UNRAVELING THE GENETICS OF HUMAN DISEASES BY INTEGRATING PATTERNS FOR EPISTASIS DETECTION

AN HONORS THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
OF STANFORD UNIVERSITY

Hyunghoon Cho
Principal Adviser: Daphne Koller
May 2012

Abstract

The role of epistasis, or gene-gene interactions, is essential in many non-Mendelian human diseases that arise from the sophisticated interplay of various genes. However, due to the difficulties in conducting direct assays on humans, the methods for studying epistatic interactions in complex human diseases have been limited to the analysis of natural variations in genomic measurements. Further, the overwhelmingly large numbers of possible interactions to consider lead to computational difficulties and more importantly, a significant decrease in statistical power due to the multiple hypothesis testing correction. In the research presented in this thesis, we developed a general system that efficiently prioritizes candidate interactions using various types of prior information, including previously studied gene networks, evidence of individual gene's correlation with the phenotype, and network topology. The experimental results confirmed that the statistical power of this system is far superior to other approaches. In particular, we found 17,644 significant epistasis using our method as opposed to 471 using a naive method in a glioblastoma multiforme gene expression dataset obtained from the Cancer Genome Atlas. The validity of the discovered interactions was supported by permutation and biological analyses.

Acknowledgements

First and foremost, I would like to thank my advisor, Daphne Koller, for her guidance and support during the past three years of my time at Stanford. Not only she has been a great source of personal inspiration, her thoughtful advice guided me through several important life decisions. Her classes were one of the most difficult, but enjoyable courses I have taken at Stanford, and they have been the prime source of my passion to pursue the field of Artificial Intelligence.

The work presented in this thesis was in collaboration with Alexis Battle. I am very grateful for the countless hours she has patiently spent on providing me with her guidance. I admire her ability to quickly analyze a given phenomenon and design the next steps of investigation, without which this thesis would not have existed. I learned from her the fundamental tools of scientific research, and her teachings will guide me through the steps of my career to follow.

I would like to thank my good friend, David Wu, for many enlightening discussions. He always provided moral support when I needed one and also helped me tremendously as a personal english tutor while I was writing this thesis.

I would also like to thank Irene Kaplow and Yi Liu for kindly letting me stay in their office and being a good source of inspiration and life advice. I also thank Mary McDevitt for reviewing the draft of this thesis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, remarkable advances in molecular measurement techniques such as whole-genome sequencing and gene expression microarray has enabled the large-scale assessment of the molecular states of human genome [9]. This has taken us one step closer to fully understanding the genetic causes of human diseases. By studying the association between genomic measurements and complex phenotypes, we can broaden our understanding of the biological processes that govern disease development and enable more accurate prediction of disease risks and personalized treatments for patients.

Despite the advances in the field and many success stories [2, 18, 21, 23, 41], researchers have found that the current genetic models of complex traits can only explain a small fraction of familial clustering of the traits, an issue most commonly known as the *missing heritability* problem [31]. Many explanations for the missing heritability have been suggested, including large numbers of variants of smaller effect yet to be found and rare variants with larger effect not represented in current datasets due to small sample sizes. Among them, one of the most challenging and important problems is the poor detection of *epistasis*, or gene-gene interactions.

Epistasis is generally defined as the interaction between two or more genes with respect to a phenotype of interest [12]. In other words, if the effects of a particular gene on a phenotype are modulated by another gene, we say there is an epistatic interaction between the two genes. It is important to note that the term epistasis refers to the effects on the phenotype only, and does not necessarily imply a particular physical interaction between proteins or genetic elements.

Although there has been some notable successes in modeling Mendelian disorders where

a single gene accounts for most of the observed variance in the phenotype [14, 29], many genetic human diseases are complex traits involving large numbers of genes and their complicated interactions. This is precisely why incorporating epistatic interactions is believed to significantly improve existing genetic models of human diseases [16, 34]. However, gene-gene interaction networks for complex human diseases is far from being fully characterized due to many difficulties in epistasis detection.

1.1 Challenges of Epistasis Detection

Unlike a number of microorganisms such as budding yeast, for which gene-gene interactions have been thoroughly studied through synthetic experiments [15], humans are much more difficult subjects to study. This is primarily due to the higher number of genes involved and the impracticality of conducting direct assays on humans (we cannot directly modify a subject's DNA or force mating, for instance). As a result, analyzing natural variation in genomic measurements has been the most feasible method for finding epistasis in humans. However, this approach also poses a number of challenges, as described below.

In order to detect epistasis in a given molecular dataset, one needs to find a pair or a set of genes that shows statistically significant interaction. However, the number of combinations to consider is often very large: evaluating all n-way interactions among N genes requires us to test $\binom{N}{n}$ gene combinations. In the case of pairwise interactions ($n = 2$), this typically amounts to hundreds of billions of tests to perform, which is estimated to take over 3 years to run on a single processor [40]. While this number can be reduced to a reasonable level with the use of efficient graphics processing units (GPUs) and/or parallelization, the runtime is further multiplied when we consider interactions among three or more genes. Thus, even with the help of large computing clusters, exhaustively searching for epistasis is often computationally infeasible.

In addition to computational difficulties, testing a large number of hypotheses leads to a critical reduction in the statistical power of epistasis tests. *Multiple testing correction* is a common procedure that proposes stricter significance thresholds to counteract the rise of false positives caused by testing a large number of hypotheses. In epistasis detection, however, the number of hypotheses is so large that any attempt to control the number of false positives comes at a rather high cost of possibly losing many epistatic interactions.

1.2 Related Work

There has been active development of epistasis detection methods over the last few years [13]. In order to address the challenges of having too many candidate interactions, some methods aggregate multiple genetic variants into *risk groups*. A prime example of this is multi-factor dimensionality reduction (MDR, [37]) where subsets of variants are iteratively merged if the predictability of the phenotype can be improved. Even though this method allows us to exhaustively search for epistasis using all of the provided variants, it suffers from some major drawbacks, including that important interactions could be missed due to merging too many variants and that additional efforts are required to understand the interactions within each risk group.

In contrast, many approaches instead focus on restricting the search to a subset of possible interactions, thus decreasing the computational burden and the effect of multiple testing correction. One class of filtering methods use a variety of techniques including logic regression [39], MCMC logic regression [27], random forests [17], and random jungle [38] to stochastically explore the space of possible interactions. Another class of methods are based on a greedy approach, where some knowledge about the genetic variants, which could be either generic or dataset-specific, guides the filtering process. For instance, some methods use the evidence of lower order effects to filter candidate interactions [32].

Recently, a growing number of studies have been focusing on *biological plausibility* to reduce the search space [7, 19]. The intuition behind these studies is that epistatic interactions are more likely to be present between genes that are connected in a *biological network*, which encodes previously studied associations between pairs of genes. For example, Emily et al. consulted a set of directly connected gene pairs in a functional protein-protein interaction network, STRING, to reduce the number of epistasis tests from 125 billion to 71,000 in a case-control genotype dataset [19]. They were able to find significant pairwise interactions for Crohn's disease, bipolar disorder, hypertension, and rheumatoid arthritis using this method.

