first ranking) 96% of the time, but identifies all four only 10% of the time.

Finally, we compared our performance to that of a naïve approach that ranks proteins according to their sum of distances to the input disease-related genes. As can be seen in Figure 4.5, our method is considerably more accurate when searching the entire network, achieving performance gains of more than 70%.

4.3 Validation on real data

To further evaluate the algorithm's performance, we applied it also to real data sets in which a causal gene is known. To this end, we used gene expression data for diseases where the

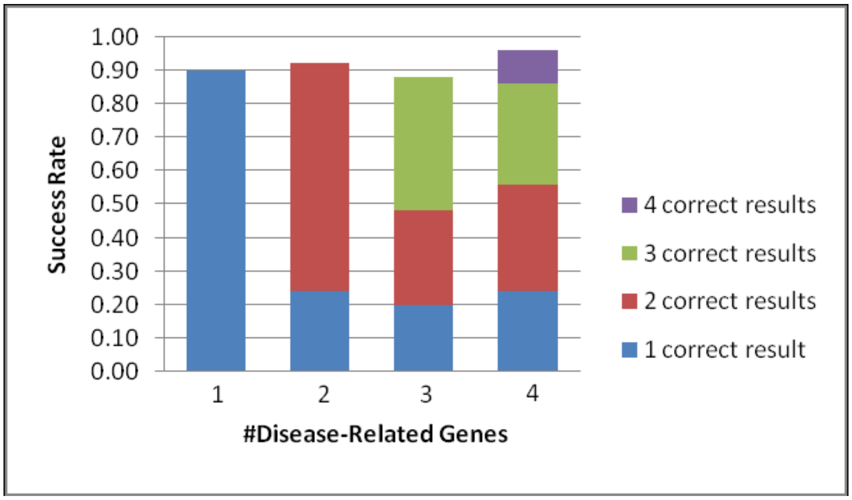

Figure 4.4: Performance on simulated data with multiple disease-causing genes. The success rate measures the percentage of runs where the simulated causal genes were ranked first, for 1 to 4 target genes.

genetic origin is known, or knockout data sets where a gene was knocked out and as a result other genes changed their (wild-type) expression levels. To simulate partial knowledge on the location of the causal gene we used information on the chromosomal segment in which the gene is located from the OMIM database [17].

The first set of experiments was performed on knockout data from [11], where the knockout effect of several genes was investigated under DNA damage conditions. In response to knocking out the transcription factor NFκB, 48 genes changed their expression levels. This gene is located in chromosomal segment $4q24$ which contains 31 genes, 15 of which appear in the PPI network. Reassuringly, NFκB was ranked first in all our tests. When knocking out the signaling protein ATM, 47 genes changed their expression levels. ATM lies within segment $11q22$ which contains 75 genes, 31 of which appear in our network. Overall, ATM ranked third with an average rank of 3.12.

Next, we used data on Acute Lymphoblastic Leukemia (ALL) [2], consisting of expression profiles for a subset of acute leukemias involving chromosomal translocation of the mixed leukemia gene (MLL). Overall, 67 genes were found to be differentially expressed and appeared in our network. The MLL gene is located at segment $11q23$, which contains 168 genes,

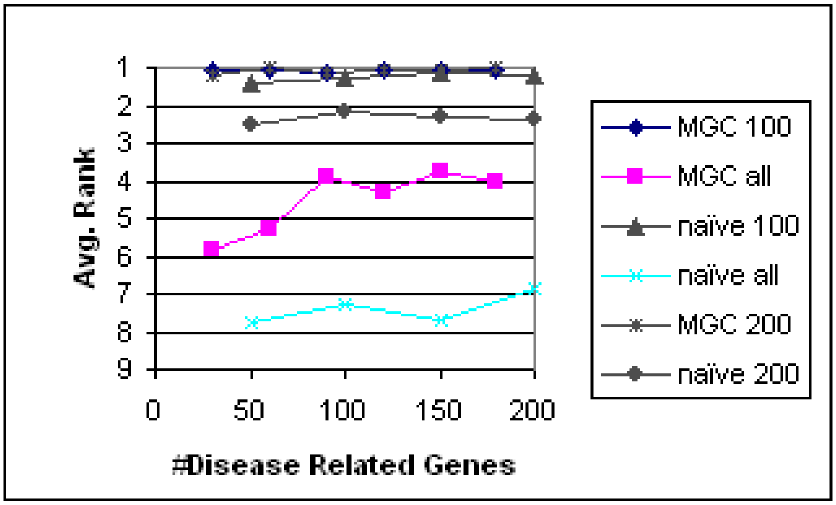

Figure 4.5: Performance on simulated data in comparison to a naïve approach.

