Enhancing the prioritization of disease-causing genes through tissue-specific protein-protein interaction networks

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

by
Oded Magger

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

October, 2011
Acknowledgments

Many people had a part in the successful completion of this thesis, and I would like to express my gratitude to all of them.

First and foremost to my advisor, Prof. Roded Sharan, for teaching me what it means to be a professional researcher and a scientist and for guiding me throughout this project.

I would like to thank Prof. Eytan Ruppin for his contagious passion for research, for his great advice, and for numerous one-on-one research meetings despite the fact that I was not officially his student. These are not taken for granted.

I would also like to thank my co-authors – Oron Vanunu, Yedael Waldman and Tomer Shlomi, and also the rest of my lab-mates – Liat Perlman, Nir Atias, Yael Silberberg, Assaf Gottlieb, Eyal Cohen, Shiri Freiilich, Misael Mongiovi, Dana Silverbush and Liram Vardi.

I would like to thank my Computer Architecture students for enriching my graduate studies experience.

And finally, Liraz, thank you for making my time out of the lab so enjoyable throughout this degree.

This thesis is dedicated to my late grandmother, Judith Rosenblith, whose final words were ‘Hard work doesn’t kill you’.

This study was supported by a fellowship from the Edmond J. Safra Bioinformatics Program at Tel Aviv University.
Abstract

The prioritization of candidate disease-causing genes is a fundamental challenge in the genomics of Human health, making it more feasible for researchers to develop new diagnosis methods and treatments for hereditary diseases. Current state of the art prioritization methods exploit a protein-protein interaction (PPI) network for this task, based on the observation that genes causing phenotypically-similar diseases tend to lie close to one another in a PPI network. However, to date, these methods have used a static picture of human PPIs, whereas diseases impact specific tissues in which the actual PPI networks may be dramatically different.

Here, for the first time, we assess the contribution of tissue-specific information to gene prioritization. By integrating tissue-specific gene expression data with PPI information, we construct tissue-specific PPI networks for 60 tissues and investigate their prioritization power. We find that tissue-specific PPI networks considerably improve the prioritization results compared to those obtained using a generic PPI network. We also provide rigorous analysis of the hypothesis underlying our network construction methods – that disease-genes tend to be expressed in the disease-related tissues under regular conditions. Furthermore, we utilize tissue-specific PPI networks to infer novel disease-tissue associations, pointing to sub-clinical tissue effects that may escape early detection.
# Table of Contents

1 Introduction ........................................................................................................................................... 5

2 Background ............................................................................................................................................. 8
   2.1 Discovering disease-causing genes ................................................................................................. 8
   2.2 Computational candidate gene prioritization .................................................................................... 9
   2.3 Tissue-specific protein interaction networks .................................................................................. 10
   2.4 Large-scale biological data sets ..................................................................................................... 11
      2.4.1 Generic data sets ....................................................................................................................... 11
      2.4.2 Tissue-specific data sets .......................................................................................................... 12

3 Algorithmic approach ......................................................................................................................... 14
   3.1 Tissue-specific PPI networks construction ....................................................................................... 14
   3.2 PRINCE: A network-based prioritization algorithm ...................................................................... 16
   3.3 Disease-tissue association inference ............................................................................................... 17

4 Results .................................................................................................................................................. 19
   4.1 Tissue-specific expression of disease-causing genes ...................................................................... 19
   4.2 Properties of Node-Removal Tissue-Specific networks .................................................................. 21
   4.3 Tissue-specific prioritization of candidate disease-causing genes .................................................. 24
   4.4 Evaluation of disease-tissue association inference ........................................................................ 29

5 Conclusions .......................................................................................................................................... 32
Chapter 1

Introduction

A fundamental challenge in human health is elucidating the molecular basis of genetic diseases. Contemporary methods for discovering disease-causing genes usually consist of two steps: first, genome-wide association studies identify genomic intervals that are linked to a disease of interest. Second, the genes within these intervals are examined for their causal relation to the disease [1-3]. Experimentally verifying that a gene is associated with a disease is an expensive and time-consuming process, calling for the prioritization of candidate causal genes. A plethora of computational methods were developed to meet this challenge. These methods are often based on system-wide data, for example protein interaction networks of gene expression. Many state of the art algorithms for the gene prioritization problem are network-based, exploiting the tendency of genes causing similar diseases to lie close to one another in Protein-Protein Interaction or Functional-Linkage networks (PPI and FL, respectively) [4-6]. However, these methods do not take into account the fact that the vast majority of genetic disorders tend to manifest only in a single or a few tissues [5,7,8]. Typically, the same data sets are used to prioritize genes for a liver disease or a brain disease, even though the molecular landscapes of a Hepatocyte and a neuron are quite dissimilar.

The current study is the first aiming to enhance the accuracy of existing network-based gene prioritization algorithms by taking into account tissue-specific information. We incorporate tissue-specific gene expression data into the prioritization process and demonstrate its impact on the prioritization results. The integration is achieved by constructing tissue-specific protein-protein interaction networks and employing them in the prioritization. The rationale behind this approach is that many disorders involve a disruption of the ‘molecular fabric’ of different, healthy tissues. From a protein interaction network point of view, this disruption can be often characterized as a perturbation of a gene, corresponding to node removal, or the perturbation of an interaction between two gene products, corresponding to an edge removal [9]. In the context of genetic disorders, even though the underlying harmful mutation exists in all the cells of our body, it most often wreaks havoc only in a few tissues. This tissue selectivity is likely to emerge
due to differences in the functionality of the mutated protein within these tissues, its tissue-specific interacting proteins, its abundance and the abundance of its interactors. Thus, the hypothesis underlying the current work is that a tissue-specific network, which better represents the actual disease-related tissue, is likely to yield more accurate prioritizations of the diseases it manifests.

This is achieved by constructing tissue specific PPI networks and utilizing them for gene prioritization instead of the standard, generic PPI network. First, we examine the hypothesis that a gene is likely to be expressed in a healthy tissue for its mutation to clinically manifest in that tissue. Indeed, a large majority (71–83%) of the known disease-causing genes are significantly expressed in the corresponding disease-associated tissue. However, not all disease-associated genes are significantly expressed in the tissues where the disease is manifested. Interestingly, as shown in chapter 4.1, we find that most of the remaining genes either have a low expression level across all tissues, or are involved in either mediating a response to external stimulus or multi-cellular developmental processes, and as such are not expected to have high expression under steady-state conditions in the adult tissue.

Focusing on the cases where the disease-related gene is expressed in the associated tissue, we show that integrating tissue specific expression information into a gene prioritization scheme markedly improves its prediction accuracy. Specifically, we generate tissue-specific PPI networks for 60 healthy human tissues using gene expression data from those tissues [10]. We then apply the same candidate prioritization algorithm for both the original and the tissue-specific PPI networks, and compare the performance of each in a cross-validation setting. We find that the tissue-specific variant of the algorithm yields higher area under the receiver-operator characteristics curve (AUC) and gives the correct gene a higher ranking than the original variant more often than not. Finally, we extend our method to infer new disease-tissue associations.