1.3 Our Approach

In the research presented in this thesis, we further developed the use of biological networks as a form of prior knowledge by introducing a *network connectivity* score that represents how connected two genes are in a given biological network. This gives us the ability to distinguish

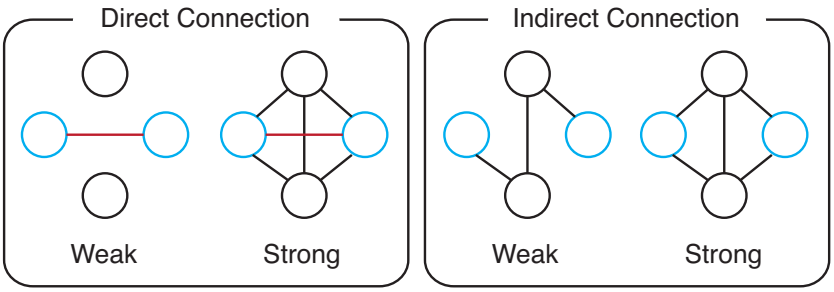

Figure 1.1: *Illustration of different types of connections in a biological network.* The connectivity between the two blue genes in a network of four genes varies with respect to how many connecting paths there are between the two genes. By using the network connectivity score, we can discover strong *indirect* connections that appear even stronger than some weak *direct* connections.

strong indirect connections from weak indirect connections (Figure 1.1), thus allowing us to use the biological networks more effectively. In addition, we have determined that other patterns of epistasis, including strong marginal effects and the existence of *hub genes* (group of genes with dense interactions), can also be valuable for predicting the likelihood of each candidate interaction. Further, instead of committing to a particular type of prior knowledge to use a priori for epistasis detection, we believe that it is better to *let the data speak for themselves*. That is, we want to learn the usefulness of each prior in the context of each dataset of interest. Based on this motivation, we developed an adaptive method that iteratively learns the applicability of different sources of prior information and prioritizes the remaining, untested candidate interactions using this weighted information. One important aspect of this method is its malleability; any additional sources of information can easily be incorporated. This thesis demonstrates the strong potential of this adaptive framework in pairwise epistasis detection setting and invites further efforts for improvements.

1.4 Outline

The remainder of this thesis is organized as follows. In Chapter 2, I provide relevant theoretical background and methodology for the key components of our detection method, including the construction of epistasis test and the computation of network connectivity score. In Chapter 3, I present empirical evidence for three patterns of epistasis (network

connectivity, marginal effect, and hub genes), which provide the foundation for our detection method. In Chapter 4, I describe in fine detail our adaptive method that combines these patterns of epistasis to prioritize candidate interactions. In Chapter 5, I provide performance evaluation of our method with statistical and biological evidence of the validity of the discovered interactions. Finally, Chapter 6 summarizes this thesis and discusses some limitations of our proposed method and future plans.

Chapter 2

Background & Methodology

In this chapter, I delve into some technical details of this study. These details will provide background necessary to an understanding of our epistasis detection method and statistical tools used for data analysis.

2.1 Statistical Test for Epistasis

In a typical disease association study, we are given a set of molecular measurements for a group of genes taken from subjects that exhibit varying degrees of the disease-related phenotype. In such setting, the concept of gene-gene interaction can be mathematically described in terms of the strength of *multiplicative* effects in a linear model of the phenotype. The following characterization proposed by Cordell [12] has been the most widely used statistical definition of epistasis.

Assume we are given a case-control data for a specific disease with each subject's DNA genotyped at multiple loci. Let p be the probability of being affected by the disease. Let x_i^j be an indicator variable representing the underlying genotype j at locus i. The log odds ratio of disease risk can be modeled with multiplicative terms as

$$\begin{aligned} \log\left[\frac{p}{1-p}\right] &= \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + \beta_3 x_2^1 + \beta_4 x_2^2 \\ &\quad + \beta_5 x_1^1 x_2^1 + \beta_6 x_1^1 x_2^2 + \beta_7 x_1^2 x_2^1 + \beta_8 x_1^2 x_2^2, \end{aligned} \tag{2.1}$$

where β_0 corresponds to the mean effect, β_{1-4} correspond to the additive effects, and β_{5-8} correspond to the multiplicative effects. If there is no epistasis between the two loci being

tested, then the interaction model is equivalent to the null model

$$\log\left[\frac{p}{1-p}\right] \quad = \quad \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + \beta_3 x_2^1 + \beta_4 x_2^2,$$

where the multiplicative terms have been removed. By comparing these two models using a likelihood ratio test (with four degrees of freedom), we can statistically test whether two loci are epistatic or not.

Now we slightly modify the above definition for the gene expression data with *continuous* variables as opposed to discrete genotypes. Assume we are given a disease-related phenotype data (e.g., number of days to death) with each subject's gene expression profile (which contains a real-valued expression level for each gene). Let y be the phenotype value, and z_1 and z_2 be the expression levels of the two genes. The modified interaction model for gene expression data can be written as

$$y \quad = \quad \beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 z_2 + \beta_4 z_2^2 + \beta_5 z_1 z_2, \tag{2.2}$$

where we have added the quadratic terms z_1^2 and z_2^2 to provide a correct baseline for the non-linear interaction term $z_1 z_2$. We can similarly test the significance of the multiplicative effect (corresponding to β_5) by comparing this model to the null model

$$y \quad = \quad \beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 z_2 + \beta_4 z_2^2,$$

using F-test [20], which is a statistical test for comparing two linear regression models.

2.2 Weighted False Discovery Rate Control

As mentioned briefly in Section 1.1, extremely large numbers of statistical hypotheses exacerbate the trade-off between statistical power (true positives) and Type 1 error (false negatives). Introduced by Benjamini & Hochberg [5], the false discovery rate controlling procedure gives an effective way to increase statistical power while maintaining the same overall error rate. The procedure is described below.

Let $p_{(1)} < \cdots < p_{(m)}$ be the ordered p-values from m hypothesis tests, where $p_{(0)}$ is defined as 0. The original false discovery rate controlling procedure rejects any null

hypothesis i for which $p_{(i)} \leq \tau$ with

$$\tau = \max \left\{ p_{(i)} \mid p_{(i)} \leq \frac{\alpha i}{m} \right\},$$

where α is a free parameter. This procedure controls the false discovery rate at level $\alpha m_0/m$, where m_0 is the number of true null hypotheses. This implies that the false discovery rate is bounded above by α.

With some information about the plausibility of hypotheses being tested, we can further increase statistical power by weighting the p-values. This effectively allows us to use a different significance threshold for each hypothesis with respect to our prior beliefs, while controlling the overall error rate. The following procedure for weighted false discovery rate control was introduced in [22]:

1. Assign weights $w_{(i)} > 0$ to each $p_{(i)}$ such that $\frac{1}{m} \sum_{i=1}^{m} w_{(i)} = 1$.
2. For each $i = 1, \ldots, m$, compute $q_{(i)} = p_{(i)}/w_{(i)}$.
3. Apply Benjamini & Hochberg's false discovery rate controlling procedure to $q_{(1)}, \ldots, q_{(m)}$ at level α.

This controls the false discovery rate at the same level as Benjamini & Hochberg's procedure. Note that the weights $w_{(1)}, \ldots, w_{(m)}$ are to be chosen independently from the corresponding p-values.

2.3 Network Connectivity Score

The connectivity of two nodes in a network (as seen in Figure 1.1) can be computed using a biased random walk model. The basic intuition is that if two nodes are highly connected, then a random walk starting on each node will have similar properties. This can be formulated as a special application of topic-sensitive PageRank algorithm introduced in [24].

Let $v^{(t)}$ be a vector of probability distribution over all nodes in the network, representing the state of the random walk at time t. Let M be a transition matrix for the given network, reflecting the edge structure and the weights of each edge (if applicable). Starting with $v^{(0)} = e_j$ (i.e., the random walk begins at node j), we can compute the stationary distribution for

node j using the following update rule until convergence:

$$v^{(t+1)} := \beta M v^{(t)} + (1 - \beta) e_j,$$

where $0 < \beta < 1$ and $1 - \beta$ represents the level of *taxation* — where the random walker teleports back to node j with probability $1-\beta$ at each time step. The personalized stationary distribution for node j, which we denote by u_j, represents the *reachability* of each node from node j. After computing the personalized stationary distribution for every node, we compute the cosine similarity between every pair of nodes (u_i, u_j). This gives us the network connectivity score for every pair of nodes.

In our research, we applied this method to a biological network where the nodes represent genes and the edges encode some type of direct interaction between pairs of genes. The network connectivity score in this setting reflects how strong the association between two genes is in a given network, taking indirect interactions into account.