61 of which are in the network. When applying the algorithm to this data, MLL scored best with an average rank of 1.5. The second highest ranking gene, with an average rank of 2.86, was matrix metallopeptidase 7 (MMP-7), a member of the matrix metalloproteinase family. This gene has been linked before to leukemia [26], and many other forms of cancer (see,e.g., [25, 37]). Three additional proteins that ranked among the top 10 are involved in phosphorylation signaling cascades known to be involved in the leukemic processes [35].

Finally, we applied our algorithm to input sets from multiple expression studies on breast cancer [46, 43, 33, 48]. We tested the rate at which our algorithm manages to recover BRCA1 (breast cancer 1, early onset) or BRCA2 (breast cancer 2, early onset), two of the major causal genes known. Here our search was conducted on 114 genes of the BRCA-1 associated segment ($17q21$) and 32 genes of the BRCA2 associated segment ($13q12$). The results, summarized in Table 4.1, show that at least one of BRCA1 or BRCA2 was recovered in each of the data sets. Intriguingly, the top 10 results searching a BRCA1 segment from both [33] and [48] were enriched with genes involved in transcriptional regulation, whereas searching a BRCA2 segment from [43, 46, 48], the results were rich in phosphorylation-associated genes, in accordance with the distinct functions of these two genes.

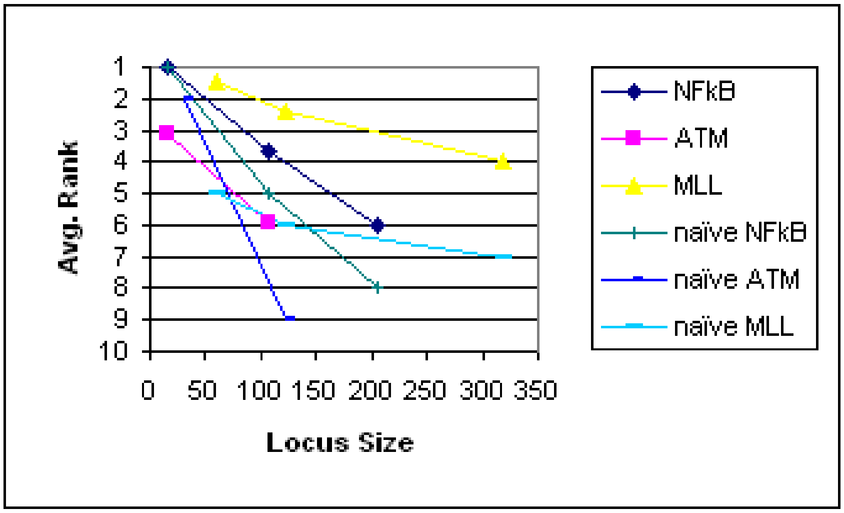

Figure 4.6: Comparison of our approach and a naïve one on three real data sets.

Data Source	BRCA1	BRCA2
[33]	5.72	> 10
[48]	2.88	1.76
[43]	> 10	2.5
[46]	> 10	3.06

Table 4.1: Average ranks of BRCA1 and BRCA2 on breast cancer data from different studies.

4.3.1 Application to Myasthenia Gravis

We study in detail Myasthenia Gravis (MG), a neuromuscular autoimmune multi-factorial disease for which the causal genes are largely unknown. We used data from [14], which contains a list of genes that are significantly expressed in the thymus of patients with mild and severe cases of the disease. First, we applied the algorithm to each severity class separately, using 391 genes for mild MG and 354 genes for severe MG. Next, we composed a list of 63 genes which appear in the severe cases but not in the mild ones. In all these applications the search for causal genes was conducted on the entire set of proteins in the network. The results are summarized in Table 4.2.