This thesis is organized as follows: **Chapter 2** provides the scientific background of this research, including an introduction to the challenge of prioritizing candidate disease-causing genes, a review on recent work on tissue-specific biological networks and a survey of the large scale biological data sets used in this work. **Chapter 3** introduces our method in detail, starting from the construction of tissue-specific protein interaction network, moving to a summary of the
prioritization algorithm which uses these networks as an input, and concluding with our novel method of inferring disease-tissue associations based on this pipeline. Chapter 4 summarizes the experimental results of this work: First, an analysis of the expression of disease genes in the disease’s associated tissue and of the topological properties of tissue-specific protein interaction networks. Then, the contribution of tissue-specific networks to the prioritization task is evaluated in depth by comparing tissue-specific and generic variants of the prioritization algorithms. Finally, our disease-tissue association inference method is evaluated by seeing how well the predicted associations correlate with associations known by the literature. Last, Chapter 5 concludes this work by summarizing it, discussing some of its finer points and raising several issues for future research.
Chapter 2

Background

In this chapter we discuss some of the fundamental concepts of this work, review relevant past research and survey the main large-scale data sets used herein.

2.1 Discovering disease-causing genes

A genetic disease is a disorder that is caused by abnormalities in genes or chromosomes. It might be passed down from parents to their offspring, or caused by mutations or DNA damage in the individuals themselves. Some of the disorders are monogenic, caused by a single mutated gene. Other disorders are polygenic, and can be caused by mutations in multiple genes or the combination of several mutated genes.

On the single-gene level, this abnormality might result in the gene being underexpressed, overexpressed or not expressed at all; It may also result in the gene being transcribed and translated onto ‘faulty’ proteins who might lose some of the ‘healthy’ proteins’ functionality. On the large-scale level, these local changes might have far-reaching consequences on the cell, such as causing it to multiply uncontrollably (as often happens in cancer), the shape of the cell (e.g. Sickle cell disease), or blocking a crucial metabolic pathway. These effects in turn lead to the visible symptoms of the disease.

For example, Cystic Fibrosis is a hereditary disease caused by a mutation in a gene called CFTR, which produces a transmembrane protein responsible for transporting chloride ions through the cell membrane. There are ~1500 known harmful mutations that could hinder the function of this gene. For example, ΔF508 – a deletion of three nucleotides leading to the deletion of the phenylalanine amino acid at the 508th position in the protein - cause the protein to fold improperly. If there aren’t enough functional copies of the CFTR protein in the cell, the concentration of chloride and several other ions in the cell diverges from normal, leading to severe effects such as clogging of the airways (due to mucosal inflammation), pancreatic insufficiency, endocrinal perturbations, male infertility and more.

In order to diagnose and treat genetic disorders, researchers need to first identify the causal genes and elucidate the molecular cause of the disease. Modern methods in this field rely heavily on computational tools, and consist of two steps:

The first step is to perform a **genome-wide association study** (GWAS). In GWAS, a cohort of patients (usually of the same ethnic origin) is taken, and their genome is mapped for known polymorphisms. A group of healthy people (again, usually of the same ethnic origin) is used as control. The purpose of this mapping is to find a polymorphism that differentiates between the two groups in a statistically significant manner. This polymorphism could be a difference in a single nucleotide (SNP), an altered number of DNA repeats, and more. However, the polymorphism is not necessarily the harmful mutation causing the disorder. More likely, the disease-causing gene is located near the polymorphism. This is due to the fact that two close positions on the chromosome tend to stay linked despite chromosomal recombination.

Thus, the output of GWAS is a **linkage interval**, a region of the genome where the disease-causing gene is likely to reside. The second step in the process is selecting genes from within this interval and **experimentally validating** their association to the disease. However, the experimental process is expensive and time consuming, and a linkage interval may contain dozens or even hundreds of disease gene candidates. It is infeasible to examine all of the genes in the linkage interval. Thus, the researcher needs to **prioritize** the genes in the interval and start examining the most promising candidates.

### 2.2 Computational candidate gene prioritization

Formally, given a disease of interest $d$ and a set of genes $G = \{g_1, \ldots, g_n\}$, a prioritization function $F$ assigns a strength-of-association score for every $g_i$ such that higher score means higher likelihood that $g_i$ is the disease-causing gene. Typically, a researcher will select the $k$ genes with the highest score and try to validate their causality experimentally.

An automatic tool for candidate gene prioritization has several advantages over manually selecting genes from the interval. First, it is able to perform the task much faster. Second, it can take into account very large volumes of data. And third, it is as unbiased as the data used as input.
Plenty of prioritization algorithms and tools were published in recent years. These methods are often based on system-wide data such as protein interaction networks [11-16], gene expression [15,17-19], sequence similarity of genes [20,21], functional similarity and annotation [15,19,20] and more (for a review on these methods see [6,22]). Many state of the art algorithms for the gene prioritization problem use protein interaction or functional linkage networks [15,23], exploiting the tendency of genes causing similar diseases to lie close to one another in the network [4-6].

Notable recent works include Linghu et al[15], which combined several measures of gene-gene similarity to construct a functional linkage network, and then used a naïve Bayes classifier to prioritize the candidates according to their direct neighbors’ association with diseases of the same class as the query disease. Wu et al developed AlignPI [24], a tool that aligns the protein interactional network and a phenotype-similarity networks, identifies small bi-clusters of diseases and genes, and utilizes these clusters to prioritize genes. A previous work by the same group, CIPHER[13], scored a candidate gene $g$ for a certain disease $d$ based on the correlation between the vector of similarities of $d$ to diseases with known causal genes, and the vector of closeness in a protein interaction network of $g$ and those known disease genes. Kohler et al. presented a prioritization function based on a Random Walk over the protein interaction network. Finally, the author of this thesis has been a co-author of PRINCE [11], a prioritization method the performs network propagation in order to obtain a smooth scoring function over the network (see section 3.2 for detailed description of PRINCE). A recent paper by Navlakha and Kingsford [23] compared the performance of several network-based prioritization approaches, and clearly showed the superiority of global methods such as Random Walk and PRINCE.

### 2.3 Tissue-specific protein interaction networks

A key novelty in this work is employing a tissue-specific protein interaction network as input to a network-based prioritization algorithm. The concept of tissue-specific protein interaction networks is relatively unexplored. Most previous studies in this domain focused on investigating the properties of direct, local interactions in the context of tissue specific expression data. Bossi and Lehner [25] analyzed human PPIs in a tissue-specific context, showing that many housekeeping proteins interact with highly tissue-specific proteins, which in turn implies that housekeeping proteins may have tissue-specific roles. Emig and Albrecht [26] expanded this
analysis to identify functional differences between tissues, showing that tissue-specific protein interactions are often involved in transmembrane transport and receptor activation. Lin et al [27] analyzed the topological properties of housekeeping and tissue specific proteins within the generic (non tissue-specific) PPI network. A few studies generalized these concepts to a global investigation, utilizing the entire network as a whole, and not just direct local interactions. For example, Waldman et al. [28] analyzed translation efficiency in humans using PPIs. Using tissue specific PPI networks, they showed that proteins whose genes are translated more efficiently in a specific tissue tend to have more connections within this tissue as compared to other proteins in the same tissue. Of note, the lack of tissue specific PPI networks stands in marked difference from the existence of many tissue- and cell-specific variants of other types of biological networks, such as regulatory networks [29-31], functional linkage networks [32,33] and metabolic networks [34-36].

2.4 Large-scale biological data sets
All of the prioritization algorithms presented in Chapter 2.2, as well as the biological networks presented in Chapter 2.3, rely heavily on system-wide, large-scale data sets made available by the latest technological advancements in life sciences. In this section we will present the data sets used in this work, and the data structures derived from them. The data sets are partitioned into those that were used in the generic variant of PRINCE (2.3.1), and tissue-specific data used to augment PRINCE for this work (2.3.2).

2.4.1 Generic data sets
We constructed a weighted Human PPI network with 9,998 proteins and 41,702 interactions. The network is based on three high throughput experiments [37-39] and the HPRD database [40]. The interactions were assigned confidence scores based on the experimental evidence available for each interaction using a logistic regression model adapted from [41].