2.4 Datasets

2.4.1 Disease Association Data

We replicated the experiments discussed in Chapters 3 and 5 with two fundamentally different molecular measurements. One dataset associates the disease status with genotypes at multiple loci (a typical setting of genome-wide association studies), and the other dataset associates complex phenotypes (e.g., number of days to death or age at initial pathological diagnosis) with gene expression levels. These datasets are described in more detail below:

- *Glioblastoma multiforme gene expression dataset (GBM).* This dataset contains expression levels of 11925 genes in 451 subjects, provided by the Cancer Genome Atlas[1]. After evaluating the predictability of a number of complex traits, we chose the number of days to death as the target phenotype for epistasis tests.

- *Type I diabetes case-control genotype dataset (T1D).* This dataset is obtained from Wellcome Trust Case Control Consortium [2]. It contains a case group of 1,766 individuals and a control group of 1,331 individuals. Originally, the genotypes were assayed at 500,000 different single nucleotide polymorphism (SNP) loci, but we ran the data

[1] http://cancergenome.nih.gov/
[2] http://www.wtccc.org.uk/

through a quality control filter and an individual correlation filter for computational reasons. In the end, we considered all pairwise interactions of 9,668 loci as candidates.

2.4.2 Biological Networks

Four different kinds of biological networks were used in our epistasis detection system, each drawing from different sources of information:

- *Protein-protein interaction network (HPRD).* Obtained from the Human Protein Reference Database[3], this network depicts manually curated scientific information including protein-protein interactions, post-translational modifications, and enzyme-substrate relationships. We augmented this network with another network of a similar origin, provided by InnateDB[4].

- *Pathway network (PWC).* Obtained from Pathway Commons[5], this network contains an edge between every two genes that are associated with a molecular pathway (a series of interactions in a cell that leads to a certain cell function).

- *Functional protein association network (STRING).* Obtained from STRING[6], this network contains known and predicted protein interactions. The interactions include both physical and functional associations, derived from various sources including genomic context and conserved co-expression.

- *Gene ontology network (GO).* The annotations for each gene representing its cellular component, molecular function, and biological process were obtained from the Gene Ontology[7]. Then this network was constructed by connecting two nodes if they share a Gene Ontology category.

[3]http://www.hprd.org/
[4]http://www.innatedb.ca/
[5]http://www.pathwaycommons.org/
[6]http://www.string-db.org/
[7]http://www.geneontology.org/

Chapter 3

Patterns of Epistasis

In this chapter, I discuss several patterns of epistasis that are informative of the presence of epistatic interaction between two genes. To analyze these patterns, we used Wilcoxon's rank-sum test [42] to test whether the distribution of epistasis p-values (see Section 2.1) in a subset of candidate interactions selected according to a particular pattern is significantly different from the distribution of epistasis p-values in the candidates outside of this subset. If the epistasis p-values in the subset tend to be lower (more significant) than those not in the subset, we say that there is a significant *enrichment* of epistasis in the subset. In this chapter, this statistical test is referred to as the *enrichment test.*

3.1 Network Connectivity

Previously studied gene-gene interactions, such as physical interaction between two proteins or co-existence in a pathway, yield more epistatic effects in a variety of genomic assays. In particular, a subset of candidate interactions that are connected in STRING network (Subsection 2.4.2) is significantly enriched for epistasis (Figure 3.1). Further, enrichment of epistatic interaction is not restricted to genes with direct associations, but rather is influenced by the connectivity of genes within the graph. For instance, two genes with numerous paths between them are more likely to be epistatic than genes with a single connection, even if direct. To quantify this property, we computed a *network connectivity* score for every pair of genes, based on a random walk model of the interaction graph (Section 2.3). This score incorporates several desirable properties, including number of paths, length (number of hops) of each path, and strength of each edge (if available). Evidence suggests

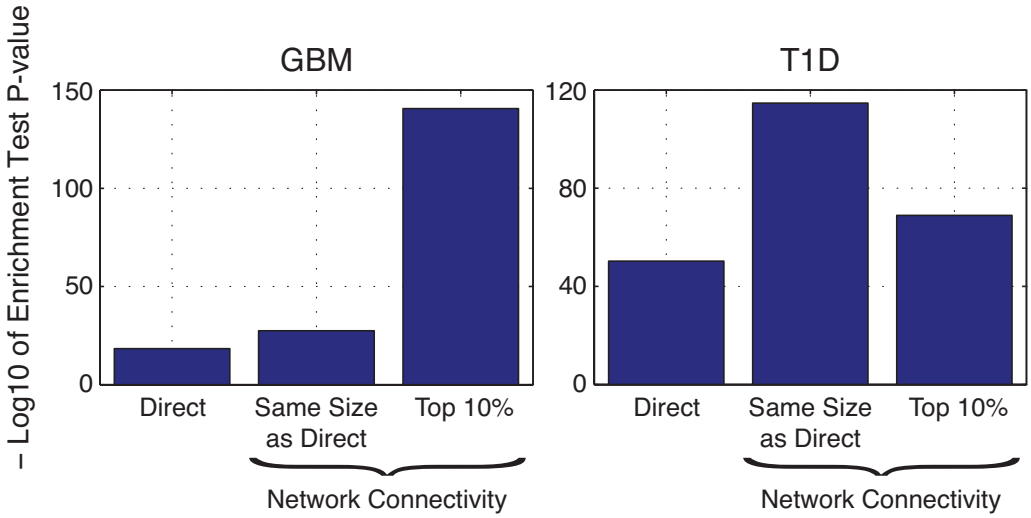

Figure 3.1: *Comparison of direct connections and strong indirect connections in STRING network.* The graphs show the results of enrichment test on three groups of pairs: (1) pairs with direct connection in STRING, (2) same number of pairs selected according to network connectivity score, and (3) top 10% pairs according to network connectivity score. The results suggest that, while direct connections indeed display significant enrichment, taking strong indirect connections into account leads to a much higher enrichment of epistasis.

that this serves as a better predictor of epistasis than direct connections (Figure 3.1). After evaluating four biological networks presented in Subsection 2.4.2, we found significant enrichment of epistasis in high ranges of network connectivity scores for HPRD, PWC, and STRING, but not for GO (Figure 3.2).

3.2 Marginal Effect

Genetic variants that individually affect the phenotype to a significant degree tend to have more epistatic interactions among them. This tendency has been used to guide the prioritization of candidate interactions in several studies [26, 32, 33]. Figure 3.3 demonstrates significant enrichment of epistasis among pairs of genetic variants with strong marginal effects. Note that the level of marginal effect for each pairwise interaction was defined as $-\log(\sqrt{p_1 p_2})$, where p_1 and p_2 are the two p-values of the individual variant's correlation with the phenotype. This pattern is very strong for the gene expression data, but we note that in the genotype data, the group with strongest marginal effects exhibit relatively less

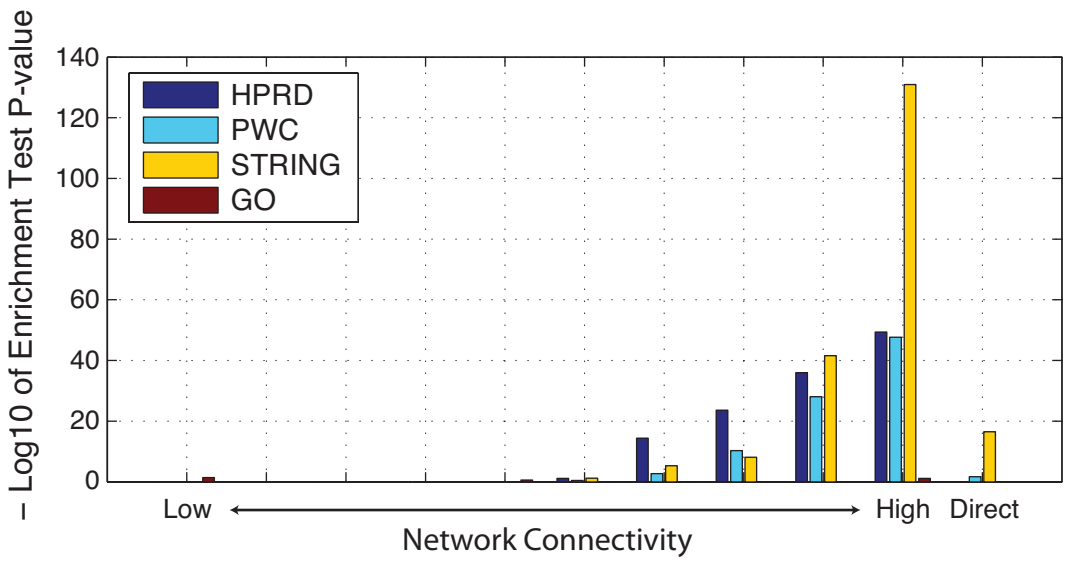

Figure 3.2: *Comparison of enrichment of epistasis in different biological networks.* Network connectivity scores were computed for each of the four biological networks (Subsection 2.4.2) and all of the candidate interactions in the gene expression dataset were divided into subgroups according to their connectivity. Each subgroup was then tested for enrichment. The graph shows clear correlation between the network connectivity score and the enrichment of epistasis for HPRD, PWC, and STRING. However, no significant enrichment was observed for GO.

