DATA AND NETWORK SCIENCE FOR NOISY HETEROGENEOUS SYSTEMS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Andrew Kent Rider

Nitesh V. Chawla , Co-Director

Scott J. Emrich , Co-Director

Graduate Program in Computer Science and Engineering
Notre Dame, Indiana
April 2013

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

DATA AND NETWORK SCIENCE FOR NOISY HETEROGENEOUS SYSTEMS

Abstract

by

Andrew Kent Rider

Data in many growing fields has an underlying network structure that can be taken advantage of. In this dissertation we apply data and network science to problems in the domains of systems biology and healthcare. Data challenges in these fields include noisy, heterogeneous data, and a lack of ground truth.

The primary thesis of this work is that the application of data mining and network science to data with these challenges must be carefully joined with domain knowledge. In the fields of systems biology and healthcare, data mining is increasingly being used to create models that represent the current state of understanding of important problems. These models are used to determine the direction of future work and to evaluate novel approaches. Therefore, any systematic bias in the models can be detrimental to scientific progress. For these same reasons, data mining has enormous potential to contribute to advances in our understanding.

Through our study of data and network science in this context we innovate new methods and highlight open and important problems. We *strongly advocate* the use of multiple measures for relationships in data in addition to heterogeneous data for the construction of network models, as relationships are often a matter of degree and no single measure or data set can capture everything about a problem.

To my family. For everything.

CONTENTS

FIGURES

TABLES

ACKNOWLEDGMENTS

I would like to thank my advisors, Nitesh Chawla and Scott Emrich for providing the opportunity to research important and fascinating problems. Their guidance and insight have been invaluable to me. I would also like to thank my committee members: Prof. Kevin Bowyer, Prof. Michael Ferdig, and Dr. Ashok Srivastava for their questions, suggestions, and guidance, which have improved this document and my work as a whole. My thanks also to my many collaborators.

I am also grateful to the Eck Institute for Global Health, which has provided both financial support and a vibrant community for learning about problems and research that impact our world.

Finally, I would like to thank my friends and family whose support has meant the world to me.

CHAPTER 1

INTRODUCTION

Interpretable models are essential for aiding data understanding in any domain. Networks, clustering algorithms, and topic models provide intuitive representations of data for domain experts. By understanding the factors that contribute to creating meaningful models in systems biology we can improve understanding of the domain, which can in turn improve our knowledge of how to create better models for this data.

For any data mining task, it is important to understand the domain and the goal of the analysis. The key to a successful model is the use of an appropriate measurement for relationships within the data. We demonstrate this in Chapter 3, in which we describe a novel approach to identifying overlooked genetic associations in *Plasmodium Falciparum*.

Although it is not always clear how to measure relationships in data, it is an essential step in the creation of network models and clusters. When the goal of a study is an exploratory analysis, the appropriate measure may be unclear. In Chapter 4 we propose a method to learn ensemble similarity measures in order to improve clustering algorithms without knowledge of what might constitute a good distance measure.

The use of a network representation of data can be beneficial even when the underlying problem is not necessarily inherently a network. Most approaches that utilize such a representation use it to replace the "normal" representation where features are assumed to be independent. In Chapter 5 we describe an ensemble topic modeling approach that explicitly utilizes disease co-occurrence and frequency to predict disease risks for patients.

As more data is collected and more methods of collecting data are invented, methods to integrate heterogeneous data and to extract knowledge from large scale data data become increasingly important. Unfortunately, the rate at which data is collected far outstripes the rate at which high-confidence knowledge is produced. One of the fundamental problems in data mining is how to evaluate models when there are few classified examples or uncertainty in the labeled data. In Chapter 6 we study the reliability of common performance metrics in the presence of mislabeled negative class data. This is especially important in systems biology and healthcare, where a pair of non-interacting proteins may simply not have been observed interacting or a patient may have a disease but may not have been tested for it.

In Chapter 7 we briefly survey current integrative network approaches to model genetic interactions. These approaches utilize heterogeneous data sources to construct models of the entire genome of an organism. The use of heterogeneous data enables such approaches to determine relationships between genes with greater confidence than alternative approaches. Even when integrating heterogeneous data, the combination of the way relationships are measured in the models and the data type the measure is applied to may have a significant impact on the resulting model. We demonstrate that network models are strongly influenced by the combination of data type and distance measure in Chapter 8.

One theme throughout this work is that data does not exist in a vacuum. Domain knowledge can be critical to the design of meaningful and interpretable data mining models. We validate this view once more in Chapter 9, in which we describe a collaborative effort and and approach to uncovering relationships between promoter sequences and gene expression that won a global competition in 2011.

1.1 Contributions

- A novel alternative approach to identifying overlooked genetic associations in *Plasmodium Falciparum* (Chapter 3).
- A novel approach to create ensemble distance measures that utilizes the relatively few available labeled gene interactions to improve unsupervised clustering of genes (Chapter 4).
- An ensemble topic modeling approach that explicitly utilizes disease co-occurrence to predict disease risks for patients while preserving data privacy (Chapter 5).
- An evaluation of the reliability of common performance metrics in the presence of mislabeled negative class data (Chapter 6).
- A survey of integrative network approaches to model genetic interactions (Chapter 7).
- A thorough investigation into how distance measures and data types interact to create a “good” biological network model (Chapter 8).
- A prize winning innovative approach to uncovering relationships between promoter sequences and gene expression (Chapter 9).

CHAPTER 2

PROBLEM SETTING

In this chapter we outline the data mining challenges in the domains of systems biology and healthcare and introduce the terminology used throughout this dissertation.

The goal of systems biology is to gain a more complete understanding of biological systems by viewing their components and the interactions between them simultaneously. In this context the components of a biological system can be genes, proteins, or any other distinct unit that is a part of a cell.

This setting contains a unique combination of challenges.

- Ground truth knowledge is frequently challenged and is far from complete for any organism.
- The experimental conditions under which data are gathered has a strong effect on the phenomena that occur in the data. The result is that data sets ostensibly about the same kind of information can be very different.
- Interactions in a biological system can be measured in numerous ways, each of which contains a distinct piece of information about the system and none of which capture all of the information in the system.
- Many different biological properties can be measured but are seldom measured on the same data in the same experiment. Therefore, heterogeneous data about a single organism is usually also from different experiments and individuals.

In data mining terms, what this means is that we have sparsely labeled, noisy, heterogeneous data with unknown biases. This combination of challenges can make the learning and evaluation of models difficult.

Here we outline many of the general approaches to machine learning in this context as well as their benefits and drawbacks for these problems.

2.0.1 Supervised learning

In supervised learning problems, one has a set of observations or *instances* with *labels* that describe group membership or class. Each instance is a collection of *features* that describe measurements or aspects of the instance. Given class labels, the task is to form a model that can predict the class of unlabeled instances. There may be labels for one or many groups. The label of interest is typically called the *positive class* and everything else is called the *negative class*. A supervised approach learns a *decision boundary* that marks the separation between classes with a combination of specific feature values or a function.

In a setting where the available class labels do not contain enough information or are noisy, learning a useful decision boundary can be difficult. It is often the case with biological data that one class contains far fewer instances than another. A common task in systems biology is to identify interacting genes. However, even for an organism with relatively few genes, the number of known interactions is much less than the number of known interactions. This problem is known as *class imbalance*. Algorithms such as NaïveBayes and Hellinger Distance Trees are relatively robust to this problem, but it can also pose problems for evaluation regardless of the underlying algorithm, as we describe in Section 2.0.3.

In addition to the relative lack of class labels, there are often far more features available in biological data sets than instances. This problem is known as *the curse of dimensionality*. Explicitly modeling the data as relationships can alleviate this problem somewhat by, for example, transforming a 10 by 1000 matrix of instances and features into a 1000 by 1000 matrix of similarities.

2.0.2 Unsupervised learning

In unsupervised learning problems, instances have no class labels. Given data without class labels, the task is to extract meaningful patterns from the data that reveal insights about the underlying system. Clustering is a common approach to unsupervised learning, which simply organizes instances into groups based on a *similarity measure* or *distance metric*. Similarity measures or distance metrics may utilize any number of measures or statistical tests to determine the extent to which concepts are related to each other. Two of the most commonly used similarity measures are pearson's correlation and mutual information.

Networks also often rely on similarity measures. Learning groups of similar objects can be informative but may not be specific enough for the intended purpose. The focus on explicit relational information in the network approach can be both constraining and can highlight aspects of the problem that might not be apparent otherwise. Networks are composed of nodes and edges. An example network could be composed of people (*nodes*) and their friend relationships (*edges*) in a social network. In systems biology a network often views genes as nodes and physical, genetic, or other interactions between them as edges. The *topology* of a network can reveal the general structure of relationships in the data and is often used as a means to compare networks [6]. Local measures of the topology surrounding individual nodes have been used to identify specific types of interactions in networks [71]. We will discuss current network models in systems biology in depth in Chapters 7 and 8.

2.0.3 Evaluation

These different approaches require different performance measures. For problems with class labels, a common approach is to rank instances by how confident a classifier is that they belong to the positive class. Many performance measures are based on precision, recall (or true positive rate), and false positive rate. These are described in equations 2.1 - 2.3. These measures and more are defined in terms of true positives (tp), false positives (fp), and false negatives (fn). A true positive is an instance that is predicted to belong to the positive class and in fact does. A false positive is an instance that is predicted to belong to the positive class but actually belongs to the negative class. A true negative is an instance that is predicted to belong to the negative class and does. Finally, a false negative is an instance that is predicted to belong to the negative class but actually belongs to the positive class.

$$\text{precision} = \frac{tp}{tp + fp} \tag{2.1}$$

$$\text{recall} = \frac{tp}{tp + fn} \tag{2.2}$$

$$\text{fpr} = \frac{fp}{fp + tn} \tag{2.3}$$

The *Receiver Operating Characteristic* (ROC) curve is often plotted to evaluate the trade-off between the true positive rate and the false positive rate as the ranked instances are iterated through. Alternatively, the area under the ROC curve (AUROC) is often used as a single summary number for the ROC curve. A second common rank-based performance measure is the precision-recall (PR) curve, which shows the trade-off between precision and recall as ranked instances are iterated through. Similarly to ROC curves, the PR curve can be summarized by it's area (AUPR).

AUROC and AUPR are commonly used to evaluate classifiers but are not always appropriate for problems that lack labeled data. The usefulness of labeled instances depends on how well the available labels represent the actual classes. In cases where many of the instances are unlabeled, there may simply not be enough labeled instances to characterize the class boundary well. This is one focus of Chapter 6.

Similarly, AUROC and AUPR may not be appropriate for models of *heterogeneous data*, where class labels may represent one kind of data and the features represent many different types. In such situations, the use of labels biases the interpretation of the model towards the data type represented by the label. This is a good thing if the goal is to identify additional instances of a specific type, but may not be a useful way to interpret more exploratory or unsupervised approaches.

CHAPTER 3

DISCOVERING A MEASURE TO IDENTIFY INTERESTING GENETIC ASSOCIATIONS

We begin our contributions with a measure for interesting genetic associations using expert domain knowledge. Detailed background knowledge and an understanding of the purpose of the study were the keys to discovering this novel approach. This underscores a theme of this dissertation: that data mining must be applied with careful consideration of domain knowledge to achieve maximum impact.

3.1 Introduction

Determining regulatory interactions between genes and the relationship between genotype (the set of genes belonging to an individual) and phenotype (measured as the amount of gene expression) is a fundamental step towards understanding biological systems. An expression quantitative trait loci (eQTL) study is an approach using gene expression data and genetic variation between individuals to calculate the association between expression and genotypes. In the context of eQTL studies an "expression trait" refers to the quantity of labeled (c)DNA hybridizing to a single probe on a microarray. An eQTL is a strong association between one locus in the genome and one expression trait. eQTLs describe the global relationships, or regulatory architecture between expression levels and genotypes in an organism [79]. Genotype is identified by specific "markers" or unique sequences of DNA that are inherited.

Complexity and noise in eQTL studies make it difficult to distinguish potential regulatory relationships among the many interactions. The predominant method of identifying eQTLs finds associations that are significant at a genome-wide level. The vast number of statistical tests carried out on the data make false negatives very likely. Corrections for multiple testing error render genome-wide eQTL techniques unable to detect modest regulatory effects. Modest effects should not be overlooked as much of the behavior of a cell may emerge from simultaneous modest effects.

We propose an alternative method to identify eQTLs that builds on the strengths of traditional approaches. In contrast to genome-wide techniques, our method determines the significance of an association between an expression trait and a locus with respect to the set of all associations to the expression trait. The use of this specific information facilitates identification of expression traits that have an expression profile that is characterized by a single exceptional association to a locus.

Our approach identifies expression traits that have exceptional associations regardless of the genome-wide significance of those associations. This property facilitates the identification of possible false negatives for genome-wide significance. Further, our approach has the property of prioritizing expression traits that are affected by few strong associations. Expression traits identified by this method may warrant additional study because their expression level may be affected by targeting genes near a single locus.

We demonstrate our method by identifying eQTL hotspots in *Plasmodium falciparum* (malaria) and *Saccharomyces cerevisiae* (yeast). We demonstrate the prioritization of traits with few strong genetic effects through Gene Ontology (GO) analysis of Yeast. Our results are strongly consistent with results gathered using genome-wide methods and identify additional hotspots and eQTLs.

New eQTLs and hotspots found with this method may represent regions of the genome or biological processes that are controlled through few relatively strong genetic interactions. These points of interest warrant experimental investigation.

3.2 Background

eQTL studies determine the associations between expression traits and loci on a genome-wide scale, often involving millions of statistical tests [59, 19]. This process leads to a multiple testing problem, where as the number of statistical tests increases, more exceptionally unlikely observations are seen purely by chance. eQTL studies are particularly susceptible to this problem, especially when larger genomes, marker sets, or sets of individual genotypes are considered.

It is common among eQTL studies to compensate for multiple testing by using a permutation test [23, 70]. A permutation test enables measurement of genome-wide significance for associations in eQTL studies by simulating the null hypothesis of no differentially expressed genes. For each iteration of the permutation test each expression trait is associated with a random genotype and the association between the genotype and expression trait is recalculated. The maximum value for an iteration of the permutation test is an estimate of the maximum association that is expected purely by chance when there is no significant association between genotype and expression level. After a number of repetitions of this process the maximum value for each repetition is used as one element in the null distribution. This distribution represents the relationship between genotype and expression level under the assumption that there is truly no significant association between genotypes and expression traits. The intuition behind this process is that a truly differentially expressed expression trait will have a stronger association than even the largest associations that occur by chance in the null distribution.

The stringent thresholds imposed by error correcting methods such as the permutation test limit the ability of traditional eQTL techniques to identify moderate genetic effects. Finding false negatives by simply lowering the threshold for significance would undermine the error correction so we focus our approach on measuring significance at the individual expression trait scale. Our approach capitalizes on the fact that false negatives are most likely to occur near the cutoff for significance and should therefore be very significant relative to the vast majority of observations. This information allows us to create a model distribution that we expect describes the 'association profile' of an expression trait with interesting genetic effects. Our approach uses Hellinger distance to determine which traits most closely match this model distribution. The approach builds on genome-wide techniques by measuring the similarity between distributions of genome-wide corrected p-values and allows us to simultaneously utilize corrections for multiple testing and detect associations that are moderate on a genome-wide scale but significant for individual expression traits.

3.3 Results and Discussion

The foundation for this study was the work of Gonzales *et al.* who performed eQTL analysis across the progeny of the Hb3 drug resistant and the Dd2 drug sensitive malaria parasites 18 hours post erythrocyte invasion [59]. Expression levels were measured using microarray analysis. The specific probes used in the microarray analysis and the corresponding Hellinger distances are available as Additional File 1. A permutation test was used to transform the LOD scores for each marker/expression trait combination into genome-wide corrected p-values. False discovery rates of 24% and 14% were reported for genome-wide significance levels of 5% and 1%, respectively. Regulatory hotspots were determined by comparing the number of expression traits mapping to each locus with a genome-wide corrected significance of 0.05 to the simulated null distribution. Regions surpassing the $95th$ percentile frequency in the null distribution were considered regulatory hotspots.

3.3.1 *Plasmodium falciparum* hotspot analysis

We obtained the data used in Gonzales *et al.* and repeated the eQTL mapping and calculation of hotspots [59]. We used the results as a baseline for comparison with our Hellinger distance method. *Plasmodium falciparum* is a relatively understudied organism so we prioritize identification of false negatives and report the hotspots identified using the Hellinger distance statistic without GO analysis.

Expression traits with very significant lowest corrected p-value show a great deal of variation in Hellinger distance in Figure 3.1. Variation in Hellinger distance decreases as p-value increases. This trend shows that while the genome-wide and Hellinger distance methods tend to disagree about which traits are most interesting, there is a much higher degree of agreement about which traits are not interesting.

Figure 3.2 shows the distribution of Hellinger distances for all 7665 expression traits. The small tail contains expression traits that closely match the model distribution. The large tail contains expression traits that have no association that distinguishes itself significantly from the rest. We will use expression traits from both tails of this distribution to demonstrate the significance of the priority assigned to traits with different expression profiles.

At a 0.05 significance level, only 914 expression traits had eQTLs. We measured the overlap between these expression traits and an equal number of expression traits from the small tail of the distribution and then from the large tail. We found that 292 or 31.9% of the expression traits with the 914 smallest Hellinger distance statistics also had eQTLs. We calculated the overlap for the 914 expression traits with the largest Hellinger distances and found that there were only 30 traits (3.28%) with eQTLs. The large difference in overlap between the traits with eQTLs and traits in either tail of the Hellinger distance distribution demonstrates that the Hellinger distance does provide a distinct ordering of traits. The relatively small overlap among traits with significant Hellinger distance shows that many of the expression traits

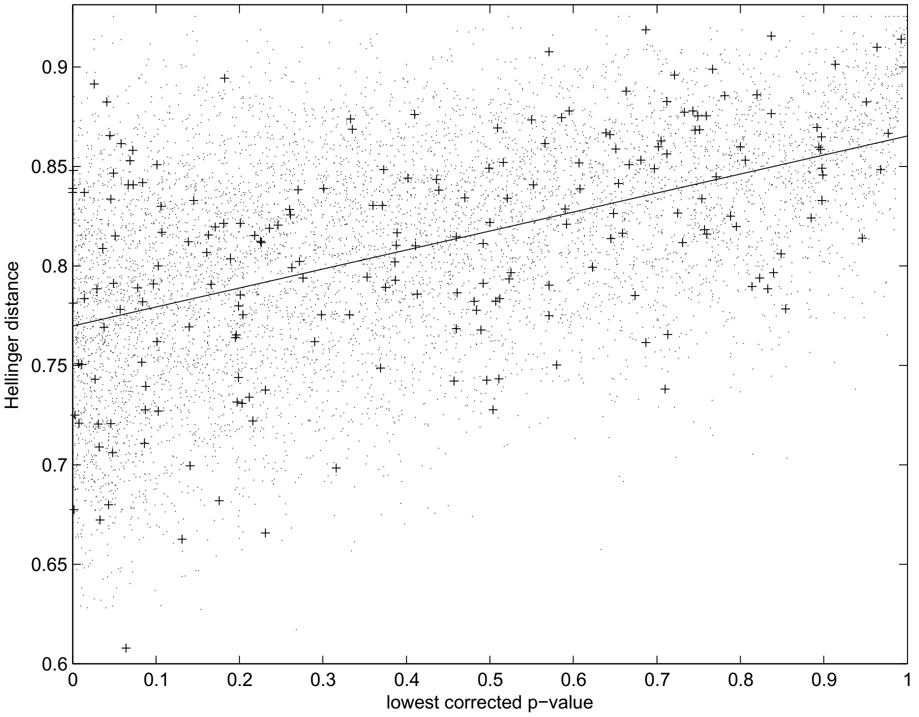


Figure 3.1. The Hellinger distance and p-value have a weak correlation, indicating that genome-wide significance is not a major consideration in the calculation of Hellinger distance. The r^2 value for the linear regression model on this data is 0.248. The pluses represent expression traits for which the strongest association to a locus is on the same chromosome as the trait.

without significant eQTLs nevertheless have an exceptional association with at least one locus.

Expression traits with 95*th* percentile Hellinger distance values were assigned to hotspots at the locus with the smallest genome-wide corrected p-value. We identified twenty-two Hellinger distance hotspots and eleven of the twelve hotspots reported by Gonzales *et al*, shown in Table 3.1. The table lists hotspots on each chromosome found using Hellinger distance (HD) and the genome-wide approach (GW) and the proportion of cis-acting eQTLs in hotspots on the chromosome. Cis-acting eQTLs were defined as those which are most strongly associated to markers on the chromosome they appear in.

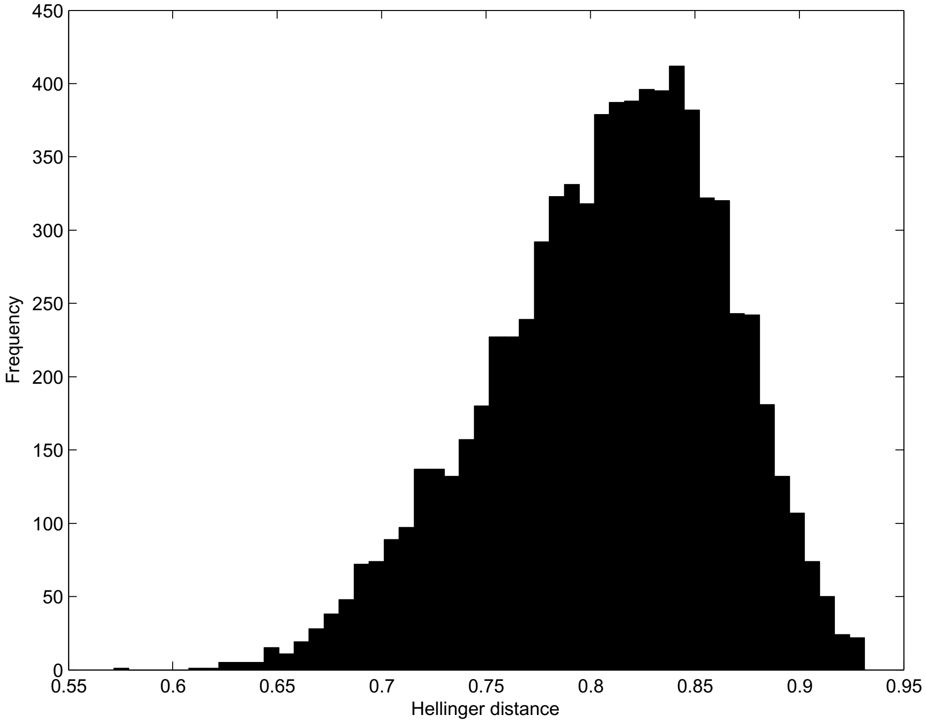


Figure 3.2. A histogram displaying the distribution of Hellinger distances across all 7665 expression traits. The larger Hellinger distances represent expression traits that may be regulated equally by multiple loci while the smaller Hellinger distances correspond to expression traits which have single or few exceptionally strong associations.

We compared the Hellinger distance hotspots in the small tail of the distribution to the genome-wide hotspots at the marker level (Figure 3.3). The majority of the hotspots found were consistent, verifying that hotspots found using Hellinger distance strongly correspond to genome-wide hotspots. While the Gonzales paper did not report marker locations of eQTL hotspots, our results indicate that nine hotspots also match at the marker level.

We also compared Hellinger distance hotpots in the large tail of the distribution to eQTL hotspots. These hotspots only overlap with four of the previously reported eQTL hotspots. Hellinger distance hotspots in the large tail should contain expression traits that have no single exceptionally strong association to a locus. Traits with

TABLE 3.1

HOTSPOTS ON EACH CHROMOSOME FOUND USING HELLINGER DISTANCE (HD) AND THE GENOME-WIDE APPROACH (GW).

chromosome	HD	cis HD eQTLs	GW	cis GW eQTLs
3	5	0/27	1	0/29
4	2	5/9	0	0
5	6	6/191	8	12/439
7	2	0/11	0	0
8	2	0/14	0	0
9	1	0/7	1	1/12
10	1	0/4	0	0
12	2	0/12	1	0/18
14	1	0/4	0	0

Hotspots on each chromosome found using Hellinger distance (HD) and the genome-wide approach (GW) and the proportion of cis-acting eQTLs in hotspots on the chromosome. Cis-acting eQTLs were defined as those which are most strongly associated to markers on the chromosome they appear in.

multiple eQTL are expected to occur in the large tail of the Hellinger distribution. We see this expectation fulfilled in Figure 3.4, in which there are many hotspots that do not agree with previously identified hotspots. The few hotspots that overlap eQTL hotspots contain few traits compared to the overlapping hotspots from the small tail.

A significant difference between our method and the genome-wide approach is that the genome-wide approach provides multiple statistics relating to each expression trait. An expression trait may have multiple associations with genome-wide significance but the Hellinger distance provides only one statistic that measures the

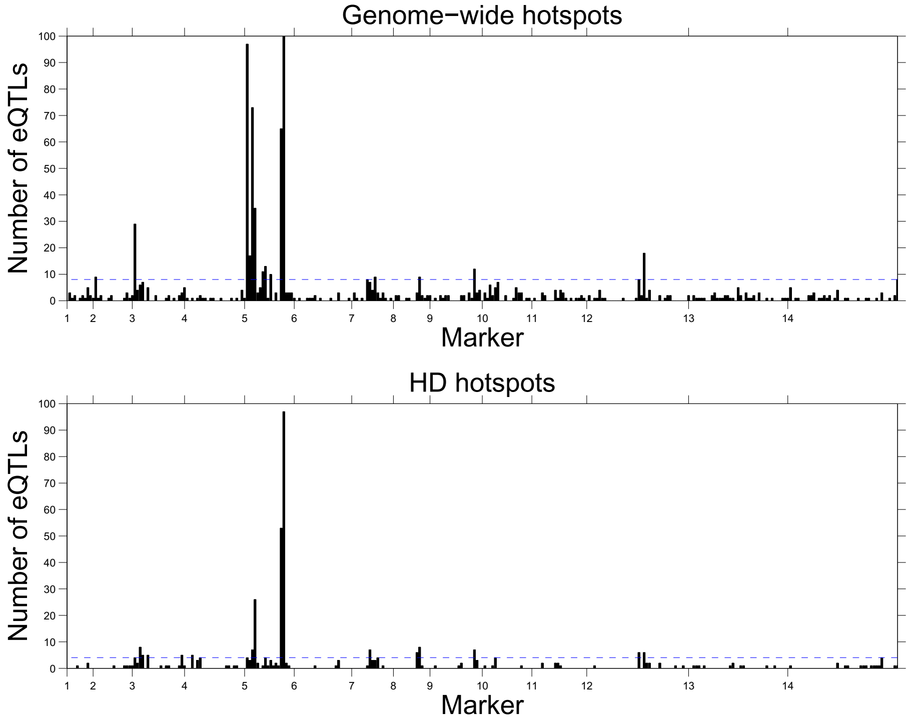


Figure 3.3. Marker positions on the genome versus the frequency of significant associations or eQTLs in the small tail of the distribution. The dashed line represents the cutoff for significance. The first bar graph shows the frequency of eQTLs by the genome-wide method while the second shows the frequency of expression traits with small Hellinger distance. Traits with significant Hellinger distance are assigned to the marker they are most strongly associated with.

extent to which the smallest p-value is exceptional among the expression trait's associations. The result is that there are less total Hellinger distance statistics than p-values and the cutoff for significant hotspots by Hellinger distance is lower that the cutoff for genome-wide significant hotspots. While the scale considered in the two approaches differs, the trends are similar.

At the chromosome level, all but one of the hotspots found in the Gonzales study were identified as hotspots in the small tail of the Hellinger distance distribution. We found multiple additional hotspots on chromosomes 3, 10, 11, 12, and 14. Each new

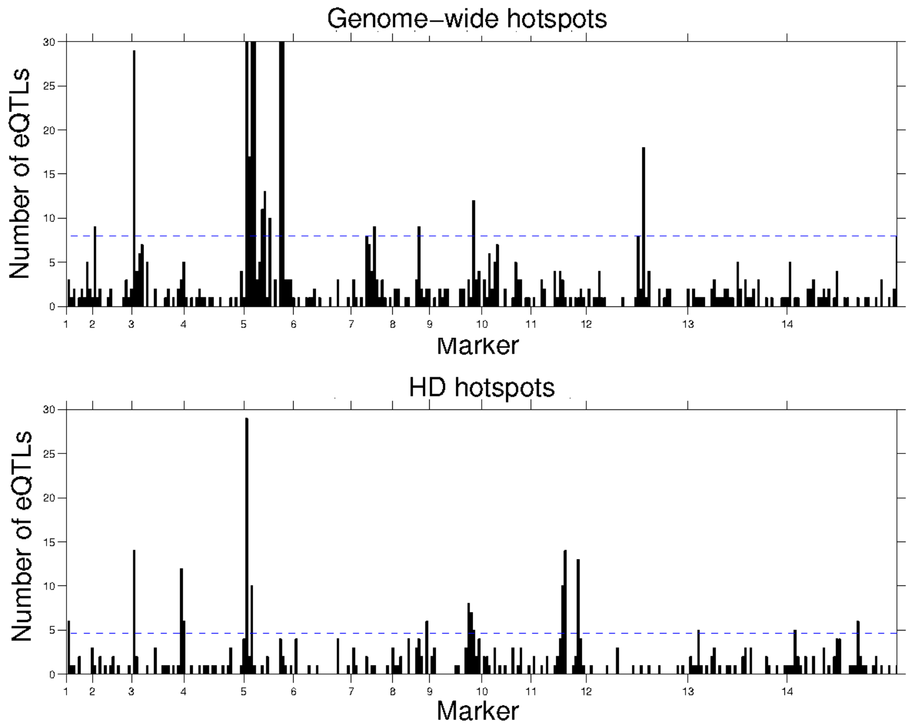


Figure 3.4. Marker positions on the genome versus the frequency of significant associations or eQTLs in the large tail of the distribution. The dashed line represents the cutoff for significance. The first bar graph shows the frequency of eQTLs by the genome-wide method while the second shows the frequency of expression traits with small Hellinger distance. Traits with significant Hellinger distance are assigned to the marker they are most strongly associated with.

hotspot has the interesting property of being the locus most strongly associated with a significant number of expression traits. These may be regulatory hotspots with significant regulatory effects that are unrecognized because of the low genome-wide significance of the individual associations.

Our results are similar to those found in a malaria study by Huang *et. al.* in which a graph theoretic approach is used as an alternative to traditional eQTL mapping [73]. The authors use a tripartite graph to model the relationships between genes, strains, and genotype. Their approach identifies eQTLs by finding maximal bipartite cliques associated with a loci to the number in random cliques. While the underlying method is different, the general approach is the same; they consider the available data in a novel way to identify additional hotspots. They use the Gonzales *et al.* data and identified seventeen hotspots. The positions of the hotspots identified with their new method appear largely consistent with the hotspots found using our method.

3.3.2 Yeast Gene Ontology analysis

To more thoroughly examine the significance of the new associations identified as significant by our approach, we applied the above experiment to the well studied organism yeast. We used expression and genotype data from Brem *et. al* to perform linkage analysis and calculation of hotspots [12]. Yeast has the advantage of having a thoroughly annotated genome. Therefore, in addition to performing the steps covered in our examination of *Plasmodium* we performed GO enrichment analysis and compared the GO terms found in expression traits with small Hellinger distances to the terms found in expression traits with genome-wide eQTLs. We used GO::TermFinder, an open-source GO term analysis tool introduced in Boyle *et. al* [9].

Using the same eQTL mapping methods and permutation test we used for *Plasmodium falciparum*, at cutoff for significance of 0.05, we identified 2719 expression traits with significant eQTLs. We repeated the procedure used to analyze the *Plasmodium* data. Again, we compared expression traits in the small tail of the distribution of all Hellinger distances to those in the large tail. As seen in the *Plasmodium falciparum* analysis, more expression traits in the small tail of the Hellinger distance distribution overlapped expression traits with eQTLs than those in the large tail. We found that 62.15% of the expression traits in the small tail also had eQTLs while 29.82% of those in the large tail had eQTLs.

We found a similar trend for GO terms enriched in traits with significant Hellinger distance. We found that 43 of the 102 process GO terms found among the 2719 expression traits with the smallest Hellinger distance were not enriched in expression traits with eQTLs. In contrast, there were a total of 8 terms enriched for the expression traits with the 2719 largest Hellinger distance statistics. We list the number of GO term results for process, function, and component terms in Table 3.2. Columns **small tail** and **large tail** indicate the number of total GO terms found for expression traits in the denoted tail of the Hellinger distance distribution that are not enriched in expression traits with eQTLs. The small tail contained 28 cis-acting eQTLs and the large tail contained 12. Each tail contained 318 eQTLs. We expected and found a fairly large number of new process terms enriched among the expression traits identified with small Hellinger distance. It is interesting to note that although we found many new process terms, we only found 5 new function and 4 new component terms. However, because a single expression trait may be related to multiple process, function, and component categories, it is very difficult to determine the importance of the few additional function and component terms. Regardless, the expression traits identified by Hellinger distance are enriched for many processes that are not enriched within expression traits with eQTLs but are associated with many of the same functions and components.

We identified cis and trans-acting eQTLs in both tails of the Hellinger distance distribution. We defined cis-acting eQTLs as those which are most strongly associated to markers on the chromosome they appear in. Conversely, trans-acting eQTLs appear on a different chromosome than the one they are most strongly associated with. We use this definition because it is a definitive and non-arbitrary cutoff. In the small tail there were 28 cis-acting expression traits out of 318. The large tail contained 12 cis-acting expression traits out of the total 318 in the tail.

TABLE 3.2

THE NUMBER OF TOTAL GO TERMS FOUND.

GO category	small tail	large tail
process	43	8
function	5	7
component	4	27

Columns **small tail** and **large tail** indicate the number of total GO terms found for expression traits in the denoted tail of the Hellinger distance distribution that are not enriched in expression traits with eQTLs. The small tail contained 28 cis-acting eQTLs and the large tail contained 12. Each tail contained 318 eQTLs.

3.3.3 GO similarity and gene essentiality analysis

We analyzed the GO term similarity and essentiality for terms enriched in sets of traits identified with both approaches.

GO similarity (or semantic similarity) measures the similarity of pairs of terms by the distance between them in a tree describing the hierarchy of GO terms. The semantic similarity of GO terms was computed by the Lin method via the GOSim package [91, 54]. We used t-tests to compare the GO similarity of a random set of 1000 Yeast GO terms and the GO similarity for the traits with eQTL as well as the traits with the 5% smallest Hellinger distances. The distribution of GO similarities in both sets of expression traits were significantly different from the random set at a significance level of 0.0001. We determined that the distributions of GO similarity between the Hellinger distance set and the eQTL set of traits were significantly different from each other (p = 2.2e-16) with a two-sample Kolmogorov-Smirnov test. We used the same test between the set of traits with eQTLs and large Hellinger distance and the set of traits with the 5% smallest Hellinger distance but without eQTLs and found that they were significantly different at a p-value of 1.059e-13.

Gene essentiality refers to the necessity of a gene for the survival of the organism [153]. We used a hypergeometric test to determine that the traits with eQTLs but without significant Hellinger distance had a marginal enrichment of essential genes with a p-value of 0.0113. Traits with small Hellinger distance but without eQTLs were more strongly enriched for essential genes at $p = 0.0002$.

3.4 Conclusions

We have demonstrated a novel approach to interpretation of eQTL data that builds on traditional approaches to identify possible false negatives and new points of interest for researchers. Our approach provides a statistic that describes the extent to which the distribution of associations connecting an expression trait to every loci matches the distribution we expect for expression traits with significant genetic effects. Expression traits identified through this method have associations which are exceptional within the scope of all associations to that expression trait. These associations may not be statistically significant at the genome-wide level but an exceptional association is very likely to indicate an interesting regulatory relationship regardless of the p-value.

Our approach addresses two potential sources of error in conventional genome-wide association studies. Expression traits that are not typically identified in eQTL studies may still have some associations that are exceptional among that expression trait's associations. Such a case may represent a false negative because, while an association may not be statistically significant in a genome-wide scope, its exceptional strength in the context of a single expression trait may indicate an interesting and overlooked regulatory effect. These expression traits may be identified by inspecting those with associations near the cutoff for genome-wide significance that also have a significant Hellinger distance.

A second potential source of error in eQTL studies comes from expression traits that are associated strongly with multiple loci. Due to the chaotic nature of recombination and uncertainty in linkage analysis, it is often the case that an expression trait is found to be strongly associated with multiple adjacent loci. Our approach minimizes the impact of this uncertainty by providing a single statistic per expression trait.