Mild	Severe	Severe but not mild
LGALS3BP	INSR	HLA-B
PEX6	ZMYM2	PRKACA
INSR	GJA1	PPP1R2
CD46	GUCY2C	HLA-A
POU4F2	CD46	CALCOCO1
ITGAL	EBF1	GRLF1
ZDHHC4	GATA3	GYS1
CHST9	STAT3	HIST1H2AJ
PRKCH	CEBPA	HMMR
CD44	MET	RBBP9

Table 4.2: A summary of the results obtained on Myasthenia Gravis data.

In mild MG, the highest ranking proteins contribute to general housekeeping functions: cell growth and cell-cell interactions, transcriptional activity and peroxisome properties. In Severe MG, we also observed impairments in hematopoietic differentiation compatible with the lymphocytic hyperproliferation which is characteristic of MG thymuses [14].

When looking at the set of genes that were differentially expressed in the severe cases, but not in the mild ones, the highest ranking protein was major histocompatibility complex, class I, B (HLA-B). This protein is part of the HLA class I heavy chain paralogs, which play a central role in the immune system and are expressed in nearly all cells. The linkage between HLA and MG is supported by previous studies [47, 18, 9]. The second highest ranking protein was cAMP-dependent protein kinase catalytic subunit alpha isoform 1 (PRKACA). This protein is known to phosphorylate and inhibit acetylcholine receptor functioning and affect the disease [24, 36].

Chapter 5

Large Scale analysis

For a systematic analysis of causal genes as a function of disease type and stage, we created a compendium of human disease gene expression data by mining the Gene Expression Omnibus (GEO) [10, 5]. We recovered 229 samples from 119 diseases (see description below) and applied the algorithm to this data. These samples are studied separately, compared to other samples from the same disease or from different disorders, and combined to create a gene-disease bipartite graph.

5.1 Data Processing

Each sample was downloaded from the GEO database. It holds gene expression results from diseased cases and expression results from normal ones. The data were then normalized and a t-test was applied in order to find which genes were differentially expressed ($p-value < 0.05$) between the diseased and the normal cases. This list of genes, termed *sample*, was then input to the algorithm. Following our experience in section 4 that on real data, the true gene ranked in the top 10, we took the top 10 results from each iteration as the output, while the top result was used for the next iteration. We ran each sample 50 times and averaged the results of each iteration.

5.2 Gene-Disease Network

In order to allow a global view of the connection between genes and diseases we constuct a gene-disease bipartite graph. To this end, we used genes and samples as nodes and connected a gene to a sample if the result of the algorithm on the sample included the gene as causal.

To validate the obtained network, we evaludated it against diseases in OMIM that have a known genetic origin. Out of the 229 samples, 16% (37 of 229) did not have an entry in OMIM, and for 69% (157 of 229), there is no known genetic origin. Of the remaining 35, in 20% (7 of 35) of the cases, our algorithm identified the correct causal gene.

From this graph, we created two projections, a human disease network (HDN) and a disease gene network (DGN). The HDN is created by using diseases as vertices and edges are added between two vertices if the corresponding diseases share a causal gene in the bipartite graph. The DGN is similarly constructed with genes as vertices, where an edge connects two genes associated with the same disease in the bipartite graph.

We then compared our results to the work of [15], who analyzed a gene-disease network that was created from the OMIM database [17]. We use this work as a reference because it relies on highly validated data. We first compare the HDNs, where there exists a central component of similar size: $516/1284 = 0.401$ compared to $107/219 = 0.489$. Similar hubs were found in both networks, e.g., breast cancer (17 neighbors), and leukemia (12 neighbors). This result is logical, as there are some tumor repressor genes that have been linked to many forms of cancer. Interestingly, out of the top 15 most connected diseases, 10 were cancer related ($p < 0.01$). When comparing DGNs, there is also a giant component of similar size, $903/1777 = 0.508$ compared to $177/337 = 0.525$. In our DGN, prominent hubs appear, Dynamin (29 neighbors) and CDC2L1 (24 neighbors), but these could be artifacts as they are housekeeping genes. In the work by [15] different hubs appeared, such as TP53, but these, due to high connectivity in the PPI network were descriminated against in our method.