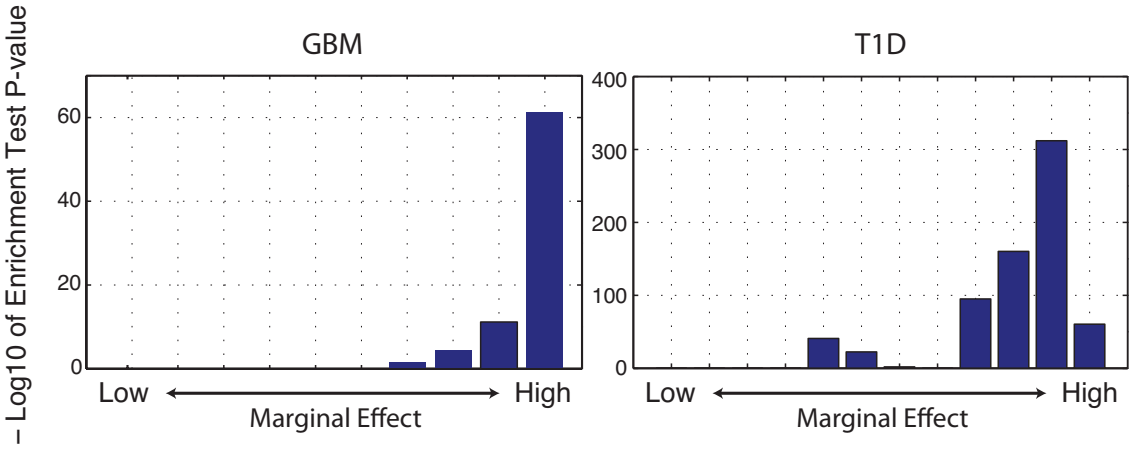

Figure 3.3: *Enrichment of epistasis in varying ranges of marginal effect.* The pairs of genes or SNPs were divided into subgroups according to their marginal effects and each subgroup was tested for enrichment of epistasis. The results demonstrate significant enrichment for higher ranges of marginal effect in both datasets.

epistasis than other groups with strong marginal effects. One explanation for this observation would be that individual variants that are highly correlated with the phenotype might not leave much room for improvement to the multiplicative terms in Equation 2.1, in terms of the predictability of phenotype.

3.3 Hub Genes

Epistasis hubs, or groups of genes with many epistatic interactions, are prevalent in many interaction networks. This is generally because there are several genes with central roles in the development of a particular disease. The existence of hub genes has been observed in genetic interaction assays for model organisms, such as budding yeast (*Saccharomyces cerevisiae*), where 1% of all genes tested contributed to almost 6% of all genetic interactions [15]. Figure 3.4 demonstrates that we do in fact see hub genes in our datasets. We found that the majority of nodes in the epistasis network has a small number of neighbors, while a few nodes have a very large number of them. In contrary, a randomly generated network with the same number of edges has a more balanced distribution with the majority of nodes having medium degree. This suggests that edges are concentrated around a small number of genes in the epistasis network, which indicates the existence of hubs.

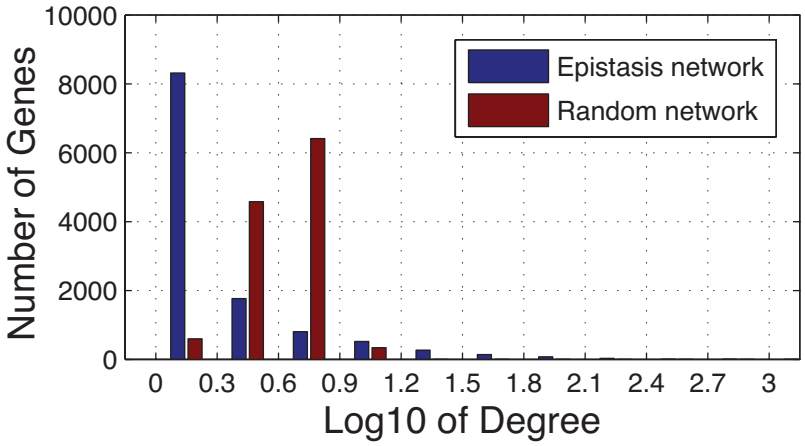

Figure 3.4: *Histogram of degrees in the epistasis network of the gene expression dataset.* The epistasis network was constructed from the set of statistically significant interactions found by our adaptive method. The degree of each node, which corresponds to the number of epistatic interactions each gene has, was counted and summarized as a histogram. A random network with the same number of edges was generated for comparison. The result suggests the existence of hub genes, indicated by edges being concentrated around a small number of genes.

Chapter 4

System Overview

In this chapter, I describe our adaptive method for epistasis detection that combines the three patterns of epistasis discussed in the previous chapter to repeatedly prioritize remaining candidate interactions.

4.1 Epistasis Prediction Model

I first present the description of a model that integrates patterns of epistasis to predict the likelihood of a candidate interaction. More specifically, this model (denoted by h) predicts the negative logarithm of the epistasis p-value of a given pair of genetic variants. The features of this model are explained below.

Let $\mathcal{X}$ be a set of all genetic variants, $\mathcal{N}$ be a set of biological networks, $\mathcal{C}$ be a set of Gene Ontology (GO) categories, and $\mathcal{G}$ be a set of all genes. Let $G : \mathcal{X} \mapsto \mathcal{G}$ be a function that maps individual genetic variant to a corresponding gene. Let $f_n : \mathcal{G} \times \mathcal{G} \mapsto \mathbf{R}_+$ be a function that takes two genes and returns the network connectivity score using network $n \in \mathcal{N}$ (Section 2.3). Let $C : \mathcal{G} \mapsto \mathcal{P}(\mathcal{C})$, where $\mathcal{P}(\mathcal{C})$ is the power set of $\mathcal{C}$, be a function that takes a gene and returns a subset of GO categories associated with that gene. Let $I : \mathcal{X} \times \mathcal{X} \mapsto \mathbf{R}_+$ be a function that takes two genetic variants and returns the level of marginal effect, which is defined as $-\log(\sqrt{p_1 p_2})$, where p_1 and p_2 are the two p-values of the individual variant's correlation with the phenotype. Given a pair of genetic variants

$(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$, our epistasis prediction model $h(x_1, x_2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \epsilon)$ is defined as

$$\begin{aligned} h(x_1, x_2|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \epsilon) &= \alpha_{G(x_1)} + \alpha_{G(x_2)} + \sum_{n \in \mathcal{N}} \beta_n f_n(G(x_1), G(x_2)) \\ &\quad + \sum_{c \in C(G(x_1)) \cup C(G(x_2))} \gamma_c + \delta I(x_1, x_2) + \epsilon \end{aligned} \tag{4.1}$$

where α_g is a weight for gene g, β_n is a weight for network n, γ_c is a weight for GO category c, δ is a weight for marginal effects, and ϵ is an intercept term. This is a fairly simple linear regression model that combines all of the aforementioned patterns: $\boldsymbol{\alpha}$ corresponds to knowledge about hub genes, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ correspond to knowledge about biological networks, and δ corresponds to knowledge about marginal effects.