We have demonstrated a strong agreement between our method and traditional genome-wide techniques for hotspot and GO analysis. Even more interesting are the points of disagreement between the two methods. New hotspots and GO terms found with this method may represent regions of the genome or processes which are controlled through few relatively strong genetic interactions. These points of interest warrant experimental investigation.

3.5 Methods

We use the Hellinger distance statistic to measure the similarity between a model distribution and the distribution of associations linking an expression trait to each locus.

Hellinger distance is a nonparametric statistical test for distributional divergence [25]. It carries the following properties: dH(P, Q) is in $[0, \sqrt{2}]$. Hellinger distance is symmetric and non-negative, implying that dH(P,Q) = dH(Q,P). Finally, squared Hellinger distance is the lower bound of KL divergence. Hellinger distance essentially compares the shape but not the scale of the magnitude of the two distributions.

This is achieved by first splitting each distribution into an equal number of bins. This step is essentially building a histogram of each distribution. Each bin contains some proportion of the total values in one distribution. The next step compares the proportion held in each bin to the proportion held in the corresponding bin in the other distribution. This proportional comparison is how Hellinger distance measures divergence without regard to scale. A more precise definition follows:

$$HD = \sqrt{\sum_{a}^{b} (\sqrt{P_a/|P|} - \sqrt{Q_a/|Q|})^2} \tag{3.1}$$

Where P_a and Q_a are the counts for corresponding bins for the two distributions and $|P|$ and $|Q|$ indicate the total number of values in the distributions.

eQTL mapping calculates the association between each expression trait and each locus. The result is a set or distribution of associations for each expression trait. The permutation test provides a genome-wide corrected p-value for each association. Our method is based on calculating the Hellinger distance between each expression trait's p-value distribution and a reference distribution. As Hellinger distance does not make any assumptions about the shape or scale of the distributions being compared any reference distribution can be used while preserving the meaning of the statistic.

However, the fact that false negatives are more likely to occur near the cutoff for significance allows us to tailor the reference distribution to reflect our expectation for false negatives. Associations near the cutoff for significance, while not statistically significant, are still a great deal more significant than the vast majority of the associations. Therefore we expect there to be a large, relatively empty range between the strongest association and the majority of the associations. We use the reference distribution of values defined by $y = x^3$ over the integers from 1 to the number of loci to model this expectation. This reference distribution provides a balance between linear ordination and ease of interpretation. It allows the Hellinger distance statistic to be interpreted as evidence that a trait is controlled by a single locus or few loci.

We calculated the Hellinger distance using numbers of bins ranging from 10 to 100 in intervals of 10. Over that interval there are between 30 and 3 observations in each bin for the Plasmodium data. As the bin number approaches either extreme of the interval the hellinger distance becomes less able to reliably distinguish differences between distributions. This occurs because either too many or too few observations fall in each bin. The number of bins used did not make a significant difference to the results. We use thirty bins to provide an empirically acceptable binning granularity. The bin width is calculated as:

$$binwidth = (max(distribution) - min(distribution))/30 \tag{3.2}$$

The bin-width for each distribution is calculated separately.

This approach to determining the number of bins must be repeated for each additional data set. A potential alternative and more general method of determining the number of bins would be to use a kernel density bandwidth optimization technique [144].

The large difference between exceptional p-values and typical p-values causes the bulk of the values in the distribution to appear in the smaller bins in the histogram. The effect is that a greater difference between the most significant association and the bulk of the associations results in a lower Hellinger distance. The degree to which the strongest association in the distribution is exceptional is the primary factor in the shape of the distribution and therefore the Hellinger distance. In other words, the Hellinger distance statistic describes the extent to which an expression trait's most significant association is exceptional among all of its associations. Our choice of reference distribution reflects the expectation that the highest frequency occurs in the smallest bins and tapers off towards the largest association. Because we chose a reference distribution that we expect describes an expression trait that is controlled primarily at a single locus, the Hellinger distance measures the importance of that locus to the trait's expression.

One key difference between the genome-wide approach and our Hellinger distance based approach is that they measure significance on different scales. The genome-wide approach, in combination with a permutation test, provides statistics measuring the significance of each trait/locus association with respect to all associations between expression traits and loci. Our approach differs in that, for each expression trait, the Hellinger distance approach measures significance with respect to all the associations between a single trait and every locus. The result is that the genome-wide approach provides a statistic for each individual trait/marker association while the Hellinger distance provides a single statistic per expression trait. However, in the interpretation of Hellinger distance statistics it is important to consider that the calculation is based on distributions of genome-wide corrected p-values. Though Hellinger distances measure significance at an expression trait level, the elements of the underlying distribution are already corrected for multiple testing error at a genome-wide level.

CHAPTER 4

ENSEMBLE SIMILARITY MEASURES FOR CLUSTERING

In the previous chapter we proposed an approach to measuring relationships between genotype and phenotype that drew on domain knowledge of what constitutes interesting genotype-phenotype relationships. This is an informative approach in cases where domain knowledge is well established enough to determine an appropriate measure for "interestingness." However, it is more common that we lack the information necessary to choose a single informative measure. This is often the case in more exploratory and unsupervised approaches, where the objective of the study is not well defined. Even in cases where the objective is to discover a specific type of relationship in the data, there is often no clear choice for a single measure to identify it. This is a fundamental concern in networks and clustering algorithms, as each algorithm relies on the underlying similarity measure.

In this chapter we introduce the second argument of our thesis. Network approaches to data mining stand to benefit a great deal from the use of not only heterogeneous data, but —we emphasize— multiple distance measures. In this Chapter we describe an ensemble approach that uses expression data and expert curated gene interaction data to learn a combination of diverse similarity measures that produces more biologically meaningful clustering results across a wide variety of clustering algorithms.

4.1 Introduction

A primary goal of systems biology is to uncover the mechanisms underlying the behavior of a cell. Relationships between genes encode most of this information and are often discovered and represented as pathways that lead to essential products. Understanding these relationships is a very challenging problem as even the simplest organisms contain a multitude of genes that interact in complex combinations to deal with environmental conditions. An additional complicating factor is that current high-throughput technology used to measure the activity level of genes is notoriously noisy [29]. As there are very few well understood genetic interactions, clustering is a common first step to understanding this data [65, 30, 42].

We present an approach to clustering that utilizes known gene-gene interaction data to improve results for already commonly used clustering techniques. The approach creates an ensemble similarity measure that can be used as input to any clustering technique and provides results with increased biological significance while not imposing any constraints on the clustering method. We posit that an intelligent combination of multiple statistics can describe the extent to which two genes are similar more precisely than any single statistic. Our approach uses supervised learning to build an ensemble statistic from any number of descriptive statistics.

We focus on clustering microarray data. Microarrays enable simultaneous high-throughput measurement of the expression level of genes. Our approach leverages the expression data of genes that are known to interact to obtain additional information about relationships between less well understood genes. In contrast to the typical clustering approach, in which a single clustering algorithm uses a single similarity measure, this method has the potential to recognize any relationship that can be described by a statistic.

We apply our approach to the model organisms *Saccharomyces cerevisiae* (yeast) and *Escherichia coli*. Both are ideal organisms to consider given the availability of annotation and experimentally derived gene interaction data [22, 82].

4.1.1 Overview

Two general observations about data mining guide our approach. First, noisy data complicates identification of interesting patterns. We approach this problem by using experimentally derived gene interaction data and the random subspaces method. Second, even uninformed ensemble models tend to outperform more straightforward approaches [17]. This principle supports our assessment that even weakly descriptive statistics contain information that is missed by stronger predictors

The approach can be described roughly in four steps.

- Calculate descriptive statistics on the microarray data for each gene pair.
- Train C4.5 decision trees on random subspaces of the features using experimentally derived positive and negative interacting gene classes [118].
- Calculate a measure of feature importance based on the structure of the trees in our model.
- Weight each statistic by its feature importance and create ensemble similarity measures.
- Cluster the ensemble data.

First we calculate descriptive statistics on the microarray data for each gene pair. Each statistic describes a different type of relationship between a pair of genes. In order to demonstrate the success of our approach we use a set of statistics that, with the exception of correlation, we believe will result in poor clustering results. Second we train C4.5 trees on random subspaces of the features using experimentally derived positive and negative interacting gene classes from gold standard data sets. The random subspaces approach builds classifiers using subsets of the available statistics. The use of random subspaces allows the classifiers to investigate how different combinations of statistics work together to predict gene interactions. Some statistics may act in combination to improve classification whereas others may interfere with classification or obscure the positive effects of less reliable predictor variables. We overcome this issue by evaluating our approach across all possible subspace sizes. Next we calculate a measure of feature importance based on the structure of the trees in our model. C4.5 trees were chosen because of the conceptual ease of determining feature importance as a function of tree structure. We use the feature importance to weight the individual statistics and combine them into an ensemble statistic. The use of feature importance as a weighting mechanism in combination with the random subspaces method has the effect of increasing the weight of statistics that were good predictors of gene interactions. Finally, we cluster the ensemble data. Figure 4.1 depicts the full experimental design. The individual steps are explained in detail in the methods section.

4.2 Data

4.2.1 Gene Expression Data

We considered yeast expression data from a line cross experiment composed of a 131 strains and 5979 probes [12]. Genetic line cross experiments have great potential to elucidate the causative agents of drug resistance and can shed light on the intricate relationships between genes and ultimately targets for drug design [79]. We also considered expression data from an *E. coli* experiment studying the effect of oxygen deprivation [27].

4.2.2 Positive and Negative Gold Standards

Positive and negative gold standard sets of gene interaction data for yeast were obtained from a manually curated set of GO terms, which was balanced in terms of functional classes of genes [104]. Interacting genes were selected by voting results from a team of six expert biologists. Gene pairs were said to be interacting if each

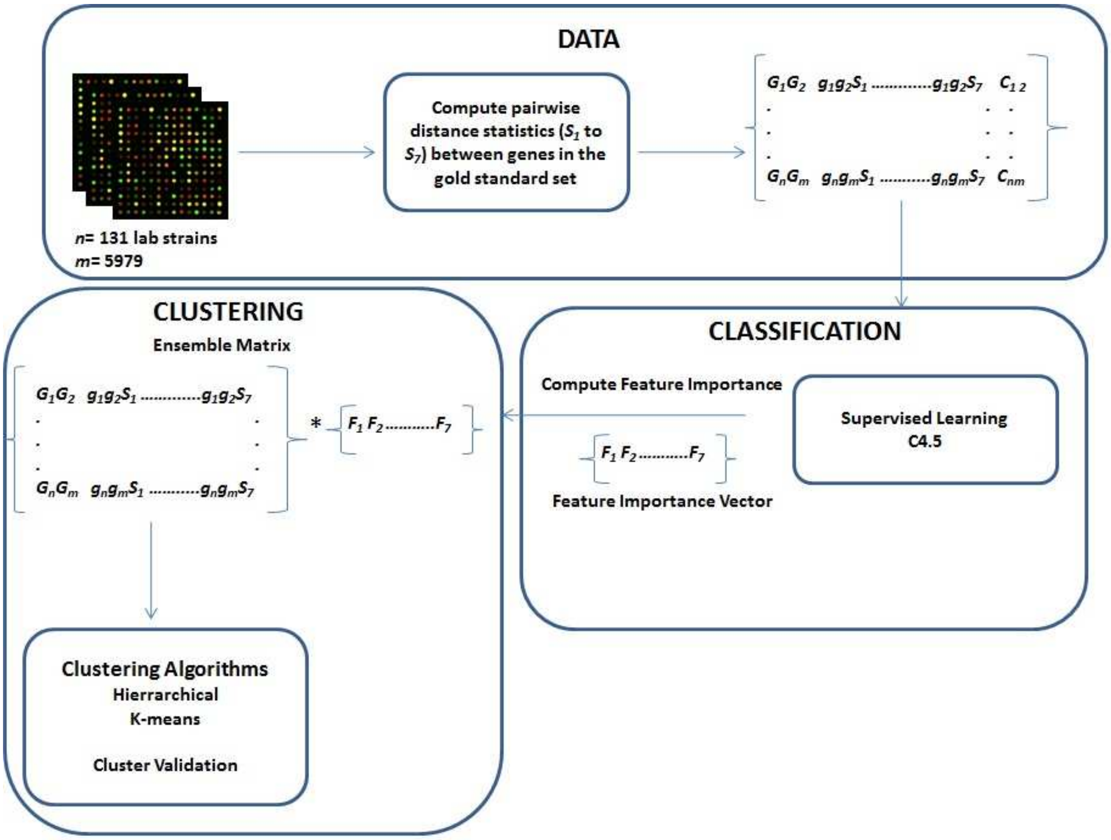

DATA
n= 131 lab strains
m= 5979
Compute pairwise distance statistics (S_1 to S_7) between genes in the gold standard set
G_1G_2 $g_1g_2S_1$ ....$g_1g_2S_7$ C_{12}
G_nG_m $g_ng_mS_1$ ....$g_ng_mS_7$ C_{nm}
CLUSTERING
Ensemble Matrix
G_1G_2 $g_1g_2S_1$ ....$g_1g_2S_7$
G_nG_m $g_ng_mS_1$ ....$g_ng_mS_7$
* F_1 F_2..........F_7
Clustering Algorithms
Hierrarchical
K-means
Cluster Validation
CLASSIFICATION
Compute Feature Importance
Supervised Learning
C4.5
F_1 F_2..........F_7
Feature Importance Vector

Figure 4.1: Overview of the approach. The method begins by computation of pairwise distance statistics. We calculated seven statistics (S1 to S7) between each interacting gene pair (e.g G1G2) in both the positive and negative gold standards data set. Many gene pairs have a class label C that can either be positive or negative as defined in the gold standards data set. C4.5 was then used to estimate feature importance (F1 to F7) as the sum of information again across all splits of a given feature.

gene shared a GO term specific enough to imply functional association. Positive and negative sets consist of pairs of genes that have been confirmed or refuted as interacting through laboratory experiments rather than computational approaches. Six expert biologists voted on whether each of a large set of GO terms should be considered interacting. Terms with many votes were considered interacting while terms with one or less vote were considered non-interacting. The use of this data allowed us to create an ensemble statistic while minimizing functional bias due to an unbalanced hierarchy of GO terms.

Interacting *E. coli* genes were derived from gene pathway data. Pathway data were gathered from the Ecocyc database [82]. Ecocyc is a comprehensive database of the current knowledge about *E. coli*. Genes that occur in the same pathway were considered interacting. Negative interactions were simulated by randomly selecting pairs of genes that did not share a pathway.

4.3 Methods

4.3.1 Similarity measures

Our approach is flexible in that the number of statistics that can be combined into an ensemble is only limited by the computational burden in the classification step. Individual statistics are weighted by the amount they contribute to predicting interactions. This approach reduces the effects of statistics that do not contribute to the identification of interacting gene pairs. We demonstrate this property by using a collection of statistics, most of which we do not expect to discriminate well between interacting and non-interacting genes. These weakly predictive statistics may have regions in which they are locally good predictors or they may be good predictors in combination with other statistics. Our approach is designed to take advantage of these effects.

We calculated seven similarity measures on the expression data for each gene pair, shown in Table 4.1.

4.3.2 Classification

Each statistic describes a different relationship in the data. Our goal is to combine the strengths of all the statistics into a single ensemble. We do this by leveraging patterns in the expression of gene pairs that are known to be interacting.

The greatest challenge in evaluating the usefulness of each statistic is that they can interact in complex ways. A classifier built on a pair of statistics may have additional predictive power over a single statistic classifier. However, a different pair of statistics

TABLE 4.1

FEATURES USED FOR SUPERVISED LEARNING.

Feature	description
City block distance	Distance along a grid
Correlation	The extent to which two variables are linearly related
Cosine similarity	The cosine of the angle between two vectors
Covariance	Amount a pair of variables change together
Hellinger distance	The similarity in shape of marginal distributions
Kolmogorov-smirnov	Similarity in shape and magnitude of two distributions
Mutual information	Mutual dependence of two variables

may be less useful if they both contain similar information or are poor predictors. An additional challenge is that statistics can be locally strong predictors. Locally strong predictors may be overshadowed by generally better predictors. This is a loss because each statistic, even a generally poor predictor, contains some information about a relationship in the data. Our approach seeks to weight statistics in proportion to their overall usefulness in identifying interacting genes.

We trained classifiers on random subsets of similarity measures. Using a single random subset of two similarity measures, we might train a classifier on only correlation and mutual information data. This approach is known as the random subspaces method [72]. It allows us to investigate the effects of various combinations of similarity measures on prediction of gene interaction. We used C4.5 decision tree classifiers on random subspaces of similarity measures. The C4.5 algorithm splits data into subsets by an information gain criterion [118]. Because of this, a similarity measure that is a locally good predictor of interactions may not be split on in a C4.5 tree that has access to similarity measures with more predictive power. Random subspaces allow even poor predictors to be built into a tree. This allows our approach to recognize similarity measures that are good predictors of specific small subsets of interactions even when they are poor predictors in general.

The original yeast interaction data contained a large imbalance towards non-interacting gene pairs. In order to create classifiers with an emphasis on identifying positive interactions, we took a random samples composed of equal numbers of positive and negative interactions for the training set. The *E. coli* data was also balanced in terms of positive and negative interactions.

4.3.3 Feature importance

We calculated feature importance as the sum of information gain across all splits in decision trees for each similarity measure. We believe that this is an informative metric because information gain depends on the amount of data split as well as the usefulness of the split for prediction. Splits further down the tree typically affect less data and have lower information gain. This trend agrees with the intuition that splits lower in the tree are less important to overall tree structure.

The feature importance was measured as the mean of the feature importance for each similarity measure from all classifiers. Because the classifiers were used exclusively to derive feature importance, to validation was necessary. Finally, we transformed each feature importance measure into the proportion of total feature importance across all similarity measures. Table 4.2 contains the similarity measures used in classifiers and the scaled feature importance for both data sets. The statistic with the largest feature importance for the yeast data set is correlation closely followed by covariance. In contrast, mutual information had the largest feature importance by far in the *E. coli* data set.

TABLE 4.2

SIMILARITY MEASURES USED FOR SUPERVISED LEARNING AND THE FEATURE IMPORTANCE ASSIGNED TO THEM.

Similarity measure	Yeast feature importance	*E. coli* feature importance
City block	0.1850	0.1386
Correlation	0.2089	0.1001
Cosine	0.1744	0.1931
Covariance	0.2021	0.1060
Hellinger Distance	0.1073	0.1053
Kolmogorov-smirnov	0.0592	0.0208
Mutual information	0.0627	0.3357

4.3.4 Ensemble similarity measure

We used three approaches to build ensemble similarity measures. All component similarity measures were range standardized such that all elements fell between zero and one. Each similarity measure was weighted by multiplying all values in the similarity matrix with the corresponding feature importance. A weighted sum ensemble was created by computing the sum of each corresponding element from all similarity measure matrices. Similarly, weighted min and max ensembles were created by taking the min and max respectively for each element of the matrix.

4.3.5 Clustering Algorithms

Hierarchical clustering, k-means clustering, and the Walktrap clustering algorithm were applied to the ensemble similarity measures.

The k-means clustering algorithm attempts to identify the best fit clusters by minimizing the within cluster sum of squared distance from cluster centers [67]. Given unlimited iterations, the k-means algorithm attempts to optimize globally on its clustering criterion and tends to result in clusters with spherical shape and size. We used k-means clustering with five random restarts and a voting process for cluster membership to reduce the possibility of the algorithm converging to locally optimized clusters.

We report results for two agglomerative hierarchical clustering criterion: UPGMA and Ward's method. UPGMA groups clusters by the mean distance between elements of each cluster [132]. This results in a tendency to group clusters with small variance. Ward's method groups clusters explicitly with regard to cluster variance by joining two clusters based on the minimum increase in variance when two groups are merged [152]. This approach tends to result in equal sized spherical clusters. In contrast to K-means clustering, Ward's method and UPGMA both optimize locally on their clustering criterion [46].

The walktrap algorithm is designed to capture community structure by simulating random walks in networks [114]. Walktrap creates a similarity measure based on the probability that random walks from each node end at each other node. Communities are merged using Ward's method.

We tested our method with two additional hierarchical clustering criterion, including single linkage and median linkage. We found that Single linkage and median linkage produced very poor clustering results. Our findings with respect to Single linkage agree with results reported in [56]. Additionally, we found that Markov Clustering produced results similar to single and median linkage. We focus here on the results that best demonstrate the differences between clustering with a single similarity measure and clustering with an ensemble statistic.

4.3.6 Cluster validation

There are two general approaches to validation of microarray cluster results: validation based on internal measures and validation based on additional biological knowledge. [104, 64] We use both approaches, using the F-measure to evaluate interactions present in clusters and the Biological Homogeneity Index to measure the validity of cluster results.

The F-measure is a measure of accuracy based on the trade-off between precision, the proportion of gene pairs in a cluster that are known positive interactions (Equation 8.5), and recall, the proportion of the known interactions that are in the cluster (Equation 8.6). The F-measure is defined in Equation 4.3.

$$precision = tp/(tp + fp) \tag{4.1}$$

$$recall = tp/(tp + fn) \tag{4.2}$$

$$F - measure = \frac{precision * recall}{(precision + recall)} \tag{4.3}$$

The Biological Homogeneity Index (BHI) measures cluster validity based on the proportion of genes in each cluster that share at least one GO annotation [31]. Each pair of annotated genes x and y in cluster D that share at least one GO term (C(x)=C(y)) in Equation 4.4 increases the proportion of total genes with shared terms.

$$\frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\varepsilon D} I(C(x) = C(y)) \tag{4.4}$$

Where k is the number of clusters.

4.3.7 Statistical comparison of results

We utilized the Wilcoxon signed-rank test to compare pairs of cluster results. A signed-rank test is a non-parametric analog of a t-test. It compares the difference between tied pairs of items by ranking the differences into positive and negative sets of ranks. [37]

$$W_+ = \sum_{d_i>0} rank(d_i) + 1/2 \sum_{d_i=0} rank(d_i) \tag{4.5}$$

$$W_- = \sum_{d_i<0} rank(d_i) + 1/2 \sum_{d_i=0} rank(d_i) \tag{4.6}$$

Where d_i is the distance between tied pair i. The smaller of the two values, T, is given a z-score as follows:

$$z = \frac{T - \frac{1}{4}N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}} \tag{4.7}$$

Where N is the number of observations.

We used the Friedman test to rank the performance of ensembles across different clustering algorithms. The Friedman test is a non-parametric equivalent of ANOVA. In contrast to ANOVA, the Friedman test does not make the assumption that the sample means being tested have related means or that the underlying variables have equal variance. Instead, the Friedman test assumes that the data come from populations with the same continuous distributions and that all observations are mutually independent. These assumptions are desirable for our data because clustering results from separate algorithms may be extremely variable.

The Friedman test compares multiple treatments across multiple data sets under the hypothesis that all treatments are equivalent and should have the same rank. The test compares the mean rank of all combinations of sample i (of N total data sets) and algorithm j (of k total algorithms) by first calculating the mean performance of each algorithm across samples in Equation 4.3.7 then comparing the mean performance of algorithms in in Equation 4.3.7.

$$R_j = \frac{1}{N}\sum_i r_i^j \tag{4.8}$$

Where R_j is the rank of algorithm j.

$$\chi_F = \frac{12N}{k(k+1)}[\sum_j R_j^2 - \frac{k(k+1)^2}{4}] \tag{4.9}$$

4.4 Results

In previous work we trained 50 C4.5 trees on random subspaces of every size, from subspaces using single similarity measures to subspaces using every similarity measure [120]. Observations made during that analysis lead us to extend the approach to take all random subspace sizes into account simultaneously. Therefore we trained classifiers on 100 random subspaces of random sizes, utilizing anywhere from a single feature to all features. A set of ensemble similarity measures was created for each data set. In the following sections we compare the effects of using these ensembles as the basis of clustering to the effect of using the correlation alone.

4.4.1 Yeast interaction based validation

We performed Friedman's tests to determine if any of the ensembles significantly affected the cluster results. Table4.3 shows the Friedman's test results comparing the effect of different similarity measures on each clustering algorithm. The table shows that there are significant differences in results depending on the similarity measure used for all algorithms except k-means.

TABLE 4.3

FRIEDMAN'S TEST RESULTS COMPARING THE EFFECT OF SIMILARITY MEASURES ON EACH CLUSTERING ALGORITHM.

Algorithm	p-value
UPGMA	**1.38e-06**
Ward	**0.00129**
K-means	0.07717
Walktrap	**0.00825**

Each p-value represents a comparison of clustering results gathered using each ensemble and correlation with the algorithm named in the row. P-values significant at $\alpha = 0.05$ appear in bold.

We used signed rank tests to determine which ensembles have the greatest effect on the clustering results in Table 4.4. We tested the hypothesis that the median F-measure of ensemble-based results is greater than the median F-measure of correlation-based results. The min ensemble is the only similarity measure with a statistically significant effect on cluster results at $\alpha = 0.05$. We also tested the opposite hypothesis, that the median F-measure of correlation-based results is greater than the corresponding ensemble-based results, and found no significant effects. In absolute terms, the best BHI results were achieved by the min ensemble in combination with either the Walktrap algorithm or UPGMA.

Figure 4.2 shows the F-measure versus the number of clusters for all similarity measures and all clustering algorithms. As indicated by Table 4.4, the min ensemble performs very well in combination with UPGMA and Walktrap. Regardless of clustering algorithm, all ensembles appear to always do at least as well as correlation alone and noticeably better in UPGMA and Walktrap results.

TABLE 4.4

SIGNED RANK TEST RESULTS TESTING THE HYPOTHESIS THAT THE MEDIAN F-MEASURE OF ENSEMBLE-BASED RESULTS IS GREATER THAN THE MEDIAN F-MEASURE OF CORRELATION-BASED RESULTS.

	Algorithm			
Ensemble	UPGMA	Ward	K-means	Walktrap
min	**0.0177**	0.5147	0.4267	**0.0057**
max	0.3696	0.5147	0.5147	0.4267
sum	0.1965	0.3152	0.4852	0.4267

For each clustering algorithm and ensemble, the vector of F-measures corresponding to the number of clusters were compared to the corresponding set of F-measures from correlation-based clustering results. P-values less than 0.05 appear in bold.

4.4.2 Yeast BHI validation

We took the same approach to validation with the BHI. Friedman's test results in Table 4.5 showed that there are significant differences ($\alpha = 0.001$) between cluster results for Ward's method, and Walktrap depending on which similarity measure is used.

We further investigated the inconsistencies by performing signed rank tests comparing the clustering results given by one clustering algorithm and each ensemble to the same clustering algorithm and correlation. Table 4.6 shows that ensemble-based results were always significantly better than correlation-based results with the exception of Ward's method and the combination of UPGMA and the min ensemble. Correlation and Ward's method was significantly better than all ensembles at

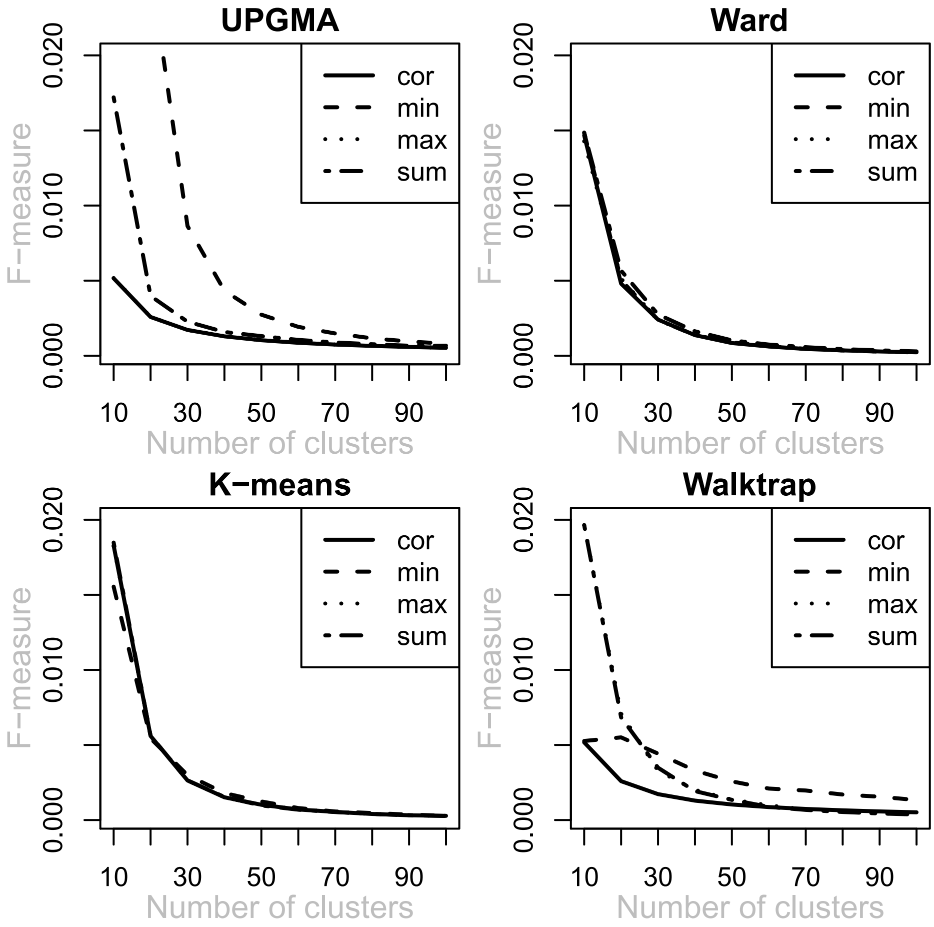


Figure 4.2: F-measure versus number of clusters.

$\alpha = 0.0001$. The best BHI was observed for the Walktrap algorithm with the min ensemble. It was marginally greater than the second best combination, the sum ensemble and UPGMA with a p-value of 0.0525.

Figure 4.3 shows the BHI across numbers of clusters produced. The figure further supports the signed rank test results. The BHI appears much more erratic in UPGMA results than in any other algorithm. Most of the variation in all algorithms appears to occur in the smaller numbers of clusters and level off as the number increases.

Comparing Figure 4.2 and Figure 4.3 we see that both the min and sum ensembles provide the best clustering results overall and provide the best clustering

TABLE 4.5

FRIEDMAN'S TEST RESULTS COMPARING THE EFFECT OF SIMILARITY MEASURES ON EACH CLUSTERING ALGORITHM.

Algorithm	p-value
UPGMA	0.0771
Ward	**0.0001**
K-means	**1.32e-5**
Walktrap	**0.0001**

Each p-value represents a comparison of clustering results gathered using each ensemble and correlation with the algorithm named in the row. P-values significant at $\alpha = 0.05$ appear in bold.

results according to both annotation-based and interaction-based validation methods in UPGMA and Walktrap clusters.

Viable clusters are those that contained enough genes with GO terms and interactions for analysis. With the exception of UPGMA and Walktrap, all clustering experiments resulted in precisely the number of desired viable clusters. Table 4.7 shows the number of viable clusters produced by UPGMA and walktrap with all similarity measures across all numbers of clusters. Clustering experiments that resulted in few viable clusters may be finding interesting clusters of genes that simply do not have the necessary annotation or studied interaction data for validation. Results with very small numbers of viable clusters should be considered suspect because of the lack of validation data. In such cases, validation measures may appear artificially high because all known data about genes occurs in the same few clusters. In this light, the combination of min ensemble and UPGMA should be considered suspect as well as the combination of correlation and Walktrap. However, the results from

TABLE 4.6

SIGNED RANK TEST RESULTS TESTING THE HYPOTHESIS THAT THE MEDIAN BHI OF ENSEMBLE-BASED RESULTS IS GREATER THAN THE MEDIAN BHI OF CORRELATION-BASED RESULTS.

	Algorithm			
Ensemble	UPGMA	Ward	K-means	Walktrap
min	0.1237	1	**5.41e-6**	**5.41e-6**
max	**0.0376**	1	**5.41e-6**	**0.0002**
sum	**0.0177**	1	**5.41e-6**	**0.0007**

For each clustering algorithm and ensemble the vector of BHIs corresponding to the number of clusters were compared to the corresponding set of BHI from correlation-based clustering results. P-values less than 0.05 appear in bold.

combination of the min ensemble and Walktrap are still greater than all other results except for the min ensemble and UPGMA.

4.4.3 *E. coli* interaction based validation

Friedman's test results comparing the effect of different similarity measures on each clustering algorithm showed significant ($\alpha = 0.001$) differences in results depending on the similarity measure used for all algorithms.

We tested the hypothesis that the median F-measure of ensemble-based results is greater than the median F-measure of correlation-based results in Table 4.8. The max and sum ensembles performed significantly better for UPGMA than correlation at $\alpha = 0.05$. All ensembles appear to not quite significantly outperform correlation for Walktrap clustering. Tests on the opposite hypothesis, that the median F-measure of correlation-based results is greater than the corresponding ensemble-based results,

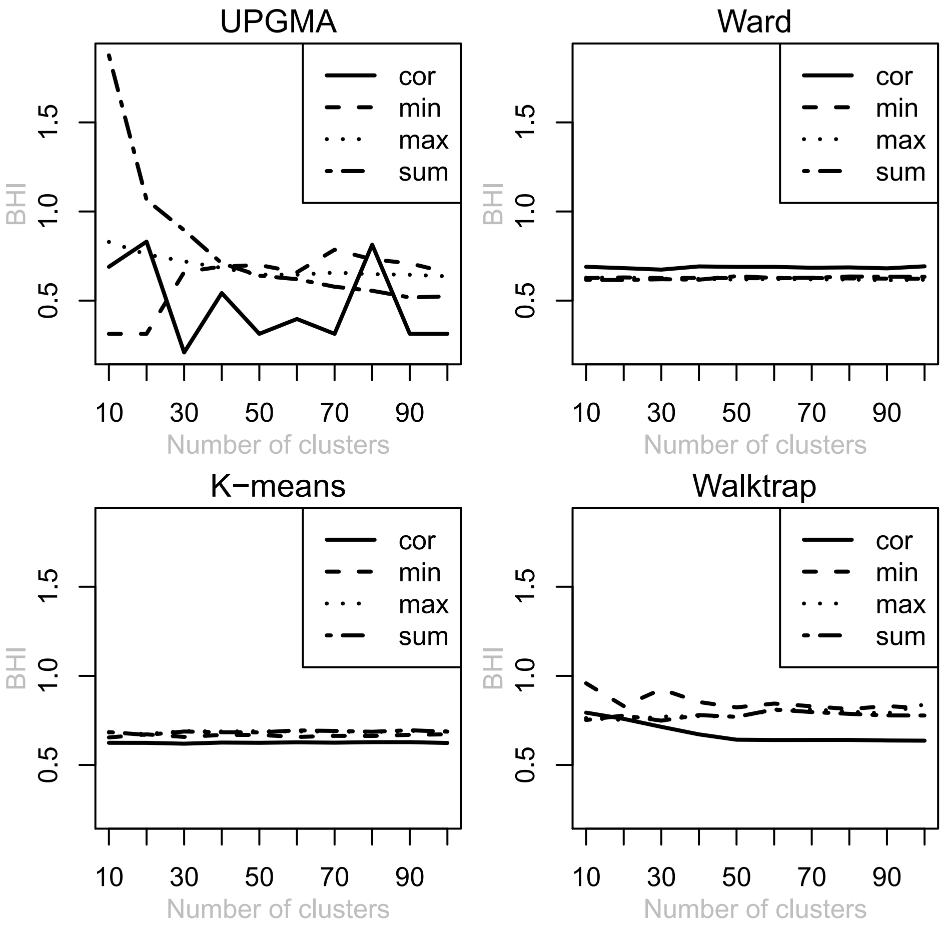


Figure 4.3: BHI versus number of clusters.

found no significant effects.

Figure 4.4 shows the F-measure versus the number of clusters for all similarity measures and all clustering algorithms. The max ensemble does not appear in the UPGMA figure because it is nearly ten times better than the next best results. A large part of UPGMA's success is due to the tendency to produce a single large cluster containing the vast majority of observations. As a result, the number of viable clusters (those containing known interactions and at least two genes) was less than the number produced with other algorithms. As indicated by Table 4.8, the min, max, and sum ensembles perform very well in combination with both UPGMA and

Walktrap. The max and sum ensembles also performed well compared to correlation with Walktrap clustering.