We used two major online databases for disease-related information: OMIM[42] and GeneCards[43]. The first, OMIM, acronym for Online Mendellian Inheritance In Man, is an online database of genetic disorders and their associated genes. In October 2011, this database contained entries for more than 5000 known phenotypes and more than 13000 disease-related genes. In order to produce a disease-disease phenotype similarity network, van-Driel et al[44] text-mined OMIM phenotype entries and created a feature vector for each one, composed of the
occurrences of several keywords. Similarity of two disorders was computed as the correlation between their feature vectors. van-Driel’s disease similarity network consists of 5040 disorders.

van-Driel et al note in their paper that disease-disease similarities with a score of < 0.3 tended to be non-informative, whereas disease-disease similarities with a score of > 0.6 were often validated by the high functional similarity of the diseases’ causal gene. As a result, in [11], we performed a logistic transformation on the weights of the disease-disease similarity network with coefficient c=-15, so that similarities in range [0,0.3] are assigned a very low probability, and conversely similarities in range [0.6,1] are assigned a very high probability. We use this logistic transformation also in the current work.

We used GeneCards [43], an online encyclopedia of Human genes, to generate disease-gene associations. The entry of every gene contained in our PPI was text-mined for a particular pattern suggesting a causal link between the gene and an OMIM entry. We extracted 1599 known disease-gene spanning 1369 diseases and 1043 proteins. While some of the other works text-mined OMIM for this purpose, we used GeneCards because the association set appears to contain more causative and less circumstantial associations compared to the OMIM-based association set[16].

2.4.2 Tissue-specific data sets
The Novartis Research Foundation Gene Expression Database (GNF) tissue-specific gene expression data set [10] is an extensive gene expression profile covering 79 Human tissues. We used 60 non-diseased, mature tissues for which disease-tissue association data existed (out of 79 tissues). We averaged replicas and for each gene took the probe with the maximal expression level. Following [25,45], a gene was considered expressed in a tissue if its expression in that tissue was at least 200 Affimatrix average-difference (AD) units. Similarly, a protein was defined to be present in a tissue if the corresponding gene was expressed there. We downloaded the GNF data set from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) (GDS596), a public repository of curated gene expression profiles.

It should be noted that this assumption is far from accurate. The measurement of mRNA levels, like every experimental procedure on this scale, may contain errors. Some genes are expressed even when their mRNA levels are below 200 AD. Furthermore, mRNA expression and protein
abundance are not always correlated. However, despite the best efforts of the scientific community, compendiums of human tissue-specific protein abundance levels across multiple tissues are not nearly as comprehensive as the mRNA expression dataset we use, both in tissue scope, gene coverage and quantitative resolution [46].

To the best of my knowledge, there are no manually curated disease-tissue association databases. Thus, in order to link diseases with the tissues they affect, we used a computationally-generated disease-tissue association matrix which was contributed by Kasper Lage [47]. Lage et al. estimated the association of a tissue and a disease by measuring their co-occurrences in PubMed abstracts. Specifically, the association score was computed using Ochiai’s coefficient (OC) [48], and then normalized by the sum of all OCs for the same disease. Subsequently, each disease was associated with the tissue that attained the highest association score, hereafter denoted MAS (Maximal Association Score). Lage et al validated their matrix by sampling randomly selected associations for literature evidence, and found that a threshold of 8% association score yields an accuracy level of 80%, and a threshold of 40% results in an accuracy level of ~90%.
Chapter 3

Algorithmic approach

In this chapter we present in detail the computational methods employed in this work: models for constructing tissue-specific networks, a network-based candidate gene prioritization algorithm, and our novel approach for computationally inferring disease-tissue associations.

3.1 Tissue-specific PPI networks construction

We considered two methods for converting the generic PPI network into a tissue-specific network using a given tissue-specific expression profile. These methods are summarized in Figure 3.1 and discussed below.

![Diagram of tissue-specific PPI network reconstruction methods](image)

**Figure 3.1:** A summary of tissue-specific PPI network reconstruction methods. First, we determine the set of expressed genes in a given tissue based on an expression cutoff of 200 Affimatrix AD units. The set of expressed genes is then superimposed on the general PPI using one of two strategies: (a) Node Removal – removing genes which are considered unexpressed from the network. (b) Edge Reweight - Reducing the weight of an edge connecting one or two unexpressed genes. This results in a tissue specific PPI network.
A naive method, titled “Node Removal” (NR), was used previously [25,28]. By this method, a tissue-specific PPI network is generated by removing from the network proteins that are not expressed in the relevant tissue, along with the proteins’ adjacent edges. As seen in chapter (4.2), this method drastically alters the topology of the protein interaction network, making it less interconnected.

The second tissue-specific network reconstruction method, novel to this work, is titled ‘Edge Reweight’ (ERW). By this method, we do not alter the topology of the generic network, but rather modify the edge weights to reflect the probability that the corresponding interactions take place at the specific tissue.

Our underlying assumption was that an interaction between proteins $P_1$ and $P_2$ occurs at a specific tissue $t$ if only if $P_1$ and $P_2$ interact in the general network and are both expressed at tissue $t$. Denote the event that proteins $P_i$ and $P_j$ interact in the generic network as $I_{i,j}$, and the event that protein $P_i$ is expressed in tissue $t$ as $X(i,t)$. Now, a gene is considered expressed in a given tissue if its measured expression level in that tissue is above 200 AD units. However, expression data is often noisy [45,49] so there is a chance that a gene not passing this cutoff is still expressed (we assume that if a gene passes the threshold then it is indeed expressed in the given tissue). If we denote this probability by $rw$, then

$$w'_{i,j} = P(P_i, P_j \text{ interact} \mid \text{Tissue} = t) = P(I_{i,j} \mid t) * P(X(i,t) \mid t) * P(X(j,t) \mid t) = w_{i,j} * rw^n$$

where $w_{ij}$ is the original weight of the interaction and $n$ is the number (0-2) of lowly-expressed genes in tissue $t$ out of \{P$_1$,P$_j$\}. Thus, conversion of the generic PPI weight to a tissue specific PPI weight using the edge reweight method involves multiplying an edge’s weight by $rw$ if one of its adjacent genes is not expressed in the tissue, and by $rw^2$ if neither of the edge’s adjacent genes are expressed in the tissue.

$rw \in \{0,1\}$. Note that when $rw = 0$, the ERW network becomes the NR network; conversely, when $rw = 1$, the ERW network is identical to the original PPI network. Thus, varying values of $rw$ allow us to control just how tissue specific the network is.
3.2 PRINCE: A network-based prioritization algorithm

In order to prioritize candidate disease genes, we used the PRINCE network-based prioritization algorithm. While in a previous work we compared PRINCE to other state-of-the-art algorithms, this time we compare PRINCE’s performance when given different PPI networks as input.

Given a query disease, PRINCE assigns a prior score to genes associated with known diseases that are phenotypically similar to the query. This score is then propagated through a PPI network in an iterative process, culminating in a smooth scoring function where the score of a node tends to be similar to the scores of its neighboring nodes.

In detail, denote the PPI network as $P = \langle V, E, w \rangle$ where $V$ are the proteins / nodes, $E$ is the set of protein-protein interactions / edges, and $w$ is the interaction confidences / edge weights. $w : V \times V \to [0,1]$, $E \subseteq V \times V$. The query disease is denoted $q$, and we wish to compute a prioritization function $F : V \to \mathbb{R}$ such that the higher $F(v)$ is, the higher the probability that $v$ is associated with $q$. $Y : V \to [0,1]$ is the prior knowledge function assigned to every $v \in V$. If $v$ is associated with a disease $d$, $Y(v) = disease\_similarity(q, d)$. Otherwise $Y(v) = 0$.