4.2 Iterative Learning Procedure

Algorithm 4.1 describes the iterative learning procedure that repeatedly learns the parameters of the epistasis prediction model from the newly tested interactions and further prioritizes untested pairs. The step-by-step description of this procedure is as follows. First, we begin by choosing an initial batch of candidate interactions according to marginal effects. Note that one can use other heuristics (e.g., network connectivity) for this initial selection of pairs. After testing the initial batch, we learn the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta$ and ϵ described in Section 4.1 in a ℓ_1-regularized linear regression setting. Then we use Equation 4.1 to compute the estimated epistasis p-value for each of the untested pairs. After sorting the untested pairs by their estimated epistasis p-value, we select the top few most likely candidates for epistasis testing. Once we obtain more tested interactions, we update the model parameters and repeat this process. Once we run out of pairs to test or have exhausted the allotted time (recall that there often too many pairs to consider), we terminate the procedure.

4.3 Multiple Testing Correction

In order to declare a subset of interactions tested during the iterative learning procedure *statistically significant*, we use weighted false discovery rate control (weighted FDRC, Section 2.2) to correct for multiple hypothesis testing. This allows us to use a different significance threshold for each candidate according to our prior belief, while maintaining the overall

Algorithm 4.1 Pseudocode of iterative learning procedure with multiple testing correction

```
Test initial batch of candidates selected by marginal effects
while there are remaining untested candidates and the allotted time is not over do
    Learn model parameters α, β, γ, δ, ϵ from all previously tested interactions
    Estimate epistasis p-values of untested candidates using h(x1, x2|α, β, γ, δ, ϵ)
    Store current p-value estimates of untested candidates (for weighted FDRC)
    Test next batch of most likely candidates selected by estimated p-values
end while
Perform weighted FDRC on tested interactions using their stored estimates for weighting
Report statistically significant interactions
```

false discovery rate. Note that our prior belief about each candidate interaction is reflected by the p-value estimated by the trained epistasis model. Thus, we use the most recent p-value estimate prior to the point when each candidate was tested as the hypothesis weight (Algorithm 4.1). The weights among the tested candidates are then normalized to have mean one to qualify for weighted FDRC. This process produces a final list of statistically significant epistatic interactions found by our method.

We might note that there is an ambiguity regarding which function of the estimated p-value we should use as the weight. A simple choice would be $(-\log(\text{estimated } p\text{-value}))^c$, where c is a positive integer reflecting how confident we feel about our estimates. Developing an elegant method for selecting c is an area for further investigation. In the results, discussed in the following chapter, we used $c = 3$ for the gene expression data and $c = 1$ for the genotype data.

Chapter 5

Results & Discussion

In this chapter, I present and discuss experimental results of our epistasis detection method. First, we compared the statistical power of our method against other prioritization of candidates. Then, we conducted a permutation analysis and a biological analysis of the significant gene-gene interactions discovered in the gene expression data to present stronger evidence of their validity.

5.1 Statistical Power Evaluation

Applying our adaptive method to both genotype and gene expression datasets, we observed that the number of epistatic interactions discovered by our method is clearly superior to the following two baseline schemes: random prioritization of candidates and prioritization according to marginal effects (Figure 5.1). In particular, for the gene expression dataset, our method achieved a 364.31% increase in the number of significant interactions after every pair has been tested (Table 5.1). This is a huge improvement in statistical power, while the false discovery rate was controlled at the same level. Furthermore, even though the increase in the final number of significant interactions for the genotype data was not as dramatic, Figure 5.1 shows that our method identified the same number of epistatic pairs much earlier than the other methods.

It is interesting to note that the number of significant interactions started to decrease after a certain point in the iterative learning procedure in the gene expression dataset (as seen in the left graph in Figure 5.1). This observation can be attributed to the multiple

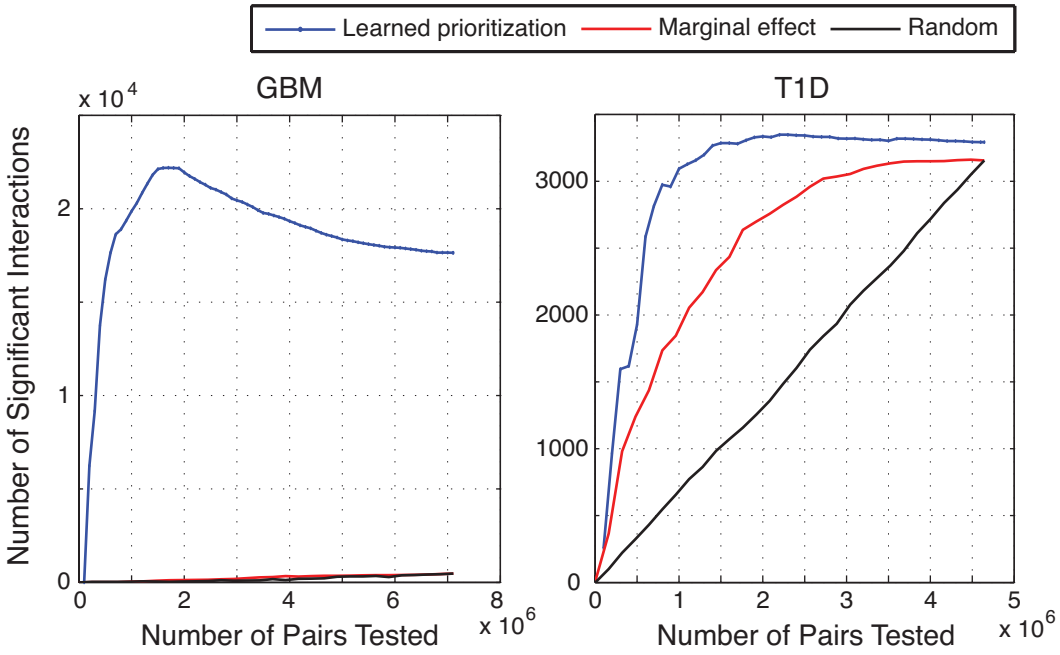

Figure 5.1: *Comparison of number of significant interactions discovered versus number of pairs tested.* These graphs compare how many significant interactions were found as we incrementally tested more candidates according to (1) our iterative learning procedure, where every batch of newly tested candidates informed the prioritization of the remaining candidates, (2) a descending order of marginal effects, and (3) random permutation of candidates. The false discovery rate was controlled at 0.05. The results show that our method finds more significant interactions and finds them more quickly.

	Number of Significant Interactions	
	GBM (Gene expression)	T1D (Genotype)
Baseline (FDRC)	471	3,155
Learned Prioritization (Weighted FDRC)	17,644	3,239

Table 5.1: *Comparison of final number of significant interactions discovered.*

testing correction (the more hypotheses we test, the stricter the significance threshold becomes). Naturally, one might want to terminate the iterative learning procedure at this point so as not to lose any more significant interactions. However, we note that new interactions are still being discovered after this point in spite of the fact that the overall number of significant interactions is declining. Thus, there is a tradeoff between finding possibly stronger (and more useful) interactions later on and retaining a larger number of significant interactions. The optimal choice for this tradeoff may vary depending on the nature of specific application.

The following sections focus more closely on the final set of 17,644 significant interactions found in the gene expression dataset to assess their validity.

5.2 Permutation Analysis

The statistical significance represented by p-values produced from a statistical test generally relies on several assumptions about the underlying data (e.g., normality of variables). For this reason, permutation-based p-values, computed by generating a large number of permutations of the data and observing how often the phenomenon of interest appears by chance, often provide a more accurate way of estimating the statistical significance of the results. To check the validity of the p-values computed by our epistasis tests, we generated 1,000 permutations of phenotypes across the subjects and re-tested all 17,644 significant interactions for each permutation. After noting that each interaction had a similar null distribution, we combined all 17,644,000 p-values to fit a Weibull distribution using the negative logarithm of the p-values. Using this as the empirical null distribution of p-values, we recomputed the significance of all 17,644 interactions. We found that these permutation-based p-values in fact match almost perfectly with the p-values from the epistasis tests (Figure 5.2). Also, controlling the false discovery rate at 0.05, all 17,644 empirical p-values were still deemed significant. This demonstrates the validity of the p-values computed by our epistasis tests.