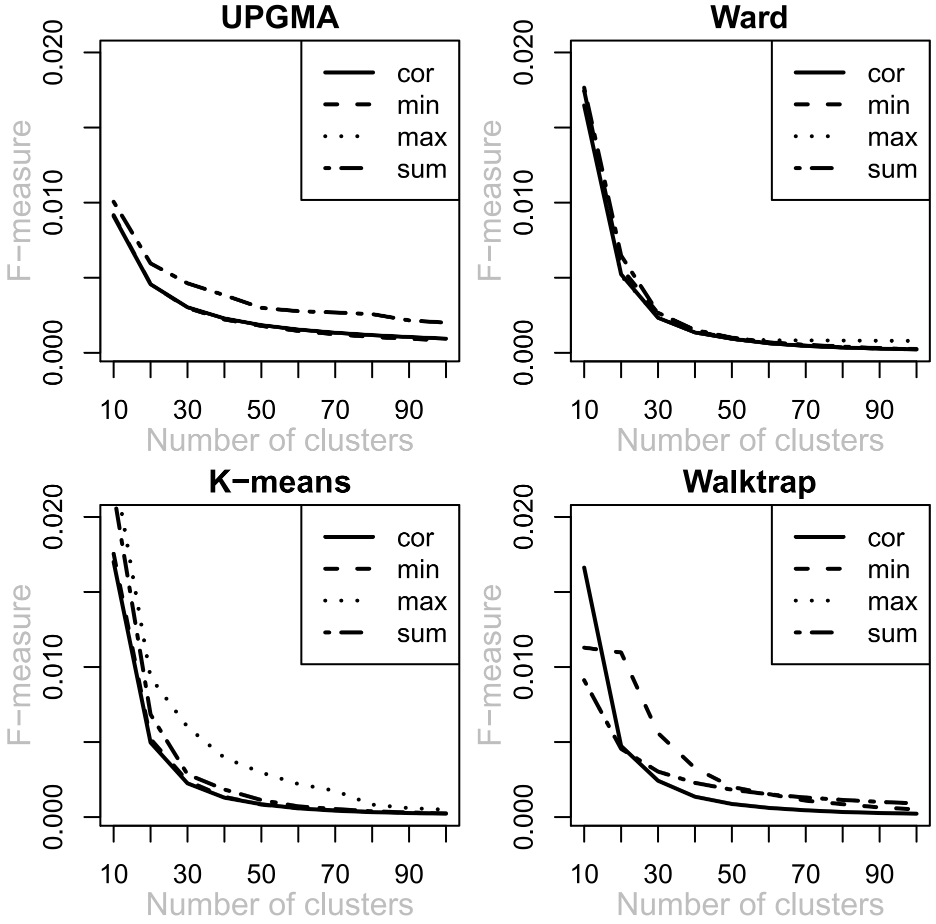


Figure 4.4: F-measure versus number of clusters.

4.4.4 *E. coli* GO annotation based validation

Friedman's test results for the BHI were similar to the F-measure results in that the effect of different similarity measures on each clustering algorithm showed significant differences in results depending on the similarity measure used for all algorithms at $\alpha = 0.01$.

Table 4.9 shows results for tests of the hypothesis that the median BHI of ensemble-based results is greater than the median BHI of correlation-based results. All ensembles significantly outperformed correlation in UPGMA results at $\alpha = 0.001$. The sum ensemble with Ward's method significantly outperformed correlation and all other ensembles. The min ensemble with the K-means algorithm significantly outperformed correlation and all other ensembles. The best overall BHI was a result of the min ensemble and Walktrap. Signed rank tests of the hypothesis that the median BHI of ensemble-based results is less than the median BHI of correlation-based results showed that all other ensembles performed significantly worse ($\alpha = 0.05$) than correlation except Ward's method with the min ensemble.

Figure 4.5 shows the BHI versus the number of clusters for all similarity measures and all clustering algorithms. As indicated by Table 4.9, the min, max, and sum ensembles perform very well in combination with UPGMA.

Table 4.10 shows the number of viable clusters produced by UPGMA and walktrap with all similarity measures across all numbers of clusters. In UPGMA results the max ensemble produced a single viable cluster. Therefore the overwhelmingly positive F-measure results for this combination should be considered highly suspect. Walktrap and the sum and max ensembles produced only single viable clusters for most numbers of desired clusters. This data explains why their results according to both validation methods are the same. Regardless of these considerations, clustering with ensemble similarity measures outperform clustering with correlation in all cases.

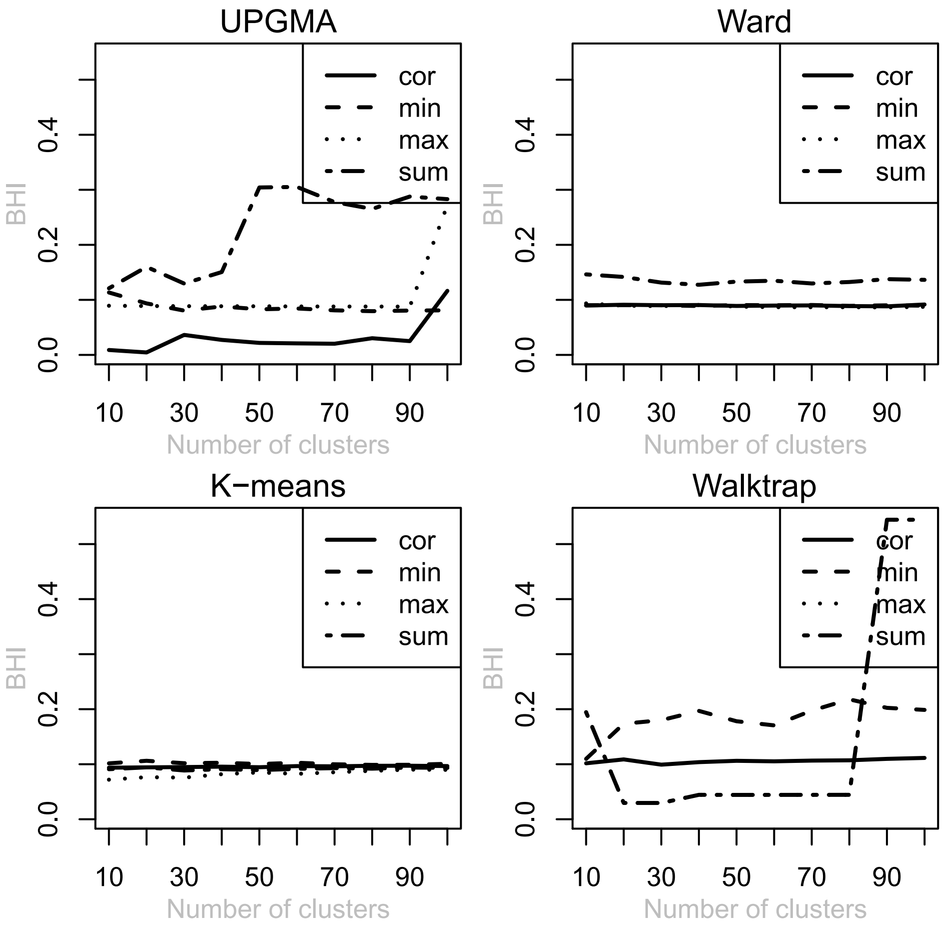


Figure 4.5: BHI versus number of clusters.

4.4.5 Comparing validation measures across organisms

Overall, the interaction-based validation in Figures 4.2 and 4.3 show similar trends. In both figures the F-measure decreases as the number of clusters increases, indicating that the precision and recall grow more unbalanced. In both organisms ensembles perform significantly better than correlation in UPGMA and Walktrap results. The best F-measure in each organism was achieved by using an ensemble similarity measure.

The annotation-based validation in Figures 4.4 and 4.5 was similarly uniform across organisms. BHI results for Yeast and *E. coli* agree about the success of ensembles in UPGMA clustering results and the min ensemble in K-means clustering results. They also agree about the poor performance of ensembles in Ward's method. The best BHI was achieved with Walktrap in both organisms.

4.5 Discussion

We have presented strong evidence that an ensemble can out-perform correlation across different clustering algorithms already in use for analysis of microarray data. Furthermore, we have shown that our ensemble approach consistently produces better clustering results than correlation alone across organisms and validation types. An important observation to come out of this study is that the most commonly used clustering algorithms in microarray analysis were consistently outperformed by Walktrap. Although UPGMA and K-means clustering are the most commonly used algorithms for exploratory analysis of microarray data, we found that the absolute best results with either validation approach and across organisms were most often achieved by use of the Walktrap algorithm.

The few apparent discrepancies between the interaction based validation results and the GO similarity validation results may be attributed to the underlying differences in what is measured by the two approaches. The interaction based validation considered only gene pairs which shared a GO term in the case of Yeast or a shared pathway in the case of *E. coli.* The GO similarity validation on the other hand used all available GO terms and measured the distance between all pairs on the GO hierarchy. Not only did the GO similarity method have more data available but it considered less specific relationships. The GO analysis of *E. coli* was particularly affected by a lack of annotation as there were only slightly more annotated gene products than there were genes. In light of the differences between the results, the combinations of clustering algorithm and similarity measure that performed well by both measures are particularly interesting. Poor agreement between gold standards and validation methods is a pervasive problem in biological validation but it should not imply that the approaches are unreliable [104]. Although each gold standard or validation method has its own bias it can still be informative.

The importance of the data set used in this procedure cannot be overstated. We attempted the same experiment with Biogrid Yeast and *E. coli* interaction data [135] with less definitive but still encouraging results. The clustering results using ensembles built on the Biogrid interaction set tended to be statistically indistinguishable from clustering results using correlation alone. Although the Biogrid database contains a larger collection of curated interaction data, we feel that the data sets we used were important contributing factors to the success of this approach.

4.6 Conclusions

We have described a method that provides a number of advantages over typical approaches to gene clustering: i) it intelligently weights similarity measures by their predictive power, allowing a number of statistics to be utilized regardless of their individual usefulness. ii) The method employs prior biological knowledge in the form of known gene to gene interactions represented by positive and negative gold standards and integrates this into the similarity measure. iii) It complements and improves upon existing common and successful methods of analyzing high-throughput biological data. iv) Because it creates an ensemble similarity measure rather than altering a clustering approach, it could be used with clustering methods beyond those discussed here.

TABLE 4.7

THE NUMBER OF VIABLE CLUSTERS FOR CLUSTERING EXPERIMENTS USING UPGMA AND WALKTRAP AND ALL ENSEMBLE SIMILARITY MEASURES AND CORRELATION IN THE YEAST DATA SET.

	UPGMA				Walktrap			
Number of clusters	cor	min	max	sum	cor	min	max	sum
10	6	1	2	2	1	6	10	10
20	14	1	9	6	1	13	20	20
30	21	2	18	9	1	13	30	30
40	31	3	28	12	1	18	40	40
50	39	3	37	14	1	21	50	50
60	49	5	47	19	1	28	60	60
70	59	8	57	20	1	29	70	70
80	69	8	66	20	2	33	80	80
90	79	11	76	22	1	34	90	90
100	89	15	86	29	1	38	100	100

Viable clusters are those that contained enough genes with GO terms and interactions for analysis. All clusters produced by Ward's method and K-means clustering were viable.

TABLE 4.8

SIGNED RANK TEST RESULTS TESTING THE HYPOTHESIS THAT THE MEDIAN F-MEASURE OF ENSEMBLE-BASED RESULTS IS GREATER THAN THE MEDIAN F-MEASURE OF CORRELATION-BASED RESULTS.

	Algorithm			
Ensemble	UPGMA	Ward	K-means	Walktrap
min	0.6303	0.4267	0.3979	0.0827
max	**5.41e-6**	0.1763	0.0525	0.0715
sum	**0.0262**	0.3696	0.3696	0.0715

For each clustering algorithm and ensemble the vector of F-measures corresponding to the number of clusters were compared to the corresponding set of F-measures from correlation-based clustering results. P-values less than 0.05 appear in bold.

TABLE 4.9

SIGNED RANK TEST RESULTS TESTING THE HYPOTHESIS THAT THE MEDIAN BHI OF ENSEMBLE-BASED RESULTS IS GREATER THAN THE MEDIAN BHI OF CORRELATION-BASED RESULTS.

	Algorithm			
Ensemble	UPGMA	Ward	K-means	Walktrap
min	**0.0007**	0.3979	**5.41e-6**	**1.08e-05**
max	**0.0005**	0.9855	1	0.9384
sum	**5.41e-06**	**5.41e-06**	0.9999	0.9384

For each clustering algorithm and ensemble the vector of BHIs corresponding to the number of clusters were compared to the corresponding set of BHIs from correlation-based clustering results. P-values less than 0.001 appear in bold.

TABLE 4.10

THE NUMBER OF VIABLE CLUSTERS FOR CLUSTERING EXPERIMENTS USING UPGMA AND WALKTRAP AND ALL ENSEMBLE SIMILARITY MEASURES AND CORRELATION IN THE *E. COLI* DATA SET.

	UPGMA				Walktrap			
Number of clusters	cor	min	max	sum	cor	min	max	sum
10	1	7	1	2	10	7	2	2
20	1	17	1	8	20	16	2	2
30	2	27	1	11	30	26	1	1
40	5	37	1	12	40	35	1	1
50	7	47	1	16	50	41	1	1
60	8	57	1	17	60	48	1	1
70	10	67	1	17	70	55	1	1
80	13	77	1	20	80	63	1	1
90	13	87	1	25	90	70	1	1
100	16	97	1	27	100	78	1	1

Viable clusters are those that contained enough genes with GO terms and interactions for analysis. All clusters produced by Ward's method and K-means clustering were viable.

CHAPTER 5

ENSEMBLE TOPIC MODELS FOR PRIVACY PRESERVING DATA SHARING ACROSS DISTINCT POPULATIONS

Throughout this dissertation, we have considered data mining challenges in the domain of systems biology. Namely, the problems of integrating heterogeneous and noisy data with little ground truth. These same problems exist in the healthcare domain. Healthcare data is additionally very sensitive as patients have privacy concerns. We present an ensemble topic model approach to estimate patient disease risk while ensuring the privacy of patient healthcare records. In this chapter we utilize one of the central themes of this work, the use of an explicit representation of data as relationships, in order to improve modeling in the healthcare domain. Our approach is also in support of our assertion that the use of multiple measures is important for models: we utilize both an explicit co-occurrence representation of the data and a "flat" or transactional representation.

5.1 Introduction

With the recent signing of the Patient Protections and Affordable Care Act into law the use of electronic medical data is set to become ubiquitous in the United States. This presents an unprecedented opportunity to use data mining for the benefit of patient health. However, there are two major hurdles to utilizing this wealth of data. First, medical data is not centrally located but is often divided into regions by companies that warehouse the data for local hospitals or by research organizations. Second, federal and state laws prevent sharing specific or identifiable information. It would be of great benefit to share information between these distinct populations, as a data set concerning cancer patients for example could greatly increase the accuracy of related computationally derived disease risks for everyone else. Still, organizations may have a vested interest in keeping their data sets private as they may have been gathered and curated at great cost. We propose an approach to allow the sharing of beneficial information while staying within the bounds of the law and maintaining the privacy of the data. We show that the use of a novel probabilistic graphical model can can facilitate effective transfer learning between distinct healthcare data sets by parameter sharing. Our method utilizes aggregate information from distinct populations in order to improve the estimation of patient disease risk.

The growing and mandated use of electronic medical records will allow scientists to unveil new discoveries about human health. An enormous quantity of healthcare data is created every year and there are vast amounts of past medical records that are being imported into electronic formats. A Center for Disease Control study estimates that patients in the United States made 1.2 billion visits or an average of 4.05 visits per patient to physicians' offices in 2007 [123]. The rate of visitation increased 11 percent since 1997.

Aspirin is one exemplary case of the utilization of available medical data. Multiple studies have found that Aspirin reduces the long term risk of colorectal cancer, the progression of cardiovascular disease and the likelihood of stroke [1, 5]. These studies relied on the wealth of data already available about Aspirin. Similarly, we hope to utilize existing electronic health records (EMRs) to discover relationships between diseases and to improve disease risk prediction for patients.

While recently enacted healthcare laws mandate the use and utilization of EMRs, they do not specify how they should be stored or who should store and maintain the records in sufficient detail. At present this lack of centralization is impeding the meaningful use of this data. Each hospital and medical research organization may have their own data set, which they are compelled by law not to share due to privacy concerns. Regional data warehousing organizations have arisen to consolidate and store EMRs, but they are equally subject to restrictions on sharing the data. Additionally, EMRs have become a commodity, as the maintenance and security of the storage systems can be costly. Therefore, organizations may have incentive to protect their data sets.

Sharing complete EMRs would be the best means of promoting beneficial and meaningful use but there are obstacles to full disclosure of the data. However, the Health Insurance Portability and Accountability Act of 1996 stipulates that aggregated information can be shared freely [60]. We propose an approach that allows aggregated information to be shared between distinct organizations with EMR data in a way that increases the accuracy of computational prediction of disease risk. We utilize an ensemble approach to gain more predictive accuracy with little information. We posit that this approach is mutually beneficial to all organizations warehousing EMRs and maintains the privacy of the patients while protecting any potential interests in keeping valuable data sets private.

We further propose the use of learned parameters of topic models as an alternative approach to creating interpretable network models from EMR data. Networks have been an intuitive and useful approach to modeling complex data and presenting a representation that domain experts can understand. Perhaps the most closely related work in the healthcare domain utilizes collaborative filtering to create a disease-gene network based on some similarity criterion between diseases [32]. While network models can be very effective at identifying disease risk, many network approaches utilize different edge weighting methods, which may lead to different interpretations of the data [136]. Furthermore, many approaches to integrating distinct networks are computationally intractable. We view the network approach as a bottom-up construction of a relational model by inspecting individual health records. We propose a top-down topic modeling approach that begins with a partitioning function we wish to optimize on the data. By creating a topic model that explicitly measures disease co-occurrence we simultaneously learn the network that best models the data according to our criterion and partition it into meaningful groups with co-occurring diseases. The use of this approach simultaneously creates an interpretable model and allows easily computed solutions for combining information that creates a network. To this end we propose a novel extension to a well known probabilistic graphical model that optimizes the grouping of medical records on occurrence and co-occurrence of disease.

This is to our knowledge the first use of topic models to infer network structure and the first application to EMR data. However, topic models have been used to identify topics in medical documents and public health topics in Twitter [155, 111, 39]. Ensemble topic models have also been studied, although not in this context [128].

This study makes three contributions to the medical domain:

- A novel application of topic modeling to the analysis of disease risk.
- A novel topic modeling approach for the study of relational data.
- We support our proposition that distinct EMR warehousing organizations should share general information with evidence that it can improve the utility of disease risk prediction on each individual data set.

5.2 The model

Our approach extends the well studied Dirichlet Process Mixture Model (DPMM), which is depicted in Figure 5.1. DPMM is a non-parametric approach that learns an unspecified number of groups with distinct distributions over features.

5.2.1 The Dirichlet Process

The Dirichlet distribution is the multivariate generalization of the beta distribution. One formulation of the DP is described in the stick-breaking process. Imagine that a stick is broken repeatedly such that the first section has a length dependent on the Beta distribution: $\beta'_1 \sim Beta(1, \alpha)$. The remainder of the stick is broken in the same way such that $\beta_k = \beta'_k * \prod_{i=1}^{k-1}(1-\beta'_i)$. It has been shown that if $G \sim DP(\alpha_0, G_0)$,

$$G = \sum_{k=1}^{\infty} \beta_k \gamma_{\phi_k} \tag{5.1}$$

Where β_k are stick-breaking weights depending on the parameter α_0, γ_{ϕ_k} is an atom at ϕ_k, each representing an independent random variable [125]. Using this stick-breaking approach, any measure can be used to determine a set of discrete weights. DPs are often used to set priors for components of mixture models [126].

5.2.2 Dirichlet Process Mixture Model

We focus here on the use of multinomial base distributions as the multinomial is appropriate for the measurement of binary and count data.

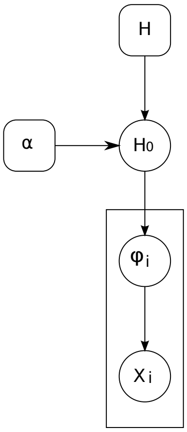


Figure 5.1. The Dirichlet Process Mixture Model. H is a base distribution from which weights are drawn as described in Equation 5.1. The mixing proportions of the components are specified by H_0. The parameters of the base distributions are specified by ϕ_i. X_i represents an observed instance.

The model is described by the hierarchical specification:

$$
\begin{aligned}
\phi_i | H_0 &\sim H \\
H &\sim DP(H, \alpha) \\
x_i | \phi_i &\sim F(\phi_i)
\end{aligned}
\tag{5.2}
$$

In the standard DPMM, F is the multinomial probability mass function. In the context of EMRs, the DPMM with a multinomial base distribution over disease occurrence optimizes for groups of patients that have received the same diagnoses.

5.2.3 Our approach

We propose an alternative formulation of a DPMM in which F is a function of both the multinomial over the diseases and a second multinomial over disease co-occurrence. This model allows us to explicitly construct a network representation of the data by utilizing co-occurrence counts explicitly. It is inspired partly by the effectiveness and generality of gaussian mixtures. A multivariate Gaussian is parameterized by a mean vector and a covariance matrix. The covariance matrix specifies not only the spread of values around the mean but the relationship between features. This is much more specific information than is captured by a multinomial. However, inferring the parameters of a multivariate Gaussian can be much more complex than inferring the parameters of a multinomial. Furthermore, a multivariate gaussian is not appropriate for binary values as a Gaussian distribution only accurately models a set of binary values in the edge cases where all values are 1 or 0.

Our approach finds a balance between parsimony and specificness by placing equal weight on the co-occurrence of diseases and the presence of disease. We call our approach Co-occurrence Based Clustering (CBC) for its focus on explicitly learning the co-occurrence of diseases. In CBC, the multinomial over the diseases is analogous to the mean vector of a multivariate Gaussian. The multinomial over the disease co-occurrences is analogous to the covariance between each pair of diseases. This formulation allows CBC to capture co-occurrence explicitly while maintaining generality. This is essential as many patients may be lacking multiple appropriate diagnoses The model is learned by Gibbs sampling in which the likelihood function gives equal weight to the two multinomials, as shown in Equation 5.3.

$$F(\mathbf{X}, \mathbf{X}', \phi, \phi') = \frac{\frac{[\sum_{\mathbf{i}}^{\mathbf{k}} \mathbf{X_i}]!}{\prod_{\mathbf{i}}^{\mathbf{k}} \mathbf{X_i}!} \prod_{\mathbf{i}}^{\mathbf{k}} \phi_{\mathbf{i}}^{\mathbf{X_i}} + \frac{[\sum_{\mathbf{i}}^{\mathbf{k}} \mathbf{X'_i}]!}{\prod_{\mathbf{i}}^{\mathbf{k}} \mathbf{X'_i}!} \prod_{\mathbf{i}}^{\mathbf{k}} \phi'^{\mathbf{X'_i}}_{\mathbf{i}}}{\mathbf{2}} \tag{5.3}$$

Where $\mathbf{X}$ is the matrix of diagnoses for all patients, ϕ is the matrix of disease occurrence parameters for each component, ϕ_i is the probability of observing a disease (alternatively $\phi_i = \mathbf{X_i}/\sum \mathbf{X_i}$), and $\mathbf{X}'$ and ϕ' are the analogous parameters for disease **co**-occurrence.

Relationships between diseases may not be apparent from examining their frequency separately. The use of a relational representation of the data —the co-occurrence of diseases— allows even simple models to take into account this more specific information.

While DPMM can be applied to the co-occurrence of diseases, the reliance on co-occurrence alone can undermine the generality of the model. If a patient has a disease that has not been diagnosed, then all 252 (in our data) potential co-occurrences will be missing, whereas in the flat occurrence representation, only a single value will be missing. Thus a little noise can have an overwhelming effect on the model. CBC is more tolerant to this source of noise by virtue of considering both co-occurrence and frequency.

In our analysis we demonstrate these differences by comparing disease ranking results across three formulations of DPMMs: DPMM is a DPMM trained on disease occurrence data. COOC is a DPMM trained on co-occurrence data, and CBC is our model utilizing both representations of the data.

5.2.4 Markov Chain Monte Carlo inference

We utilize a version of Gibbs sampling with auxiliary parameters [106]. This approach allows us to sample the component membership of the model without having to integrate with respect to the prior distribution H. Algorithm 1 describes the steps in our sampler.

Algorithm 1 Gibbs sampler with auxiliary parameters

```
For $i = 1, ..., n$: Let $k^-$ be the number of components $c_j$ for $j \neq i$, and $h = k^- + m$, where $m$ is the number of auxiliary variables. If $c_i = c_j$ for some $j \neq i$, draw values independently from the base distribution $H$ for the parameters of components $\phi_c$ (and correspondingly $\phi'_c$) for which $k^- < c \leq h$. If $c_i \neq c_j$ for all $j \neq i$, let $c_i$ have the label $k^- + 1$, and draw values from H for those $\phi_c$ for which $k^- + 1 < c \leq h$. Draw a new value for $c_i$ from $1, ..., h$ with the following probabilities:

    $$P(c_i = c | c_{-i}, y_i, {y'}_i, \phi_{-c}, {\phi'}_{-c}) = \begin{cases} b\frac{n_{-i,c}}{n-1+\alpha} F(y_i, {y'}_i, \phi_c, \phi'_c) & for\, 1 \leq c \leq k^- \\ b\frac{\alpha/m}{n-1=\alpha} F(y_i, {y'}_i, \phi_c, \phi'_c) & for\, k^- < c \leq h \end{cases} \quad (5.4)$$

    where $y_i$ is an instance, ${y'}_i$ is the corresponding set of co-occurrences, $n_{-i,c}$ is the number of $c_j$ for $j \neq i$ that are equal to $c$, and $b$ is a normalizing constant. Remove $\phi_c$ that are not associated with at least one observation.
For all $c \in c_1, ..., c_n$: draw a new value from $\phi | y_i$ such that $c_i = c$.
```

In our experiments we used the parameters $m = 1$ and $\alpha = .01$. The algorithm specifically describes the sampler for CBC, but the samplers for DPMM and COOC are the same with the exceptions that DPMM uses ϕ_c and y_i exclusively and COOC uses ${\phi'}_c$ and ${y'}_i$ exclusively.

5.2.5 Ensemble learning

The goal of ensemble learning in this context is to allow models to achieve performance increases through the use of data from distinct sources. Examples include data from different domains such as healthcare and genomics, data with different distributions such as healthcare data from different ethnic or socioeconomic groups, or even data with different feature spaces if for example there are no occurrences of a disease in one group that is present in another group. We utilize an approach that is common in transfer learning, known as parameter passing [108]. In this approach base models are learned on distinct data sets. The base models are then combined in a separate step by joining the parameters of the models. Wang et al. propose a similar approach in which different topic models are combined by running an additional clustering step on the component *labels* from base topic models [151].

Figure 5.2 outlines the process used to create ensembles. We build ensembles by training base models on each demographic data set. The ensemble step combines the occurrence parameters ϕ of the base models into a single matrix. A DPMM is trained on this matrix to form a ensemble-level model. Disease risk is assessed for an individual patient by first finding the component of this ensemble model that best fits their disease profile, then combining the parameters of every component from the base models whose parameters are in the ensemble-level component. The base-model parameters are averaged to form the parameters of a consensus model.

5.3 Data

We test our approach on a data set of EMRs from medicare patients. Our data set contains information from 13,039,018 elderly patients with a total of 32,341,348 medical records. The data originates from claims data for medicare beneficiaries who were at least 65 years old in 1993 [21]. As co-occurrence is a focus of our approach, we use

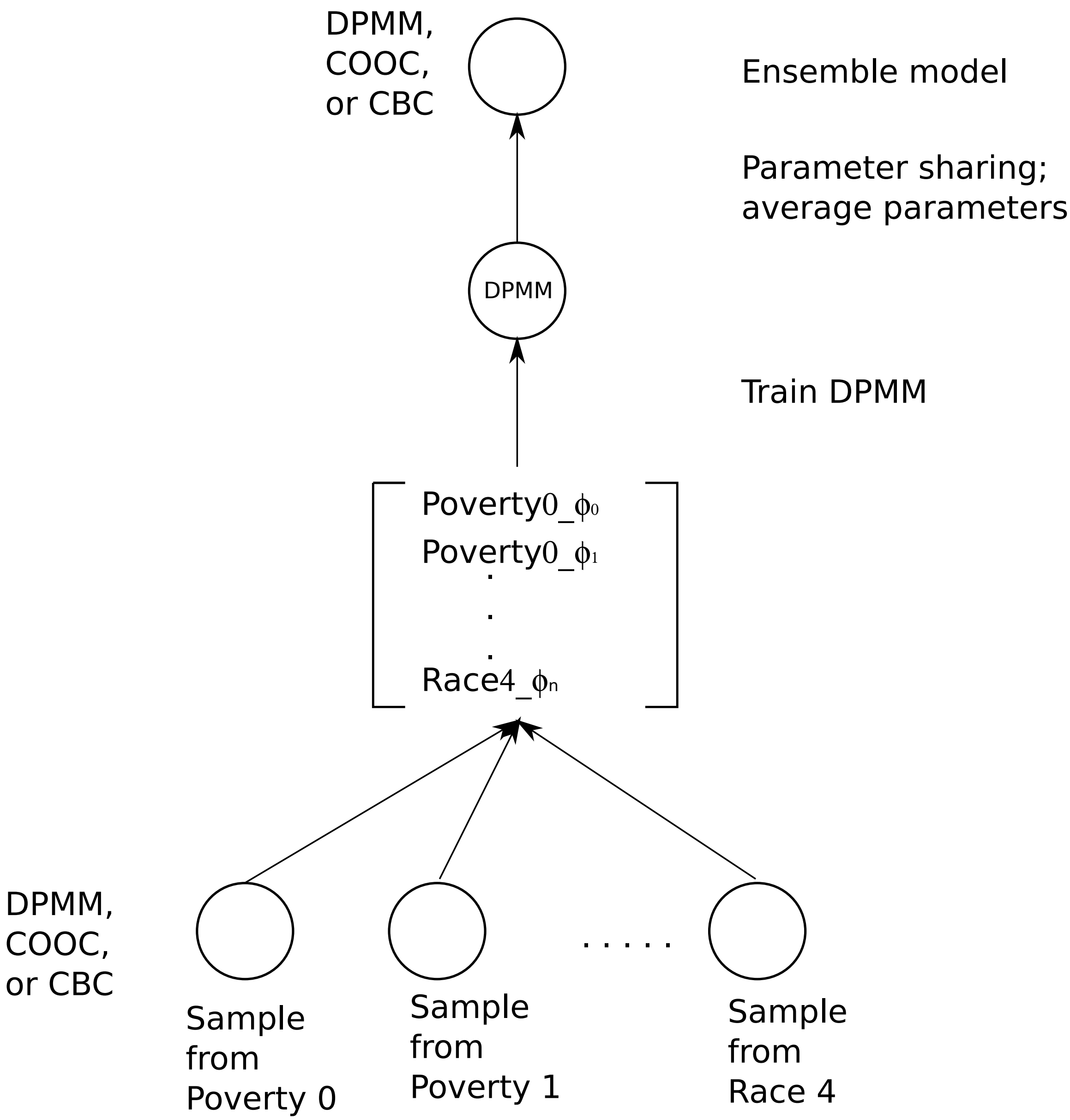


Figure 5.2. The ensemble takes base models, each consisting of the results from a single model trained on one data set, then combines the parameters of each component from the base mixture models into a single matrix.

a subset of the data containing records from 7,895,283 individuals with three or more diagnoses. The raw data set contains ICD-9 codes for describing the diagnoses that apply to each patient. ICD-9 codes exist in a hierarchy of disease that can complicate analysis [33]. Collapsed ICD-9 codes provide a mapping from specific diagnoses to general diagnoses. For example, ICD-9 codes “9843” and “9845” correspond to pneumonia from whooping cough and pneumonia from anthrax, respectively. Both can be described by the shortened ICD-9 code “984” or by the CCS code “122”. CCS codes provide a standardized coding system based on the ICD-9 specification and is designed to be clinically meaningful and more useful for statistical analysis. Therefore, we utilize the Clinical Classifications Software (CCS) codes to provide a more general non-hierarchical classification of disease than ICD-9 codes [28].

We approach our goal of demonstrating the effectiveness of learning across distinct healthcare data sets by splitting the data into distinct populations by demographics based on poverty level, gender, race, and age. Table 5.1 shows that these groups tend to have similar numbers of disease diagnoses. The most significant difference appears to exist between the two poverty groups, in which the variance and kurtosis are strikingly different. This indicates that there is a wider range of number of diagnoses among these two groups.

TABLE 5.1: DESCRIPTIVE STATISTICS FOR THE NUMBER OF DISEASES PATIENTS SUFFER FROM IN EACH DEMOGRAPHIC.

	All	Poverty 0	Poverty 1	Gender 0	Gender 1	Race 0	Race 1	Race 2	Race 3	Race 4
# of patients	7895283	7050861	844421	3247168	4648114	83500	7050861	652158	79181	8075
Maximum	66	66	63	66	57	47	65	51	51	39
Mean	9.25	9.12	10.36	9.18	9.30	8.77	8.21	8.69	8.28	7.12
Variance	20.87	19.97	26.98	20.31	21.25	23.00	20.47	24.55	22.04	14.43
Skewness	1.61	1.63	1.38	1.62	1.60	1.43	1.62	1.51	1.66	1.96
Kurtosis	3.31	3.42	2.34	3.34	3.28	2.43	3.36	2.75	3.60	5.31

Individual disease prevalence is much more strikingly different between these groups. The top 20 most common diseases in the original data set are listed in Table 5.2. A patient who is in demographic poverty 1 is twice as likely to be diagnosed with a cognitive disorder as a patient on the other side of the poverty line. A patient of gender 0 is nearly three times as likely to suffer from genitourinary symptoms as a patient of gender 1. There are many additional differences between races and other demographics that demonstrate the distinctions between these populations. We were surprised to find that the disease prevalence in the 10% youngest patients in the data set was very similar to the disease prevalence in the 10% oldest patients. However, this may be explained by the fact that the distribution of patient age is strongly skewed towards the younger patients and that the data set consists entirely of patients with at least 65 years of age. Given the lack of interesting differences between the age groups, we focused on the poverty, gender, and race demographics for our analysis. Among the race demographics, only 8,075 patients were of Race 4, providing a relatively small sample. As such, Race 4 was not used for analysis.

5.4 Evaluation

We trained DPMM, COOC, and CBC on fifty random samples of 4,500 instances from each poverty, race, and gender dataset. CBC models trained on a single sample from each demographic data set were combined as described in Section 5.2.5. Ensembles of DPMM and COOC models were constructed in the same way. Gibbs sampling was carried out for an average of 996 iterations (1000 with few exceptions for technical reasons) for each base model and for the ensemble models.

The accuracy of disease risk was measured by holding out a test set of instances of 500 patients from each demographic and calculating the likelihood of each disease given all but one of the observed diseases in the test instance. This was repeated by withholding each observed disease for every test instance. A test set of 500 instances with an average of 5 diseases per patient would result in 2500 individual rankings.

The ranking of the diseases was evaluated as in previous work by calculating the proportion of predicted disease rankings that were in the top ranks [33]. The lower

TABLE 5.2: PERCENTAGE OF PATIENTS WITH TOP 20 MOST PREVALENT DISEASE BY DEMOGRAPHIC.

	All	Poverty 0	Poverty 1	Gender 0	Gender 1	Race 0	Race 1	Race 2	Race 3	Race 4
Essential Hypertension	43.78	43.88	42.92	39.10	47.04	33.34	36.28	46.10	36.88	37.32
Fluid and electrolyte disorders	39.48	38.18	50.30	34.19	43.18	38.09	31.64	39.29	35.85	29.91
Coronary atherosclerosis	38.69	38.99	36.16	43.39	35.40	34.57	37.07	29.97	31.21	31.45
Cardiac dysrhythmias	32.74	32.94	31.02	35.98	30.47	34.03	32.53	25.93	27.38	25.67
Congestive Heart Failure	27.84	27.09	34.09	27.50	28.08	31.77	25.27	26.11	25.01	17.15
Urinary tract infections	26.58	25.17	38.38	18.16	32.46	24.72	21.03	26.11	21.29	17.23
Bronchitis	25.42	25.13	27.83	31.68	21.05	25.81	25.48	18.73	19.85	16.27
Anemia	20.72	20.25	24.62	18.88	22.00	17.15	14.08	22.18	17.51	15.30
Diabetes w/o complication	18.80	18.31	22.88	18.74	18.84	14.15	13.76	20.57	19.74	17.36
Pneumonia	18.40	17.50	25.90	19.98	17.29	21.32	16.42	15.96	19.92	14.19
Surgical complications	16.67	17.22	12.15	19.70	14.56	13.66	16.50	12.63	15.61	15.62
Osteoarthritis	14.61	14.46	15.82	10.30	17.61	10.45	11.52	10.61	6.67	6.29
Bacterial infection	13.44	12.68	19.76	10.26	15.66	14.01	12.49	13.63	13.06	10.86
Heart valve disorders	12.41	12.70	09.97	11.73	12.88	12.68	12.29	10.33	10.23	9.79
Cerebrovascular disease	11.80	11.38	15.24	11.77	11.81	12.82	10.49	14.37	14.03	13.60
Pneumothorax	11.60	11.57	11.85	12.10	11.25	12.77	11.18	10.51	11.28	9.04
Genitourinary symptoms	11.51	11.56	11.08	17.99	6.98	12.03	11.31	11.36	11.51	11.41
Cystic fibrosis	11.07	11.14	10.54	10.85	11.23	9.52	10.17	7.95	10.67	10.41
Cognitive disorders	10.89	9.76	20.39	8.89	12.29	14.48	10.54	13.16	8.68	6.83
Gastrointestinal hemorrhage	10.69	10.49	12.33	11.07	10.42	8.63	7.33	8.21	9.60	9.18

ranks are more important as a medical professional reviewing a list of predicted diseases is much more likely to read predictions early in the list.