Next, we examined the hypothesis that genes linked by disorder associations tend to have similar functions and are likely to be connected in the PPI network. To do so, we laid the DGN over the PPI network and looked for overlaps. Our results showed a 6-fold increase over random, similar to the 10-fold increase reported in [15].

When looking at the percentage of essential genes within the list of causal genes, the results are similar (31% in our study vs. 22% in [15]). The percentage of housekeeping and non-housekeeping genes participating in the networks are also similar (12% and 11% in our results compared to 9.9% and 13.5% in [15], respectively).

Overall, we find a high level of agreement between our results and those of [15].

5.3 Disease Progression

We assume that complex diseases usually have more then one causal gene and thus their disease related genes are more widespread in the PPI network then in a a monogenic disease. Also, as a disease progresses, secondary and tertiary effects cause gene expression changes. To measure this, we propose measuring the number of iterations the algorithm undertakes when analyzing the disease gene expression, i.e, using the number of iterations as a measure of the dispersion of the disease related genes in the network.

We first examined the distribution of the number of iterations needed by the algorithm to cover the genes associated with all diseases and with different types of diseases (Figure 5.1). It can be seen that the subgroup of cancerous diseases has a substantially higher mean then the rest of the diseases ($p < 10^{-4}$ via Wilcoxon rank sum test). Monogenic diseases (with known genetic origins) can almost always be explained in a single round (mean 1.3), while unknown diseases tend to have a higher mean then known diseases (2.2 and 2.8 respectively, $p < 0.01$).

Specific examples reveal that when analyzing diseases such as Duchenne muscular dystrophy the algorithm covers all the disease related genes in one iteration. Whereas in more complex diseases, such as breast cancer, the number of iterations is 3.11 on average (over the different samples), reaching a maximum of 6 iterations.

The progression of diseases can also be seen to influence the iteration number. Cases classified as severe tend to require a higher number of iterations compared to mild ones. Out of the 119 diseases, 25 have more then one sample which can be categorized by severity level. When analyzing these data sets, the severe cases show a mean of 3.3 iterations and the mild 2.7 ($p < 0.01$). For example, when comparing early Parkinson's disease patients [39] to a more general study [23], the early patients display a local change in gene expression pattern, which only takes one iteration to decipher; in the general study 3 iterations are required. In Alzheimer's disease, even for incipient Alzheimer's disease patients [6] it takes two iterations while severe cases require three (see Figure 5.2).