5.3 Biological Analysis

Here, I present some biological evidence that support the plausibility of the significant interactions discovered. Looking at the genes associated with 17,644 significant interactions, we found that the top three most represented genes — *LGALS8*, *SPP1*, and *PMF1*, each

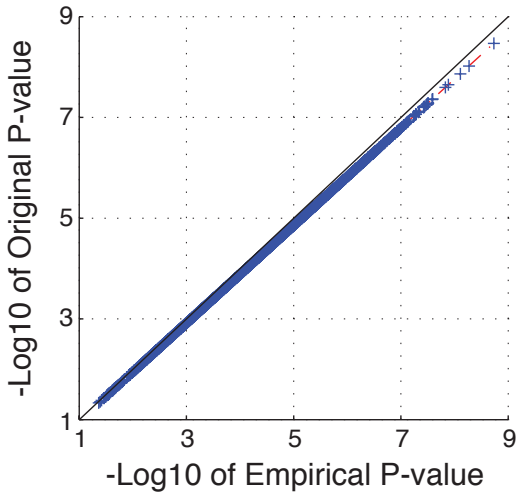

Figure 5.2: *QQ-plot of original p-values versus empirical p-values of significant interactions discovered in the gene expression dataset.* A total of 17,644,000 null hypotheses were generated from 1,000 permutations of phenotypes across the subjects. Using the Weibull distribution fitted from the set of epistasis p-values of these null hypotheses, we computed the empirical p-value of each of our 17,644 significant interactions found in the gene expression dataset. This graph demonstrates that the p-values from the epistasis tests match well with the empirical p-values.

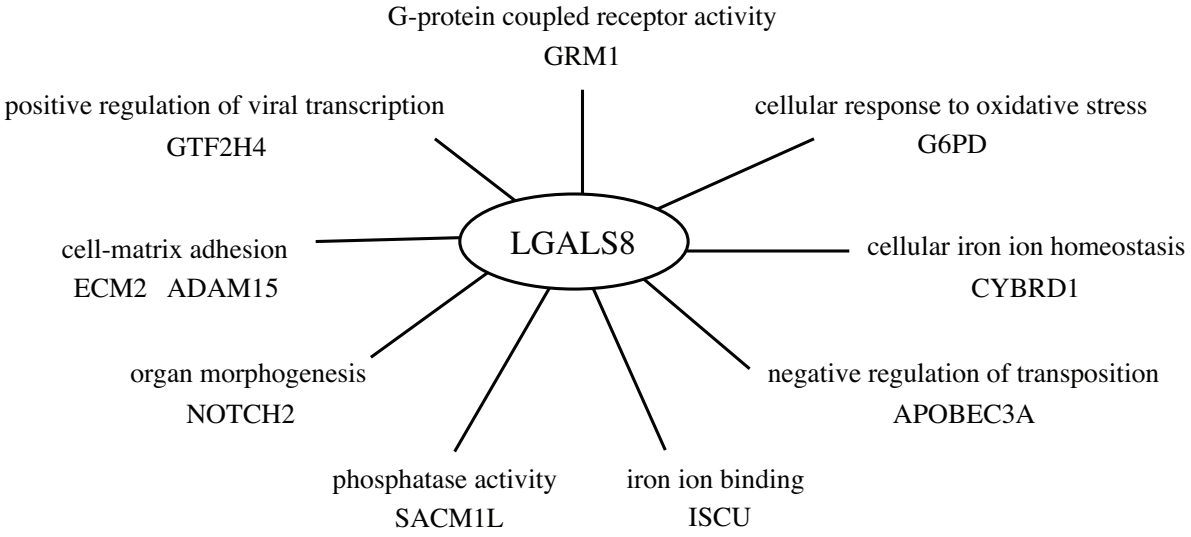

Figure 5.3: *Network of significant epistasis for glioblastoma multiforme involving* LGALS8.

involved in 724, 435, and 364 epistatic interactions, respectively — have been the subject of published studies suggesting significant association with a type of malignant cancer (recall that this dataset is on glioblastoma multiforme (GBM), which is a type of malignant brain tumor) [1, 10, 11]. In particular, the most represented gene, *LGALS8,* encodes a protein called galectin-8, which is known to be a modulator of multiple cell functions characteristic of tumor cells including cell growth [3] and cell migration [35]. Further, galectic-8's specific association with GBM has been characterized in [10].

In addition, we found that some epistatic interactions associated with *LGALS8* are also biologically meaningful. Figure 5.3 presents a number of significant interactions associated with this gene along with their corresponding Gene Ontology categories. As mentioned earlier, galectin-8 is involved in cell migration, which is closely related to cell-matrix adhesion (*ECM2* and *ADAM15*) and phosphatase activity (*SACM1L*) [28]. The interaction between *NOTCH2* and *LGALS8* is also biologically intriguing; the galectin family has been shown to have differential expression in different tissues during embryo development in mice [36] and *NOTCH2* also shows differential expression in the development of the mouse brain [25].

Last, we looked for biologically interesting clusters of genes, grouped according to their coefficients of interaction with other genes (corresponding to the multiplicative term in Equation 2.2). This coefficient represents the *directionality* and *strength* of the interaction

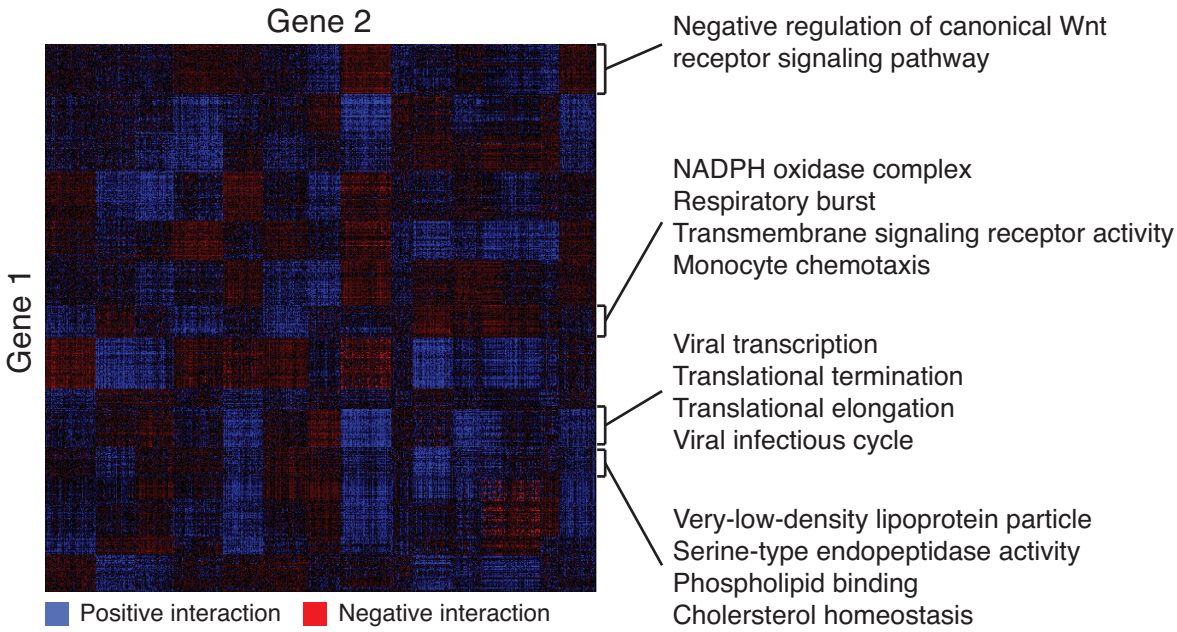

Figure 5.4: *Clustering based on the coefficient of interaction.* The figure visualizes all pairwise interactions among the 3,607 genes represented in the list of 17,644 significant interactions of the gene expression dataset. Each row and column in the matrix corresponds to the interaction profile of a gene. The color of each cell represents the *directionality* (blue is positive and red is negative), and the brightness represents the *strength* of interaction (brighter means stronger). The 20 clusters found using k-means algorithm are grouped together in the matrix. A number of Gene Ontology categories significantly represented in a cluster are shown.

between the two genes. The motivation behind this experiment is that clustering could reveal important *functional* groups of genes that tend to have similar interactions with other genes. We chose a total of 3,607 genes that were represented at least once in our final set of significant interactions and computed the interaction coefficient for every pair of genes. Figure 5.4 presents a visualization of 20 clusters obtained by the standard k-means algorithm [6]. As shown in the figure, we found a number of interesting clusters with significant enrichment of a few Gene Ontology categories. The association with GBM is supported by prior research for some of these categories, including "negative regulation of canonical Wnt receptor signaling pathway" [8], "monocyte chemotaxis" [4], and "very-low-density lipoprotein particle" [30].