5.5 Results

The log-likelihood plot in Figure 5.3 shows that the likelihood of the algorithms appear to be in stable states after 1000 iterations on the gender demographic.

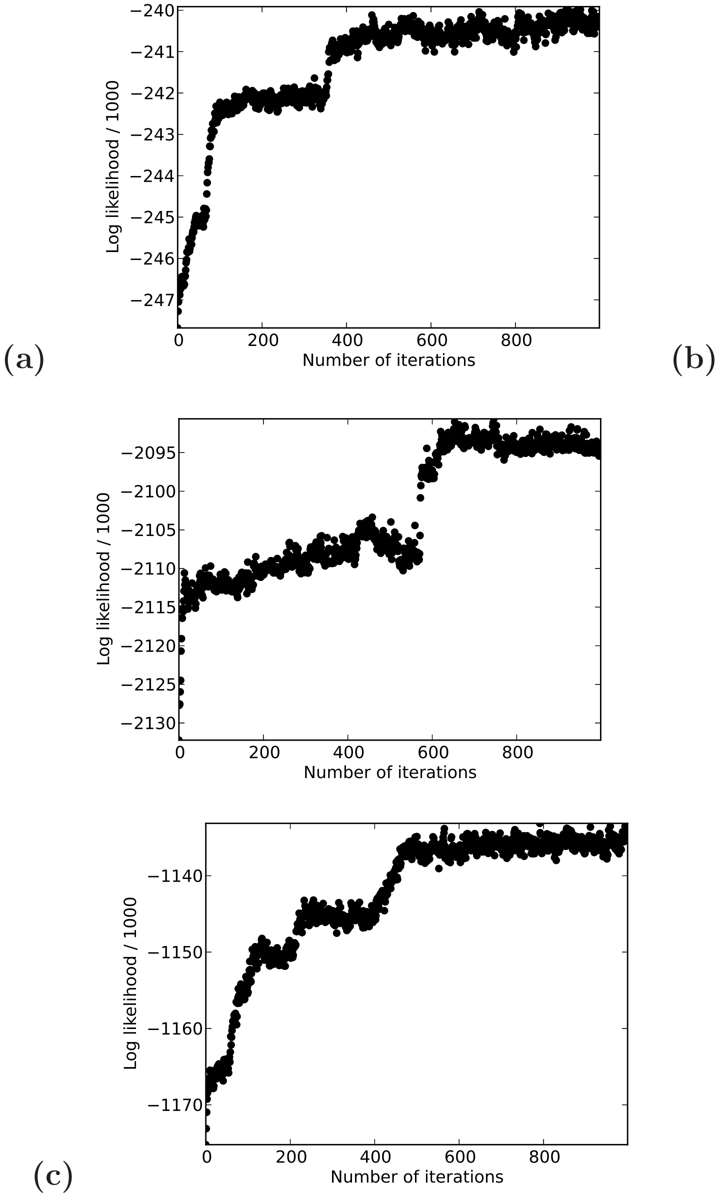


Figure 5.3. The log-likelihood of all three algorithms over 1000 iterations of Gibbs sampling. All log-likelihood values are divided by 1000 for readability. Panel **(a)** shows the log likelihood of DPMM. Panel **(a)** shows the log likelihood of DPMM. Panel **(b)** shows the log likelihood of DPMM using the co-occurrence data. Panel **(c)** shows the log likelihood of CBC.

The proportion of missing diseases that were ranked as disease risks for patients is shown in Figure 5.4. The proportions were determined by averaging across the ranking for patients in test sets from all demographics and all experiments. The ensemble of CBC models provides the best rankings in the highest ranks, with the most accurate predictions for 8 of the first 10 ranks, 18 of the first 20 ranks, 27 of the first 50 ranks.

This approach ranks the missing disease from a patient's diagnosis in the first 10 listed diseases 47.5% of the time and the first 20 ranks 76.5% of the time. The ensemble of DPMM models performs best when considering ranks greater than 29. CBC is expected to perform better at lower ranks as the components in CBC utilize more specific co-occurrence data. All of the methods place nearly 100% of the missing diseases in the first 50 ranks. Notably, base-CBC performs next to worst, whereas the ensemble-CBC performs the best. This indicates that the base CBC models are diverse; they capture differences in the separate demographics.

The nearest comparison to this study utilizes collaborative filtering on collapsed ICD-9 diagnoses to rank the likelihood of diagnoses in the last visit based on patients' medical history [33]. Where that method identifies 54.7% of future diagnoses in the top 20 ranks, our approach identifies 76.5% of held out diagnoses. While these methods utilize the same data, it is important to note that the approach of Davis et al. relies on temporal data and collapsed ICD-9 codes instead of CCS codes, making direct comparison problematic.

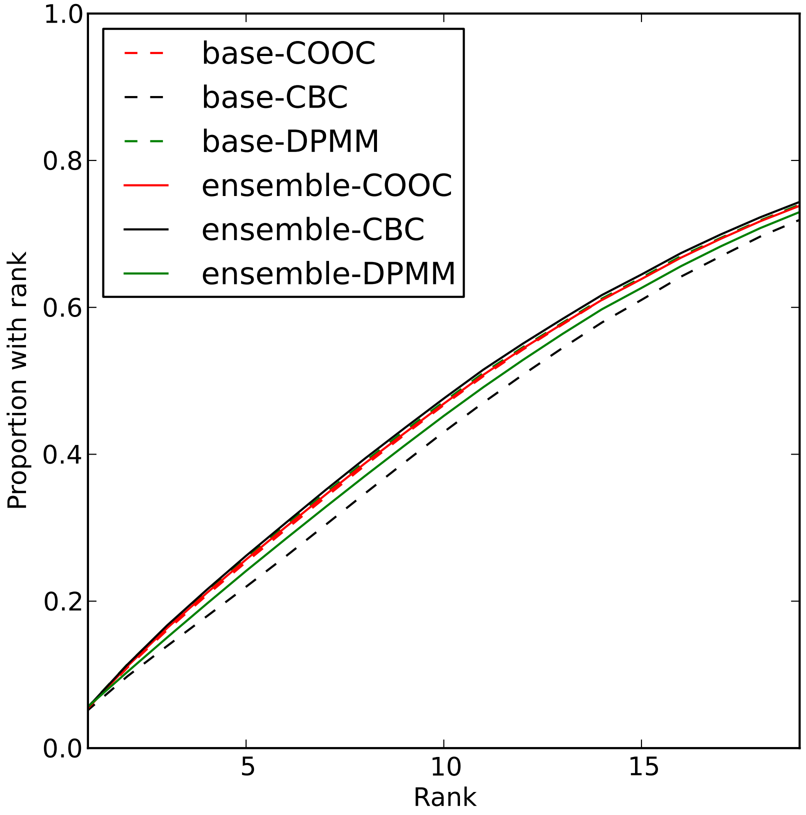


Figure 5.4. The proportion of held-out diseases given rank less than or equal to the value on the x-axis. Labels “base” and “ensemble” correspond to ranks given by the algorithms trained on a single demographic data set and ranks given by the ensemble across demographics.

Different patients may have very different histories of diagnosis. The specific diagnoses that an individual patient has may be more or less predictive than others. Figure 5.5 shows the relationship between the number of diagnoses and the mean rank of diagnoses for individual patients. As expected, the variance in the accuracy of the model decreases sharply as the number of available diagnoses increases.

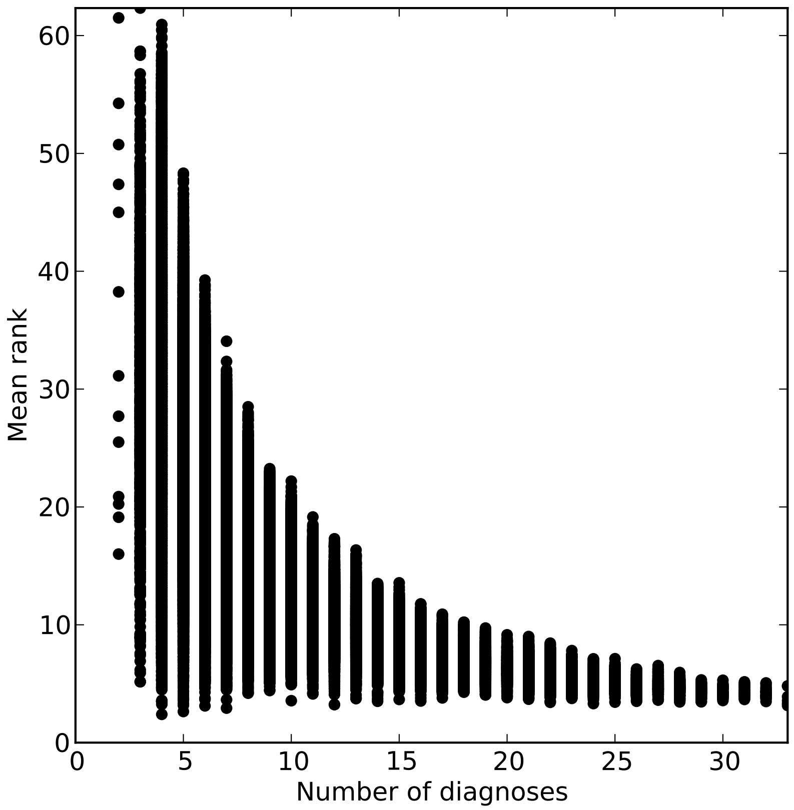


Figure 5.5. Mean rank for individual patient diagnoses versus the number of diagnoses available based on an ensemble of CBC models.

5.6 Interpreting the model

In addition to ranking individual patient disease risks, we are interested in creating a global model of disease relationships. Figure 5.6 shows a network constructed from one ensemble CBC model. Nodes represent diseases. Their groups (signified by color) are determined by the component in the ensemble that contains the most patients with the given diagnosis. Edge weight was determined by averaging disease co-occurrence across all components. Therefore, edges may tend to represent *global* co-occurrences rather than within component co-occurrences. The figure shows 64 distinct disease groups, determined by the number of patients in the base model components contributing to each ensemble component. Many of these groups contain diseases which are clearly similar. For example, the yellow group in the top row and fifth from the left, contains 9 cancer diagnoses. The remaining three are "gastritis and duodenitis," "intestinal infection," and perhaps oddly, "deficiency and other anemia." Other clusters appear less specific to the layman, but still contain common sense groups. For example, the larger red group, bottom left and three from the left, contains 9 pregnancy related diagnoses and 4 abdominal pain related diagnoses. Edges in the network represent the mean edge strength from across all components in the ensemble. We used the 99th percentile edges to form this network. These weights represent strong relationships that are not strong enough to determine component membership alone. Some of the strongest edges join two groups, the pink group third from the right in the middle row, and the red group third from the left in the third row. These join chemotherapy related issues, dizziness or vertigo, nervous system anomalies, unspecified circulatory disease, and unspecified eye disorders. A thorough analysis of this network requires medical expertise, however it is clear that there is meaningful structure to be investigated here. This model is provided with CCS designations as a Cytoscape file at http://www.cse.nd.edu/∼arider1/cbc_meta_meanedge.cys.

Figure 5.7 shows a spring-layout view of the same network. This figure shows

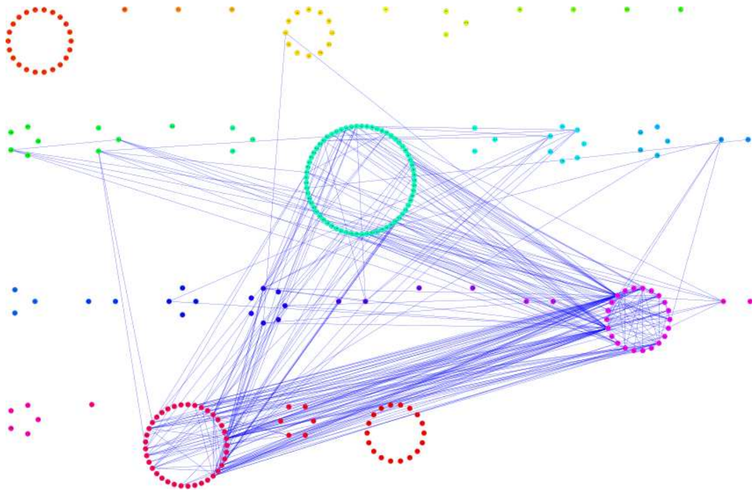


Figure 5.6: A network constructed from a single CBC ensemble. Nodes represent disease diagnoses and edges represent co-occurrences. Node groups are determined by the component in the model with the most diagnoses. Edge weight was determined by averaging disease co-occurrence across all components. Therefore, edges may tend to represent *global* co-occurrences rather than within component co-occurrences. Only edges in the 99th percentile weight category are shown.

that the clustering as determined by edge weight is also informative. The five of the six nodes in the group nodes on the rightmost side of the central cluster concern birth related diagnoses. The group of four nodes at the bottom most edge of the central cluster contain “OB-related trauma to perineum and vulva,” “fetal distress and abnormal forces of labor,” “Cardiac and circulatory congenital anomalies,” and “acquired foot deformities.” This layout additionally highlights two hubs, “disorders of lipid metabolism” and “coma; stupor; and brain damage.”

We refrain from making a full enumeration of interesting clusters, but we have found that various edge weight based layouts provide additional clusters that seem to make sense. We encourage the reader to investigate these clusters by downloading the provided network.

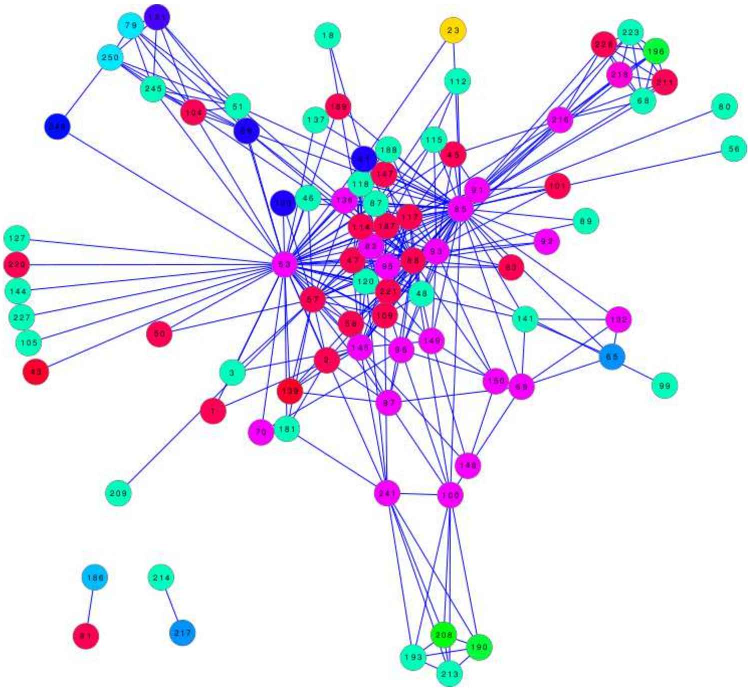


Figure 5.7: Another view of Figure 5.6 using the spring-layout based on edge weight. Edge weight was determined by averaging disease co-occurrence across all components and reflects global trends. This view reveals clusters and hubs in the network.

5.7 Discussion

We set out with the goal to provide an approach that would allow and encourage EMR warehousing organizations and research centers to share EMR data for their mutual benefit and the benefit of patients. Our analysis demonstrates that the proposed use of aggregate data improves ranking across diverse patient populations. Therefore we strongly recommend that EMR warehousing organizations share this aggregate data both as an act of good will and as an act of self interest, as more available data will improve modeling on individual data sets.

We additionally sought to provide a means to create an interpretable model from disparate aggregate data. We proposed a method that explicitly utilizes co-occurence data to learn a network while simultaneously providing imroved disease risk predictions. We demonstrated that the network constructed contains comprehensible groupings of disease occurrence, based both on the component labels in our model and on the global mean edge weights used to construct the network. Although it can be exceedingly difficult to quantify the utility of a network model, the provided examples do indicate that this model may contain useful medical information.

CHAPTER 6

EVALUATING MODELS TRAINED WITH MISLABELED NEGATIVE CLASS INSTANCES

In Chapter 4 we proposed an approach to improve exploratory analysis of noisy data. Our approach focused on improving unsupervised learning in part because of a lack of good or complete ground truth data in systems biology. In this chapter we confront this problem and study how the state of ground truth data in systems biology affects classification. While the relatively small amount of labeled data available is a problem, a more subtle challenge is that the negative class is not well defined in for many problems with underlying network structure in the data. Classifiers trained on genetic or other interactions are commonly trained using a negative class made up of interactions that are simply not known to be positive [109, 116, 147]. In this chapter we examine the behavior of standard evaluation approaches for classifiers trained on data with mislabeled negative class instances. We generalize this study to many different domains and algorithms as this problem is not limited to network representations of data or unsupervised methods.

The problem of mislabeled negative class instances is ubiquitous in systems biology and healthcare. The work in this chapter underscores the primary thesis of this work: that proper utilization and understanding of domain knowledge is key to success in data mining. Understanding fundamental problems in the data can lead to better data mining.

6.1 Introduction

The traditional concept of a negative class does not apply to many problems for which classification is increasingly utilized. When the negative class contains mislabeled instances that are truly positive, the ambiguity in the decision boundary complicates the classification task and confounds attempts to accurately evaluate classifier performance. While there have been attempts to design algorithms specifically for this problem, the task of evaluation in this scenario is not well understood. In this chapter we seek to answer the following questions: How reliable are evaluation metrics when the negative class contains an unknown proportion of mislabeled positive class instances? What can evaluation metrics tell us about potential systematic biases in the data? Can evaluation metrics give us further insight into the *data* when it contains mislabeled positive class instances? These questions deserve careful investigation when faced with this scenario. In the pursuit of answers we provide a motivating real world case study and provide a general framework for approaching evaluation when the negative class contains mislabeled positive class instances. We show that the behavior of evaluation metrics is unstable in the presence of uncertainty in class labels. Furthermore, the stability of evaluation metrics depends on the kind of bias in the data. Finally, we investigate the effects of the amount of bias on these metrics. We show that the type and amount of bias present in the data can have a significant effect on the ranking of evaluation metrics and the degree to which they over or underestimate the true performance of classifiers.

Standard classifiers are often applied to data with a poorly defined negative class [109, 116, 147]. In many cases, there is an implicit assumption that data are mislabeled completely at random. This is common even among algorithms that are designed for mislabeled positive class data [43, 92]. This assumption is unrealistic in real world scenarios where there may be multiple sources of different systematic biases in experimentation and data collection. Furthermore, the proportion of true negative class instances to mislabeled positive class instances is often expected to be overwhelmingly large. While this would seem to validate confidence that the effect of mislabeled positive instances will be minimal, it has not been shown to be a safe assumption for an unknown proportion of mislabeled instances with unknown bias.

We motivate this study through the analysis of an experiment that is actually used to try to answer some of the most pressing issues in biology today. In performing the study we uncover additional critical questions that must be answered in order to answer our motivating question, "How reliable are evaluation metrics when the negative class contains an unknown proportion of mislabeled positive class instances?"

6.2 Case Study

Physical interactions between proteins are one of the primary mechanisms by which a cell carries out its function. While there are high-throughput methods to measure protein-protein interactions (PPI), expense, noisy measurements, and the sheer number of possible interactions in even relatively simple organisms renders complete tests for all interactions infeasible. The identification of interacting proteins based on known interactions and related information is a common classification task in the biological domain [116].

The discovery of unknown protein interactions can have significant impact in pharmaceuticals and biology. With this in mind, we trained NaïveBayes classifiers on incremental updates to known protein interactions in Yeast. We collected this data from BIOGRID, a curated repository for protein interaction data sets from multiple organisms. This data is a real world case in which mislabeled positive class instances (unknown protein interactions) were incrementally revealed to be positive class with each update of the system [10]. Features consisted of expression data, Gene Ontology information, and known pathways. Each of these types of data have been previously used in the classification of protein interactions [78].

Expression data measures the amount of gene product (i.e. RNA) produced from each gene. It is an indirect way to measure the amount of protein produced by a cell. We gathered two features from expression data: one from a line cross experiment, in which two strains of yeast were bred, and one from a compendium of treatments in which yeast were exposed to chemicals and given mutations before measurements were taken [11, 74]. We collected a third feature based on the gene ontology (GO). The GO is a hierarchy of categories that describe the function, process, and biological components that genes are involved in. This feature was created by counting the number of GO slim terms (a high level set of GO terms) shared between each pair of genes. We used the number of shared pathways between genes as the fourth feature [22]. Pathways describe a series of interactions that lead to a product or change in a cell. Yeast has approximately 6000 genes, translating to roughly 18 million unique protein interactions. We trained NaïveBayes classifiers on this data set for five versions of BIOGRID. There was an average difference of about 20,000 interactions between each version of BIOGRID. Each data set contained 8 million instances with all positive protein interactions from that version of BIOGRID. The remainder of the instances were randomly under-sampled from the remaining potential protein interactions.

In order to evaluate classifier performance, we measured both the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR) of models trained on data using five versions of BIOGRID. We were interested in how accurate the evaluation metrics were in measuring classifier performance when many of the positive class instances were mislabeled. To this end, we measured the AUROC and AUPR based on the class labels from each given version of the BIOGRID database and the class labels from a more recent version of BIOGRID (version 3.1.85). We call the AUROC and AUPR that are based on the class labels from earlier versions of BIOGRID the “bias class AUROC and bias class AUPR” because of the presence of mislabeled instances. Similarly, we call the AUROC and AUPR that are based on the class labels from the most recent version of BIOGRID the “true class AUROC and true class AUPR” because of the additional positive class instances that are correctly labeled.

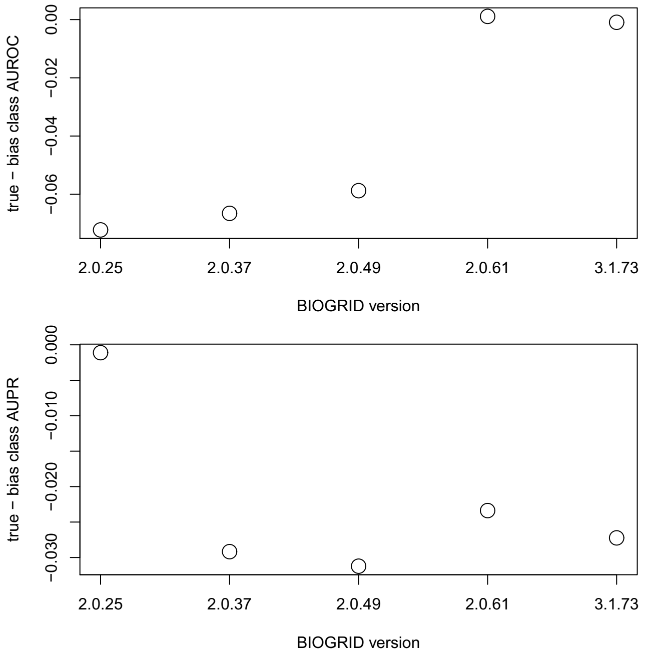


Figure 6.1. AUROC and AUPR of classifiers trained to predict protein interactions. The x-axis shows the BIOGRID update used to label positive interactions.

Figure 6.1 shows the difference between the true and bias class AUROC and the true and bias class AUPR of classifiers trained on the PPI data sets. Both the bias class AUROC and the bias class AUPR tend to overestimate classifier performance. The fact that the difference between the true class and bias class for both metrics does not improve reliably suggests that additional correctly labeled positive class instances are not giving the classifier enough information about the remaining mislabeled instances. In other words, the decision boundary remains noisy regardless of the smaller amount of mislabeled positive class instances. Figure 6.2 further supports this assessment.

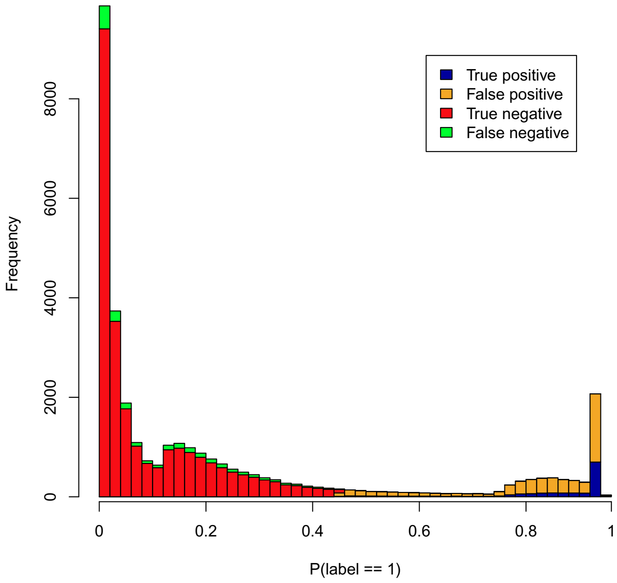


Figure 6.2. Histogram of positive class (interacting protein) probabilities based on known interactions from BIOGRID version 2.0.25. Only the 47 smallest bins shown for clarity.

Figure 6.2 shows a bar chart of the instances colored according to their class labels. For clarity the three largest bars are not shown as they contain the vast majority of instances. True positives, false positives, true negatives, and false negatives were identified by comparing the predicted class labels from classifiers trained on known interactions from the earliest BIOGRID version to the “true class” labels from the latest version of BIOGRID. For example, we identified true positives as instances that are known to be positive protein interactions in the latest version of BIOGRID that were also predicted as positive class by classifiers trained on known protein interactions from the first version of BIOGRID. The distribution appears multimodal, indicating that there is information within the given features that clearly separates many protein interactions into distinct groups. False negatives appear randomly spread throughout the true negatives. This may indicate that the protein interactions classified as false negatives are not related within the features in this data set to the protein interactions identified correctly as interacting.

The utility of the ranking for the identification of mislabeled positive instances is directly addressed in Figure 6.3. The figure shows the log probability of randomly obtaining the observed number of true instances in the bins in Figure 6.2 over the probability given by the classifier that an instance belongs to the positive class. We calculated the probability of randomly observing the number of positive class instances observed in each bin using the hypergeometric distribution. As with Figure 6.2, the mislabeled positive class instances were identified by comparing labels from the first version of BIOGRID to the latest version. A more negative number indicates a lower probability. The figure shows that the ranking of positive class instances becomes decreasingly informative as the probability assigned by the classifier decreases until about $p = 0.08$ where it levels off. Between 0.15 and 0.2 there is a large dip in the log probability that corresponds to one of the modes in Figure 6.2. This may indicate that there is something distinguishable about the instances that fall in this group.

Figures 6.1-6.3 demonstrate that our evaluation of the classifier is optimistic and that the addition of correctly labeled proteins does not seem to reliably affect classifier performance. This may indicate that the mislabeled positive class instances are

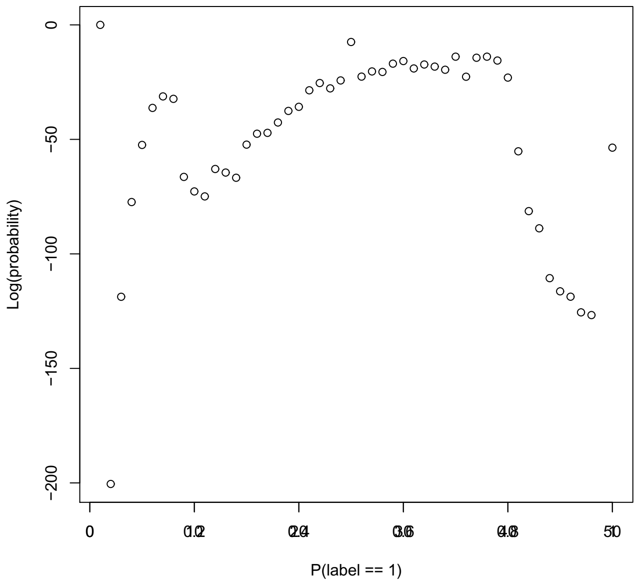


Figure 6.3. Log probabilities of randomly observing greater than the number of interacting proteins observed in the bins of figure 6.2.

mislabeled completely randomly. However, there is in fact at least one known systematic bias in the data used for this study. The Gene Ontology contains many more annotations for genes that are known to be related to heavily researched topics than for genes related to less interesting biological functions or processes [104]. Might there be a latent variable that captures this notion of "interestingness?" Is the absence of a latent variable or the presence of sufficient information within our data suggested by the evaluation metrics? Does the lack of reliable improvement as more mislabeled interactions were corrected suggest that the proportion of mislabeled instances does not affect the classifier or does the slight improvement at the more recent BIOGRID version indicate that there is some important threshold? There may be specific answers to these questions for this data set, but we attempt to answer these questions more generally in the following sections.

6.3 Generalizing the problem

Many of the questions brought up by the case study concern whether the mislabeled positive class instances are mislabeled systematically. In real world problems we often know that there is bias in the data but we do not know what kind of bias exists. In biology, there is a bias in the well studied protein interactions that is related to how interesting the protein's function is. As a result, the poorly understood proteins may be poorly characterized in the data, confounding attempts at classification. In such cases we often know *that* instances may be mislabeled, but are unable to ascertain *how* the data is mislabeled. Bias in the data may be systematic or random. Furthermore, it may be expressed as mislabeled instances or missing data. While bias in the data is a commonly studied problem in the literature, the focus has been on learning in biased data sets [61, 157]. It is equally important to study the effect of bias on the performance metrics used to evaluate the performance of learning algorithms.

Generally speaking, data can be missing in three ways, mirroring the missingness mechanisms set out in Allison et al.: missing at random (MAR) when values are missing in a way that is explained within the data; missing not at random (MNAR) when values are missing in a way that could be explained by a latent variable to which a learner does not have access; and missing completely at random (MCAR) when values are missing and there is no variable, latent or observed, that explains the missing values [2]. In this work we consider an analogous problem in which the bias takes the form of mislabeled instances in the data rather than missing instances. We term these cases BAR, BCAR, and BNAR for this type of bias. These three cases may have marked effects on the evaluation of classifier performance.

In a typical supervised learning scenario, classifiers can be trained and ranked by any of a large number of evaluation metrics [48]. This situation is complicated by the presence of bias in the data. Not only can different evaluation metrics give conflicting rankings, but they may react to the presence of different types of bias in different ways. We focus on the AUROC and the AUPR. These measures are commonly used as a single representative number to describe classifier performance. AUROC and AUPR have been studied in the context of class imbalance and in comparison to each other [58, 34]. However, AUROC and AUPR have not been studied in the context of mislabeled bias. A primary goal of this work is to empirically determine how these metrics behave in the presence of different types and levels of bias.

6.4 Systematic Bias in Class Labels

We consider class labels to be *poorly defined* if one class (which we will call the *positive* class) contains only correctly labeled instances, whereas the other class (which we will call the *negative* class) contains both correctly and incorrectly labeled instances.

While many data sets can be considered poorly defined, the underlying cause can vary greatly from data set to data set. In particular, depending on how the data is collected, different types of biases may be injected into the mislabeling of instances in the data set (i.e., a positive class instance may not have a completely random chance of being mislabeled). Therefore, in this section we discuss the various types of biases that can be found in real world data sets, and the way in which we simulate each of the types of bias. Note that in each of the bias injection mechanisms, only one class (the positive class) can have its labels flipped.

6.4.1 Injecting bias

We modeled each type of bias by injecting it into data sets. This approach may compound existing bias in the data sets, but our assumption is that the data sets are correctly labeled.

Completely random bias (BCAR) was injected into data sets by changing the label of positive class instances uniformly at random. We injected random bias (BAR) into data sets by sorting the data by a single feature and flipping the class label of the first X% of the positive class instances. Data sets were made to be biased not at random (BNAR) by sorting the instances by a single feature, flipping the class label of the first X% of the positive class instances, and removing the feature that was used to sort the data.

In order to isolate the effect of correlated features on the bias, we injected bias into data sets based on the most independent feature f as defined in Equation 6.1.

$$f = \arg\min_i \sum_{j \in X, i \neq j} |corr(X_i, X_j)| \tag{6.1}$$

This equation minimizes the absolute value of the correlation between each pair of features, where X is the set of feature vectors and $corr(X_i, X_j)$ is the Pearson correlation coefficient computed between features i and j.

6.5 Experimental Design

It is difficult to separate the behavior of an evaluation metric from specific classifiers. To approach this problem we observe how AUROC and AUPR behave over multiple classifiers trained on the same data sets. To preserve the validity of comparisons, we trained classifiers on the same folds with the same randomly permuted data with precisely the same biased instances.

In order to highlight differences between the two evaluation metrics, we measure both using the true class labels and the flipped class labels. This allows us to measure the AUROC and AUPR under two common scenarios in the practice of data mining: one in which classifiers are trained on data with an unknown bias and one in which classifiers are trained on data with an unknown bias but true class labels are discovered afterwards. By observing how the AUROC and AUPR behave under these two scenarios we may be able to identify important differences in how these performance metrics behave in the presence of bias.

We simulate these two scenarios by measuring the AUROC and AUPR with the flipped class labels (the first scenario) and the true class labels (the second scenario). Classifiers were trained on data with varying levels of bias. We used the probability estimates output by classifiers to rank the instances. We then used the ranking and the biased class labels to calculate the "bias class" AUC and the true class labels to calculate the "true class" AUC. This enables us to measure the effects of bias on the performance measures, and how robust each of the metrics and classifiers are to varying degrees of bias. If the performance on the "true" labels is much worse than that of the performance on the "biased" labels, the classifier metric combination is not effective at ascertaining the true performance of the classifier on the problem. Similarly, if the "true" class performance is much better than the "biased" class performance then the metric is overly pessimistic, and not suitable for cases where noise in one class label is prevalent.

6.5.1 Evaluation Metrics

ROC curves compare the true positive rate and the false positive rate while precision-recall curves compare the precision to recall (or true positive rate). ROC curves measure the "completeness" of predictions as the amount of false positives increases while precision-recall curves measure the "purity" of predictions as the amount of captured true positives increases. This difference underlies some of the observed strengths and weaknesses of using the area under both types of curve.

AUROC can be overly optimistic in cases of imbalanced data while making fewer assumptions about misclassification costs than other metrics such as accuracy [41, 115]. This makes sense in the context of viewing ROC as a measurement of "completeness", as a model may have a low precision but a high recall. AUPR has been used to overcome this concern in highly skewed data sets [84, 85]. It has been shown that AUPR and AUROC can give conflicting rankings for different classifiers trained on the same data [34]. We will demonstrate that this occurs across data sets and at various levels of bias.

6.5.2 Classifiers

To minimize the likelihood of sampling error, we trained classifiers on 100 random permutations of each data set in Table 6.1 using 10-fold cross validation. Classifiers included C4.5 trees (C4.5), NaïveBayes (NB), 5-nearest neighbors (NN), support vector machines (SVM), and Multilayer Perceptrons (MLP). We used unpruned and uncollapsed C4.5 trees with Laplace smoothing at the leaves. These are common parameters for C4.5 when used in imbalanced problems [24]. Unspecified parameters remained as their default in WEKA [63]. These algorithms were chosen to provide a range of classification approaches. AUROC and AUPR calculations were averaged across folds and permutations of the data.

6.5.3 Data sets

We selected 27 real data sets from the UCI repository, and 1 artificial data set [8]. The real data sets were selected to maximize diversity, allowing us to draw conclusions based on a wide range of evidence. These data sets were considered ground truth data, with accurately labeled instances, thereby allowing us to construct the "true" baseline performance. Regardless of the accuracy of this assumption, the availability of the original class labels allows us to calculate performance metrics with both true and biased data. Combined with the injection of different types of bias, this allows us to evaluate the stability of performance metrics. All data sets are listed in Table 6.1.