PRINCE’s prioritization function $F(v)$ is designed to both take into account both the prior knowledge $Y(v)$ and to be smooth over the network, meaning that adjacent node have similar score. To achieve this goal, $F(v)$ is calculated as a linear combination of both $Y(v)$ and the scores of $v$’s neighboring nodes, $F(N(v))$ (where $N(v) = \{u | u \in V \land < u, v > \in E\}$).

$$F_{prince}(v) = \alpha \left[ \sum_{u \in N(v)} F_{prince}(u) \cdot w(u, v) \right] + (1 - \alpha) \cdot Y(v)$$

$\alpha \in [0,1]$ determines which of the two considerations (prior knowledge vs. neighbourhood score) is more dominant. The equation has a deterministic solution, but it is slow to calculate for large data sets as it requires heavy operations such as matrix inversion. Fortunately, the solution can be easily and efficiently approximated using an iterative process where in every step the following equation is applied for every node $v$:

$$F_{prince}^t = \alpha \cdot W \cdot F_{prince}^{t-1} + (1 - \alpha) \cdot Y$$
Where \( t \) is the index of the current step. Note that this equation is written in matrix notation, but is essentially the same as the previous equation, with \( F_{prince}^{t} \) being calculated based on \( F_{prince}^{t-1} \). The starting condition is \( F_{prince}^{1} = Y \). We’ve shown previously that this iterative solution converges within a few iterations[11].

PRINCE is affected by several parameters. For this work, we used the same parameters as in[11], namely 10 iterations, \( c = -15 \) and \( \alpha = 0.9 \).

### 3.3 Disease-tissue association inference

Having the ability to predict the effects of disease genes on specific tissues, naturally gives rise to the question: What tissues are most likely to be affected by a harmful mutation in a disease-causing gene? This is of particular interest, since while the overt clinical manifestations of a disease are usually well-known, in many cases it may have more subtle, sub-clinical tissue effects that may escape early detection. Such alterations may manifest themselves at later stages of the disease, and may be wrongly attributed to other potential complications and confounding factors, instead of the original disease, which can serve at least as an important predisposing factor.

To investigate this potential scenario in depth, we developed a method to computationally infer disease-tissue associations using the tissue-specific networks and their prioritizations as described in the previous sections.

Given a disease \( d \), it’s causal gene \( g \) and a set of tissues \( T = \{\text{tissue}_1, \ldots, \text{tissue}_n\} \), we wish to rank the tissues according to their relevance to the disease. The ranking of \( T \) in relation to \( d \) is performed as follows:

First, we remove the association between \( d \) and \( g \), similar to the cross-validation procedure used to assess PRINCE’s performance (see chapter 4.3), and run PRINCE for every tissue-specific network \( t \in T \). Then, we evaluate how well PRINCE ranked the actual causal gene \( g \) in every tissue and rank the tissues accordingly. We used two different measures for appraising PRINCE’s success:

- **Relative PRINCE Rank**: the tissues are sorted according to the rank PRINCE assigns to \( g \), when prioritizing an artificial interval of 100 genes which includes \( g \) (see chapter 4.3).
For example, if PRINCE ranks $g$ 3rd when using the Heart PPI as input and ranks $g$ 5th when using the Liver PPI as input, then heart has better tissue-relevance rank than liver.

- **Absolute PRINCE Score**: the tissues are sorted according to the score PRINCE assigns to $g$. For example, if PRINCE assigns $g$ a score of 0.6 when using the Heart PPI as input and a score of 0.8 when using the Liver PPI as input, then liver has better tissue-relevance rank than heart.

These two methods are not redundant. It is entirely possible for one tissue to be ranked above another using one method, and the other way around when using another. For example, let’s assume that $G = \{g, y1, y2\}$ is our linkage interval, and that when using PRINCE with the Heart PPI the resulting scores are (0.5, 0.2, 0.1) and when using the Liver PPI the resulting scores are (0.8, 0.9, 0.7). If we use the Relative Rank method then the Heart gets a better rank since $g$ is ranked 1st in the interval (whereas when using the liver PPI $g$ is ranked 2nd). However, if we use the Absolute Score method then liver gets a better rank since 0.8 > 0.5.

Both methods have their shortcomings. Relative Rank is prone to ties between tissues (e.g. when PRINCE ranks $g$ in 2nd place for 3 different tissues), whereas Absolute Score is affected by whole-network biases (e.g., all of the genes could get high scores relative to other tissue-specific PPIs).
Chapter 4

Results

In this chapter we present the result of the analyses and experiments performed in this work. We start with examining the expression of disease-causing genes in their disease’s assigned tissue, and elucidating interesting topological properties of Node Removal tissue-specific protein interaction networks. Then, we evaluate the performance of our novel prioritization algorithm as well as its application to infer disease-tissue associations.

4.1 Tissue-specific expression of disease-causing genes

The underlying hypothesis of our tissue-specific network construction models (presented in Chapter 3.1) is that disease-causing genes are expected to be expressed in the disease-affected tissue when it is in a healthy state. The rational is that in order to a mutation to have an effect on a cell, the mutated gene must have some function in the cell under normal conditions and is therefore expected to be expressed. Below we describe our validation of this assumption.

For every disease, we assigned the tissue that had the maximal association score (MAS) with that disease, and filtered diseases whose MAS was below a predefined threshold. For most of the analyses in this chapter, we used two thresholds: MAS > 8% was the cutoff used by Lage et al., estimated by them to provide 80% assignment accuracy. Filtering by this threshold produced a set of 920 disease-gene associations, spanning 729 diseases and 632 genes. The second threshold, MAS > 40%, was estimated by Lage et al to provide ~90% accuracy. This threshold yielded 349 associations spanning 290 diseases and 269 genes.

Next, we constructed binary tissue-specific gene expression profiles for 60 healthy tissues based on the Novartis Research Foundation Gene Expression Database (GNF) [10] (Chapter 2.3.2). Out of 9998 proteins composing the generic (not tissue-specific) protein-protein interaction network, the number of proteins expressed in each tissue varied between 1322 (13%) to 7113 (71%; mean=4500.8, standard deviation=1399.3; Table 4.1).

For each gene-disease association, we checked whether the causal gene was expressed in the tissue assigned to the disease. Interestingly, we found that a considerable fraction of the causal
genes were not expressed in their assigned tissue, ranging between 29% and 17% from 
MAS>8% to MAS>60%, respectively (Figure 4.1). Importantly, this fraction is significantly 
smaller than that expected by chance (38.25% lowly-expressed genes are expected on average 
across all MAS thresholds, \( p < 1E^{-5} \); see the end of this section for a detailed description of the 
statistical significance tests used).

**Figure 4.1:** The fraction of disease-causing genes expressed in the tissue of their pertaining 
disease, compared to the random expectation (obtained through a permutation test), for 
different MAS thresholds. The error bars represent the minimal and maximal fraction of 
expressed genes observed at random (over 10,000 permutations) for each MAS threshold.