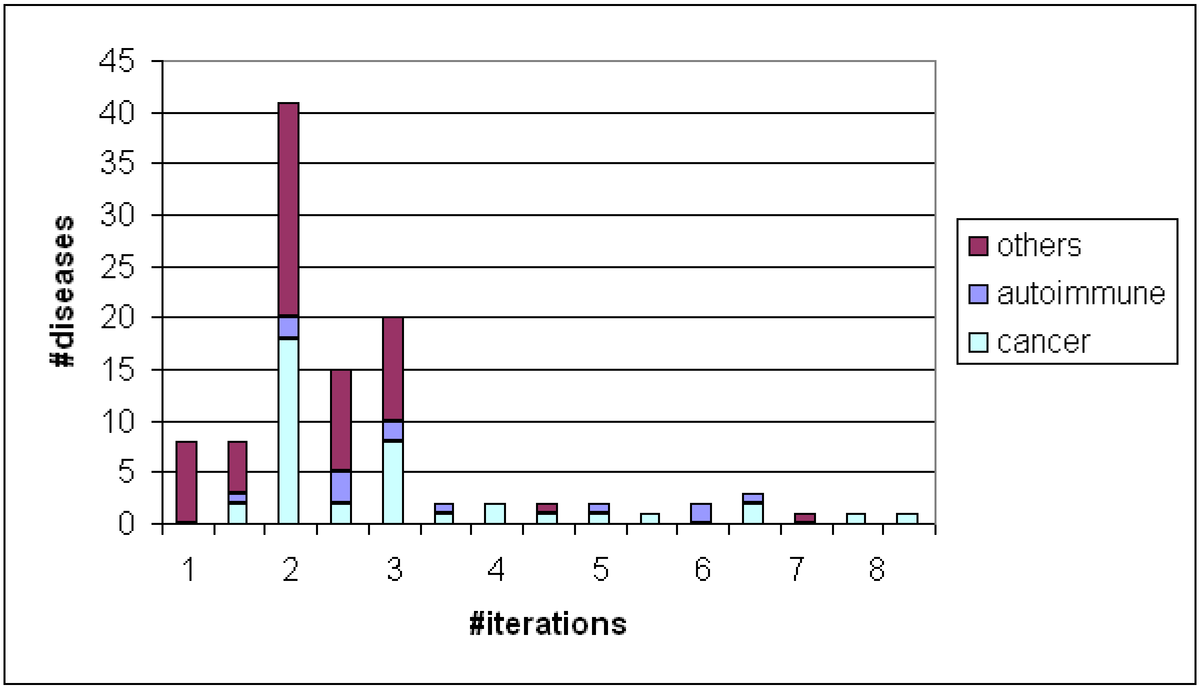

Figure 5.1: The distribution of the number of iterations executed by the algorithm on different disease types.

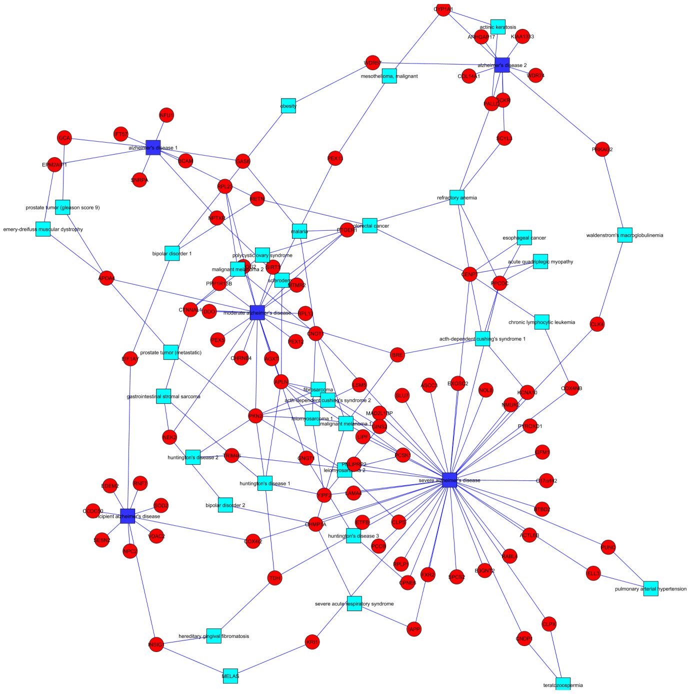

Figure 5.2: The Alzheimer's disease subnetwork. A partial view of the gene-disease network, centered around 5 Alzheimer's disease samples. One can see the variation in the number of causal genes, where the severe and the moderate cases require more iterations and hence have more genes linked to them.

Chapter 6

Conclusions

We presented a new approach to causal gene prediction that is based on integrating protein-protein interaction network data with gene expression data under disease conditions. The latter are used to highlight a set of disease-related genes that are assumed to be in close proximity to the causal genes in the PPI network. Based on this assumption, we apply a greedy heuristic that recovers putative causal genes as those admitting pathways to a maximal number (in expectation) of disease-related genes. We comprehensively validated the accuracy of our algorithm in pinpointing causal genes, both in simulations and on real network data. Having applied our algorithm to a large set of diseases, we were able to show that the effects of diseases change as the disease progresses and that multifactorial diseases show much wider changes to gene expression then simple monogenic diseases. By applying our algorithm to data on Myasthenia Gravis, we were able to suggest candidate causal genes and gain insights about their roles in the progression of the disease.

While our results are encouraging, several enhancements could be introduced to our framework. First, it would be revealing to integrate protein-DNA interactions into the network and study the impact of such interactions on the identified genes and pathways. To date, no experimental large-scale transcriptional network is available for human, although recent computational efforts have aimed at inferring it (see, e.g., [1]). Second, it could be beneficial to analyze data from multiple related diseases, assuming that in such cases the causal genes for each of the diseases lie in close proximity to one another within the network(see, e.g., [21]).