6.6 Results

In order to determine how AUROC and AUPR behave under different levels and types of bias, we used signed rank tests to evaluate the hypothesis that the mean rank of a classifier as given by the true class AUC was less than or equal to the mean rank of the classifier as given by the bias class AUC. Tied ranks corresponded to data sets. This test was done for each classifier and with each type and level of bias. Significant values indicate that the bias class AUC overestimates performance. We also tested the opposite hypothesis, that the mean rank of a classifier as given by the true class AUC was greater than or equal to the mean rank of the classifier as given by the bias class AUC. This hypothesis corresponds to the bias class AUC underestimating performance. P-values shown in Table 6.2a and Table 6.2b reflect tests of the first hypothesis and numbers in bold indicate significance at a level of $\alpha = 0.01$ for either test. Values in bold that are greater than 0.01 indicate that the second hypothesis was rejected.

Most of the significant differences occur in data that is BAR but some are present in BNAR data sets. Some differences are consistent between BAR and BNAR for C4.5, NB, and NN in both Table 6.2a and Table 6.2b. Comparing the two tables, we see that the bias class AUROC for C4.5 classifiers tends to overestimate performance but the bias class AUPR underestimates performance. NB classifiers show the opposite trend, where the bias class AUROC underestimates performance but the bias class AUPR overestimates performance. It is interesting to note this statistically significant difference in light of the fact that the AUROC and AUPR both overestimated classifier performance in the case study.

6.7 Comparison of AUROC and AUPR across data sets

In the previous section, we considered how AUROC and AUPR respond to different levels and types of bias. We now compare the reliability of AUROC and AUPR in ranking classifiers with the same level and type of bias. We used signed rank tests to evaluate the hypothesis that the mean rank of a classifier as given by the true class AUROC was less than or equal to the mean rank of the classifier as given by the true class AUPR. Again, tied ranks corresponded to data sets. The test was done for each classifier and with each type and level of bias. Significant values indicate that the classifier's mean ranking across data sets according to AUROC was greater than the mean ranking across data sets according to AUPR. We also tested the opposite hypothesis, that the mean rank of a classifier as given by the true class AUROC was greater than or equal to the mean rank of the classifier as given by the true class AUPR. This hypothesis corresponds to the ranking of a classifier by true class AUROC being less than its ranking by true class AUPR. P-values shown in Table 6.3 reflect tests of the first hypothesis and numbers in bold indicate significance at a level of $\alpha = 0.01$ for either test. Values in bold that are greater than 0.01 indicate that the second hypothesis was rejected.

Significant changes in the p-values in Table 6.3 indicate that the ranking of a classifier according to the AUROC and AUPR is different. The more significant p-values correspond to more reliably different rankings across the data sets. The rank of C4.5 trees according to the true class AUPR tends to be greater than the rank according to the true class AUROC in BCAR and BAR data sets and to a lesser extent, BNAR data sets. The rank of NB classifiers according to the true class AUROC tends to be greater than the rank according to the true class AUPR in BAR data sets and to a lesser extent in BNAR data sets. It is interesting to note that C4.5 and NB were also the classifiers that had the most significant values in Table 6.2a and Table 6.2b. The AUROC and AUPR both overestimated performance in the case study. The results in this table further confirm that the agreement between the AUROC and AUPR in the case study is unusual.

6.8 Case study revisited

Now that we have observed how AUROC and AUPR behave with a variety of classifiers trained on data with different systematic biases and different levels of bias, we can make better informed conclusions about where to look for bias and what type of bias to expect. These observations may guide us to improve the performance of classifiers on this data.

It is important to note that the ranking rewarded by AUROC and AUPR are different. The fact that both overestimate classifier performance in the case study indicates that the ranking is neither optimizing completeness nor precision in the mislabeled positive class instances. Recall that there is a known bias in the GO feature related to how interesting researchers find particular genes or functions. Given the behavior of AUROC and AUPR for NB classifiers in Tables 6.2 and 6.3, if the bias in the data were BAR we would expect the AUROC and AUPR to under and overestimate classifier performance respectively. However, both AUROC and AUPR overestimated performance in Figure 6.1. This suggests a few possibilities. First, the data may not be BAR. This is strongly suggested by the results in Tables 6.2-6.3 and by our use of a reduced set of GO terms. Second, there may be a latent variable, either "interestingness" of particular proteins to researchers or something else that could provide the classifier vital information to improve the ranking. This may be further suggested by the middle mode in Figure 6.2 and its correspondence with additional mislabeled instances in Figure 6.3. Third, and most likely of all, there may be a combination of systematic biases in the data. Each feature was drawn from data gathered through experiments with their own biases and may combine to create data that seems BCAR. From this analysis, we can conclude first, that the data is not simply BCAR, and second, that the first place to start looking for additional features that explain the mislabeled positive class instances is the middle mode in Figure 6.2.

6.9 Discussion

An understanding of the strengths and limitations of evaluation metrics can allow us to use and interpret them more effectively. Knowing the expected behavior of a performance metric under specific conditions can facilitate the detection of anomalous behavior and help to more accurately measure performance. The expected behavior of any combination of evaluation metric and classifier does not mean the same behavior will be observed on a specific data set. However, it can be used to guide further investigation and identify potential sources of systematic bias.

The approach taken in this study can be used more generally as a framework to approach the analysis of data with a poorly defined negative class. If researchers have access to a data set with incremental updates as we did in our case study, then the ideas of “true class” and “bias class” can be used to make an educated guess about what kind of bias is being added to the data set or whether multiple sources of bias may be present. Additionally, the use of multiple evaluation metrics helped to identify anomalous behavior while at the same time their agreement in our case study allows us to more confidently assess the usefulness in the ranking of false negatives. Each figure gave us further insight into the data. Namely, how the evaluation metrics were over or underestimating performance (Figure 6.1) how the classifier grouped the data (Figure 6.2) and how informative the ranking was about mislabeled positive class instances (Figure 6.3.

In this work we sought to address the question “How reliable are evaluation metrics when the negative class contains an unknown proportion of mislabeled positive class instances?" We discovered that there is much that we can uncover about the nature of bias in the data and the reliability of our evaluation. We addressed two key questions in our investigation. First, "how do AUROC and AUPR behave under varying levels of bias in the data set?" Our experiments show that the trend to over or underestimate classifier performance (Table 6.2a and Table 6.2b) is fairly stable across levels of bias. A second question addressed is, "What is the effect of different types of bias in the data on AUROC and AUPR?" Table 6.2a and Table 6.2b indicate that the type of bias does have an effect on whether the class AUROC and class AUPR tend to under or overestimate the performance of NB and C4.5 classifiers. Of course, it is difficult to observe the behavior of an evaluation metric outside of the context of classifiers. Indeed, we found that different combinations of classifier and evaluation metric have discernibly different behaviors.

One concern that arose while studying how the amount of mislabeled data affects evaluation was that the class imbalance rose with the proportion of mislabeled instances. A data set with evenly balanced classes would end up with a 19:1 class imbalance ratio when 90% of the class labels were flipped. The added effects of the imbalance problem could have a confounding effect on the evaluation metrics. Even so, NB, a skew insensitive classifier, was one of the few classifiers that significantly differently ranked by AUROC and AUPR. Regardless, because we observed changes in AUROC and AUPR across all proportions of mislabeled instances, we feel that the effect of the class imbalance problem is controlled in our experiments.

This study relied on an idealized scenario in which only one type of bias affected a data set at a time through a single feature. The combinatorial problem of applying each type of bias to each feature was prohibitive both in terms of time as well as complexity of analysis. However, we showed that in many data sets, even if data is mislabeled with respect to the least dependent feature—the best case scenario for bias—AUROC and AUPR can over or underestimate classifier performance.

We focused on AUROC and AUPR, but it is reasonable to expect still more different behaviors from additional evaluation metrics. We intend to extend this work to cover additional metrics, as well as a more thorough evaluation of the class imbalance problem in this context. One future direction might be to investigate the use of combinations of evaluation metrics to overcome individual biases. Perhaps the tendency of AUROC to overestimate performance and the tendency for AUPR to underestimate performance for C4.5 (and the opposite tendencies for NB) can be used together to get a measure that is more robust to mislabeled instances. On the other hand, variations of the parameters for these two classifiers may alleviate the problem.

TABLE 6.1

DATA SETS USED IN THIS STUDY.

Name	Features	Feature type	Instances
letter	16	continuous	20000
ism	6	continuous	11180
page	10	continuous	5473
estate	12	continuous	5322
krkp	36	discrete	3196
hypo	25	mixed	3163
SVMguide1	4	continuous	3089
segment	19	continuous	2310
artificial	8	continuous	2000
splice	60	continuous	1000
tic-tac-toe	9	discrete	958
oil	49	continuous	937
pima	7	continuous	768
breast-w	9	continuous	699
credit-a	15	mixed	690
crx	15	mixed	690
vote	16	discrete	435
vote1	15	discrete	435
horse-colic	22	mixed	368
ion	34	continuous	351
bupa	6	continuous	345
heart-c	12	mixed	303
threenorm	19	continuous	300
twonorm	20	continuous	300
heart-h	13	mixed	294
breast-y	9	mixed	286
sonar	59	continuous	208
heart-v	13	mixed	200

TABLE 6.2

TRUE CLASS VERSUS BIAS CLASS AUC.

(a) True class AUROC versus bias class AUROC. Signed rank tests compared the rank of classifiers across data sets to determine if the mean rank given by the true class AUROC was less than or equal to the mean rank given by the bias class AUROC.

True class AUROC versus bias class AUROC											
Bias	Classifier	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
BCAR	C4.5	1.000	0.977	0.386	0.681	0.986	0.682	0.293	0.212	0.074	0.120
	MLP	1.000	1.000	0.681	0.807	0.044	0.386	0.807	0.981	0.978	0.681
	NB	1.000	1.000	0.681	0.977	0.977	0.977	0.825	0.117	0.963	0.979
	NN	1.000	0.977	0.681	0.074	0.579	0.383	0.425	0.579	0.579	0.960
	SVM	1.000	0.173	0.977	0.977	0.579	1.000	1.000	0.500	1.000	0.049
BAR	C4.5	1.000	0.026	**0.003**	**0.000**	**0.000**	**0.000**	**0.001**	**8e-05**	**0.000**	**8e-05**
	MLP	1.000	0.033	0.021	**0.004**	0.028	0.035	0.015	0.559	0.822	0.740
	NB	1.000	0.982	**0.999**	**1.000**	**0.999**	**0.999**	**0.998**	0.991	**0.997**	**0.992**
	NN	1.000	0.932	0.426	0.911	0.986	**0.999**	0.987	**0.999**	0.975	0.719
	SVM	1.000	0.977	0.978	0.991	0.975	0.956	**0.995**	0.954	0.912	0.918
BNAR	C4.5	1.000	0.388	0.579	0.152	0.133	0.196	0.297	0.755	0.951	0.519
	MLP	1.000	0.681	0.330	0.027	**0.003**	**0.009**	0.014	0.087	0.138	0.784
	NB	1.000	0.970	0.936	0.974	**0.998**	**0.998**	**0.997**	0.920	0.836	0.943
	NN	1.000	0.286	0.283	0.548	0.666	0.813	0.500	0.696	0.529	0.529
	SVM	1.000	0.977	0.500	0.579	0.500	0.087	0.500	0.173	0.060	0.153

(b) True class AUPR versus bias class AUPR. Signed rank tests compared the rank of classifiers across data sets to determine if the mean rank given by the true class AUPR was less than or equal to the mean rank given by the bias class AUPR.

True class AUPR versus bias class AUPR											
Bias	Classifier	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
BCAR	C4.5	1.000	0.273	0.536	0.500	0.029	0.586	0.623	0.099	0.370	0.777
	MLP	1.000	0.035	0.546	0.304	0.932	0.372	0.793	0.589	0.537	0.682
	NB	1.000	0.133	0.060	0.120	0.286	0.867	0.286	0.536	0.030	0.021
	NN	1.000	0.967	0.669	0.931	0.396	0.010	**0.003**	**0.006**	0.039	0.231
	SVM	1.000	0.931	0.809	0.802	0.870	0.972	0.985	0.990	0.991	0.972
BAR	C4.5	1.000	0.952	**0.996**	**1.000**	**1.000**	**1.000**	**0.998**	**1.000**	**0.998**	**0.995**
	MLP	1.000	0.812	0.992	**0.994**	0.982	0.625	0.749	0.625	0.571	0.401
	NB	1.000	**0.003**	**0.001**	**0.000**	**0.000**	**0.001**	**0.002**	0.054	0.122	0.018
	NN	1.000	0.762	0.606	0.323	0.151	0.025	**0.003**	**0.005**	**0.001**	0.416
	SVM	1.000	0.204	0.627	0.404	0.518	0.580	0.658	0.102	0.292	0.187
BNAR	C4.5	1.000	0.647	0.897	0.792	0.860	0.853	0.329	0.240	0.554	0.918
	MLP	1.000	0.637	0.964	0.810	0.500	0.841	0.970	0.988	0.837	0.935
	NB	1.000	**0.007**	0.015	0.015	**0.006**	**0.004**	0.017	0.040	0.018	**0.008**
	NN	1.000	0.986	0.585	0.156	0.314	0.095	0.076	0.445	0.663	0.750
	SVM	1.000	0.411	0.420	0.981	**0.994**	0.980	0.993	0.963	0.862	0.802

TABLE 6.3

SIGNED RANK TESTS COMPARING CLASSIFIER RANK.

True class AUROC versus true class AUPR											
Bias	Classifier	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
BCAR	C4.5	0.908	0.983	**0.995**	0.948	0.977	**0.994**	**0.998**	**0.994**	**0.998**	0.964
	MLP	0.790	0.401	0.388	0.725	0.187	**0.010**	0.411	0.246	0.027	0.637
	NB	0.425	0.461	0.609	0.630	0.586	0.425	0.430	0.962	0.543	0.069
	NN	0.521	0.722	0.596	0.500	0.691	0.226	0.023	0.058	0.412	0.187
	SVM	0.234	0.076	0.036	0.139	0.102	0.102	0.016	0.170	0.087	0.864
BAR	C4.5	0.908	0.990	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**
	MLP	0.790	0.790	0.871	0.982	0.973	0.953	0.987	0.564	0.384	0.267
	NB	0.425	0.045	**0.004**	**0.001**	**0.001**	**0.001**	**0.000**	**0.006**	**0.005**	0.030
	NN	0.521	0.733	0.416	0.054	0.040	**0.002**	0.014	**0.001**	**0.009**	0.196
	SVM	0.234	0.084	0.417	0.173	0.069	0.037	0.011	0.252	0.166	0.084
BNAR	C4.5	0.710	0.942	0.992	**0.999**	**0.996**	0.983	0.951	0.796	0.132	0.519
	MLP	0.884	0.596	0.630	0.950	0.991	0.985	0.991	0.848	0.834	0.536
	NB	0.427	0.172	0.163	0.057	**0.005**	**0.006**	0.022	0.152	0.065	0.241
	NN	0.658	0.962	0.673	0.066	0.513	0.168	0.072	0.124	0.517	0.304
	SVM	0.171	0.039	0.075	0.118	0.098	0.608	0.294	0.658	0.944	0.730

Signed rank tests compared the rank of classifiers across data sets to determine if the mean rank given by the true class AUROC was less than or equal to the mean rank given by the true class AUPR.

CHAPTER 7

CURRENT INTEGRATIVE NETWORK APPROACHES

Network approaches utilize a diverse set of techniques including as components elements of statistical and machine learning techniques. We discuss many of these techniques in order to give context for the problem setting, the current state of the field, and the studies in following chapters. This chapter sets the stage for the further investigation of our primary thesis. It additionally gives context for a second argument in this dissertation: that multiple measures should be used in addition to heterogeneous data sets. Many recently developed techniques utilize heterogeneous data but rely on a single measure or algorithm to identify relationships in the data. The complexity of these approaches and the diversity in the underlying techniques for measuring relationships in data underscore the importance of our focus in previous chapters on how relationships in networks are measured and utilized.

7.1 Introduction

A goal of systems biology is to gain a more complete understanding of biological systems by viewing all of their components and the interactions between them simultaneously. Until recently, the most complete global view of a biological system was through the use of gene expression or protein-protein interaction data. With the increasing number of high-throughput technologies for measuring genomic, proteomic, and metabolomic data, scientists now have the opportunity to create complex network-based models for drug discovery, protein function annotation, and many other problems. Each technology used to measure a biological system inherently presents a limited view of the system. However, the combination of multiple technologies can provide a more complete picture. Much recent work has studied integrating these heterogeneous data types into single networks. Here we provide a survey of integrative network-based approaches to problems in systems biology. We focus on describing the variety of algorithms used in integrative network inference. Ultimately, the survey of current approaches leads us to the conclusion that there is an urgent need for a standard set of evaluation metrics and data sets in this field.

The history of genetics has been a process of uncovering increasing amounts of complexity and depth in biological systems. In the past, we knew that DNA was transcribed into RNA and then translated to proteins. Our growing knowledge of alternative splicing and other post-transcriptional regulation complicated this view. We knew that transcription factors were the primary regulators of gene expression. This view became complicated by our increasing knowledge of the regulating effect of phosphorylation on transcription factors. Given the complexity of biological systems and the certain knowledge that we do not fully understand fundamental aspects of biology, it is important to carefully consider how prior knowledge and diverse data types are incorporated into computational models.

As we learn more about genetics, it is becoming increasingly clear that the traits and behaviors of organisms are emergent: they are the product of complex interactions between numerous biological components. In systems biology, networks are used to capture this complexity by modeling an entire biological system. This approach gives scientists a global view of a biological system that can enable further understanding of the nature of human disease as well as new tools to understand the processes driving life [122].

Networks are a versatile tool that have been used to model interactions between numerous different biological concepts. Nodes can be used to represent genes, proteins, metabolites, or any other discrete biological component or concept. Edges in the network may represent the relationship between a gene and a protein, similarity of function between genes, or any other pair of biological concepts. Edges may represent multiple types of relationships simultaneously. Each type of relationship reveals unique information about an organism. For example, protein-protein interaction (PPI) data reveals which proteins can physically interact, but alone it does not impart knowledge about how an organism will react to stimuli. Similarly, gene expression data can reveal how an organism responds to stimuli in terms of the amount of RNA produced but it does not impart any knowledge about the physical mechanisms that cause change in the organism's behavior. Therefore, the key to furthering our understanding of biological systems the integration of diverse data types.

Differences in the underlying architecture of networks can affect their utility. Directed networks such as Bayesian networks or networks that use asymmetric edge weighting metrics implicitly contain some indication of causality [53, 119]. These methods are well suited for making specific inferences about how the effects of a perturbation to one or more genes will propagate through the network. Undirected networks make fewer assumptions about how nodes are connected and are often less computationally demanding to construct but may yield less specific information.

7.1.1 Contributions

There are a number of review articles that cover network inference. De Smet et al. reviews network inference and integrative methods in the context of how they approach the problem of underdetermination [36]. Sharan et al. reviews several integrative network methods in the context of clustering [127]. Hecker et al. covers network models for time course behavior of gene expression data and integration of heterogeneous data sources. They discuss a wide range of network inference algorithms both within and outside of the context of integrative approaches. They cover the inclusion of previous biological knowledge such as expected network topology.

In terms of the integration heterogeneous data types, they primarily cover Bayesian networks [68]. Gitter et al. survey a number of approaches to integrate time series data with various heterogeneous data types gathered from single time points to create dynamic regulatory networks [57]. Califano et al. review a number of integrative networks approaches in terms of the combinations of data used [14]. They describe how different approaches use different combinations of data types to uncover specific relationships in the data. They also address the need for more focus on awareness of context specific regulation in network models. Bebek et al. focus on integrative approaches specifically used for the identification of biomarkers and the betterment of clinical science [7]. Here, we focus on presenting a wide range of integrative model types and exclusively on the integration of heterogeneous data. Our purpose is to provide a familiarity with the variety of algorithms used for integration in network models.

In Section 7.2 we discuss some of the most commonly used data types in integrative network models. Section 7.3 covers the problem of network inference in the abstract. In Sections 7.3.1 through 7.3.4 we discuss various approaches to network inference and cover examples of each in some detail. Section 7.3.1 and 7.3.2 cover Bayesian and other probabilistic networks. Section 7.3.3 discusses integration methods based on machine learning techniques. In Section 7.3.4 we cover techniques that rely on identifying modules in networks and context specific regulatory patterns. Finally, in Section 7.4 we discuss some of the patterns that emerge from examining the variety of methods discussed in the previous sections. We conclude that there is an urgent need for consensus about how to evaluate and compare models.

7.2 Data

Each data type is collected in a unique way. Additionally, two data sets describing the same type of data may not be comparable due to differences in scale, noisy data, or measurement errors. Therefore, normalization and the use of well curated data are essential for meaningful comparisons between data sets and the integration of diverse data types. For example, microarray results have been shown to vary based on the location of probes on the chip, complicating comparisons between results gathered with different chips [15]. Even assuming identical chips, different approaches to normalization can have significant impact on the meaning of the data and on the validity of comparisons between data sets [117]. Another complication is that there are multiple methods to measure the same data type. There are many distinct methods to measure PPI, each with different strengths and weaknesses [22]. For example, Affinity Capture-MS protein interactions are determined by using a "bait" protein that is "captured" by a polyclonal antibody or an epitope tag. The associated partner is then identified by mass spectrometry. An alternative approach, co-immunoprecipitation, isolates a protein with antibodies. Interacting partner proteins are then detected with western blotting. Different methodologies in data collection add noise or bias in different ways that must be accounted for in analysis.

Precursors to integrative networks used microarray expression data alone to infer regulatory and other types of relationships between genes. Microarrays enable high-throughput measurement of the expression level of genes. Expression levels measure the relative amount of RNA produced from the transcription of genes. RNA levels give some indication about the amount of protein that is expected to be produced. Since proteins are the primary causes of change in a cell, expression data can give an indirect evidence towards answering many different questions in systems biology. Many studies have relied on clustering and network models to identify functionally similar genes or infer regulatory networks based on expression data [65, 30, 42, 96, 100].

Protein-protein interactions provide more direct information in the form of which proteins physically interact. Like expression data, PPI data is commonly used to cluster genes or proteins or to infer networks in order to identify novel interactions or determine function [141, 127].

Some data types are themselves integrative. The ChIP-chip technique combines microarrays with chromatin immunoprecipitation to allow the identification of protein binding sites on DNA [3]. This is particularly useful for the study of transcription factors (TFs) which are proteins that transcribe DNA into RNA and are thought to play a major role in the regulation of gene expression. Motifs or identifiable strings of DNA can also be located computationally from sequence data to identify potential transcription factor binding sites (TFBS). Expression quantitative trait loci (eQTLs) use genetic variation between individuals in combination with gene expression data to measure the association between expression levels and genotypes. An expression trait refers to the amount of RNA produced by a gene. Each eQTL represents a strong association between a position or locus in the genome and the expression level of a gene. eQTLs describe the relationship between genotype and phenotype and enable inferences about the regulatory interactions between genes [79].

Annotation data can come from many sources and can describe experimentally or computationally derived knowledge such as functions associated with biological components or pathways that components are a part of. The Gene Ontology (GO) keeps curated functional annotations for genes [4]. Annotations exist in a hierarchy such that a gene my have a number of general and specific functions. Pathway information describes the chain of biological components involved in causing some event or fulfilling some function in a cell. Many databases exist to curate pathway and other data types, often for specific organisms [22, 81, 82].

One consideration that affects many data types is the experimental conditions under which measurements are taken. For example, the expression level of genes can change drastically based on environmental and genetic conditions [79, 74]. Common genetic conditions include gene knockout experiments, in which a gene is made inoperative, and chemical or environmental treatments. Measurements may also involve an element of time under a condition or after a treatment.

7.3 Network Inference

The basic problem of network inference is to create a network that has a meaningful topology. Ultimately this means creating a sparse network in which only important edges are present. This is accomplished in various ways by different algorithms. In the abstract, there are two general types of networks: distance-based networks and probabilistic networks.

Network inference algorithms universally depend on some measure of dependence or distance between biological components. The approach used to calculate edge weight can have a significant effect on what is contained in the resulting network [136, 35]. Mason et al. compared co-expression networks based on Pearson's correlation to co-expression networks based on the absolute value of Pearson's correlation and showed that modules in the signed network are more biologically coherent [98]. Probabilistic network inference faces a similar problem in that conditional probabilities can be calculated in a number of different ways.

The fundamental assumption in relevance networks and other distance-based networks is that relationships between biological components can be accurately ranked in some meaningful way. Once the relationships between all components have been quantified, edges are removed from the network. This results in a sparse network with some meaningful topology that is determined in part by the edge weighting method and in part by the pruning criterion. Each approach makes different underlying assumptions that can impact the information contained in the network. Relevance networks make inherent assumptions in the choice of weighting method and the pruning approach. The underlying assumption is that the weighting method correctly ranks edges in terms of importance. Zhou et al. use Pearson's correlation to infer a co-expression network for yeast [158]. They use the shortest paths between all nodes in the network to identify functionally related genes. This approach assumes that transitive relationships that are represented in the network may be as important to understand relationships between genes as direct relationships. ARACNE makes the opposite assumption and explicitly disallows triangles in the network, assuming that all triangles contain an indirect relationship that should not be explicitly represented in the network [96].

Additional approaches in this category use LASSO (Least Absolute Shrinkage and Selection Operator) or related linear methods to explicitly penalize and eliminate weak relationships [142]. LASSO and related approaches optimize the parameter vector of linear equations such that their ℓ_p norm is less than or equal to a given value. LASSO's constraint is based on the ℓ_1 norm whereas other approaches may use different norms or a combination of norms as is the case with elastic nets [88]. Such approaches are used to discover a sparse topology and replace an arbitrary threshold with a more principled one [51, 129, 62]. One recent related integrative approach uses Multi-Block Partial Least Squares (sMBPLS) to find sets of input variables from multiple data types including copy number variation, DNA methylation, and microRNA expression that together explain the gene expression in cancer data [90]. Wenyuan et al. use a tensor-based approach to identify sets of recurring subgraphs from large sets of heterogeneous biological networks [89]. This approach is similar to LASSO and similar approaches in that the sparseness of the resulting networks are controlled primarily through the choice of ℓ_p norm in the objective function.

Other approaches that can be described by the category of distance-based networks focus on machine learning techniques such as feature selection and decision trees. MRNET uses a maximum-relevance minimum-redundancy feature selection method to identify important neighbors for every node. After the pairwise mutual information between expression levels of all genes is calculated, edges are effectively pruned by the feature selection algorithm. For each node, the algorithm selects the neighbor with the highest mutual information that has the lowest redundancy with the neighbors already selected. Neighbor selection stops when the score of the next best neighbor is below a threshold [100].

Probabilistic or graphical models represent the dependence between random variables as nodes in a network. Edge weights represent conditional probabilities. This approach naturally captures the noise and stochastic nature of biological data.

7.3.1 Bayesian networks

Bayesian networks are one of the most commonly used methods of integrating diverse biological data types. They describe biological data as random variables. Using this approach, measurements of a gene's expression levels may be interpreted as samples from a random variable. Relationships in Bayesian networks are directed, reflecting the conditional dependence between variables. As such, they are often interpreted as causal. This interpretation allows Bayesian networks to represent pathways and to be used to predict the effect of perturbations to the system. Bayesian networks can be discrete, continuous, or a mixture of both.

Discrete Bayesian networks model the probability of discrete states. For example, an edge between nodes A and B can indicate the probability that gene B is highly expressed given the state of gene A. Discrete Bayesian networks may require that each node have a prior distribution to represent the possible prior states of the variable. A model relying only on the frequency of observed values may be unable to assign a probability to new observations if they do not fall within the observed range. Discrete Bayesian networks can model relationships in the data relatively concisely with a conditional probability table for each node that lists the probability of each state given the inputs. One drawback is that discretization of the data may lead to information loss. Bayesian networks that use continuous variables rely on conditional probability densities instead of conditional probability tables. Continuous variables may also be modeled using linear conditional densities, in which the conditional density of a node X is dependent on its parents as shown in Equation 7.1. The equation shows that the conditional density of X given its parents p is linearly dependent on the values of the parents. It is common to use a normal distribution in this approach. Continuous Bayesian networks do not lose information due to discretization but it is more computationally complex to infer the continuous model than the discrete model.

$$P(X|p_1, ..., p_n) = N(\beta_0 + \sum_{i}^{n} \beta_i * p_i, \sigma^2) \tag{7.1}$$

There are three major steps in Bayesian network inference. First, a structure must be proposed. Second, the parameters or probabilities associated with edges and nodes must be set. Third, networks must be evaluated to determine how well they model the data. These steps are commonly used iteratively to propose a structure and parameters, then evaluate it against further structural changes. This process allows a search through potential Bayesian network models.

Identifying edges in the network is a critical step in Bayesian network inference, as the direction of edges can greatly affect the interpretation of the model. The presence or lack of edges between nodes can also have a large effect as it determines the conditional relationships between variables. The most straightforward method to infer network structure would be to exhaustively compare every possible network. This approach is prohibitively expensive, as the number of possible networks grows super exponentially with the number of nodes [107]. Practical methods rely on sampling or heuristics to reduce the search space dramatically.

The sparse candidate algorithm relies on simple local statistics such as correlation to identify potential parents for each gene [53]. It greatly reduces the search space by evaluating edges only between a node and its candidate set. The algorithm can then use hill-climbing or a divide and conquer approach to determine edges. Choices made early in the assignment of edges can result in a restricted search space. Therefore, the algorithm iteratively creates a network then updates the candidate parent sets for each node by replacing nodes in node X's candidate set with a transitive relationship with nodes that had a weaker dependency with X.

Sampling methods such as the Metropolis-Hastings algorithm can be used to reduce computational cost of structure learning at the expense of an accurate description of the data. Sampling and other inexact techniques are often used repeatedly and then averaged to form a single network. Alternatively, one model or a few 'good' models can be selected as representative of all possible models. This process is called model selection when one network is chosen or selective model averaging if multiple representative networks are averaged [99].

Model parameters in Bayesian networks are conditional probability distributions or tables. A continuous node may assume that the observed data come from a normal distribution. However, the parameters of the distribution, the mean and standard deviation, may be incorrect. If the assumed distribution or prior is incorrect then the calculated probability of an observed instance and the fit of the network to the data will be incorrect. Parameter fitting is the process of calculating the priors and conditional probabilities in the network.

In Equation 7.2, D is the data, E is background knowledge, and θ is the model. $p(\theta|E)$ and $p(\theta|D, E)$ are the prior and posterior probability distributions for the model θ, respectively. The prior describes the agreement between the prior knowledge and the network. The posterior describes how well the model fits the observed data. We direct the reader to Heckerman and Needham et al. for a more thorough treatment of parameter fitting and the selection of priors [69, 107].

$$p(\theta|D, E) = \frac{p(\theta|E)p(D|\theta, E)}{p(D|E)} \tag{7.2}$$

There are two primary ways to include prior knowledge in Bayesian networks. The first is to constrain the edges in the structure learning step. This is a commonly used approach to integrate heterogeneous biological knowledge [159, 66]. The second is to update the priors in an iterative process. Often, a Bayesian network will be inferred and the parameters fitted to one type of biological knowledge, then priors are updated to take into account additional sources of data iteratively [139, 76].

Zhu et al. use a mixture of constrained and prior-updated techniques to integrate data types into a Bayesian network. They use the sparse candidate algorithm to infer structure in Bayesian networks based on only expression data, based on eQTL data, and based on expression data, eQTL data, TFBS, and PPI data [159]. For each network type, they learned 1000 networks and determined a consensus network that consisted of edges that were present in at least 30% of the networks. Loops were resolved by removing the weakest edge. Prior knowledge gained from eQTL data was incorporated by constraining edge direction such that genes with cis-acting eQTLs (as defined in Doss et. al. 2005) are considered as potential parent nodes for genes with trans-acting eQTLs in the same region of the genome [40]. Representative genes were used to incorporate TFBS and PPI data. They used a set of genes that were determined to be the most strongly associated with a transcription factor to represent each transcription factor in the network. The prior probability that the gene associated with a transcription factor is the parent of other genes that carry the TFBS was proportional to the number of expression traits correlated with the transcription factor's expression levels. The inferred networks were evaluated in terms of predicting functional categories from the Gene Ontology, predicting genes regulated by various transcription factors, and predicting the response of gene expression to gene knockout experiments.

7.3.2 Other probabilistic networks

While Bayesian networks are a popular approach to integrating diverse data types, there are many other network models that rely on a probabilistic interpretation of the data. As is the case with Bayesian networks, learning the structure of probabilistic models in general can be computationally prohibitive. Structure learning for probabilistic graphical models has been the subject of much recent research. Wainwright et al. use ℓ_1 regularized logistic regression to learn the structure of each node's neighborhood in a Markov network [150]. Other approaches make the structure learning problem tractable by restricting the model's structure. Choi et al. propose algorithms to learn tree-structured probabilistic models [20]. Srebro controls the tree-width (maximum clique size) of Markov networks in order to limit the computational cost of inferring network structure while providing a provable performance bound [134]. This is by no means an exhaustive list of approaches to infer probabilistic networks. Many approaches fit a prior distribution to the data in order to measure explanatory power. Friedman and Nachman use Gaussian processes to learn the structure of Bayesian networks [52]. Gaussian processes model the relationship between a set of variables and an output variable by defining a mean function and a covariance function over the random input variables. In this approach, response variables are modeled as mixtures of related Gaussians. In this framework, the structure of a candidate network can be evaluated by computing the marginal likelihood of the data given the structure.

Tu et al. use a stochastic network to integrate PPI, TFBS, phosphorylation, eQTL, and expression data in order to identify causal genes and regulatory pathways [143]. Their model works under the assumption that causal or regulating genes in the network regulate their targets through either direct or indirect effects on the activity of transcription factors. They take into account the possibility that transcription factors can be regulated at the protein level. They also make the common assumption that gene activity correlates with gene expression. Protein-protein interactions are represented in the network as undirected edges, protein phosphorylation and TFBS are represented as directed edges. Each node has a set of transcription factors that bind to it and a set of genes with eQTLs that are candidate regulators. For each node in the network they estimate the likelihood that every neighboring gene is the cause for its expression by calculating Pearson's correlation between the expression level of the two genes. The algorithm determines the causal regulator of gene G by taking random walks without cycles along the edges in the network until it reaches a candidate eQTL gene. They used this algorithm on subsets of expression data from specific treatments as well as with bootstrapped samples to observe variation in transcription factor activity and account for variation in expression levels. The method was evaluated by comparing predicted relationships against a compendium of gene knock-out expression data.

Lee et al. propose a method to represent functional associations between biological components. They use a Bayesian statistics approach to determine the likelihood that genes are functionally linked based on evidence from heterogeneous data sources [86]. They use microarray data, phylogenetic profiles, PPI, functional linkages from text mining, as well as four other data types. Their log-likelihood score compares the frequency of linkages in each data type between genes that share a pathway to the frequency of linkages between genes that do not share a pathway. In Equation 7.3, $P(L|E)$ is the frequency of linkages (L) in a data type (E) between genes in the same pathway, $\sim P(L\|E)$ is the frequency of linkages between genes in different pathways for the data type. P(L) and ∼ P(L) are the total frequency across data types of all linkages between genes sharing a pathway and not sharing a pathway, respectively.

$$LLS = \frac{P(L|E)/\sim P(L|E)}{P(L)/\sim P(L)} \tag{7.3}$$

This method relies on the use of the KEGG (Kyoto Encyclopedia of Genes and

Genomes) pathway and sub-cellular location data as ground truth data for the calculation of LLS [81]. The use of a common ground truth allows scores for different types of data to be meaningfully compared. The resulting integrative network showed improved accuracy in terms of linking genes that share pathways in the KEGG database over other methods.

Other methods integrate diverse data types and model the stochastic nature of biological systems use hidden Markov models, Markov networks, and naïve Bayes models [44, 38, 146].

7.3.3 Statistical and machine learning approaches

Machine learning and statistical approaches are distance based as many provide some confidence or probability that a prediction is correct. They tend to be different from other distance based methods in that the distances are often determined in a supervised manner.

SEREND is a semi-supervised network construction method that integrates TFBS, DNA sequence binding motifs, and gene expression data to predict transcription factor-gene interactions [45]. It uses a logistic regression classifier for expression data and sequence motif data, then combines the two in a hierarchical classification scheme by training a third logistic regression classifier on the output of the other two classifiers. Features for the classification of expression data were from 455 expression experiments from a compendium of treatment experiments. Each instance corresponded to a gene. Class labels were activated by a transcription factor, repressed by a transcription factor, or not regulated by a transcription factor. The motif classifier used only a single feature to classify genes as regulated by the transcription factor or not regulated by the transcription factor. If the meta-classifier found that there was enough evidence that a non-regulated gene was regulated by a transcription factor, then the algorithm would switch the label from not regulated to regulated and update the weights for all classifiers. This process allows SEREND to iteratively expand its predictions about transcription factor-gene relationships until they converge. SEREND was evaluated in terms of how well it recovered gene targets that were verified in a ChIP-chip data set.