To better understand why disease-causing genes might be lowly expressed in their associated 
tissues, we studied in detail the 76 lowly-expressed disease-causing genes under a MAS 
threshold of 40%. First, we analyzed the functional annotations of those genes using the DAVID 
web server. Notably, 44 (58%) of the genes were found to be involved in multicellular 
development processes (GO:0007275, FDR E-value: 1.8E-11), where 36 of those were directly 
involved in organ development (GO:0048513, FDR E-value: 7.1E-12). Hence, mutations in these 
genes might disrupt their early embryonic activity leading to pathologies in adult tissues 
regardless of their expression in these mature tissues. In addition, 17 (22%) of the genes were 
involved in cellular response to stimulus (GO: 0051716, FDR E-value: 1.8E-4) and, therefore, 
may not be expressed under normal conditions.
We also found that disease-causing genes that were lowly expressed in the tissue associated to the disease tended to be expressed in fewer tissues than expected (12.1 tissues on average compared to 17.5 at random, p < 1E-5). In addition, these genes exhibited lower mRNA levels across all tissues than the expected (150.4 versus 224.8 Affymetrix average difference (AD) units expected by chance, p < 1E-5). We believe that these observations may partly explain the phenomenon of low-expression of genes in the pertaining disease tissues. Henceforth, we focused on the majority of disease-causing genes where the gene is indeed expressed in its associated tissue, in order to get a better assessment of tissue-specific network’s contribution to gene prioritization.

**Statistical significance tests**

To estimate the number of disease genes that are expected to be lowly-expressed at their assigned tissue at random, we computed this quantity for 10,000 permutations of the tissue assignment vector taken for a given MAS threshold. We performed a permutation rather than assigning every disease a random tissue in order to take into account biases caused by tissues that are assigned to many diseases, such as the skin and the heart. As can be seen is Figure 4.1, the fraction of disease genes expressed in the disease’s assigned tissue expected at random tends to be similar across all MAS thresholds, and is significantly lower than the actual fraction.

We analyzed in detail a set of 76 lowly-expressed disease-causing genes under a MAS threshold of 40%. To test whether these genes are significantly lowly-expressed across all tissues, we generated a random set of genes that are lowly-expressed on the same disease tissues as the original set. I.e., for each original disease-causing gene that is lowly expressed in the associated tissue, we randomly selected another gene that is also lowly-expressed in that tissue. Next, we computed for each gene the number of tissues in which it was expressed (expression breadth) and compared the resulting distribution to that of a random set, repeating the comparison across 10,000 random sets built similarly. In the same manner we also evaluated the significance of the distribution of average expression level of the lowly-expressed genes.

**4.2 Properties of Node-Removal Tissue-Specific networks**

We constructed tissue-specific protein interaction networks using two methodologies, described in detail in Chapter 3.1. While one method, **Edge-Reweight (ERW)**, does not alter the network
topology at all, the other, **Node Removal (NR)**, involves eliminating nodes and edges from the network. Consequently, such a scheme dramatically changes the connectivity of the network: while a generic PPI network tends to have a single giant connected component containing most of the network’s nodes [50], the NR networks are shattered into a relatively small giant component and many small connected components.

Our generic human network consists of 91 connected components with the giant component covering 98% of the network (9796 proteins), yet the resulting tissue-specific NR networks have 545 connected components on average, with the average size of the giant component being 3907 proteins. In addition, on average there are 503.5 genes which are expressed but have no expressed neighbors. These ‘expressed singletons’ are disconnected from the network entirely and would be ‘dead end’ nodes for algorithms like PRINCE. Interestingly, in all of the 60 tissue-specific NR, networks, the largest connected component after the giant component is of size < 10. In fact, only 22 NR networks have a secondary connected component with size > 4. This suggests that the gigantic component remains densely interconnected, and that the new, small connected components are taken off its periphery. Table 4.1 presents to full list of properties of the NR networks.
### Table 4.1: A summary of the topological properties of Node-Removal tissue-specific networks.