Despite the fact that these biological analyses were conducted at a cursory level, the examples strongly suggest that there are many plausible (and interesting) interactions among the final set of significant interactions. We believe that a more thorough assessment could reveal new, biologically significant interactions.

Chapter 6

Conclusion

6.1 Summary

In this thesis, I have presented empirical evidence for three patterns of epistasis — network connectivity, marginal effect, and hub genes — using two different types of genomic, natural variation datasets (genotype and gene expression). In particular, I showed that our *network connectivity* score, computed based on a random walk model, significantly increases the effectiveness of various biological networks in terms of predicting pairwise interactions.

In addition, I presented an adaptive method that combined these patterns to predict the likelihood of each untested candidate interaction. Our method iteratively learned from the newly tested hypotheses and re-prioritized the remaining candidates. At the end of this procedure, the predictions made by our model for each candidate were used as hypothesis weights to produce a final list of significant interactions via the weighted false discovery rate controlling procedure.

Last, I showed that the statistical power of this procedure was evidently superior to other prioritization schemes, especially for the glioblastoma multiforme gene expression dataset, in which we found 17,644 significant interactions as opposed to 471 obtained by the standard false discovery rate controlling procedure. Further, additional evidence for the validity of this result was provided: empirical p-values computed by randomly permuting the data confirmed the statistical significance of these interactions, and evidence from published biological studies supported the plausibility of the reported interactions.

6.2 Limitations of Current Approach

There are several limitations of our adaptive epistasis detection method. First, due to our incomplete knowledge about human biological networks, the extent to which the network connectivity knowledge helps the discovery of epistasis is naturally limited by the quality and completeness of the provided networks. However, this is a problem that is hard to avoid given the limitations of human understanding; our use of network connectivity scores addresses this issue by using what is currently available in a more effective manner.

Second, because our method focuses on increasing statistical power by recognizing patterns from tested interactions, it could possibly lower the chance of discovering a novel interaction that is not expressed by any of the identified patterns. However, we note that looking for rarer interactions will directly lead to a decrease in statistical power since we would need to test more hypotheses in order to discover such interactions by chance.

Last, the effect of artifacts in genomic datasets of natural variation (e.g., biases introduced in the selection of subjects in the control group) on our statistical epistasis largely remains undetected. This limitation calls for a thorough biological analysis of our reported interactions (which I partially presented in this thesis), which would provide strong evidence for biological significance.

6.3 Future Directions

Our plans for the future work include the following. First, we plan to apply our adaptive method to various other types of data and diseases to see if we can consistently replicate our positive results. Second, we plan to assess the usefulness of discovered interactions in a phenotype prediction setting, by comparing the predictability of the phenotype using an additive model versus a multiplicative model with interaction terms. Third, we plan to extend our model to discover higher order interactions (e.g., interactions among triples of genes). This could reveal more interesting interactions that are not detected by pairwise tests. Last, in order to validate some of the reported interactions, it is our long-term goal to collaborate with a biology laboratory for experimental verification.

Bibliography

[1] Ainel Aleman, Virginia Cebrian, Miguel Alvarez, Virginia Lopez, Esteban Orenes, Lidia Lopez-Serra, Ferran Algaba, Joaquin Bellmunt, Antonio López-Beltrán, Pilar Gonzalez-Peramato, Carlos Cordon-Cardo, Javier García, Javier García Del Muro, Manel Esteller, and Marta Sánchez-Carbayo. Identification of PMF1 methylation in association with bladder cancer progression. *Clinical Cancer Research*, 14(24):8236–8243, 2008.

[2] A A Alizadeh, M B Eisen, R E Davis, C Ma, I S Lossos, A Rosenwald, J C Boldrick, H Sabet, T Tran, X Yu, J I Powell, L Yang, G E Marti, T Moore, J Hudson, L Lu, D B Lewis, R Tibshirani, G Sherlock, W C Chan, T C Greiner, D D Weisenburger, J O Armitage, R Warnke, R Levy, W Wilson, M R Grever, J C Byrd, D Botstein, P O Brown, and L M Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[3] Rinat Arbel-Goren, Yifat Levy, Denise Ronen, and Yehiel Zick. Cyclin-dependent kinase inhibitors and JNK act as molecular switches, regulating the choice between growth arrest and apoptosis induced by galectin-8. *The Journal of Biological Chemistry*, 280(19):19105–19114, 2005.

[4] T Asano, T An, S F Jia, and E S Kleinerman. Altered monocyte chemotactic and activating factor gene expression in human glioblastoma cell lines increased their susceptibility to cytotoxicity. *Journal of Leukocyte Biology*, 59(6):916–924, 1996.

[5] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, 57(1):289–300, 1995.

[6] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer, 2006.

[7] Zoltán Bochdanovits, David Sondervan, Sophie Perillous, Toos Van Beijsterveldt, Dorret Boomsma, and Peter Heutink. Genome-Wide Prediction of Functional Gene-Gene Interactions Inferred from Patterns of Genetic Differentiation in Mice and Men. *PLoS ONE*, 3(2):8, 2008.

[8] Cameron Brennan, Hiroyuki Momota, Dolores Hambardzumyan, Tatsuya Ozawa, Adesh Tandon, Alicia Pedraza, and Eric Holland. Glioblastoma Subclasses Can Be Defined by Activity among Signal Transduction Pathways and Associated Genomic Alterations. *PLoS ONE*, 4(11):10, 2009.

[9] Atul J Butte. Translational Bioinformatics: Coming of Age. *Journal of the American Medical Informatics Association*, 15(6):709–714, 2008.

[10] I Camby, N Belot, S Rorive, F Lefranc, C A Maurage, H Lahm, H Kaltner, Y Hadari, M M Ruchoux, J Brotchi, Y Zick, I Salmon, H J Gabius, and R Kiss. Galectins are differentially expressed in supratentorial pilocytic astrocytomas, astrocytomas, anaplastic astrocytomas and glioblastomas, and significantly modulate tumor astrocyte migration. *Brain pathology Zurich Switzerland*, 9(1):1–19, 2001.

[11] Juxiang Chen, Qihan Wu, Yicheng Lu, Tao Xu, Yan Huang, Judit Ribas, Xiaohua Ni, Guohan Hu, Fengping Huang, Liangfu Zhou, and Daru Lu. SPP1 promoter polymorphisms and glioma risk in a Chinese Han population. *Journal of Human Genetics*, 55(7):456–461, 2010.

[12] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, October 2002.

[13] Heather J Cordell. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[14] E H Corder, A M Saunders, W J Strittmatter, D E Schmechel, P C Gaskell, G W Small, A D Roses, J L Haines, and M A Pericak-Vance. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261(5123):921–923, 1993.