Hwang et al. use a few statistical methods to combine p-values from different data sets [75]. They use an ensemble of Fisher's weighted F, Mudholkar-George's weighted T, and Liptak-Stouffer's weighted Z where the weight is a measure of the relative statistical power for each data set. They determine a combined weight by comparing a hypothetical weight distribution to an observed distribution. The resulting integrative network has a p-value for each node and edge that indicates the confidence that the node or edge belongs in the network. Multiple approaches were tested on simulated data sets, which allowed a comparison on the basis of ground truth data.

The modENCODE Consortium is a group that collects a great deal of diverse data about the model organism *Drosophila* [103]. They use correlated activity patterns from over 700 data sets to define a functional regulatory network. They use logistic regression to classify promoters as active or inactive based on chromatin modification, TFBS, and nucleosome physical properties. The resulting probabilities are used to weight the confidence of each regulatory edge in the network. They evaluated inferred networks based on the enrichment in the network compared to randomized networks of GO terms, correlation of gene expression across time, frequency of protein-protein interactions in the network, and other metrics.

The STRING (Search Tool for the Retrieval of Interacting Genes) database is a collection of data for the understanding of functional interactions among proteins [138]. Interactions in the database come from many curated data sets from multiple organisms as well as from text mining the literature, predicted interactions from gene co-expression and cross-genome homology. Each interaction in the database has a confidence score assigned to it based on benchmarks against a trusted PPI data source, the KEGG database. Each data source is individually benchmarked and then combined in a naïve Bayesian approach by simply multiplying the normalized scores together. Interactions with more support from multiple sources of data will naturally have a higher combined score. STRING is properly a search tool rather than an integrative network inference method. As such, it does not attempt to evaluate the resulting network but provides the ability to alter the data types included, as well as access the raw data.

An alternative approach to modeling heterogeneous data in a single network is to use multiple edge types in what is called a multi-relational network. Davis and Chawla use a multi-relational network approach to make predictions about disease occurrence in patients and study the relationship between diseases and genes [32]. They combine a network of disease co-morbidity data with a network of genes related to each other by their relationship to the same disease. They then use a link prediction method that uses a triad census (counting the occurrences of sets of three nodes with each possible combination of edges) as the basis to predict unknown genetic links. Predicted links were benchmarked against a number of canonical link prediction methods and performance was measured in terms of area under the ROC curve, and the precision-recall curve.

7.3.4 Modular networks and condition specific regulators

One of the fundamental problems in creating a network model for regulatory interactions in the genome is that the regulatory program of a cell appears to change under different conditions [124]. Network modules can be viewed as discrete groups composed of many types of molecules whose function is separable from other modules. The aggregate expression of these modules may have condition specific regulators. Integrative network approaches to modeling condition specific regulatory networks rely on compendiums of expression data from different experimental conditions and commonly use TFBS, ChIP-chip, or other protein-DNA interaction data [83, 55, 94].

SAMBA integrates heterogeneous data from gene expression, PPI, phenotypic sensitivity, and TFBS sources into a probabilistic bipartite network in order to identify genes with common behavior across experiments [140]. The nodes on one side of the network are genes and the other side are properties of genes or proteins. Weighted edges in the network between node N and property P are interpreted as the probability that node N has property P. Property nodes can indicate anything from interaction with a specific protein to different levels of discretized gene expression. Subgraphs are scored based on the log ratio of the observed topology under two statistical models, a model for the dependency expected in modules and a model for the background dependency. Biclustering is used to identify gene sets that share sets of properties. Modules are evaluated in terms of functional enrichment based on the Gene Ontology. It finds complete bipartite subgraphs with high density by using a hashing technique to find ’seed’ nodes and then using a local search to identify other nodes in the module.

DISTILLER is an integrative framework to identify condition-dependent modularity and regulatory relationships [87]. It uses an efficient item set mining algorithm to identify modules. It starts with “seed” modules, consisting of a small number of genes that are co-expressed in a sufficiently large number of conditions and share motifs for the same regulators. Seed modules are expanded to nodes that do not violate the module properties. A drawback of the item set mining approach is that it can be difficult to identify the most interesting modules from the large amount of potentially redundant output. DISTILLER ranks modules by a measure that takes into account how much they help to cover the entire condition space and their redundancy with already ranked modules. DISTILLER was evaluated in terms of precision and recall on a ChIP-chip gold standard data set.

7.4 Discussion

While there are many benefits to integrating diverse data types, integration of prior knowledge may reinforce bias in network models to the detriment of new discoveries. For example, a number of networks papers have observed that many biological networks appear to have scale-free topology [80, 148]. In response, methods to infer or evaluate networks based on their topology have been developed. Networks inferred using this criterion will systematically overlook possible networks with alternative architectures [156]. There is evidence that this may be happening as many of the observed scale-free topologies in biological networks may not truly be scale free. Clauset et al. showed that the methods used to measure scale-free topology in many preceding studies of biological networks were unable to distinguish between power-law distributions (such as scale-free) and a number of other distinct distributions [26]. Bias may also enter into models through other prior knowledge. For example, Zhu et al. and Tu et al. both constrain their models to use trans-acting eQTLs to constrain edges but the definition of trans acting is different [159, 143].

Network inference methods that are constrained to include edges from PPI, TFBS, eQTL, or other data may reinforce bias in the models as they do not allow room for error in the data. Less constrained approaches avoid this problem but may add a more subtle bias to the model. Many integrative network approaches construct a single network by integrating data based on a single algorithm [159, 66, 45]. As is the case with different types of data, different algorithms contain different biases. Bayesian approaches that create an ensemble or consensus model with Monte Carlo techniques may suffer less from this type of bias but may reduce bias further by use of fundamentally different algorithms.

The problem of evaluation is made extraordinarily difficult in systems biology by the scarcity of ground truth data. Even curated data sets such as PPI data from KEGG that are used to benchmark novel methods are based on uncertain data. The problem of network evaluation has been noted before in the single data type network inference problem [95]. Marbach et al. propose a unifying approach to the evaluation of network models that includes common evaluation metrics and simulated data. While these are excellent suggestions, the problem is made much more complicated by the diversity of data involved in integrative methods.

Any single type of data presents a one-dimensional view of a biological system. Therefore, evaluation based on a single data type may not be a baseline for the performance of an integrative method. Furthermore, different approaches tend to use different amounts and types of data, making the actual methods themselves very difficult to compare. There are, of course, high-confidence experimentally derived interactions, but it can be difficult to locate and identify them. Databases such as STRING, KEGG, and modENCODE will be critical for the future progress of integrative network models because they provide this service. The creation of a common body of data for evaluation and a standard for evaluation methods for integrative network approaches would allow integrative network algorithms to be truly compared. This in turn could help us to better understand the complex interplay of diverse data types.

To gain a real understanding of biological systems it is critical to understand not only the network inference methods but how they interact with specific data types and how choices of parameters in the algorithms, such as cutoffs for edge inclusion, edge weighting measure, or the type of biological experiment that produced the data all work together. Chapter 8 discusses these concerns and some of the assumptions inherent in many current network models.

CHAPTER 8

WHAT MAKES A GOOD NETWORK?

Having discussed many current approaches to modeling biological systems as networks in Chapter 7, we now focus on how the questions that we investigated in previous chapters still apply to these recent integrative models. Namely, can we improve these models by the use of targeted measures for relationships in the data, as we did in Chapter 3? The answer to this question depends on the kind of information captured by different data types and even individual data sets. Ultimately, we need to know whether models that integrate heterogeneous data through the use of a single measure are taking advantage of domain knowledge appropriately. In this chapter we show through the evaluation of simple network models that these are important concerns for the state of the art models discussed in Chapter 7 and indeed for data science in general.

8.1 Introduction

Networks provide an intuitive representation of complex biological data. However, a scientist interested in modeling, for example, gene expression data as a network is quickly confounded by the fundamental problem: how to construct the network? Of course, it is fairly easy to construct a network, but is it *the* network for the problem being considered? This is an important problem with three fundamental issues: How to weight edges in the network in order to capture actual biological interactions? What is the effect of the type of biological experiment used to collect the data from which the network is constructed? How to prune the weighted edges (or what cut-off to apply)? Differences in the construction of networks could lead to different biological interpretations.

Indeed, we find that there are statistically significant dissimilarities in the functional content as well as topology between co-expression networks constructed using different edge weighting methods, data types, and edge cut-offs. We show that different types of known interactions, such as those found through Affinity Capture-Luminescence or Synthetic Lethality experiments, appear in significantly varying amounts in networks constructed in different ways. Hence, we demonstrate that different biological questions may be answered by the different networks. Consequently, we posit that the approach taken to build a network can be matched to biological questions to get targeted answers. More study is required to understand the implications of different network inference approaches and to draw reliable conclusions from networks used in the field of systems biology.

High-throughput biological data such as protein-protein interactions (PPIs), gene expression profiles, and metabolic interactions contain information about how different components of a cell interact in concert and can be used, for example, to elucidate potential drug targets and to further our understanding of disease [149, 147]. Because biological data are generally noisy and expensive to obtain, modern "systems biology" has produced integrative analysis frameworks to help overcome the challenges and pitfalls of these important collections.

One particular analysis is network-driven, where genes can be nodes and specific—but different—approaches are often used to infer edges that predict relationships between them [105, 77]. For example, gene co-expression networks constructed with correlation-based measures have been used to identify transitive relationships [158], gene regulatory patterns [148], and biological modules [98]. Further, they have been successfully combined with transcription factor, eQTL, and PPI data into integrative Bayesian networks [159]. Gene expression networks are therefore one commonly used example of a complex model built from the high-throughput biological data collections.

Despite their importance in systems-level analysis, however, there is currently is no consensus about how to construct biological networks. Moreover, the link between a chosen network inference method to network topology and functional content (i.e., how to generate edges that encapsulate some biological knowledge) is not well understood. If the choice of a specific network construction has significant effects on the subsequent interpretation, biological advances will be significantly challenged unless done with considerable care.

To demonstrate this point, we investigate how variables in network construction affect the content of the resulting networks. Gene co-expression networks are commonly constructed as follows. Given a gene expression data set, the strength (or weight) of the co-expression relationship between each pair of genes can be quantified in a number of ways, each of which may be revealing different biological insights [120, 98]. Then, an arbitrary cut-off is chosen in order to include only the strongest edges into the network [13, 36, 97, 6]. If the edge strength is positively correlated with the underlying biological knowledge, then the choice of a cut-off can be a major factor in determining the network's ability to accurately uncover this knowledge.

Here we report significant differences in the biological information captured using two different types of gene expression data sets and different choices of network inference variables. This work shows that substantial care must accompany network analysis and is driven by the principle of induction: to conclude that a network construction variable has an effect, it is enough to find an example of two different choices for the variable (e.g., two different edge weighting measures) that result in significant differences; this constitutes a proof that the choice for that variable matters. And given one such example, it stands to reason that there may be other choices for the variable (e.g., other edge weighting measures) that also result in network differences. Because many other network inference methods rely on choosing an edge weighting method and a cut-off, this study also has broad implications.

8.2 Our approach

In brief, we studied the effects of different edge weighting methods, data types, and edge cut-offs on the functional content and topology of resulting networks.

Edge weighting methods. To maximize the significance of the observations derived from the experiments discussed in the next sections, we chose commonly used edge weighting methods: correlation-based measures (Pearson's correlation in particular) and mutual information (MI) (see Methods). Pearson's correlation is vulnerable to perturbation from outliers and is constrained in that it captures only linear relationships. Mutual information is the common alternative, and has been used extensively in relevant frameworks such as ARACNE [96], MRNET [100], and CLR [47].

Data types. Because not all genes are active at all times, under all conditions, or in all strains [79, 74], we studied the effect of the type of biological experiment used to collect gene expression data. We constructed networks from expression data obtained from two types of biological experiments: line cross [11] and treatment. In the line cross experiment, two strains of yeast (*Saccharomyces cerevisiae*) were crossed and microarray experiments were performed on the strains and their progeny. In the treatment experiment, the same species of yeast was subjected to different chemicals or mutations before having expression levels measured through microarray experiments [74] (see Methods). These are two very different types of experiments, which could affect the functional content and topologies of gene co-expression networks created from the data.

Edge cut-offs. For each combination of edge weighting method and data type we first constructed a network containing the top k strongest edges by varying k from 0% to 100% of the strongest edges (in increments of 6,000 edges). For our more detailed analyses, we focused on stronger edges by varying k from 2,500 to 75,000 of the strongest edges (in increments of 2,500 edges).

Evaluation. To evaluate how accurately a given co-expression network captured existing biological knowledge, we tested whether edges in the network corresponded to known interactions (e.g., PPIs), as well as whether genes that were connected by an edge in the network shared Gene Ontology annotations. This is a common approach to evaluation of biological networks [36, 145, 105, 4, 47, 95] We focused on yeast data types (see above) because yeast is among the most studied and best annotated species to date.

Summary. We show that different network inference strategies result in networks that may contain answers to distinct biological questions. These results and conclusions highlight challenges in network construction, along with their impact, which most urgently require the attention of the systems biology community.

8.3 Results

We constructed networks using each combination of edge weighting method, data type, and edge cut-off. We then compared networks of a given size constructed from the same data type but using different edge weighting methods. We additionally compared networks of a given size constructed using the same edge weighting method but from different data types. This was done for each of the edge cut-offs.

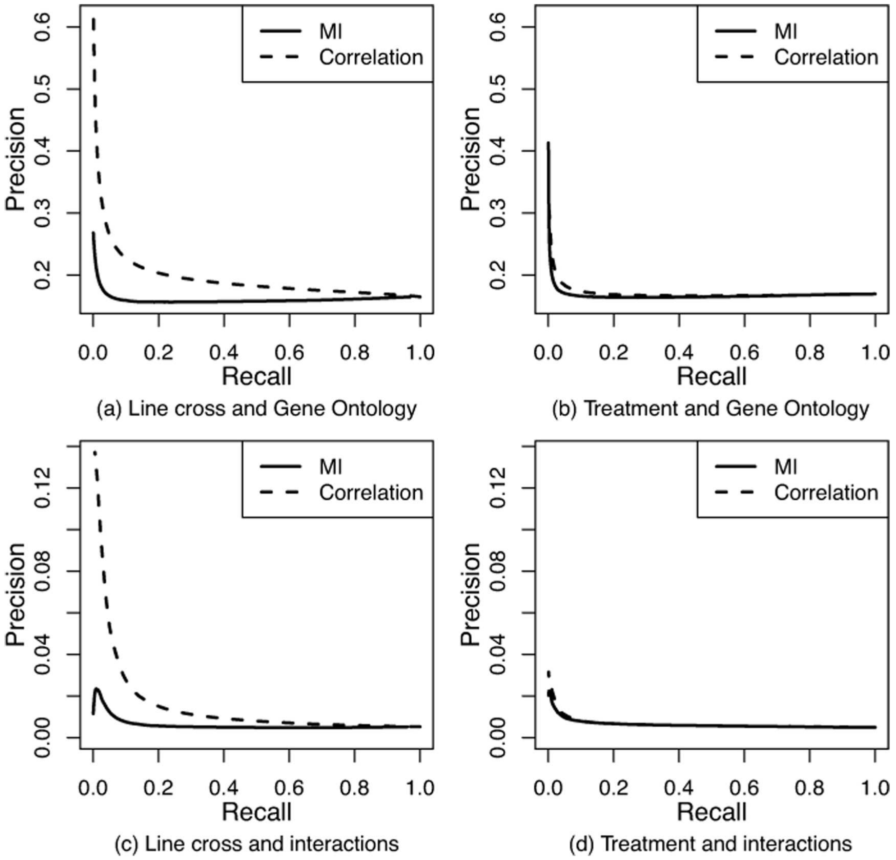
8.3.1 Does the choice of edge weighting method, data type, and edge cut-off affect the functional content of networks?

8.3.1.1 Different networks uncover different amounts of known interactions and shared Gene Ontology

We measured the functional content of networks in terms of known interactions and shared Gene Ontology (GO) terms (see Methods). These are commonly used approaches to evaluate the extent to which networks reflect existing biological knowledge [36, 145, 105, 4, 47, 95].

Specifically, we computed precision, recall, and the F-score in a network as follows. Precision is the proportion of edges in the network that correspond to known interactions or whose end nodes share a GO term. Recall is the proportion of the known interactions or gene pairs sharing a GO term that are in the network. Because there is a trade-off between precision and recall, in the sense that higher precision means lower recall and vice versa, the two measures were combined into F-score, their harmonic mean (see Methods). We computed precision, recall, and F-score for each combination of edge weighting method and data type, varying edge cut-off from 0% to 100%.

We found that the choice of *edge weighting method* affected the functional content in terms of both known interactions (when known interactions are combined independent of interaction type) and shared GO terms. Networks constructed with correlation nearly always had higher precision and recall (Figure 8.1), as well as F-score (Figure 8.2) than networks constructed with mutual information. This indicates that correlation-based networks more accurately uncover existing biological knowledge. This was much more pronounced for the line cross data than for the treatment data. The choice of *data type* also affects functional content: precision and recall were higher for networks constructed from the line cross data than for networks constructed from the treatment data. This was more pronounced for correlation than for mutual information.



(a) Line cross and Gene Ontology

(b) Treatment and Gene Ontology

(c) Line cross and interactions

(d) Treatment and interactions

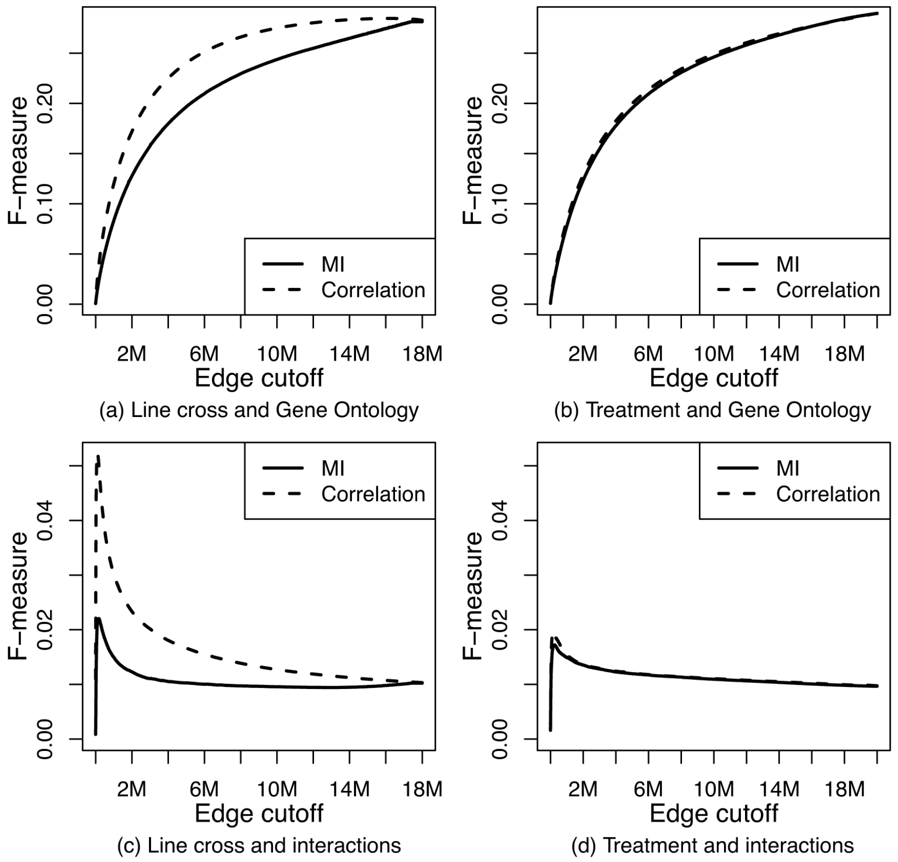
Figure 8.1. Precision-recall curves measuring how accurately networks constructed in different ways capture known biological knowledge. Panel **(a)** shows curves for networks constructed from the line cross data with respect to shared GO terms. Panel **(b)** shows curves for networks constructed from the treatment data with respect to shared GO terms. Panel **(c)** shows curves for networks constructed from the line cross data with respect to known interactions. Panel **(d)** shows curves for networks constructed from the treatment data with respect to known interactions.

As expected, precision tended to decrease and recall tended to increase with an increase in edge cut-off, independent of edge weighting method or data type. This was true for both known interactions and shared GO terms. However, at smaller cut-offs, precision tended to drop significantly. Hence, the edge cut-off just before this drop in precision could be a good cut-off for constructing the network. Alternatively, one could choose the cut-off where precision and recall cross or where F-score starts to decrease (e.g., see "peaks" in Figure 8.2 c and d). Ultimately, which edge cut-off to choose depends on one's preference for the trade-off between precision and recall or the desired network density. We conclude that the smaller the edge cut-off, the higher the precision, the lower the recall, and the sparser the network.

Note that, whereas uncovering existing knowledge is desirable for the purpose of testing the accuracy of network construction, there is no reason to assume that gene co-expression networks should fully uncover existing knowledge, as each piece of biological data could be capturing somewhat complementary functional slices of a cell.

8.3.1.2 Different methods uncover different types of known interactions

We also evaluated known interactions in greater detail by focusing on individual interaction types, instead of considering all known interactions combined as done above. For each network and interaction type, we counted how many interactions of the given type were present in the network. We then computed the probability of observing the same or higher number of interactions of the same type in the network purely by chance using the model of hypergeometric distribution (see Methods). Here, instead of varying the edge cut-off over the entire $[0\%, 100\%]$ range, as above, we varied it from 2,500 to 75,000 edges in increments of 2,500 edges, hence focusing on stronger edges only. Such a more detailed, type-specific analysis of known interactions allowed us to make more biologically relevant observations about the potential



(a) Line cross and Gene Ontology

(b) Treatment and Gene Ontology

(c) Line cross and interactions

(d) Treatment and interactions

Figure 8.2. F-scores at different edge cut-offs measuring how accurately networks constructed in different ways capture known biological knowledge. Panel **(a)** shows F-scores for networks constructed from the line cross data with respect to shared GO terms. Panel **(b)** shows F-scores for networks constructed from the treatment data with respect to shared GO terms. Panel **(c)** shows F-scores for networks constructed from the line cross data with respect to known interactions. Panel **(d)** shows F-scores for networks constructed from the treatment data with respect to known interactions.

differences in the functional content between networks constructed in different ways.

We found that the choice of *edge weighting method* affected the functional content in terms of *individual* interaction types (Figures 8.4, 8.5, and 8.6, and Figure 8.3). For example, if we focus on Affinity Capture-RNA interactions, we observe that the choice of edge weighting method made a noticeable difference for the the line cross data: the presence of interactions of this type was statistically significant in networks constructed using correlation at each of the 30 cut-offs, while it was not significant in networks constructed using mutual information at any of the 30 cut-offs. On the other hand, the choice of edge weighting method made a small difference for the treatment data when it came to this interaction type: the presence of interactions of this type was statistically significant in networks constructed using both correlation and mutual information at most of cut-offs, and the two edge weighting methods differed in only six out of the 30 cut-offs (Figure 8.6). Hence, the choice of *data type* also affected the functional content in terms of individual interaction types.

Note that the significance of the enrichment of a network constructed using one edge weighting method is often different than when using another edge weighting method. The same holds for using one data type compared to using another data type. This observation, specifically that significance may be unique to one edge weighting approach or one data type (see Methods), allowed us to quantify these differences further and determine that:

1. The difference between mutual information and correlation is more pronounced in the line cross data than in the treatment data, independent of the edge cut-off: the number of known interaction types (out of 26 of them) at which we observed a difference between the two edge weighting methods varied between 7 (27%) and 14 (54%) for line cross data and between 3 (12%) and 8 (31%) for the treatment data, depending on the edge cut-off (Figure 8.5 a).
2. The difference between the line cross data and the treatment data is more pronounced for correlation than for mutual information, independent of the edge cut-off: the number of known interaction types (out of 26 of them) at which we observed a difference between the two data types varied between 8 (31%) and 15 (58%) for correlation and between 4 (15%) and 13 (50%) for mutual information, depending on the edge cut-off (Figure 8.5 b).
3. Although there are some exceptions, for each of the two results above the differences tended to decrease with an increase of the edge cut-off used (Figures 8.5 a and b).

We reached the same conclusions when we counted, for each known interaction type, at how many of the 30 edge cut-offs the significance of the network constructed using one edge weighting method and one data type is different than the significance

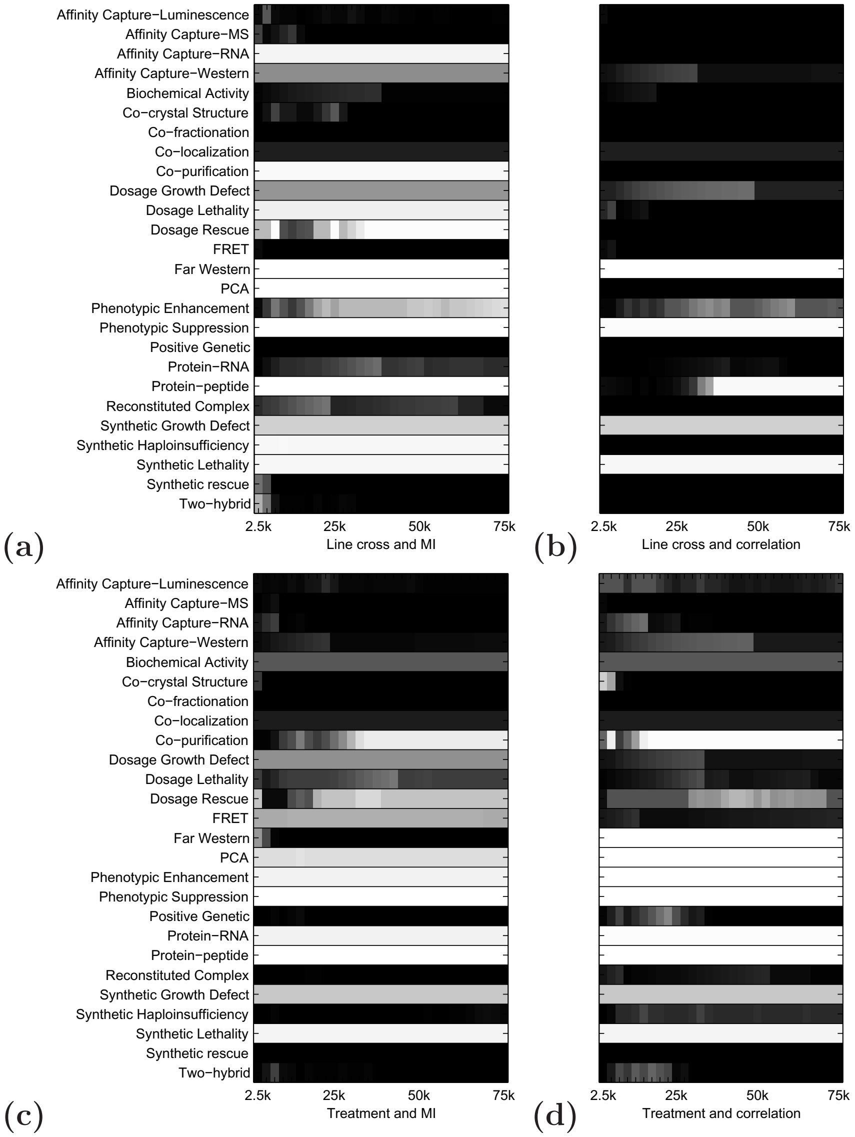


Figure 8.3. Heat maps showing the significance of the enrichment of a network of a given size (x-axis) in known interactions of a given type (y-axis) according to the hypergeometric test (see Methods). Significance is denoted by the darkness of the color, black being the most significant; significance diminishes as the color approaches white. Panel **(a)** shows enrichment results for networks constructed from the line cross data using mutual information. Panel **(b)** shows enrichment results for networks constructed from the line cross data using correlation. Panel **(c)** shows enrichment results for networks constructed from the treatment data using mutual information. Panel **(d)** shows enrichment results for networks constructed from the treatment data using correlation.

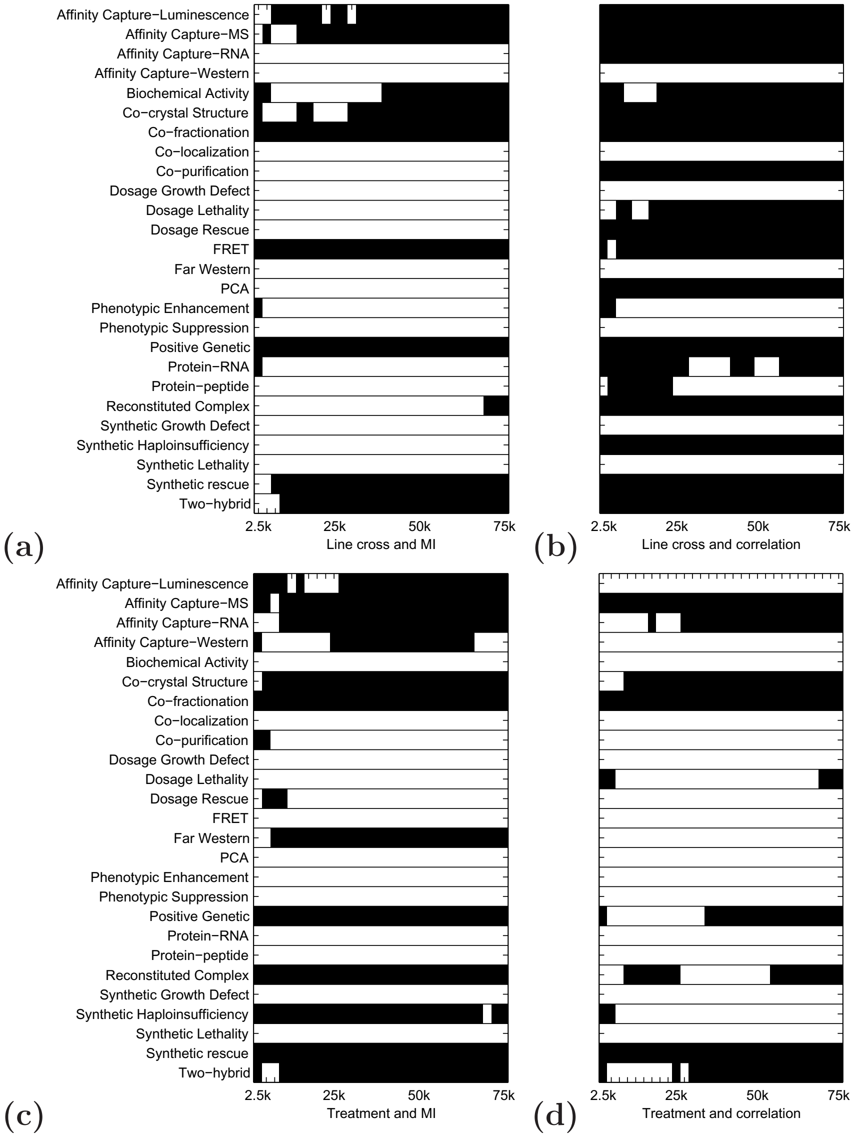


Figure 8.4. Heat maps showing where the enrichment of a network of a given size (x-axis) in known interactions of a given type (y-axis) is statistically significant (denoted by black color) according to the hypergeometric test (see Methods). The cut-off for statistical significance was computed at a False Discovery Rate (FDR) of 0.05. Panel **(a)** shows enrichment results for networks constructed from the line cross data using mutual information. Panel **(b)** shows enrichment results for networks constructed from the line cross data using correlation. Panel **(c)** shows enrichment results for networks constructed from the treatment data using mutual information. Panel **(d)** shows enrichment results for networks constructed from the treatment data using correlation.

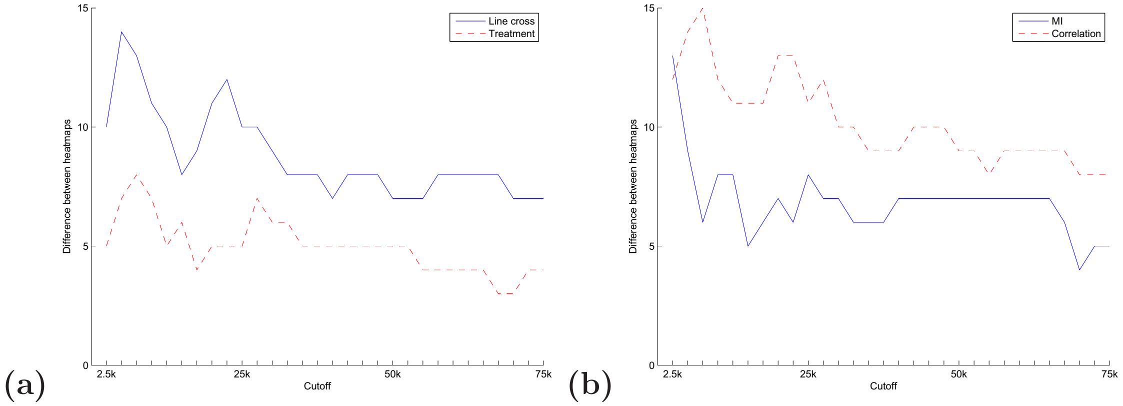


Figure 8.5. The absolute differences between networks constructed using different edge weighting methods **(a)** or data types **(b)** at different edge cut-offs (x-axis) with respect to the number of known interaction types that the networks are statistically significantly enriched in (y-axis). That is, based on Figure 8.4, for a given data type (line cross or treatment), for a given cut-off, we count for how many of the 26 known interaction types the significance of the enrichment of the network constructed using mutual information is different than the significance of the enrichment of the network constructed using correlation (panel **(a)**). By "different", we mean that the enrichment with respect to one edge weighting method is significant (denoted by black color in Figure 8.4), while it is not significant with respect to the other edge weighting method (denoted by white color in Figure 8.4). Analogously, for a given edge weighting method (mutual information or correlation) and a given cut-off, we count for how many of the 26 known interaction types the significance of the enrichment of the network constructed from the line cross data is different than the significance of the enrichment of the network constructed from the treatment data (panel **(b)**). Now, by "different", we mean that the enrichment with respect to one data type is significant (denoted by black color in Figure 8.4) while it is not significant with respect to the other data type (denoted by white color in Figure 8.4).