<table>
<thead>
<tr>
<th>Network</th>
<th>#expressed proteins</th>
<th>#interactions</th>
<th>Size of gigantic component</th>
<th>#comps with size &gt; 1</th>
<th>#comps with size &gt; 4</th>
<th>#comps with size &gt; 10</th>
<th>#comps with size &gt; 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original PPI network</td>
<td>9998</td>
<td>41049</td>
<td>9796</td>
<td>91</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>'adrenal cortex'</td>
<td>4992</td>
<td>15728</td>
<td>4307</td>
<td>56</td>
<td>2</td>
<td>1</td>
<td>561</td>
</tr>
<tr>
<td>'b cells'</td>
<td>1322</td>
<td>2026</td>
<td>890</td>
<td>22</td>
<td>1</td>
<td>1</td>
<td>380</td>
</tr>
<tr>
<td>'lymphoblasts'</td>
<td>2762</td>
<td>7215</td>
<td>2243</td>
<td>28</td>
<td>1</td>
<td>1</td>
<td>459</td>
</tr>
<tr>
<td>'CD4+T cells'</td>
<td>1980</td>
<td>4242</td>
<td>1495</td>
<td>21</td>
<td>1</td>
<td>1</td>
<td>445</td>
</tr>
<tr>
<td>'CD8+T cells'</td>
<td>1999</td>
<td>4335</td>
<td>1518</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>442</td>
</tr>
<tr>
<td>'Cardiac Myocytes'</td>
<td>4343</td>
<td>13404</td>
<td>3708</td>
<td>45</td>
<td>2</td>
<td>1</td>
<td>536</td>
</tr>
<tr>
<td>'Cerebellum Peduncles'</td>
<td>5274</td>
<td>17058</td>
<td>4651</td>
<td>52</td>
<td>1</td>
<td>1</td>
<td>512</td>
</tr>
<tr>
<td>'Medulla Oblongata'</td>
<td>5847</td>
<td>20212</td>
<td>5223</td>
<td>55</td>
<td>3</td>
<td>1</td>
<td>501</td>
</tr>
<tr>
<td>'Monocytes'</td>
<td>2005</td>
<td>4025</td>
<td>1499</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td>456</td>
</tr>
<tr>
<td>'NK cells'</td>
<td>2486</td>
<td>5735</td>
<td>1937</td>
<td>32</td>
<td>1</td>
<td>1</td>
<td>481</td>
</tr>
<tr>
<td>'Occipital lobe'</td>
<td>5499</td>
<td>18368</td>
<td>4909</td>
<td>49</td>
<td>1</td>
<td>1</td>
<td>484</td>
</tr>
<tr>
<td>'Olfactory bulb'</td>
<td>4635</td>
<td>14555</td>
<td>4022</td>
<td>37</td>
<td>1</td>
<td>1</td>
<td>532</td>
</tr>
<tr>
<td>'Pancreatic islets'</td>
<td>3237</td>
<td>7787</td>
<td>2601</td>
<td>40</td>
<td>2</td>
<td>1</td>
<td>547</td>
</tr>
<tr>
<td>'Parietal lobe'</td>
<td>6045</td>
<td>21288</td>
<td>5440</td>
<td>52</td>
<td>4</td>
<td>1</td>
<td>486</td>
</tr>
<tr>
<td>'Prefrontal cortex'</td>
<td>5605</td>
<td>18693</td>
<td>5008</td>
<td>50</td>
<td>1</td>
<td>1</td>
<td>492</td>
</tr>
<tr>
<td>'skeletal muscle'</td>
<td>6516</td>
<td>24306</td>
<td>5978</td>
<td>36</td>
<td>2</td>
<td>1</td>
<td>458</td>
</tr>
<tr>
<td>'smooth muscle'</td>
<td>2284</td>
<td>4957</td>
<td>1784</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>463</td>
</tr>
<tr>
<td>'Temporal lobe'</td>
<td>5535</td>
<td>18801</td>
<td>4950</td>
<td>44</td>
<td>1</td>
<td>1</td>
<td>491</td>
</tr>
<tr>
<td>'TestisGermCell'</td>
<td>5342</td>
<td>18331</td>
<td>4746</td>
<td>43</td>
<td>1</td>
<td>1</td>
<td>505</td>
</tr>
<tr>
<td>'TestisLeydigCell'</td>
<td>5986</td>
<td>21061</td>
<td>5396</td>
<td>45</td>
<td>2</td>
<td>1</td>
<td>493</td>
</tr>
<tr>
<td>'TestisSeminiferousTubule'</td>
<td>5805</td>
<td>20399</td>
<td>5171</td>
<td>52</td>
<td>2</td>
<td>1</td>
<td>521</td>
</tr>
<tr>
<td>'Trigeminal ganglion'</td>
<td>6694</td>
<td>25125</td>
<td>6145</td>
<td>48</td>
<td>3</td>
<td>1</td>
<td>441</td>
</tr>
<tr>
<td>'WHOLE blood'</td>
<td>2855</td>
<td>7383</td>
<td>2320</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>477</td>
</tr>
<tr>
<td>'Whole Brain'</td>
<td>4737</td>
<td>14649</td>
<td>4128</td>
<td>43</td>
<td>1</td>
<td>1</td>
<td>517</td>
</tr>
<tr>
<td>'adrenal gland'</td>
<td>4419</td>
<td>13056</td>
<td>3768</td>
<td>46</td>
<td>2</td>
<td>1</td>
<td>551</td>
</tr>
<tr>
<td>'atrioventricular node'</td>
<td>6365</td>
<td>23312</td>
<td>5786</td>
<td>53</td>
<td>2</td>
<td>1</td>
<td>457</td>
</tr>
<tr>
<td>'bone marrow CD34'</td>
<td>2147</td>
<td>4552</td>
<td>1616</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>478</td>
</tr>
<tr>
<td>'bone marrow'</td>
<td>4072</td>
<td>11883</td>
<td>3427</td>
<td>45</td>
<td>2</td>
<td>1</td>
<td>546</td>
</tr>
<tr>
<td>'caudate nucleus'</td>
<td>5543</td>
<td>18776</td>
<td>4921</td>
<td>52</td>
<td>2</td>
<td>1</td>
<td>507</td>
</tr>
<tr>
<td>'ciliary ganglion'</td>
<td>6429</td>
<td>23699</td>
<td>5874</td>
<td>53</td>
<td>2</td>
<td>1</td>
<td>436</td>
</tr>
<tr>
<td>'dorsal root ganglion'</td>
<td>7113</td>
<td>27571</td>
<td>6580</td>
<td>53</td>
<td>3</td>
<td>1</td>
<td>411</td>
</tr>
<tr>
<td>'globus pallidus'</td>
<td>6050</td>
<td>21703</td>
<td>5465</td>
<td>50</td>
<td>2</td>
<td>1</td>
<td>472</td>
</tr>
<tr>
<td>'lymph node'</td>
<td>2907</td>
<td>6956</td>
<td>2359</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td>504</td>
</tr>
<tr>
<td>'salivary gland'</td>
<td>3423</td>
<td>8875</td>
<td>2779</td>
<td>37</td>
<td>1</td>
<td>1</td>
<td>565</td>
</tr>
<tr>
<td>'spinal cord'</td>
<td>4918</td>
<td>15535</td>
<td>4312</td>
<td>34</td>
<td>1</td>
<td>1</td>
<td>533</td>
</tr>
<tr>
<td>'subthalamic nucleus'</td>
<td>5912</td>
<td>20958</td>
<td>5324</td>
<td>52</td>
<td>2</td>
<td>1</td>
<td>471</td>
</tr>
<tr>
<td>'ADIPOCYTE'</td>
<td>3718</td>
<td>10192</td>
<td>3087</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>568</td>
</tr>
<tr>
<td>'Amygdala'</td>
<td>5546</td>
<td>18643</td>
<td>4960</td>
<td>44</td>
<td>1</td>
<td>1</td>
<td>490</td>
</tr>
<tr>
<td>'Appendix'</td>
<td>5436</td>
<td>18274</td>
<td>4841</td>
<td>44</td>
<td>1</td>
<td>1</td>
<td>500</td>
</tr>
<tr>
<td>'Heart'</td>
<td>4633</td>
<td>14277</td>
<td>3992</td>
<td>47</td>
<td>1</td>
<td>1</td>
<td>539</td>
</tr>
<tr>
<td>'Hypothalamus'</td>
<td>4900</td>
<td>15202</td>
<td>4301</td>
<td>44</td>
<td>1</td>
<td>1</td>
<td>503</td>
</tr>
<tr>
<td>'Liver'</td>
<td>4291</td>
<td>11918</td>
<td>3578</td>
<td>45</td>
<td>1</td>
<td>1</td>
<td>615</td>
</tr>
<tr>
<td>Tissue</td>
<td>Genes Scored</td>
<td>Genes Scored</td>
<td>Genes Scored</td>
<td>Genes Scored</td>
<td>Genes Scored</td>
<td>Genes Scored</td>
<td>Genes Scored</td>
</tr>
<tr>
<td>------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>'Lung'</td>
<td>3740</td>
<td>10370</td>
<td>3129</td>
<td>34</td>
<td>1</td>
<td>1</td>
<td>539</td>
</tr>
<tr>
<td>'Ovary'</td>
<td>2322</td>
<td>4593</td>
<td>1767</td>
<td>23</td>
<td>2</td>
<td>1</td>
<td>505</td>
</tr>
<tr>
<td>'PLACENTA'</td>
<td>4544</td>
<td>14765</td>
<td>3965</td>
<td>38</td>
<td>1</td>
<td>1</td>
<td>498</td>
</tr>
<tr>
<td>'Pancreas'</td>
<td>3010</td>
<td>6914</td>
<td>2374</td>
<td>41</td>
<td>1</td>
<td>1</td>
<td>551</td>
</tr>
<tr>
<td>'Pituitary'</td>
<td>4208</td>
<td>12308</td>
<td>3577</td>
<td>47</td>
<td>1</td>
<td>1</td>
<td>533</td>
</tr>
<tr>
<td>'Prostate'</td>
<td>4387</td>
<td>13622</td>
<td>3801</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>504</td>
</tr>
<tr>
<td>'TONGUE'</td>
<td>4718</td>
<td>14409</td>
<td>4042</td>
<td>53</td>
<td>1</td>
<td>1</td>
<td>562</td>
</tr>
<tr>
<td>'Thalamus'</td>
<td>5114</td>
<td>16526</td>
<td>4490</td>
<td>49</td>
<td>1</td>
<td>1</td>
<td>519</td>
</tr>
<tr>
<td>'Thyroid'</td>
<td>4652</td>
<td>14512</td>
<td>4057</td>
<td>43</td>
<td>1</td>
<td>1</td>
<td>507</td>
</tr>
<tr>
<td>'Tonsil'</td>
<td>3495</td>
<td>9947</td>
<td>2870</td>
<td>38</td>
<td>1</td>
<td>1</td>
<td>546</td>
</tr>
<tr>
<td>'Uterus'</td>
<td>3998</td>
<td>11989</td>
<td>3429</td>
<td>31</td>
<td>1</td>
<td>1</td>
<td>506</td>
</tr>
<tr>
<td>'cerebellum'</td>
<td>5376</td>
<td>17500</td>
<td>4754</td>
<td>49</td>
<td>2</td>
<td>1</td>
<td>516</td>
</tr>
<tr>
<td>'kidney'</td>
<td>5025</td>
<td>16209</td>
<td>4379</td>
<td>52</td>
<td>2</td>
<td>1</td>
<td>527</td>
</tr>
<tr>
<td>'skin'</td>
<td>5228</td>
<td>17057</td>
<td>4580</td>
<td>48</td>
<td>1</td>
<td>1</td>
<td>545</td>
</tr>
<tr>
<td>'testis'</td>
<td>4922</td>
<td>15551</td>
<td>4273</td>
<td>48</td>
<td>2</td>
<td>1</td>
<td>548</td>
</tr>
<tr>
<td>'thymus'</td>
<td>3019</td>
<td>7831</td>
<td>2469</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>497</td>
</tr>
<tr>
<td>'trachea'</td>
<td>4404</td>
<td>13814</td>
<td>3808</td>
<td>46</td>
<td>1</td>
<td>1</td>
<td>504</td>
</tr>
</tbody>
</table>

Table 4.1: Continued.