Bibliography

[1] A.S. Adler, S. Sinha, T.L. Kawahara, J.Y. Zhang, E. Segal, and H.Y. Chang. Motif module map reveals enforcement of aging by continual NFkB activity. *Genes Dev*, 21(24):3244–3257, December 2007.

[2] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, January 2002.

[3] G. D. Bader and C. W. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotech*, 20(10):991–997, October 2002.

[4] S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Res.*, 16(3):428–435, March 2006.

[5] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.*, 35:D760–D765, January 2007.

[6] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield. Incipient alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc Natl Acad Sci U S A*, 101(7):2173–2178, February 2004.

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, second edition, September 2001.

[8] M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput*, pages 140–151, 2003.

[9] B. Donmez, S. Ozakbas, M.A. Oktem, M. Gedizlioglu, I. Coker, A. Genc, and E. Idiman. HLA genotypes in turkish patients with myasthenia gravis: comparison with multiple sclerosis patients on the basis of clinical subtypes and demographic features. *Human Immunology*, 65:752–757, July 2004.

[10] R. Edgar, M. Domrachev, and A.E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, January 2002.

[11] R. Elkon, S. Rashi-Elkeles, Y. Lerenthal, C. Linhart, T. Tenne, N. Amariglio, G. Rechavi, R. Shamir, and Y. Shiloh. Dissection of a DNA-damage-induced transcriptional network using a combination of microarrays, RNA interference and computational promoter analysis. *Genome Biol*, 6(5):R43, 2005.

[12] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, July 1989.

[13] L. Franke, H. Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011–1025, June 2006.

[14] A. Gilboa-Geffen, P. P. Lacoste, L. Soreq, G. Cizeron-Clairac, R. Le Panse, F. Truffault, I. Shaked, H. Soreq, and S. Berrih-Aknin. The thymic theme of acetylcholinesterase splice variants in myasthenia gravis. *Blood*, 109(10):4383–4391, May 2007.

[15] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabasi. The human disease network. *PNAS*, 104(21):8685–8690, May 2007.

[16] Tumor Analysis Best Practices Working Group. Expression profiling–best practices for data generation and interpretation in clinical trials. *Nat Rev Genet*, 5(3):229–237, March 2004.

[17] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. Mckusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl. Acids Res.*, 33:D514–517, January 2005.

[18] D.R. Huang, R. Pirskanen, G. Matell, and A.K. Lefvert. Tumour necrosis factor-alpha polymorphism and secretion in myasthenia gravis. *Journal of Neuroimmunology*, 94(1–2):165–171, February 1999.

[19] S. Karni, H. Soreq, and R. Sharan. A network based method for predicting disease causing genes. *Journal of Computational Biology*, 16(2), 2009.

[20] S. Kohler, D. Bauer, S. snd Horn, and P.N. Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958, April 2008.

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

[22] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, October 2002.

[23] T. G. Lesnick, S. Papapetropoulos, D. C. Mash, J. Ffrench-Mullen, L. Shehadeh, M. de Andrade, J. R. Henley, W. A. Rocca, E. J. Ahlskog, and D. M. Maraganore. A genomic pathway approach to a complex disease: Axon guidance and parkinson disease. *PLoS Genetics*, 3(6):e98, June 2007.

[24] Z. Li, N. Forester, and A. Vincent. Modulation of acetylcholine receptor function in TE671 (rhabdomyosarcoma) cells by non-AChR ligands: possible relevance to seronegative myasthenia gravis. *Journal of Neuroimmunology*, 64(2):179–183, February 1996.

[25] D. Liu, J. Nakano, S. Ishikawa, H. Yokomise, M. Ueno, K. Kadota, M. Urushihara, and Huang C.L. Overexpression of matrix metalloproteinase-7 (MMP-7) correlates with tumor proliferation, and a poor prognosis in non-small cell lung cancer. *Lung Cancer*, 58(3):384–391, December 2007.

[26] C.C. Lynch and S. McDonnell. The role of matrilysin (MMP-7) in leukaemia cell invasion. *Clinical and experimental metastasis*, 18(5):401–406, 2000.