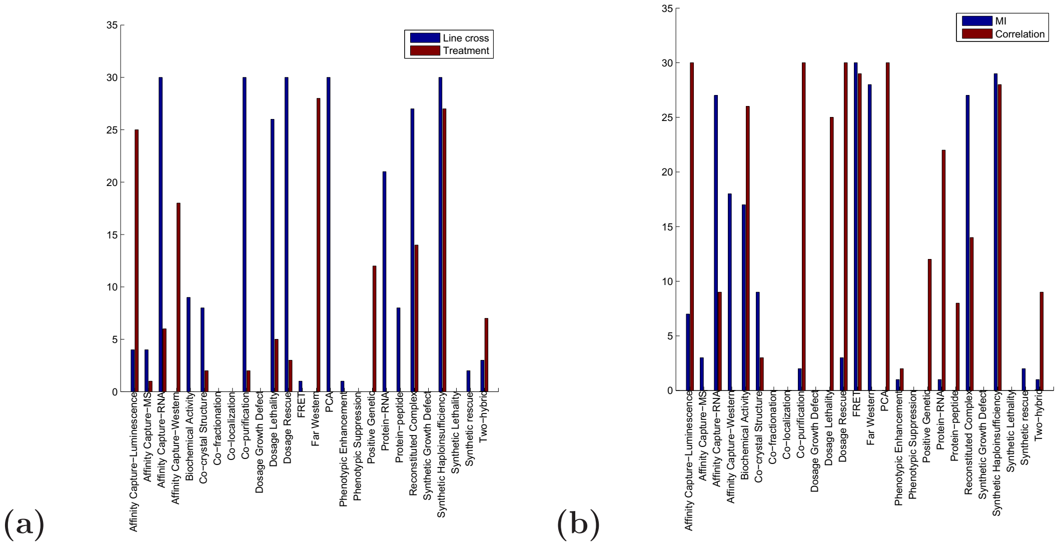


Figure 8.6. The absolute differences between networks constructed using different edge weighting methods **(a)** or data types **(b)** at different edge cut-offs (x-axis) with respect to the number of known interaction types that the networks are statistically significantly enriched in (y-axis). That is, based on Figure 8.4, for a given data type (line cross or treatment), for a given cut-off, we count for how many of the 26 known interaction types the significance of the enrichment of the network constructed using mutual information is different than the significance of the enrichment of the network constructed using correlation (panel **(a)**). By "different", we mean that the enrichment with respect to one edge weighting method is significant (denoted by black color in Figure 8.4), while it is not significant with respect to the other edge weighting method (denoted by white color in Figure 8.4). Analogously, for a given edge weighting method (mutual information or correlation) and a given cut-off, we count for how many of the 26 known interaction types the significance of the enrichment of the network constructed from the line cross data is different than the significance of the enrichment of the network constructed from the treatment data (panel **(b)**). Now, by "different", we mean that the enrichment with respect to one data type is significant (denoted by black color in Figure 8.4) while it is not significant with respect to the other data type (denoted by white color in Figure 8.4).

of the network constructed using either the different edge weighting method or the different data type:

1. The difference between mutual information and correlation is more pronounced in the line cross data than in the treatment data: in the line cross data, mutual information and correlation were different at 20-30, 10-19, 5-9, and 0-4 out of the 30 cut-offs for 8, 0, 3, and 15 out of the 26 known interaction types, respectively, while in the treatment data, mutual information and correlation were different for 3, 3, 3, and 17 out of the 26 known interaction types interaction types, respectively (Figure 8.6 a).
2. The difference between the line cross data and the treatment data is more pronounced for correlation than for mutual information: for mutual information, the line cross data and the treatment data were different at 20-30, 10-19, 5-9, and 0-4 out of the 30 cut-offs for 5, 2, 2, and 17 out of the 26 known interaction types, respectively, while for correlation, the line cross data and the treatment data were different for 9, 2, 3, and 12 out of the 26 known interaction types, respectively (Figure 8.6 b).
3. The majority of the known interaction types show disagreement between the line cross data and the treatment data: only 11 out of the 26 known interaction types showed agreement between the line cross data and the treatment data in the sense that the number of cut-offs in which correlation and mutual information differ was similar (i.e., within the same bin, where the bins are 20-30, 10-19, 5-9, and 0-4) for the two data types.
4. The majority of the known interaction types show disagreement between correlation and mutual information: only 11 out of the 26 known interaction types showed agreement between correlation and mutual information in the sense that the number of cut-offs in which the line cross data and the treatment data differ was similar (i.e., within the same bin, where the bins are 20-30, 10-19, 5-9, and 0-4) for the two edge weighting methods.

Table 8.1 further supports the above conclusions: for the majority (18-21 out of 26) of known interaction types, there is a statistically significant difference (p-value $\leq 1.7 \times 10^{-03}$) between the different edge weighting methods and data types with respect to the proportion of known interactions that they uncover across networks (corresponding to the 30 different cut-offs).

8.4 Tables

TABLE 8.1: *P*-VALUES FROM SIGNED-RANK TESTS COMPARING DIFFERENT EDGE WEIGHTING METHODS AND DATA TYPES.

	Line cross & MI v. Line cross & Correlation		Treatment & MI v. Treatment & Correlation		Treatment & MI v. Line cross & MI		Treatment & Correlation v. Line cross & Correlation	
Affinity Capture-Luminescence	$1.0\times10^{+00}$	$\mathbf{1.3}\times10^{-06}$	$\mathbf{1.5}\times10^{-05}$	$1.0\times10^{+00}$	$\mathbf{3.5}\times10^{-04}$	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Affinity Capture-MS	$1.0\times10^{+00}$	$\mathbf{9.3}\times10^{-10}$	9.8×10^{-01}	2.4×10^{-02}	9.8×10^{-01}	2.4×10^{-02}	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Affinity Capture-RNA	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$\mathbf{4.4}\times10^{-06}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{2.9}\times10^{-06}$	$\mathbf{2.9}\times10^{-05}$	$1.0\times10^{+00}$
Affinity Capture-Western	$1.0\times10^{+00}$	$\mathbf{1.1}\times10^{-04}$	3.0×10^{-03}	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{3.2}\times10^{-05}$	1.1×10^{-02}	9.9×10^{-01}
Biochemical Activity	$1.0\times10^{+00}$	$\mathbf{1.6}\times10^{-04}$	NA	NA	$\mathbf{3.6}\times10^{-04}$	$1.0\times10^{+00}$	$\mathbf{1.4}\times10^{-05}$	$1.0\times10^{+00}$
Co-crystal Structure	$1.0\times10^{+00}$	$\mathbf{9.3}\times10^{-10}$	$\mathbf{3.3}\times10^{-04}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{1.2}\times10^{-05}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Co-fractionation	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$\mathbf{3.2}\times10^{-04}$	$1.0\times10^{+00}$	$\mathbf{1.4}\times10^{-06}$	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Co-localization	NA	NA	NA	NA	NA	NA	NA	NA
Co-purification	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$\mathbf{3.6}\times10^{-04}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{3.2}\times10^{-05}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Dosage Growth Defect	$1.0\times10^{+00}$	1.9×10^{-03}	$1.0\times10^{+00}$	$\mathbf{1.6}\times10^{-04}$	NA	NA	9.9×10^{-01}	1.8×10^{-02}
Dosage Lethality	$1.0\times10^{+00}$	$\mathbf{2.0}\times10^{-06}$	$1.0\times10^{+00}$	$\mathbf{8.3}\times10^{-04}$	$1.0\times10^{+00}$	$\mathbf{1.3}\times10^{-06}$	$\mathbf{7.8}\times10^{-04}$	$1.0\times10^{+00}$
Dosage Rescue	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	9.6×10^{-01}	4.1×10^{-02}	2.8×10^{-01}	7.3×10^{-01}	$\mathbf{1.3}\times10^{-06}$	$1.0\times10^{+00}$
Far Western	$1.0\times10^{+00}$	1.9×10^{-03}	$\mathbf{1.3}\times10^{-06}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{1.3}\times10^{-06}$	$1.0\times10^{+00}$	$\mathbf{4.8}\times10^{-05}$
FRET	$\mathbf{1.1}\times10^{-04}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{2.2}\times10^{-05}$	$\mathbf{1.3}\times10^{-06}$	$1.0\times10^{+00}$	$\mathbf{2.0}\times10^{-06}$	$1.0\times10^{+00}$

Continued on next page

TABLE 8.1:

	Line cross & MI v. Line cross & Correlation		Treatment & MI v. Treatment & Correlation		Treatment & MI v. Line cross & MI		Treatment & Correlation v. Line cross & Correlation	
PCA	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$\mathbf{1.3}\times10^{-06}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{1.2}\times10^{-05}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Phenotypic Enhancement	$1.0\times10^{+00}$	$\mathbf{2.1}\times10^{-04}$	$\mathbf{1.3}\times10^{-06}$	$1.0\times10^{+00}$	$\mathbf{1.6}\times10^{-05}$	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Phenotypic Suppression	$1.0\times10^{+00}$	$\mathbf{3.0}\times10^{-06}$	$\mathbf{6.5}\times10^{-06}$	$1.0\times10^{+00}$	$\mathbf{9.7}\times10^{-05}$	$1.0\times10^{+00}$	$\mathbf{3.0}\times10^{-06}$	$1.0\times10^{+00}$
Positive Genetic	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$\mathbf{3.3}\times10^{-06}$	$1.0\times10^{+00}$	8.9×10^{-01}	1.1×10^{-01}	$\mathbf{9.3}\times10^{-10}$	$1.0\times10^{+00}$
Protein-peptide	$1.0\times10^{+00}$	$\mathbf{9.3}\times10^{-10}$	$\mathbf{1.6}\times10^{-05}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{1.4}\times10^{-06}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Protein-RNA	$1.0\times10^{+00}$	$\mathbf{4.4}\times10^{-06}$	$\mathbf{1.3}\times10^{-03}$	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Reconstituted Complex	$1.0\times10^{+00}$	$\mathbf{3.0}\times10^{-06}$	$\mathbf{6.5}\times10^{-06}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$\mathbf{4.4}\times10^{-06}$	$1.0\times10^{+00}$
Synthetic Growth Defect	NA	NA	NA	NA	NA	NA	NA	NA
Synthetic Haploinsufficiency	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$\mathbf{4.3}\times10^{-05}$	$1.0\times10^{+00}$	$1.0\times10^{+00}$	$\mathbf{9.3}\times10^{-10}$	$\mathbf{2.0}\times10^{-06}$	$1.0\times10^{+00}$
Synthetic Lethality	NA	NA	NA	NA	NA	NA	NA	NA
Synthetic Rescue	$1.0\times10^{+00}$	$\mathbf{9.3}\times10^{-10}$	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	9.5×10^{-01}	5.5×10^{-02}	$\mathbf{9.3}\times10^{-10}$	$1.0\times10^{+00}$
Two-hybrid	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$\mathbf{4.0}\times10^{-04}$	$1.0\times10^{+00}$	$\mathbf{1.2}\times10^{-03}$	$1.0\times10^{+00}$	$\mathbf{9.1}\times10^{-07}$	$1.0\times10^{+00}$
Significant differences	21		19		18		21	

P-values from signed-rank tests comparing different edge weighting methods and data types with respect to the proportion of known interactions of a given type (out of the total number of edges in the network) across 30 networks corresponding to the 30 cut-offs. For each of the 26 known interaction types, for each combination of the edge weighting method and data type, for each of the 30 cut-offs, we compute the proportion of known interactions of the given type out of all edges in the network constructed using the given edge weighting method, data type, and cut-off. Then, we compare the 30 resulting values corresponding to the 30 cut-offs between networks constructed from linecross data using correlation and networks constructed from linecross data using mutual information, between networks constructed from treatment data using correlation and networks constructed from treatment data using mutual information, between networks constructed from linecross data using correlation and networks constructed from treatment data using correlation, and between networks constructed from linecross data using mutual information and networks constructed from treatment data using mutual information. The p-value on the left of a given cell in the table tests whether the median rank of the first set of the 30 values is greater than or equal to the median rank of the second set of the 30 values. The p-value on the right of the cell tests whether the median rank of the second set of the 30 values is greater than or equal to the median rank of the first set of the 30 values. If the p-value is below a given cut-off (see below), the difference in the median ranks between two given sets is considered to be statistically significant (and is bolded in the table). We used the Šidák correction for multiple testing to identify a stringent p-value cut-off, corresponding to the 0.05 cut-off. The Šidák correction is similar to the Bonferroni correction but assumes independence of individual tests [130]. The extent to which interaction types are independent is unclear but it is common to assume independence in the case of uncertainty. Correcting for 26 tests corresponding to the 26 known interaction types, the p-value cut-off is 1.7×10^{-03}. The "NAs" correspond to no observations being made for the given interaction type. The last row counts the number of known interaction types out of 26 of them for which at least one of the two p-values is below the cut-off.

8.4.0.3 Different network construction methods target different biological questions

In addition to *quantifying* the differences in the functional content between networks constructed using different edge weighting methods or data types, as above, we also *qualitatively* studied which combination of edge weighting method and data type can (or fails to) capture a given known interaction type. For each known interaction type, we checked whether its enrichment was significant in the majority of the 30 networks (corresponding to the 30 cut-offs) constructed using the given edge weighting method and data type. We then grouped known interaction types into those that were significantly enriched for both correlation and mutual information in both the line cross data and the treatment data, those that were significantly enriched for both correlation and mutual information in the line cross data but only for mutual information in the treatment data, those that were significantly enriched for both correlation and mutual information in the treatment data but only for correlation in the line cross data, . . . , and those that were significantly enriched for neither correlation nor mutual information in neither the line cross data nor the treatment data (Table 8.2).

Only 6 out of the 26 known interaction types were found by both edge weighting methods in both data types. Specifically, if one's goal was to construct networks that would enrich for Affinity Capture-MS, Co-crystal Structure, Co-fractionation, Positive Genetic, Synthetic Rescue, or Two-hybrid interactions, one could use either of the two edge weighting methods or data types (Table 8.2). Note however, that the actual interactions uncovered by the different networks could be different, because the overlap between the networks is small (Section 8.4.1.1). Seven of the 26 known interaction types are missed by both edge weighting methods in both data types. In other words, neither the two edge weighting methods nor data types considered in this study enriched for Co-localization, Dosage Growth Defect, Phenotypic Enhancement, Phenotypic Suppression, Protein-peptide, Synthetic Growth Defect, or Synthetic Lethality interactions.

Clearly for $6+7=13$ out of the 26 known interaction types, all four combinations of edge weighting methods and data types *agree*. For the same number of known interaction types, however, at least two of the four combinations *disagree*. For example, there exist interaction types that were captured by correlation in the line cross data but not in the treatment data. Also, there exist interaction types captured by mutual information in the treatment data but not in the line cross data. Interestingly, there are no interaction types captured by mutual information in the line cross data but not in the treatment data or by correlation in the treatment data but not in the line cross data (Table 8.2).

8.4.0.4 Bottom line

All of the above results demonstrate that networks constructed in this study with different combinations of edge weighting methods or data types were enriched with different functional content. Therefore, we have strong evidence that both the edge weighting method and the type of biological experiment underlying the data affect the functional content of networks, implying that they can optimized to answer certain—but different—biological questions.

8.4.1 Does the choice of edge weighting method, data type, and edge cut-off affect the topology of networks?

8.4.1.1 Overlap of networks constructed in different ways is small

Considerable topological differences between networks constructed in different ways were immediately apparent based on the intersection of their edges (Table 8.3). Intersections were calculated between pairs of networks of a given size that shared either edge weighting method or data type. Because the size of the intersection between the networks appeared to have a linear relationship with the cut-off (networks with more strict cut-offs had nearly the same proportion of overlap as networks with less strict cut-offs), the intersections were averaged over different network sizes. Such averaged intersections were smaller between networks constructed by using the same edge weighting method but different data sets than between networks constructed by using the same data set but different edge weighting method (Table 8.3). This indicates that the choice of the data set may have a stronger effect on network construction than the choice of the edge weighting method. Nonetheless, the intersection was relatively small between all compared networks. This is an important result, since it is likely that different edges would lead to different biological interpretations, which is exactly what we demonstrated in the previous section.

8.4.1.2 Networks constructed in different ways have different topological properties

By comparing overall topological characteristics of networks constructed using different edge weighting methods or data types, we found that networks varied as follows. In general, average clustering coefficients (see Methods) were higher for networks constructed using mutual information than for networks constructed using correlation, independent of the data type, and they were higher for networks constructed from the line cross data than for networks constructed from the treatment data, independent of the edge weighting method (Figure 8.7 a). Figure 8.7 b and Figure 8.7 c further demonstrate topological differences between networks constructed with different edge weighting methods and data types. For example, networks for the line cross data and correlation had the most connected components (Figure 8.7 b) but they had the second fewest nodes involved in the components (Figure 8.7 c), indicating that these networks had many small components. On the other hand, networks for the treatment data and mutual information in general had the second fewest connected components (Figure 8.7 b) but they had the most nodes involved in the components (Figure 8.7c), indicating that these networks had few large components.

To further understand the observed topological differences, we focused on one of the 30 analyzed edge cut-offs: 25,000. We chose the cut-off of 25,000 because it is a compromise between the the peak precision in Figures 8.1 a-d. We compared the different networks corresponding to this cut-off with respect to the clustering coefficient, closeness centrality, and betweenness centrality spectra (see Methods). Topological differences were immediately apparent. For example, in the line cross data-based networks, average clustering coefficients tended to decrease as node degrees increased for mutual information, whereas this was not necessarily the case for correlation (Figure 8.8 a). Clustering spectra were also different for the line cross data-based networks (Figure 8.8 a) and treatment-based networks (Figure 8.8 b): correlation-based networks in particular appear to have almost the opposite trends for the two data types. The observed differences were statistically significant for each pair of clustering spectra, both when the data type was shared but the edge weighting method was different (p-values $< 2.2 \times 10^{-16}$) *and* when the edge weighting method was the same but the data type was different (p-values of 0.008 and $< 1.07 \times 10^{-15}$) (Figure 8.8).

Similar observations were made with respect to closeness centrality spectra. For example, in the line cross data-based networks, while average closeness centralities were higher for correlation than for mutual information for low-degree nodes, they were lower for correlation than for mutual information for high-degree nodes (Figure 8.9 a). While the observed difference between the two edge weighting methods was not statistically significant for the line cross data (p-value of 0.3511), it was statistically significant for the treatment data (p-value of 0.0063; Figure 8.9 b). Further, the difference between closeness spectra was statistically significant when the edge weighting method is the same but the data type is different, with respect to both correlation (p-value of 2.81×10^{-6}) and mutual information (p-value of 1.6×10^{-12}) (Figure 8.9).

Betweenness centrality spectra are statistically significantly different for all four combinations of edge weighting methods and data types as well (Figure 8.10).

We conclude that networks constructed using different edge weighting methods or data types have different topologies.

8.5 Discussion

The network construction problem is analogous to the problem of clustering genes based on similarity between their expression profiles [65, 30, 42]. Many clustering algorithms exist (each with its (dis)advantages [49, 133, 71]) that attempt to group together genes that have similar expression with respect to some distance metric. Hence, analogous to a network construction method, a clustering method first computes distances (or equivalently, similarities) between each pair of genes. Then, it partitions the resulting weighted fully connected network by employing a distance cut-off to determine cluster membership [67]. Examples of popular clustering methods in this context include Walktrap [114] and Markov Clustering (MCL) [154]. The wide variety of available algorithms typically result in different clustering solutions, for reasons that are analogous to those studied in this paper.

One goal of this work was to study the effect of the choice of edge weighting method on the functional content and topology of the resulting network. We note that this is not the first study to analyze how different edge weighting methods affect biological content of the networks. For example, it has been found that modules in networks constructed with signed correlation show different functional enrichment than modules in networks constructed with the absolute correlation [98], and that correlation and mutual information are strongly related in the treatment data when their values are high [136]. This is the first study, however, to analyze how the choice of edge weighting method affects networks in terms of *individual* interaction types. This study is also more comprehensive than the previous ones because it examines multiple edge weighting methods *and* data types.

Our hypothesis is that the two edge weighting methods might be capturing different types of relationships. For example, Figure 8.11 represents a known Biochemical Activity interaction for which the gene pair ranking is discordant, i.e., high with respect to one edge weighting method but low with respect to the other edge weighting method. Figures 8.12 and 8.13, on the other hand, represent known Synthetic Lethality interactions in the line cross data and the treatment data, respectively, for which the gene pair rankings are consistent across the two edge weighting methods—both rank each of these gene pairs high. Note that in our study we focus on answering whether different edge weighting methods captured different functional content to bring this important issue to the attention of the community, not on answering why different types of known interactions were captured differently by the different networks. The later is an important but complex question that is out of the scope of this study.

Another goal of this work was to study whether the type of biological experiment underlying the data affects the functional content of inferred networks because different types of biological experiments are expected to capture somewhat complementary aspects of the cell [79, 74]. To our knowledge this is the first study explicitly comparing networks constructed from line cross data with networks constructed from treatment data. We note that since we use one data set of each type, and since we do so for one species, our results should be used with caution: it is difficult to determine how much of the variation between the different networks is due to the data type versus noise in these particular data sets. Nonetheless, our analysis could be used as a *strong suggestion* that data types considered should also be carefully chosen to match the desired biological question(s).

Finally, we studied the effect of edge cut-off on resulting networks. There are a number of approaches to choosing a cut-off for network construction, including the precision-recall trade-off (Section 8.3.1.1), the approximate topology of the resulting network, or the false discovery rate (FDR) [16, 145]. Under the assumption that gene co-expression networks should capture known biological knowledge, be it interactions, GO functional similarity, or anything else, the strongest relationships should hold the most reliable known biological information (which would be reflected in higher precision at lower edge cut-offs).

We demonstrated systematically and comprehensively that the four combinations of edge weighting methods and data types disagreed with respect to the functional content and topology of the corresponding networks. Because the overlap was small between networks constructed in different ways, and since their overall topologies were different, it is no surprise that the different networks led to different biological interpretations. We showed that: (1) of the two edge weighting methods, correlation seemed to more accurately uncover existing biological knowledge, especially for the line cross data; (2) of the two data types, the line cross data seemed to more accurately uncover existing biological knowledge, especially for correlation; (3) the strongest edges indeed hold the most reliable known biological information; (4) the difference between mutual information and correlation was more pronounced in the line cross data than in treatment data, independent of the edge cut-off; (5) the difference between the line cross data and the treatment data was more pronounced for correlation than for mutual information, independent of the edge cut-off; (6) the differences in points (4) and (5) above tended to decrease with an increase of the cut-off; (7) the type of experiment underlying the data can have at least as much of an effect on the functional content of networks as the choice of edge weighting method; and (8) different types of known interactions could be uncovered with different combinations of edge weighting methods and data types.

In summary, these results demonstrate that there is no single correct way to construct a network—and different approaches can produce different networks that are better suited to answer different biological questions. For example, using correlation and treatment data results in networks that contain the most significant amounts of Synthetic Lethality interactions among all considered networks. Thus, this combination of edge weighting method and data type may be the most appropriate for studying genes that are essential to an organism's survival in combination with other genes. However, the same networks are the worst in terms of capturing Far Western interactions, which are a type of protein-protein interaction data set. As such they should probably not be used to study molecular processes that are carried out via physical interactions between proteins. We conclude that the three ubiquitous factors in network inference that were considered in this study each have significant effects on the functional content and topology of resulting biological networks and therefore do *matter* for future work in systems biology.

8.6 Methods

8.6.1 Data

8.6.1.1 Gene expression data sets

We use gene expression data resulting from two types of experiments: line cross and treatment.

In the line cross experiment, expression levels were measured from 130 segregants of a cross of two strains of yeast [11]. Just as Smith *et al.*, we used all probes for which more than 80% of the expression data was present, resulting in a total of 5,829 unique open reading frames [131]. We measured the strength of relationships between genes by relying on the normalized log ratios provided. Expression values for repeated probes were averaged. Line cross experiments can reveal relationships between heritability and expression [79].

In the treatment experiment, 300 expression profiles were generated from mutant or chemical-treated cultures for 6,314 open reading frames. We measured the strength of relationships between 6,207 unique ORFs by relying on p-values resulting from a gene-specific error model that accounts for variation in genes as well as variation across chips [74]. Repeated measurements were averaged. Treatment experiments cause perturbations (whether due to environment or mutation) that affect gene expression levels and thus enable one to elucidate relationships between genes that may not be evident when the cell is "at rest" [112].

8.6.1.2 Known interactions

By known interactions, we mean the set of interactions that are already available in public databases. Known interactions are a valuable collection of ground truth data, but the completeness of the data can vary greatly between organisms. For example, baker's yeast (*Saccharomyces cerevisiae*) is relatively well characterized, whereas non-model organisms, such as the pathogen of malaria (*Plasmodium falciparum*), are not. Hence, in our study, we focus on the well-studied yeast. Known interactions of various types were obtained from SGD (*S*accharomyces Genome Database) (Table 8.4). 17 of the 26 interaction types have more than 500 genes involved into the corresponding interactions, and 14 of the 26 interaction types have more than 1,000 genes involved into the corresponding interactions (Table 8.4). Hence, known interaction data provides a strong basis for network comparison. Different interaction types describe relationships between genes that were discovered in experiments with different biological meanings. For example, Affinity Capture-MS interactions are determined by using a "bait" protein that is "captured" by a polyclonal antibody or an epitope tag. The associated partners are then identified by mass spectrometry. Affinity Capture-MS interactions typically correspond to physical interactions between proteins. In contrast, Synthetic Lethality interactions are determined by observing when mutations or deletions in separate genes, each of which causes a minimal change in phenotype alone, result in lethality to a cell.

8.6.1.3 Gene Ontology data

The Gene Ontology (GO) assigns biological process, molecular function, and cellular component labels (i.e., terms) to genes. GO terms are arranged in a hierarchical fashion. Genes that share a GO term are typically functionally related and may thus be more likely to interact than genes that do not share a GO term. We use GO-slim biological process data [22]. GO-slim terms are a reduced set of GO terms corresponding to higher levels of the GO hierarchy.

8.6.2 Edge weighting methods

We measure the strength of the relationship between two genes by using either a signed variation of Pearson's correlation or mutual information.

8.6.2.1 Correlation

The signed variation of Pearson's correlation is given in Equation 8.1 [98].

$$signed\,correlation(x,y) = \frac{1 + correlation(x,y)}{2} \tag{8.1}$$

8.6.2.2 Estimation of mutual information

Mutual information is a measure of the mutual dependence of two random variables (Equation 8.2) [98].

$$MI(X,Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log[\frac{p(x,y)}{p(x)p(y)}] \tag{8.2}$$

Calculation of mutual information is complicated by two factors. First, there are a low number of instances from which to estimate probability distributions. Second, expression data is continuous. To overcome these issues, we utilized the Parzen window approach to density estimation [110]. Equation 8.3 estimates the density function $\hat{p}(x)$ over N samples of variable x:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x^{(i)}, h) \tag{8.3}$$

where $\delta(...)$ is the Parzen window function described in Equation 8.4, $x^{(i)}$ is the ith sample, and h is the window size. When $d = 1$, this equation returns the estimated marginal density. When $d = 2$, it gives an estimate of the joint density, $p(x, y)$, which can be used to calculate the mutual information in Equation 8.2.

$$\delta(z, h) = \frac{exp\left(-\frac{z^T \Sigma^{-1} z}{2h^2}\right)}{(2\pi)^{d/2} h^d |\Sigma|^{(1/2)}} \tag{8.4}$$

In Equation 8.4, $z = x - x^{(i)}$, h is the window size, d is the dimension of the sample, and σ is the covariance of z. Further details and the implementation used are described in [113].

8.6.3 Evaluation

Precision and recall are defined in terms of true positives (tp), false positives (fp), and false negatives (fn). In the case of the networks we consider, a true positive is an edge in the network that corresponds to a known interaction or whose end nodes share a GO term. A false positive is an edge in the network that is not among the known interactions or whose end nodes do not share a GO term. A false negative is a pair of genes that are linked by a known interaction or that share a GO term but that are not linked by an edge in the network. Equations 8.5 and 8.6 define precision and recall.

$$\text{precision} = \frac{tp}{tp + fp} \tag{8.5}$$

$$\text{recall} = \frac{tp}{tp + fn} \tag{8.6}$$

Precision-recall curves have been identified as useful alternatives to ROC curves in situations in which there is a large imbalance in the data [84, 85]. Namely, with only 220,009 known interactions and 5,892,199 pairs of genes that share a GO term out of approximately 18 million possible edges, the problem of network inference qualifies as imbalanced.

Increase in precision typically results in decrease in recall, and vice versa. The F-score reconciles precision and recall by combining them into a single score, namely their harmonic mean (Equation 8.7).

$$\text{F-score} = 2 * \frac{precision * recall}{precision + recall} \tag{8.7}$$

Given a network constructed using a given edge weighting method, data type, and edge cut-off, for each known interaction type, we counted how many interactions of the given type are present in the network. Then, we computed p, the probability of observing the same or higher number of interactions of the same type purely by chance, by using the model of hypergeometric distribution. If we denote by N the number of possible edges, by m the total number of known interactions of a given type, by n the number of edges in the network, and by k the number of known interactions of a given type that are in the network, the probability of observing exactly k known interactions in the network purely by chance is computed as shown in Equation 8.8. Then, to compute probability p, we sum Equation 8.8 over all possible numbers of known interactions equal to or greater than k.

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \tag{8.8}$$

Given a set of networks corresponding to 30 different edge cut-offs in the [2,500, 75,000] range, where the same edge weighting method and the data type are used for all 30 networks, we form a vector of 30 probabilities p for the set, where the probabilities are computed for a given known interaction type as explained above. We then compare two vectors of 30 elements corresponding two network sets constructed using the same data type but different edge weighting methods (Table 8.1). We also compare vectors of 30 elements corresponding to two network sets constructed using the same edge weighting method but different data types (Table 8.1). We do this for each known interaction type. We compare any two vectors by using the Wilcoxon signed-rank test, a nonparametric analog of the t-test [37].

8.6.4 Topological analysis of networks

We use three topological measures in our analysis: the clustering coefficient, the closeness centrality, and the betweenness centrality [101, 102].

The clustering coefficient of node v describes the proportion of v's neighbors that are connected to each other. It is computed as shown in Equation 8.9, where d is the number of neighbors of v and k is the number of connected pairs of the neighbors [93]. The global clustering coefficient of the network is the average of the clustering coefficients of all nodes.

$$\text{clustering coefficient}(v) = \frac{2k}{d(d-1)} \tag{8.9}$$

The closeness centrality of node v measures the distance from v to every other node t in the network. It is computed as shown in Equation 8.10, where $SP(v,t)$ is the length of the shortest path between nodes v and t, and V is the set of nodes of the network [121].

$$\text{closeness centrality}(v) = \frac{1}{\sum_{t\varepsilon V} SP(v,t)} \tag{8.10}$$

The betweenness centrality of node v measures the proportion of shortest paths in the network that go through v. It is computed as shown in Equation 8.11, where SP_{st} is the number of shortest paths between nodes s and t and $SP_{st}(v)$ is the number of shortest paths between s and t that pass through v [50].

$$\text{betweenness centrality}(v) = \sum_{s \neq v \neq t \varepsilon V} \frac{SP_{st}(v)}{SP_{st}} \tag{8.11}$$

For each of the three measures, we compute the corresponding “spectrum” as the average over all nodes of degree k, for each value of k. For example, clustering spectrum of a network is the average clustering coefficient of all nodes of degree k in the network (Figure 8.8). Spectra are often displayed in log scale for ease of interpretation.

TABLE 8.2

COMBINATIONS OF EDGE WEIGHTING METHODS/DATA TYPES FOR WHICH THE GIVEN INTERACTION TYPE IS SIGNIFICANTLY ENRICHED AT THE MAJORITY OF EDGE CUT-OFFS.

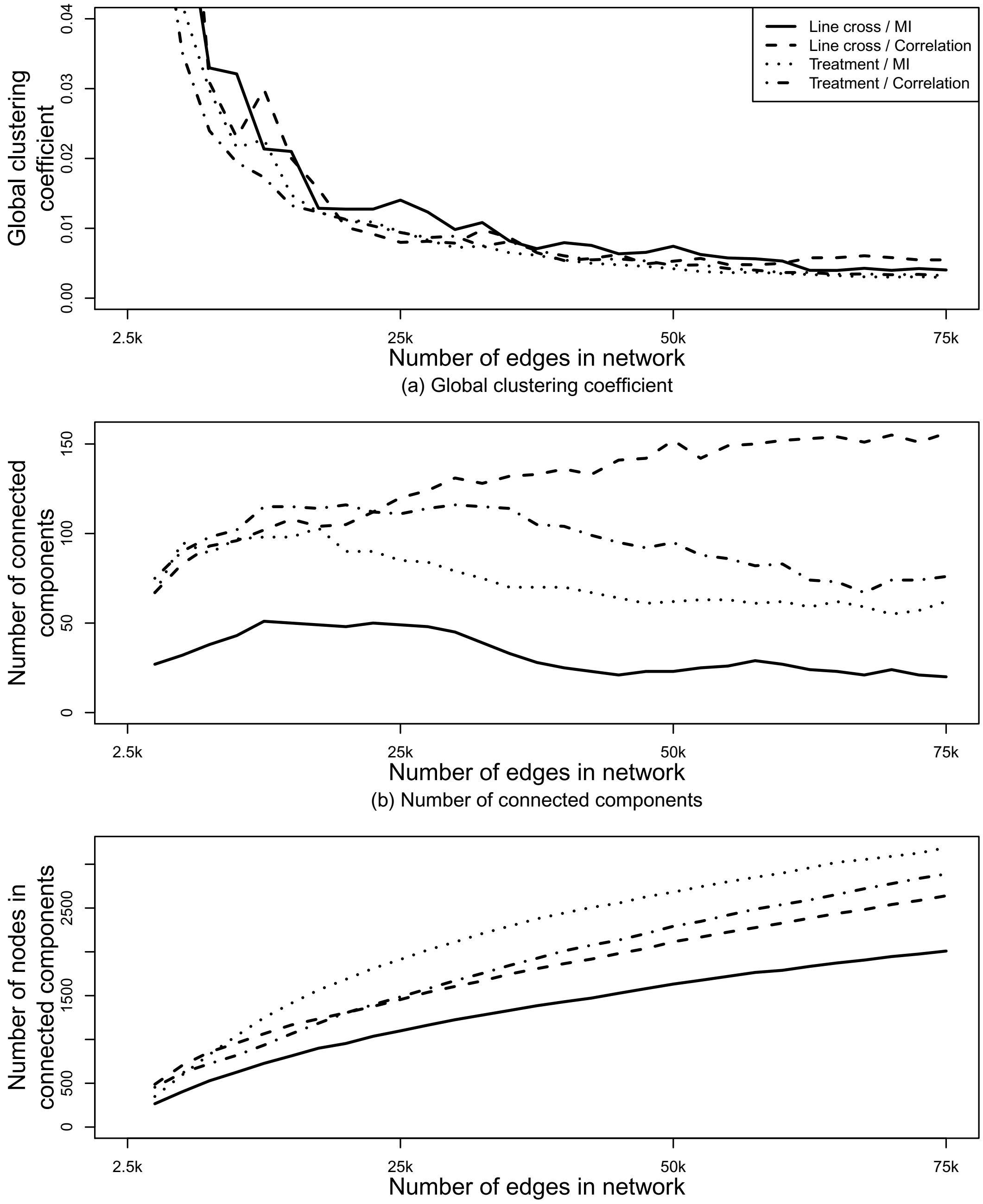
Combinations of edge weighting methods/data types	Interaction types
Found by both correlation and mutual information in both the line cross data and the treatment data.	Affinity Capture-MS Co-crystal Structure Co-fractionation Positive Genetic Synthetic Rescue Two-hybrid
Found by both correlation and mutual information in the line cross data but only by mutual information in the treatment data.	Affinity Capture-Luminescence
Found by both correlation and mutual information in treatment data but only by correlation in line cross data.	Affinity Capture-RNA Reconstituted Complex
Found by both correlation and mutual information in line cross data but none of correlation or mutual information in treatment data.	Biochemical Activity FRET
Found only by correlation in line cross data and only by mutual information in treatment data.	Synthetic Haploinsufficiency
Found by neither correlation nor mutual information in treatment data, but found only by correlation in line cross data.	Co-purification Dosage Lethality Dosage Rescue PCA Protein-RNA
Found by neither correlation nor mutual information in line cross data, but found only by mutual information in treatment data.	Affinity Capture-Western Far Western
Found by neither correlation nor mutual information in neither line cross data nor treatment data.	Co-localization Dosage Growth Defect Phenotypic Enhancement Phenotypic Suppression Protein-peptide Synthetic Growth Defect Synthetic Lethality

TABLE 8.3

EDGE OVERLAP BETWEEN NETWORKS CONSTRUCTED USING DIFFERENT DATA TYPES AND EDGE WEIGHTING METHODS.

	Line cross	Treatment	Correlation	MI
mean	30.85%	40.23%	11.98%	14.06%
min	29.41%	32.08%	9.23%	11.70%
max	33.12%	45.82%	13.37%	18.82%
stdev	0.51%	4.30%	1.35%	2.13%

Edge overlap between networks constructed using different data types and edge weighting methods, averaged over all edge cut-offs. Each column denotes the edge weighting method or data type used as the basis for comparison between networks. The column denoted by “Line cross” compares networks constructed from the line cross data using correlation to networks constructed from the line cross data using mutual information. The column denoted by “Treatment” compares networks constructed from the treatment data using correlation to networks constructed from the treatment data using mutual information. The column denoted by “Correlation” compares networks constructed from the line cross data using correlation to networks constructed from the treatment data using correlation. The column denoted by “MI” compares networks constructed from the line cross data using mutual information to networks constructed from the treatment data using mutual information.