### 4.3 Tissue-specific prioritization of candidate disease genes

In order to prioritize candidate disease genes, we used the PRINCE prioritization algorithm. We applied PRINCE to score disease causing genes using both the original PPI network and the tissue specific networks built with the NR and ERW strategies. For the latter, we optimized the $rw$ parameter by following the cross-validation process presented below for multiple values (Figure 4.2). In the following results presentation we will concentrate on three representative choices of the $rw$ parameter: 0.5, 0.1 and 0.001. As explained above, we focused the performance evaluation on the subset of disease-gene associations where the causal genes are known to be expressed in the associated (henceforth referred to as ‘expressed disease genes association set’).
Figure 4.2: Fine-tuning the RW parameter. Edge-Reweight Tissue-specific networks with varying rw parameter were generated, and used as input for PRINCE. The graph shows how PRINCE's performance, measured by its AUC in a leave-one-out cross-validation trial, changes as a function of rw. (A) Taking into account only expressed disease-genes. (B) Taking into account the entire disease-gene association set.

We assessed the performance of the different prioritizations by performing a leave-one-out cross validation procedure. In every cross-validation trial, a single disease-gene association, \(<g,d>\) was removed from the association set. In addition, any other disease-gene association involving \(g\) was removed to avoid the trivial case where mutations in the same gene cause two very similar disorders. PRINCE was then executed to score the nodes of the network. For the purpose of
performance assessment, we constructed an artificial genomic interval of 100 genes around $g$. The scores assigned to these 100 genes were compared to $g$’s score. Using an artificial linkage interval enabled us to simulate the real-life scenario where prioritization is done only on genes residing within the genomic interval association with a disease. We then generated a ROC curve by bundling together all of the scores from all of the cross-validation trials, sorting them from highest to lowest and recording true- and false- positive rates at various score cutoffs. The actual causal genes were considered positive, and the rest of the genes were considered negative. Then, we calculated the AUC to produce a single, comparable performance measurement. Also, we assess the significance of the difference between the different AUCs, by employing a 25-fold cross validation procedure. We performed random partitions to 25 equal-sized sets of disease-gene associations, and used the standard deviation of the resulting AUC values as error bars in Figure 4.3.

For MAS > 40%, the AUC of the original, generic PPI network (0.825) was lower than that of each tissue specific network (0.85-0.88). The results, summarized in Figures 4.3-4.5, point to a moderate yet considerable improvement. Among the tissue-specific networks, TS-ERW (Tissue-Specific Edge ReWeight) with $\text{rw} = 0.5$, which is the most similar to the generic network, exhibits the smallest improvement. The improvement peaks for TS-ERW with $\text{rw} = 0.001$. TS-NR (Tissue-Specific Node Removal) and TS-ERW with $\text{rw} = 0.1$ networks have comparable AUC values.

When considering the entire disease-gene association data set, we observe that the AUC of the TS-ERW networks is still higher than the AUC of the original network, but by a less decisive margin (0.845 vs. 0.83). The best performance of an TS-ERW network was observed for a parameter of $\text{rw} = 0.1$, which is higher (thus implying less tissue-specific network) than the parameter yielding the best results for the subset of associations where the causal gene is expressed in the assigned tissue. The TS-NR networks yielded a low AUC curve due to the fact that unexpressed causal genes were automatically assigned a score of 0 by PRINCE.
**Figure 4.3:** Performance comparison between the generic and different variants of tissue-specific PRINCE, according to the ROC Area under curve (AUC) of predicting the correct prioritization. Only disease-gene associations from the expressed disease genes set is taken into account. Error bars represent the standard deviation of AUC values obtained when replacing leave-one-out with 25-fold cross validation of ten random partitions. Results are for a disease-tissue MAS threshold of 40%.

**Figure 4.4:** Performance comparison between the generic and different variants of tissue-specific PRINCE, according to the Area under curve (AUC) of predicting the correct prioritization. The entire disease-gene association set is taken into account.
Figure 4.5: The ROC (Receiver-Operator Curve) yielding the results presented in (A) Figure 4.3 and (B) Figure 4.4.

We further inspected the cross-validation results by comparing the ranking of true causal genes in the generic network to the tissue-specific networks on a case-to-case basis, in order to estimate how often the tissue-specific data improves the prioritization. Given a particular tissue-specific PPI network, instead of bundling all of the cross-validation results together, we regarded every test case (disease-gene association) in the data set separately, and counted in how many cases did the tissue-specific PRINCE gave the actual causal gene a better rank compared to the entire network PRINCE, in how many cases tissue-specific PRINCE gave a worse rank, and in how many cases both input networks yielded the same rank.

We found that for every tested MAS threshold, both ERW and NR tissue-specific PRINCE ranked true associations higher than the generic PRINCE in a majority of the cases (Table 4.2). This also holds true when considering the entire association set, with the exception of NR at MAS > 8% (Table 4.3). For example, when choosing a MAS threshold of 8% (the same threshold used in [47]) and reweight parameter $rw = 0.1$, we observe that TS-ERW PRINCE gives better ranks to 288 (47%) true associations, whereas the generic network PRINCE gives better ranks to only 58 (9.5%) true associations. 266 associations are identically ranked under both network types. Note that the margin becomes smaller for lower MAS thresholds, which is likely due to the linear relationship between MAS and the fraction of expressed causal genes (Figure 4.1).
Table 4.2: Evaluation of generic and tissue-specific gene prioritization methods on a case-to-case basis, using the expressed disease genes association set.

Table 4.3: Evaluation of generic and tissue-specific gene prioritization methods on a case-to-case basis, using the entire disease-gene association set.

4.4 Evaluation of disease-tissue association inference

To evaluate the quality of our disease-tissue association inference scheme, we compared it to actual disease-tissue associations from the literature using the disease-tissue association matrix of [47].

We filtered the disease-gene association set with MAS > 40% threshold in order to retain only high-confidence associations (~90% estimated accuracy). We considered only disease-tissue associations where the causal gene is known to be expressed in the tissue assigned by Lage et al. (referred henceforth as ‘literature tissue’). We then followed the procedure portrayed in Chapter 3.3, running PRINCE with TS-ERW networks as input, using the parameter $rw = 0.1$. This value of $rw$ was chosen for its robustly positive results for both disease-gene association sets presented in this chapter.

We compared the resulting predicted disease-tissue associations to the literature tissues. For every disease, we checked what ranking was given to the tissue which was assigned the highest
association score by Lage et al (Figure 4.6). To estimate whether the high number of highly-ranked literature tissues is statistically significant, we calculated a p-value by performing a permutation test as follows: For every disease-gene association, we assigned at random a tissue to the disease, selecting from the tissues where the causal gene is expressed (to counter the bias caused from focusing on the expressed disease genes association set), and marked the ranking we give the randomly assigned tissue. We repeated this procedure 1000 times. This was repeated both for the relative rank and absolute score tissue-ranking methods.