[27] K.M. Mani, C. Lefebvre, K. Wang, W.K. Lim, K. Basso, R. Dalla-Favera, and A. Califano. A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas. *Mol Syst Biol*, 4:169, February 2008.

[28] S. A. McCarroll and D. M. Altshuler. Copy-number variation and association studies of human disease. *Nat Genet*, 39(7 Suppl):R37–42, July 2007.

[29] M.I. Mccarthy, G.R. Abecasis, L.R. Cardon, D.B. Goldstein, J. Little, J.P.A. Ioannidis, and J.N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[30] H.W. Mewes, K. Albermann, K. Heumann, S. Lieb, and F. Pfeiffer. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Research*, 25(1):28–30, January 1997.

[31] Roger L. Miesfeld. AMG lecture 23. Website. `http://www.biochem.arizona.edu/classes/bioc471/pages/Lecture23/Lecture23.html`.

[32] G. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivkumar, N. Anuradha, R. Reddy, T.M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Matthew, P. Chatterjee, K. S. Arun, S. Sharma, K. N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, K. Krishnakanth, H. Karathia, B. Rekha, N. S. Rashmi, G. Vishnupriya, H. G. M. Kumar, M. Nagini, G. S. S. Kumar, R. Jose, P. Deepthi, S. S. Mohan, T. K. B. Gandhi, H. C. Harsha, K. S. Deshpande, M. Sarker, T. S. K. Prasad, and A. Pandey. Human protein reference database–2006 update. *Nucleic Acids Res*, 34:D411–414, January 2006.

[33] Y. Pawitan, J. Bjöhle, L. Amler, A. L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E. T. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P. M. Shaw, J. Smeds, L. Skoog, S. Wedrén, and J. Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res*, 7(6):R953–964, 2005.

[34] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya, Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371, October 2003.

[35] C. Perry, A. Eldor, and H. Soreq. Runx1/AML1 in leukemia: disrupted association with diverse protein partners. *Leukemia Research*, 26(3):221–228, March 2002.

[36] C.P. Plested, T. Tang, I. Spreadbury, E.T. Littleton, U. Kishore, and A. Vincent. AChR phosphorylation and indirect inhibition of AChR function in seronegative MG. *Neurology*, 59(11):1682–1688, December 2002.

[37] C. Rome, J. Arsaut, C. Taris, F. Couillaud, and H. Loiseau. MMP-7 (matrilysin) expression in human brain tumors. *Molecular carcinogenesis*, 46(6):446–452, June 2007.

[38] J. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale

map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, September 2005.

[39] C. R. Scherzer, A. C. Eklund, L. J. Morse, Z. Liao, J. J. Locascio, D. Fefer, M. A. Schwarzschild, M. G. Schlossmacher, M. A. Hauser, J. M. Vance, L. R. Sudarsky, D. G. Standaert, J. H. Growdon, R. V. Jensen, and S. R. Gullans. Molecular markers of early parkinson's disease based on gene expression in blood. *Proc Natl Acad Sci U S A*, 104(3):955–960, January 2007.

[40] R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, 102(6):1974–1979, February 2005.

[41] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Mol Syst Biol*, 3:88, 2007.

[42] S. Sobhanifar. The yeast two-hybrid assay: an exercise in experimental eloquence. *The science creative quarterly*, 2, 2003.

[43] C. Sotiriou, S.Y. Neo, L.M. McShane, E.L. Korn, P.M. Long, A. Jazaeri, P. Martiat, S.B. Fox, A.L. Harris, and E.T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *PNAS*, 100(18):10393–10398, September 2003.

[44] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, September 2005.

[45] S. Suthram, T. Shlomi, E. Ruppin, R. Sharan, and T. Ideker. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 7:360, July 2006.

[46] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen,

A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, December 2002.

[47] C. Vandiedonck, M. Giraud, and H.J. Garchon. Genetics of autoimmune myasthenia gravis: The multifaceted contribution of the HLA complex. *Journal of Autoimmunity*, 25 Supplement:6–11, 2005.

[48] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679, February 2005.

[49] X. Wu, R. Jiang, M.Q. Zhang, and S. Li. Network-based global inference of human disease genes. *Mol Syst Biol*, 4:189, 2008.