[15] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice L Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P St Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li, Zhen-Yuan Y Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph J San Luis, Ermira Shuteriqi, Amy Hin Yan H Tong, Nydia Van Dyk, Iain M Wallace, Joseph A Whitney, Matthew T Weirauch, Guoqing Zhong, Hongwei Zhu, Walid A Houry, Michael Brudno, Sasan Ragibizadeh, Balázs Papp, Csaba Pál, Frederick P Roth, Guri Giaever, Corey Nislow, Olga G Troyanskaya, Howard Bussey, Gary D Bader, Anne-Claude C Gingras, Quaid D Morris, Philip M Kim, Chris A Kaiser, Chad L Myers, Brenda J Andrews, and Charles Boone. The genetic landscape of a cell. *Science*, 327(5964):425–31, 2010.

[16] Robert Culverhouse, Brian K Suarez, Jennifer Lin, and Theodore Reich. A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461–471, 2002.

[17] Lizzy De Lobel, Pierre Geurts, Guy Baele, Francesc Castro-Giner, Manolis Kogevinas, and Kristel Van Steen. A screening methodology based on Random Forests to improve the detection of gene-gene interactions. *European journal of human genetics EJHG*, 18(10):1127–1132, 2010.

[18] Douglas F Easton, Karen A Pooley, Alison M Dunning, Paul D P Pharoah, Deborah Thompson, Dennis G Ballinger, Jeffery P Struewing, Jonathan Morrison, Helen Field, Robert Luben, Nicholas Wareham, Shahana Ahmed, Catherine S Healey, Richard Bowman, Kerstin B Meyer, Christopher A Haiman, Laurence K Kolonel, Brian E Henderson, Loic Le Marchand, Paul Brennan, Suleeporn Sangrajrang, Valerie Gaborieau, Fabrice Odefrey, Chen-Yang Shen, Pei-Ei Wu, Hui-Chun Wang, Diana Eccles, D Gareth Evans, Julian Peto, Olivia Fletcher, Nichola Johnson, Sheila Seal, Michael R Stratton, Nazneen Rahman, Georgia Chenevix-Trench, Stig E Bojesen, Børge G Nordestgaard, Christen K Axelsson, Montserrat Garcia-Closas, Louise Brinton, Stephen Chanock, Jolanta Lissowska, Beata Peplonska, Heli Nevanlinna, Rainer Fagerholm, Hannaleena Eerola, Daehee Kang, Keun-Young Yoo, Dong-Young Noh, Sei-Hyun Ahn, David J Hunter, Susan E Hankinson, David G Cox, Per Hall, Sara Wedren, Jianjun Liu, Yen-Ling Low, Natalia Bogdanova, Peter Schürmann, Thilo Dörk, Rob A E M Tollenaar,

Catharina E Jacobi, Peter Devilee, Jan G M Klijn, Alice J Sigurdson, Michele M Doody, Bruce H Alexander, Jinghui Zhang, Angela Cox, Ian W Brock, Gordon MacPherson, Malcolm W R Reed, Fergus J Couch, Ellen L Goode, Janet E Olson, Hanne Meijers-Heijboer, Ans Van Den Ouweland, André Uitterlinden, Fernando Rivadeneira, Roger L Milne, Gloria Ribas, Anna Gonzalez-Neira, Javier Benitez, John L Hopper, Margaret McCredie, Melissa Southey, Graham G Giles, Chris Schroen, Christina Justenhoven, Hiltrud Brauch, Ute Hamann, Yon-Dschun Ko, Amanda B Spurdle, Jonathan Beesley, Xiaoqing Chen, Arto Mannermaa, Veli-Matti Kosma, Vesa Kataja, Jaana Hartikainen, Nicholas E Day, David R Cox, and Bruce A J Ponder. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–93, 2007.

[19] Mathieu Emily, Thomas Mailund, Jotun Hein, Leif Schauser, and Mikkel Heide Schierup. Using biological networks to search for interacting loci in genome-wide association studies. *European journal of human genetics : EJHG*, 17(10):1231–40, October 2009.

[20] Franklin M Fisher. Tests of Equality Between Sets of Coefficients in Two Linear Regressions: An Expository Note. *Econometrica*, 38(2):361–366, 1970.

[21] Timothy M Frayling, Nicholas J Timpson, Michael N Weedon, Eleftheria Zeggini, Rachel M Freathy, Cecilia M Lindgren, John R B Perry, Katherine S Elliott, Hana Lango, Nigel W Rayner, Beverley Shields, Lorna W Harries, Jeffrey C Barrett, Sian Ellard, Christopher J Groves, Bridget Knight, Ann-Marie Patch, Andrew R Ness, Shah Ebrahim, Debbie A Lawlor, Susan M Ring, Yoav Ben-Shlomo, Marjo-Riitta Jarvelin, Ulla Sovio, Amanda J Bennett, David Melzer, Luigi Ferrucci, Ruth J F Loos, Inês Barroso, Nicholas J Wareham, Fredrik Karpe, Katharine R Owen, Lon R Cardon, Mark Walker, Graham A Hitman, Colin N A Palmer, Alex S F Doney, Andrew D Morris, George Davey Smith, Andrew T Hattersley, and Mark I McCarthy. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–894, 2007.

[22] C R Genovese, K Roeder, and L Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[23] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[24] T H Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, 2003.

[25] M Higuchi, H Kiyama, T Hayakawa, Y Hamada, and Y Tsujimoto. Differential expression of Notch1 and Notch2 in developing and adult mouse brain. *Brain research Molecular brain research*, 29(2):263–272, 1995.

[26] J Hoh, A Wille, R Zee, S Cheng, R Reynolds, K Lindpaintner, and J Ott. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Annals of Human Genetics*, 64(Pt 5):413–417, 2000.

[27] Charles Kooperberg and Ingo Ruczinski. Identifying interacting SNPs using Monte Carlo logic regression. *Genetic Epidemiology*, 28(2):157–170, 2005.

[28] Melinda Larsen, Michel L Tremblay, and Kenneth M Yamada. Phosphatases in cell-matrix adhesion and migration. *Nature Reviews Molecular Cell Biology*, 4(9):700–711, 2003.

[29] Richard P Lifton. Genetic dissection of human blood pressure variation: common pathways from rare phenotypes. *Harvey Lectures*, 100:71–101, 2004.

[30] Lenka Maletínská, Eleanor A Blakely, Kathleen A Bjornstad, Low-density Lipoprotein Receptor-related Protein, Lenka MaletÄś, Dennis F Deen, Laura J Knoff, and Trudy M Forte. Human Glioblastoma Cell Lines: Levels of Low-Density Lipoprotein Receptor and Low-Density Lipoprotein Receptor-related Protein. *Cancer Research*, pages 2300–2303, 2000.

[31] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L

Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[32] Jonathan Marchini, Peter Donnelly, and Lon R Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417, 2005.

[33] Joshua Millstein, David V Conti, Frank D Gilliland, and W James Gauderman. A Testing Framework for Identifying Susceptibility Genes in the Presence of Epistasis. *The American Journal of Human Genetics*, 78(1):15–27, 2006.

[34] Jason H Moore. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, 56(1-3):73–82, 2003.

[35] N Nagy, Y Bronckart, I Camby, H Legendre, H Lahm, H Kaltner, Y Hadari, P Van Ham, P Yeaton, J-C Pector, Y Zick, I Salmon, A Danguy, R Kiss, and H-J Gabius. Galectin-8 expression decreases in cancer compared with normal and dysplastic human colon tissue and acts significantly on human colon cancer cell migration as a suppressor. *Gut*, 50(3):392–401, 2002.

[36] Françoise Poirier. Roles of galectins in vivo. *Biochemical Society Symposium*, (69):95–103, 2002.

[37] Marylyn D Ritchie, Lance W Hahn, Nady Roodi, L Renee Bailey, William D Dupont, Fritz F Parl, and Jason H Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147, 2001.

[38] Daniel F Schwarz, Inke R König, and Andreas Ziegler. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, 26(3):1752–1758, 2010.

[39] Holger Schwender and Katja Ickstadt. Identification of SNP interactions using logic regression. *Biostatistics Oxford England*, 9(1):187–198, 2008.

[40] Kristel Van Steen. Travelling the world of gene-gene interactions. *Briefings in bioinformatics*, 13(1):1–19, March 2011.

[41] The Wellcome, Trust Case, and Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[42] F Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.