When using relative rank, in 53% of the cases the literature tissue was ranked first (allowing ties; p<0.013). When using absolute score, we obtained a more fine-grained separation of tissues and the results seem less decisive accordingly. Nevertheless, at a considerable portion of the diseases, the tissue associated according to Lage et al was given a high ranking: in 10% of the cases the assigned tissue was ranked first (p = 0.01), and in 38% of the cases, the tissue was among the top 6 tissues (top 10%; p = 0.002). These results further show the power of tissue specific PPI approach to detect tissue specific disease involvement.
Figure 4.6: Evaluation of tissue-disease association prediction. The histogram shows the distribution of our disease-tissue ranking for the tissues assigned by Lage et al. in every test case (disease-gene association). As can be seen, in more than half of the cases the associated tissue was predicted first among all other tissues.
Chapter 5

Conclusions

In the current study we aimed to infer disease causing genes using tissue-specific PPI networks. Most previous studies that used biological networks to infer causal genes were based on generic PPI or functional linkage networks and ignored differences between tissues [4]. Nevertheless, this generic approach may be limited as there are significant differences in expression patterns between tissues, both with respect to mRNA as well as in protein levels [10,46]. These differences imply that different tissues have different active PPI networks: a specific interaction may take place in some tissues while not in others, based on the expression distribution of the interacting proteins [25]. Moreover, these differences may explain why, in many cases, a disease may affect a specific tissue and not others: the same protein may be active in specific tissues and inactive in others, or can have different function in different tissues based on its different neighbors in the different networks. Following these observations, we decided to examine the utility of building and incorporating tissue specific PPI networks in our analysis. Adding tissue specificity information for various diseases [47] we were able to perform a tissue specific inference of disease causing genes.

We used the PRINCE algorithm for gene prioritization and contrasted between generic and tissue specific PPIs. We found that the tissue specific approach enhances the performance of the algorithm. In our analysis we used two different methods for tissue specific PPI networks construction that yield different gene prioritization performance. We observed that better results were obtained when modifying the weights of the networks edges (using the ERW method) compared to following the more drastic approach of removing lowly-expressed proteins from the network (using the NR method). There may be several explanations for these differences. First, it may be related to PRINCE algorithm. A global network-based algorithm such as PRINCE is expected to be less successful when applied to a more disconnected network, such as those generated by the NR approach. Moreover, even for other algorithms that are based on local inference which is not propagated, ERW may be proven more appropriate. NR is a very strict method, eliminating every unexpressed protein, while ERW assigns a continuous value for the
interaction based on the expression of the two interacting proteins. Thus, the former is likely to be less robust to noisy data such as gene expression \([49,51,52]\)).

Interestingly, as a preprocessing step for the tissue specific PRINCE algorithm, we found that a small yet considerable fraction of disease genes are not expressed in the tissue associated with the disease. There may be several explanations for this observation. First, it may reflect an error in measurements, either of the expression microarray or the computational inference of disease-tissue association. Nevertheless, such a substantial fraction of genes is more likely to reflect a true biological observation. One possibility is that the level of gene expression does not accurately represent the activity level of the protein in the cell. A protein may be active although having lower mRNA levels. Posttranscriptional modifications or higher translational efficiency may also result in higher protein levels or longer protein half-lives \([53,54]\). In addition to putative differences between mRNA and protein levels, obviously, there may be proteins who perform their function in relatively low levels Indeed, we found that many of the genes unexpressed at their disease’s assigned tissue also have low overall expression levels, suggesting that these genes might still be expressed at functional level in the diseased tissue.

Another possibility may be that the damage to the tissue was caused by a disruption of the protein function within the tissue in earlier developmental stages. Supporting this hypothesis we found that lowly-expressed disease causing genes are enriched with developmental annotations (multicellular development processes \((\text{GO:0007275})\) and organ development \((\text{GO:0048513})\)), and with stimulus response annotations \((\text{GO:0051716})\). Hence, the protein may not be active in the adult tissues at steady state (as manifested by its expression pattern), but a mutation in the genes may alter normal development of the tissue or may prevent the normal response of the tissue to stress or other stimuli, resulting in a disease. Finally, due to the complexity and the dependencies between tissues in a multi-tissue organism, a mutation in a protein active in one tissue may result in clinical pathology in another tissue. For example, Vitamin D – dependent rickets 1A \((\text{MIM: 264700})\) is primarily a bone disorder, but it is caused by a mutation in the gene \textit{CYP27B1}, which is active in the kidney and participates in the hydroxylation of Vitamin D into its active form, Calcitriol \([55]\). Overall, the role of lowly expressed genes in causing disease in a given tissue is a rather complex one and deserves a separate analysis that is beyond the current scope.
Some limitations of the current analysis should be mentioned. First, a direct tissue specific measure of protein abundance would be more adequate than mRNA levels as a measure for the presence and hence the activation and functionality of a protein in a tissue. However, despite the best efforts of the scientific community, compendiums of human tissue-specific protein abundance levels across multiple tissues are not nearly as comprehensive as the mRNA expression dataset we use, both in tissue scope, gene coverage and quantitative resolution [46]. In addition, the mutual expression of two possibly interacting proteins does not guarantee that the interaction will take place, and there are other factors that also should be considered such as, most prominently, the proteins’ phosphorylation state. Nevertheless, even given these limitations, our tissue specific approach performs better than the generic approach. As large scale data on tissue specific protein abundance and cellular localization will become available, it will be interesting to repeat the analysis reported here to see whether it yields better predictions, as may be expected.

In recent years, PPI networks were shown to be a powerful tool in many fields of molecular biology, such as predicting protein annotation and more [4,56,57]. We hope that the results of this study will encourage future studies to integrate expression data to create tissue specific PPI networks when analyzing PPI networks, to further increase their predictive and explanatory power.
Bibliography


The abstract identifies the genes causing inherited diseases and the screening between candidate genes is a fundamental challenge in medicine in the post-genomic era. The ability to automatically screen between dozens of genes that cause potential infections allows focusing on promising candidates and, as a result, wasting less time and resources in developing treatments and diagnostic methods for the disease.

Current algorithms used in screening often rely on protein interaction networks (PPI networks), which are based on the idea that genes causing diseases with similar phenotypes tend to be close to each other in such a network. However, to date, all screening algorithms have been based on networks that present a general picture of what happens in all body tissues, while in reality, diseases affect specific tissues in which the network is active and may be very different.

In this study, we aim to estimate how much specific information about a tissue can improve gene screening algorithms. We create tissue-specific PPI networks for various tissues, by combining general PPI networks and specific information about gene expression in a tissue, and test the power of these networks in screening candidate genes. We show that tissue-specific PPI networks significantly improve the performance of the gene screening algorithm compared to the general PPI network.

Furthermore, we analyze in depth the assumption that genes causing a disease in a specific tissue tend to be expressed in a healthy tissue, a hypothesis that underlies the model of the specific tissue network. Finally, we use tissue-specific networks as a guide to predict tissues in which the disease is expected to occur at the molecular level.
זיהוי ודירוג גנים הגורמים למחלות תורשתיות
באמצעות רשתות אינטראקציות חלבונים
спектיפיות לרקמה

דיירי רדרור גנים הגרומים למחלות תורשתיות
באמצעות רשתות אינטראקציות חלבונים
спектיפיות לרקמה

זו בור והוזה חלך מחדריון לכסבלת התחנה
"מוסמר וองיברסיט" – באוניברסיטה תל אביב

על מיר
עודד מגר

העבודה הוכנה בהדרכת של פרופסור רודר שער

משורט תש"ע