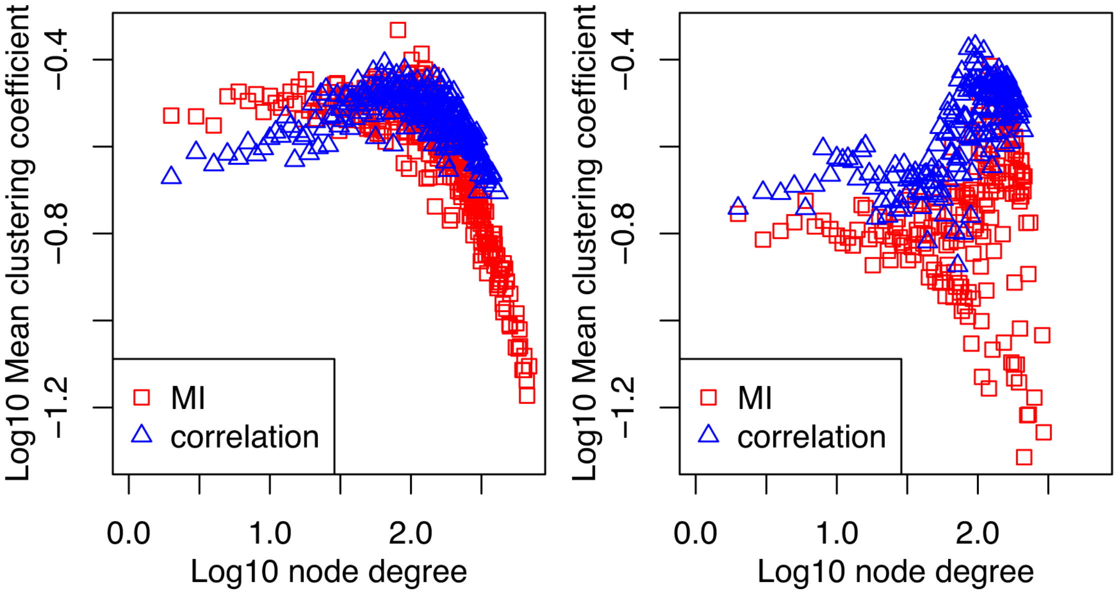


(a) Global clustering coefficient

(b) Number of connected components

(c) Number of nodes in connected components

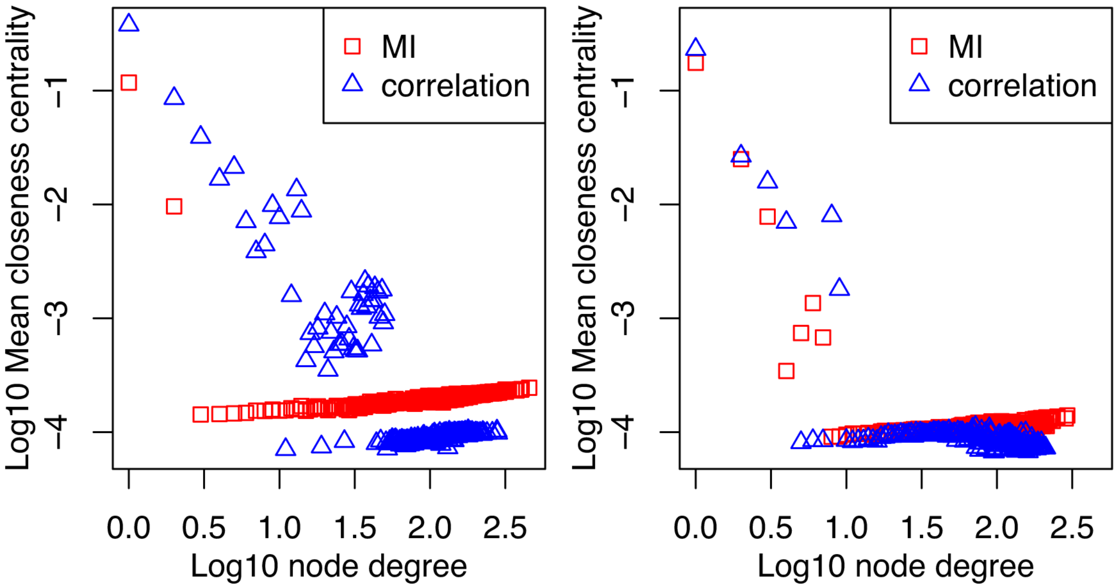
Figure 8.7. Global clustering coefficient **(a)**, the number of connected components with at least two nodes **(b)**, and the number of nodes that participate in the connected components **(c)** for networks constructed from a given data type (line cross or treatment) using a given edge weighting method (correlation or mutual information) at a given edge cut-off (x-axis).



(a) Line cross

(b) Treatment

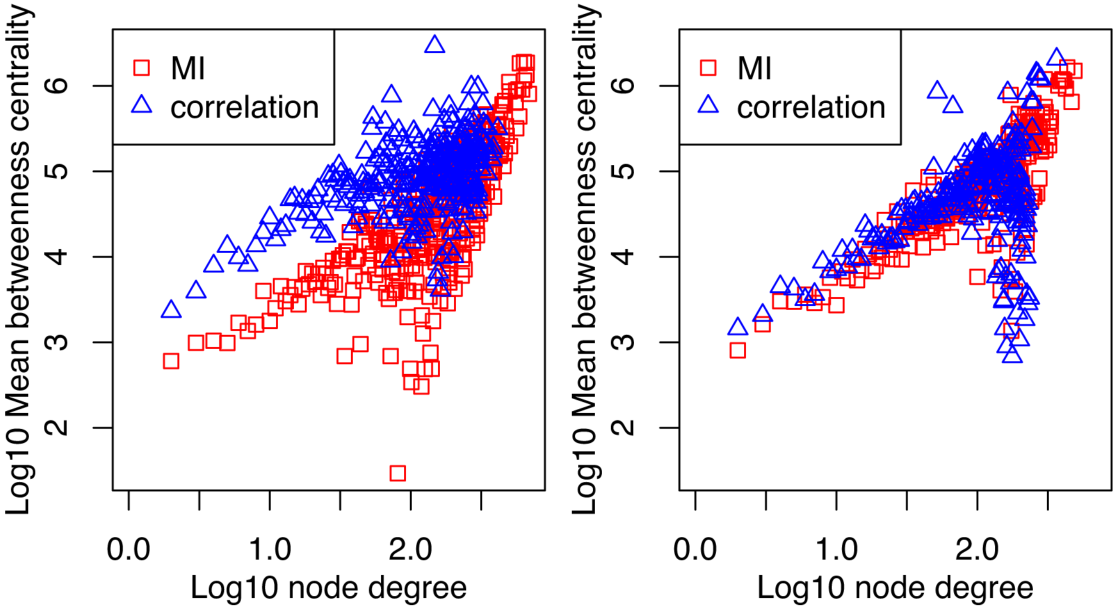
Figure 8.8. Clustering spectra for networks constructed from the line cross data **(a)** and the treatment data **(b)** at the edge cut-off of 25,000. We compared with a *t*-test pairs of blue and red spectra within panels, which share the same data type but differ in the edge weighting method. We also compared pairs of blue and blue spectra or red and red spectra across panels, which share the same edge weighting method but differ in the data type. The two spectra in panel **(a)** as well as the two spectra in panel **(b)** were statistically significantly different with p-values $< 2.2 \times 10^{-16}$. The two correlation-based spectra (blue) across the two panels were statistically significantly different with a p-value of 0.008. The two mutual information-based spectra (red) across the two panels were statistically significantly different with a p-value of 1.07×10^{-15}.



(a) Line cross

(b) Treatment

Figure 8.9. Closeness spectra for networks constructed from the the line cross data **(a)** and the treatment data **(b)** at the edge cut-off of 25,000. We compared with a t-test pairs of blue and red spectra within panels, which share the same data type but differ in the edge weighting method. We also compared pairs of blue and blue spectra or red and red spectra across panels, which share the same edge weighting method but differ in the data type. The two spectra in panel **(a)** were not statistically significantly different (p-value of 0.3511). The two spectra in panel **(b)** were statistically significantly different with a p-value of 0.0063. The two correlation-based spectra (blue) across the two panels were statistically significantly different with a p-value of 2.81×10^{-6}. The two mutual information-based spectra (red) across the two panels were statistically significantly different with a p-value of 1.6×10^{-12}.



(a) Line cross

(b) Treatment

Figure 8.10. Betweenness spectra for networks constructed from the the line cross data **(a)** and the treatment data **(b)** at the edge cut-off of 25,000. We compared with a *t*-test pairs of blue and red spectra within panels, which share the same data type but differ in the edge weighting method. We also compared pairs of blue and blue spectra or red and red spectra across panels, which share the same edge weighting method but differ in the data type. The two spectra in panel **(a)** as well as the two spectra in panel **(b)** were statistically significantly different with p-values $< 2.2 \times 10^{-16}$. The two correlation-based spectra (blue) across the two panels were statistically significantly different with a p-value of 3.3×10^{-7}. The two mutual information-based spectra (red) across the two panels were statistically significantly different with a p-value of 9.6×10^{-11}.

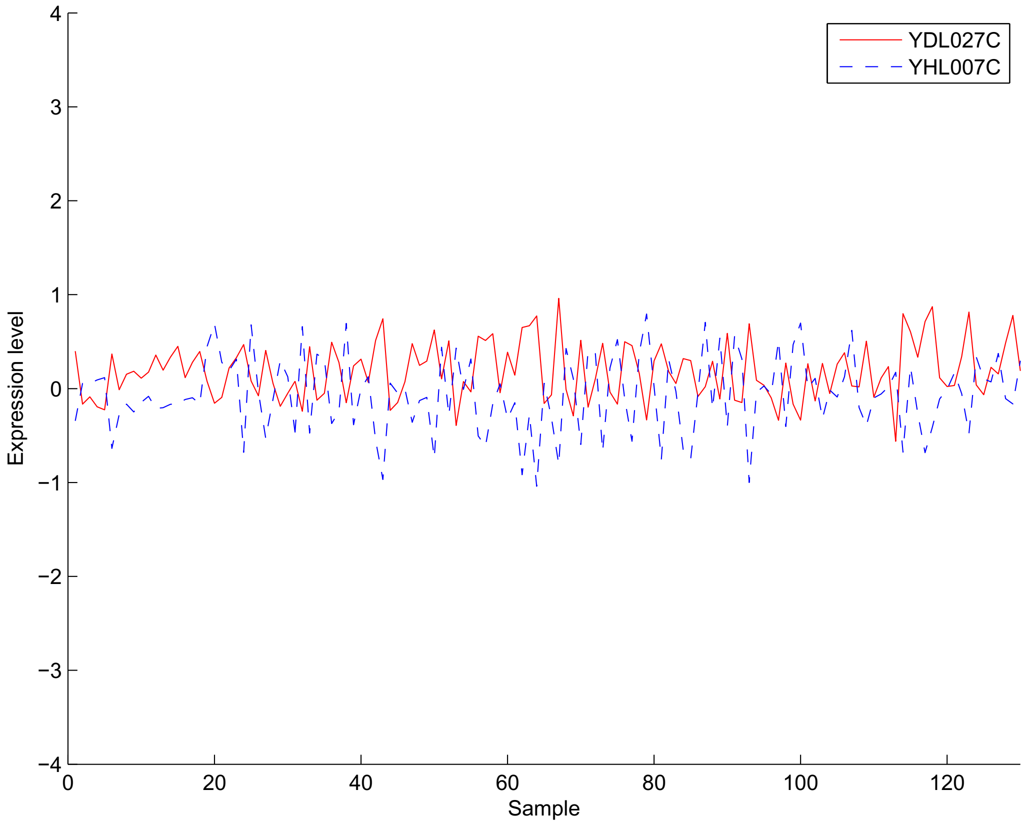


Figure 8.11. Expression levels in the line cross data of two genes that share a known Biochemical Activity interaction and for which the correlation is low while the mutual information is high. Namely, the correlation has a value of -0.742330 and the mutual information has a value of 0.157230. The correlation between these two genes is greater than the correlation between 0% of all pairs of genes in the data. The mutual information between these two genes is greater than the mutual information between 99.2% of all pairs of genes in the data.

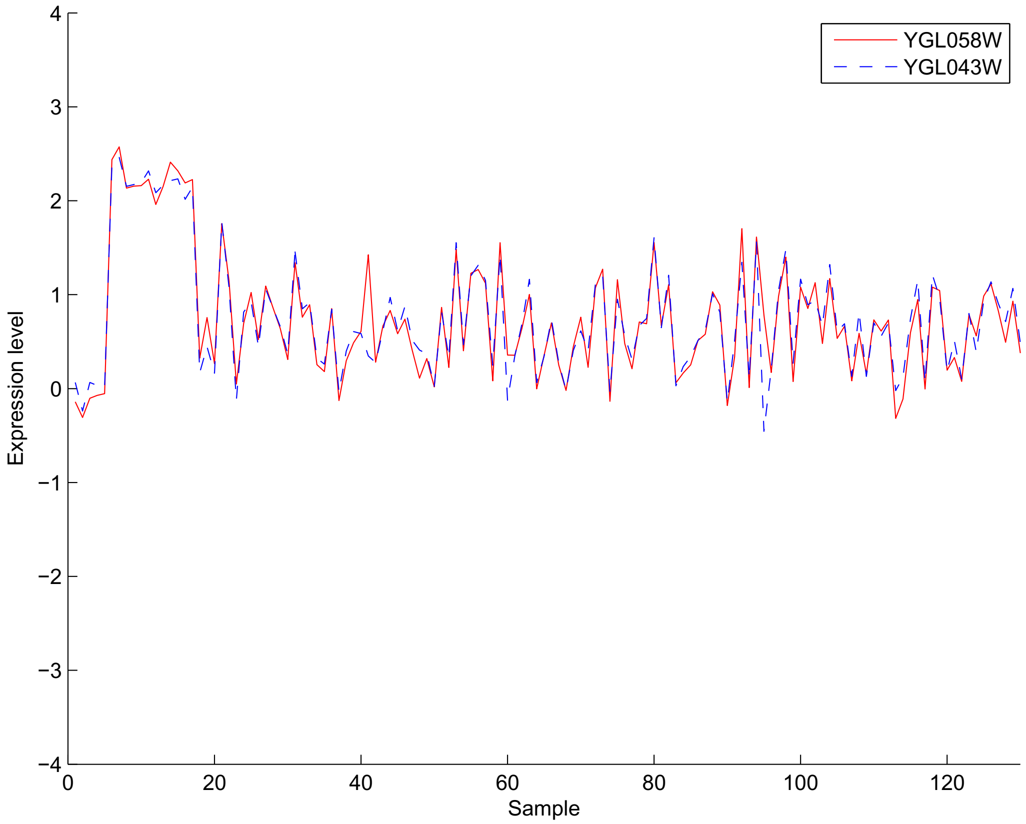


Figure 8.12. Expression levels in the line cross data of two genes that share a known Synthetic Lethality interaction and for which both the correlation and the mutual information are high. Namely, the correlation has a value of 0.958520 and the mutual information has a value of 0.863000. The correlation between these two genes is greater than the correlation of 99.8% of all pairs of genes in the data. The mutual information between these two genes is greater than the mutual information between 99.9% of all pairs of genes in the data.

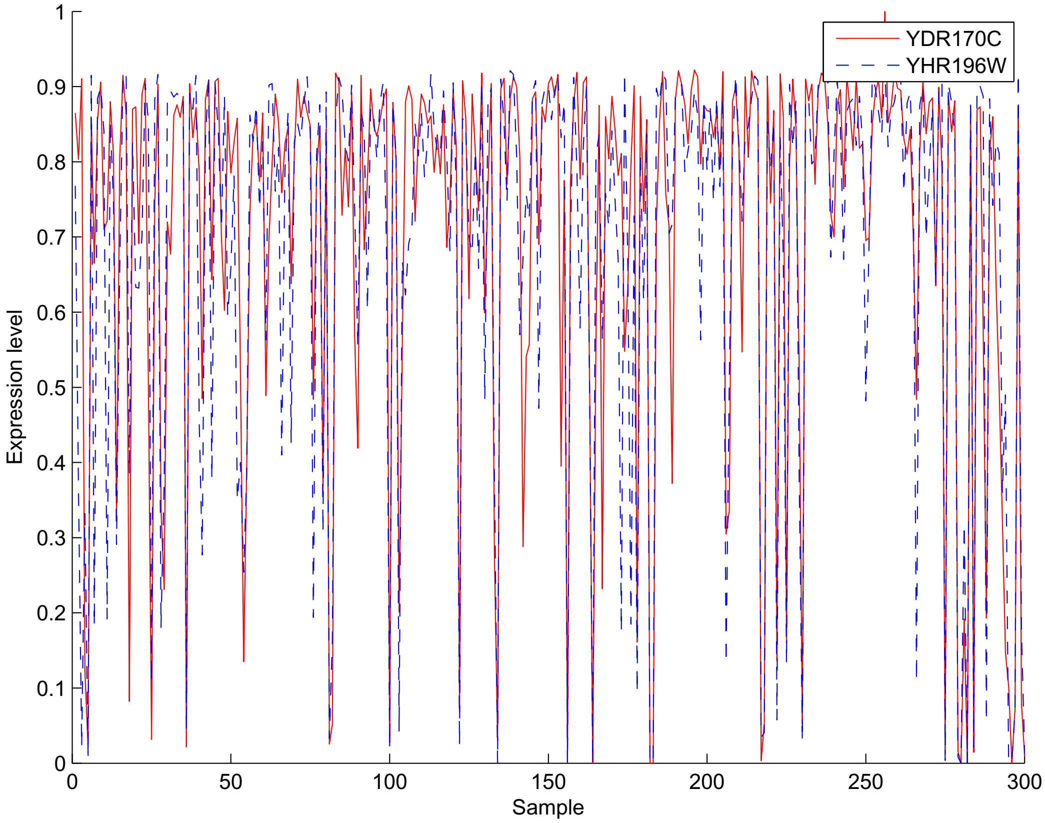


Figure 8.13. Expression levels in the treatment data of two genes that share a known Synthetic Lethality interaction and for which both the correlation and the mutual information are high. Namely, the correlation has a value of 0.858660 and the mutual information has a value of 0.347460. The correlation between these two genes is greater than the correlation of 99.7% of all pairs of genes in the data. The mutual information between these two genes is greater than the mutual information between 99.9% of all pairs of genes in the data.

TABLE 8.4

THE NUMBER OF KNOWN INTERACTIONS OF A GIVEN TYPE AND THE NUMBER OF GENES FROM EACH OF THE TWO DATA SETS THAT ARE INVOLVED IN THE CORRESPONDING INTERACTIONS.

Known interaction type	Number of interactions	Number of genes from line cross data	Number of genes from treatment data
Affinity Capture-Luminescence	32	11	11
Affinity Capture-MS	58,861	3,972	4,123
Affinity Capture-RNA	6,961	3,044	3,170
Affinity Capture-Western	12,165	2,383	2,470
Biochemical Activity	9,447	1,820	1,888
Co-purification	2,238	871	898
Co-crystal Structure	324	220	224
Co-fractionation	1,120	557	572
Co-localization	719	353	366
Dosage Growth Defect	476	258	269
Dosage Lethality	1,128	505	521
Dosage Rescue	6,554	1,930	1,999
Far Western	100	74	80
FRET	194	90	92
PCA	8,569	1,474	1,521
Phenotypic Enhancement	7,959	1,885	1,952
Phenotypic Suppression	5,672	1,363	1,406
Positive Genetic	24,810	2,806	2,915
Protein-RNA	772	380	391
Protein-peptide	208	112	116
Reconstituted Complex	3,996	1,374	1,421
Synthetic Growth Defect	26,944	2,869	2,974
Synthetic Haploinsufficiency	396	199	202
Synthetic Lethality	20,899	2,691	2,792
Synthetic Rescue	5,606	1,616	1,673
Two-hybrid	13,863	3,152	3,268

The number of known interactions of a given type and the number of genes from each of the two data sets that are involved in the corresponding interactions. There are 5,829 shared genes between the two data sets, with a total of 5,913 genes in the line cross data and a total of 6,207 genes in the treatment data.

CHAPTER 9

INTEGRATING HETEROGENEOUS NOISY DATA WITH DOMAIN EXPERTISE TO IMPROVE DATA UNDERSTANDING

As discussed in Chapter 7, many recently developed models for systems biology depend on the integration of heterogeneous data to increase predictive ability or the confidence of general biological knowledge. As we have shown in Chapters 3 and 8, it is also important to carefully consider the way in which relationships are measured. Here we present an approach that benefits from exploratory analysis and expert use of domain knowledge. It utilizes both heterogeneous data and domain-knowledge guided measurement of features. This work is a strong piece of evidence in support of our view that good data science comes from the fusion of domain knowledge and data mining. In this chapter we discuss an approach that became the winning entry to the 2011 Dialogue on Reverse Engineering Assessment Methods (DREAM) challenge.

The DREAM challenge is an annual competition in which cutting edge approaches to open problems in computational biology are pitted against each other [137]. One of the challenges for 2011 was prediction of promoter activity based on promoter sequence. Promoters are sections of DNA that occur near genes. Proteins responsible for the translation of DNA rely on promoter sequences to signal the beginning of a region that should be transcribed.

The relationship between promoter sequences and promoter activity or gene expression is not well understood. The quantitative prediction of transcriptional activity of genes using sequence information of promoters is fundamental to the understanding and engineering of biological systems. Such technology could allow the estimation of expression levels from DNA alone.

Competing teams were provided 90 donated promoters with corresponding expression levels and challenged to predict the expression of 53 held out promoters. We used a data driven approach to this problem, in which we collected multiple heterogeneous data types describing the promoter sequences. We considered a number of variables which collaborating biologists believed to be related to gene expression, including:

1. Length of the supplied promoter sequence.
2. K-mers (sequences of DNA of length k) of length one through five.
3. Mean length of k-mers composed solely of one type of nucleotide.
4. The standard deviation of the length of k-mers composed solely of one type of nucleotide.
5. Physical properties of the promoter sequences, including predicted protein deformability.

The total number of features considered was 1,641, with k-mers accounting for the vast majority. The use of a simple linear regression wrapper identified 22 important features including 18 important k-mers.

One notable absence from our feature list was transcription factor binding sites. Transcription factors and their binding sites have received much attention in the literature and have been used as a feature in many models for systems biology [139, 159]. They represent known relationships between proteins with regulatory effects and the DNA sequences that they attach to. Our success without this information suggests that other aspects of DNA sequences are important to gene expression or that much of the same information is captured by the features we used. The k-mers retained as features may also have captured the relevant transcription factor binding sites.

We trained an ensemble of 1000 support vector regressors (SVRs) on random samples of 80% of the training data using the features selected by a correlation based wrapper. Each SVR provided predictions for every instance in the test set and their predictions were averaged. By training models on random samples of the training data we accounted for the possibility that the model does could make bad predictions due to a trend in the test set that is underrepresented in the training set.

The methodology that went into building a successful model was a straightforward application of data mining, but the experimentation that allowed our model to succeed additionally lead to a better understanding of data that was incidental to the target problem.

We suspected that the entire length of the given promoter sequences may not be relevant to the promoter activity. After experimentation, we discovered that the 100 nucleotides closest to the gene held most of the information relevant to physical deformability of the DNA. We trained our model on this subset of the promoter sequences to achieve better performance. The relevance of 100 nucleotides closest to the gene was previously unknown, as demonstrated by the fact that other top placing teams used features derived from the entire length of the promoter sequences. Thus our success was enabled by expert guided feature extraction.

CHAPTER 10

CONCLUSION AND FUTURE WORK

The thesis of this dissertation was that the application of data and network science to challenges in the domains of systems biology and healthcare must carefully utilize domain knowledge. Throughout this work we have repeatedly demonstrated the validity and utility of this view. Through this process of studying data and network science in these domains we have brought forth open challenges, innovated novel methods, and highlighted important areas for future work.

Learning models that can be interpreted by domain experts from noisy and heterogeneous data is a fundamental task to the development of many developing fields. We explored how the use of domain knowledge is essential to producing a good model in Chapters 3, 4, and 8. We showed that different approaches may be appropriate depending on the level of domain knowledge of the problem and on the purpose of the analysis (Chapters 3 and 4). We demonstrated that the current approaches to integrating biological data into unified models in Chapter 7 still need to account for the effect of the measures and algorithms used to determine interactions (Chapter 8).

A second focus of this work was on the use of multiple measures for relationships in networks and other explicitly relational models. We gave context for this issue in Chapter 8 with a survey of current integrative network models in systems biology. We discussed how the use of single measures may be good for identifying specific information but is unreliable for heterogeneous data in Chapter 8. In Chapter 4 we offered a novel approach to integrate diverse distance measures for the exploratory analysis of biological data.

We also studied a common source of bias in classifiers in Chapter 6. Although we studied this problem in the context of classification, mislabeled negative class instances may be problematic when included as features in other heterogeneous models such as those described in Chapter 7.

Finally, in Chapter 9 we demonstrated how cutting edge data mining comes from the fusion of domain knowledge and data mining expertise. Understanding fundamental concerns about the data and domain leads to informed choices in modeling.

10.1 Future work

Several areas of future work are motivated by the work in this dissertation.

In Chapter 3 we proposed a way to identify interesting genetic associations that measured the behavior of genes more generally than the standard approach of simply measuring the strength of association globally. It may be valuable in general to try alternative measures for significance of genetic interactions, as it is now commonly believed that much of gene activity emerges from collections of genes working together rather than from fewer strong interactions [122, 18].

The joint approach using expert knowledge and expression data to learn an ensemble similarity measure in Chapter 4 could benefit from additional study. An ensemble similarity measure learned from more data types in addition to more similarity measures might produce even better results. Furthermore, this approach could trivially be extended to learn networks and other models.

In Chapter 5 we proposed a framework for integrating specialized knowledge from distinct models trained on distinct data sets into an interpretable ensemble. Our approach focused on populations of homogeneous data but it could easily be adapted for use on other relational data such as data in systems biology.

Our study of the effect of mislabeled negative class instances on classification in

Chapter 6 might have more impact in the field of systems biology if it were extended to include the evaluation of unsupervised methods and networks specifically.

Many of the approaches described in Chapter 7 show that the use of diverse data is an effective strategy to improve models in systems biology. We investigated the effects of using different combinations of data and similarity measures in Chapter 8. Our findings showed that more attention on the use of diverse models (relying on different similarity measures) may be a fruitful avenue of future research. Questions remain about how and why specific combinations of data type and similarity measure may result in networks with specific specializations. It seems likely that one underlying cause is the bias in a given method for taking biological measurements. This requires study in greater depth to ensure that computational models trained on these data are not systematically biased.

Our work on the DREAM project as described in Chapter 9 underscores the necessity of combining domain expertise with data mining. Feature extraction was essential to the success of this project and expert-guided feature extraction may have much more potential for the improvement of computational models. It also opens up the possibility of discovering much more about the relationship between physical properties of promoters and gene expression as the systems biology community is now aware that the relationship may be much more specific than expected.

BIBLIOGRAPHY

1. A. M. Algra and P. M. Rothwell. Effects of regular aspirin on long-term cancer incidence and metastasis: a systematic comparison of evidence from observational studies versus randomised trials. *The lancet oncology*, 2012.

2. P. D. Allison. Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, 55:193–196, 2002.

3. O. Aparicio, J. V. Geisberg, E. Sekinger, A. Yang, Z. Moqtaderi, and K. Struhl. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 21, Feb. 2005.

4. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, May 2000.

5. C. Baigent, L. Blackwell, R. Collins, J. Emberson, J. Godwin, R. Peto, J. Buring, C. Hennekens, P. Kearney, T. Meade, et al. Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials. *Lancet*, 373(9678):1849, 2009.

6. A.-L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, Feb. 2004.

7. G. Bebek, M. Koyutürk, N. D. Price, and M. R. Chance. Network biology methods integrating biological data for translational science. *Briefings in Bioinformatics*, Mar. 2012.

8. C. Blake and C. J. Merz. Uci repository of machine learning databases. 1998.

9. E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, Dec. 2004.

10. B.-J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D. H. Lackner, J. Bähler, V. Wood, K. Dolinski, and M. Tyers. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Research*, 36 (suppl 1):D637–D640, Jan. 2008.

11. R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–1577, Feb. 2005.

12. R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science*, 296(5568):752–755, Apr. 2002.

13. A. Butte and I. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Pac Symp Biocomput*, volume 5, pages 418–429, 2000.

14. A. Califano, A. Butte, S. Friend, T. Ideker, and E. E. Schadt.

15. R. D. Canales, Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, et al. Evaluation

of dna microarray results with quantitative gene expression platforms. *Nature biotechnology*, 24(9):1115–1122, 2006.

16. M. R. Carlson, B. Zhang, Z. Fang, P. S. Mischel, S. Horvath, and S. F. Nelson. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, 7, 2006.

17. N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. Learning Ensembles from Bites: A Scalable and Accurate Approach. *Journal of Machine Learning Research*, 5:421–451, Apr. 2004.

18. Y. Chen, J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts, et al. Variations in dna elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, 2008.

19. E. J. Chesler, L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. C. Hsu, J. D. Mountz, N. E. Baldwin, M. A. Langston, D. W. Threadgill, K. F. Manly, and R. W. Williams. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, 37(3):233–242, Feb. 2005.

20. M. J. Choi, V. Y. Tan, A. Anandkumar, and A. S. Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1729–1770, 2011.

21. N. A. Christakis and P. D. Allison. Mortality after the hospitalization of a spouse. *New England Journal of Medicine*, 354(7):719–730, 2006.

22. K. R. Christie, E. L. Hong, and J. M. Cherry. Functional annotations for the Saccharomyces cerevisiae genome: the knowns and the known unknowns. *Trends in Microbiology*, 17(7):286–294, July 2009.

23. G. A. Churchill and R. W. Doerge. Empirical Threshold Values for Quantitative Trait Mapping. *Genetics*, 138(3):963–971, Nov. 1994.

24. D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, pages 1–23, 2012.

25. D. D. Cieslak and N. V. Chawla. A framework for monitoring classifiers' performance: when and why failure occurs? *Knowledge and Information Systems*.

26. A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661, Feb. 2009.

27. M. W. Covert, E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96, May 2004.

28. M. E. Cowen, D. J. Dusseau, B. G. Toth, C. Guisinger, M. W. Zodet, and Y. Shyr. Casemix adjustment of managed care claims data using the clinical classification for health policy research method. *Medical care*, 36(7):1108–1113, 1998.

29. B. J. Daigle, A. Deng, T. McLaughlin, S. W. Cushman, M. C. Cam, G. Reaven, P. S. Tsao, and R. B. Altman. Using pre-existing microarray datasets to increase experimental power: application to insulin resistance. *PLoS computational biology*, 6(3):e1000718, Mar. 2010.

30. S. Datta and S. Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466, Mar. 2003.

31. S. Datta and S. Datta. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, 7(1):397, Aug. 2006.

32. D. A. Davis and N. V. Chawla. Exploring and Exploiting Disease Interactions from Multi-Relational Gene and Phenotype Networks. *PLoS ONE*, 6(7):e22670, July 2011.

33. D. A. Davis, N. V. Chawla, N. A. Christakis, and A.-L. Barabási. Time to care: a collaborative engine for practical disease prediction. *Data Mining and Knowledge Discovery*, 20(3):388–415, 2010.

34. J. Davis and M. Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

35. R. de Matos Simoes and F. Emmert-Streib. Influence of Statistical Estimators of Mutual Information and Data Heterogeneity on the Inference of Gene Regulatory Networks. *PLoS ONE*, 6(12):e29279, Dec. 2011.

36. R. De Smet and K. Marchal. Advantages and limitations of current network inference methods. *Nat Rev Micro*, 8(10):717–729, Oct. 2010.

37. J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, Dec. 2006.

38. M. Deng, T. Chen, and F. Sun. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol*, 11(2-3):463–475, 2004.

39. E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic intelligence for the crowd, by the crowd (full version). *arXiv preprint arXiv:1203.1378*, 2012.

40. S. Doss, E. E. Schadt, T. A. Drake, and A. J. Lusis. Cis-acting expression quantitative trait loci in mice. *Genome Research*, 15(5):681–691, May 2005.

41. C. Drummond and R. C. Holte. Explicitly representing expected cost: an alternative to ROC representation. In *Knowledge Discovery and Data Mining*, pages 198–207, 2000.

42. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, Dec. 1998.

43. C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 213–220. ACM, 2008.

44. J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph. Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, 3(1), Jan. 2007.

45. J. Ernst, Q. K. Beg, K. A. Kay, G. Balázsi, Z. N. Oltvai, and Z. Bar-Joseph. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS computational biology*, 4(3):e1000044, Mar. 2008.

46. B. S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Wiley, 4th edition, Jan. 2001.

47. J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biology*, 5(1):e8, Jan. 2007.

48. G. Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305, 2003.

49. S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

50. L. Freeman. A set of measures of centrality. *Sociometry*, 40(1):35–41, 1977.

51. J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

52. N. Friedman and I. Nachman. Gaussian process networks.

53. N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620, Aug. 2000.

54. H. Frohlich, N. Speer, A. Poustka, and T. BeiSZbarth. GOSim - an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, 8(1):166, May 2007.

55. F. Gao, B. Foat, and H. Bussemaker. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5(1):31, Mar. 2004.

56. F. D. Gibbons and F. P. Roth. Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Research*, 12(10):1574–1581, Oct. 2002.

57. A. Gitter, Z. Siegfried, M. Klutstein, O. Fornes, B. Oliva, I. Simon, and Z. Bar-Joseph. Backup in gene regulatory networks explains differences between binding and knockout results. *Molecular systems biology*, 5(1), 2009.

58. M. Goadrich, L. Oliphant, and J. Shavlik. Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction.

59. J. M. Gonzales, J. J. Patel, N. Ponmee, L. Jiang, A. Tan, S. P. Maher, S. Wuchty, P. K. Rathod, and M. T. Ferdig. Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS biology*, 6(9):e238, Sept. 2008.

60. U. S. Government. Health insurance portability and accountability act. *45 CFR 164.514*, 1996.

61. J. Grzymala-Busse and M. Hu. A comparison of several approaches to missing attribute values in data mining. In *Rough sets and current trends in computing*, pages 378–385. Springer, 2001.

62. M. Gustafsson, M. Hornquist, and A. Lombardi. Constructing and analyzing a large-scale gene-to-gene regulatory network lasso-constrained inference and biological validation. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 2(3):254–261, 2005.

63. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *Special Interest Group on Knowledge Discovery and Data Mining Explorer Newsletter*, 11(1):10–18, Nov. 2009.

64. J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, Aug. 2005.

65. D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(suppl 1):S145–S154, July 2002.

66. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 437–449, 2002.

67. J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

68. M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models–A review. *Biosystems*, 96(1):86–103, 2009.

69. D. Heckerman. A tutorial on learning with bayesian networks. *Innovations in Bayesian Networks*, pages 33–82, 2008.

70. J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, Feb. 2005.

71. H. Ho, T. Milenković, V. Memišević, J. Aruri, N. Pržulj, and A. K. Ganesan. Protein interaction network uncovers melanogenesis regulatory network components within functional genomics datasets. *BMC Systems Biology*, 4(84), 2010.

72. T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, Aug. 1998.

73. Y. Huang, S. Wuchty, M. T. Ferdig, and T. M. Przytycka. Graph theoretical approach to study eQTL: a case study of Plasmodium falciparum. *Bioinformatics (Oxford, England)*, 25(12):i15–20, June 2009.

74. T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, July 2000.

75. D. Hwang, A. G. Rust, S. Ramsey, J. J. Smith, D. M. Leslie, A. D. Weston, P. De Atauri, J. D. Aitchison, L. Hood, A. F. Siegel, et al. A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17296, 2005.

76. S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proceedings / IEEE Computer Society Bioinformatics Conference. IEEE Computer Society Bioinformatics Conference*, 2:104–113, 2003.

77. A. E. Ivliev, P. AC't Hoen, and M. G. Sergeeva. Coexpression network analysis identifies transcriptional modules related to proastrocytic differentiation and sprouty signaling in glioma. *Cancer Research*, 70(24):10060–10070, 2010.

78. R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*, 302 (5644):449–453, Oct. 2003.

79. R. C. Jansen and J.-P. Nap. Genetical genomics: the added value from segregation. *Trends in Genetics*, 17(7):388–391, July 2001.

80. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct. 2000.

81. M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue):D109–D114, Jan. 2012.

82. I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res*, 33(Database issue), Jan. 2005.

83. H. Kim, W. Hu, and Y. Kluger. Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae. *BMC Bioinformatics*, 7(1):165, 2006.

84. S. Kok and P. Domingos. Learning the structure of Markov logic networks. In *Proceedings of the 22nd international conference on Machine learning*, pages 441–448, New York, NY, USA, 2005. ACM.

85. T. C. Landgrebe, P. Paclik, R. P. Duin, and A. P. Bradley. Precision-recall operating characteristic (P-ROC) curves in imprecise environments. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 4, pages 123–127. IEEE, 2006.

86. I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A Probabilistic Functional Network of Yeast Genes. *Science*, 306(5701):1555–1558, Nov. 2004.

87. K. Lemmens, T. De Bie, T. Dhollander, S. De Keersmaecker, I. Thijs, G. Schoofs, A. De Weerdt, B. De Moor, J. Vanderleyden, J. C. Vides, K. Engelen, and K. Marchal. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. *Genome Biology*, 10(3):R27, Mar. 2009.

88. C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

89. S. Li, L. Hsu, J. Peng, and P. Wang. Bootstrap Inference for Network Construction. *Arxiv preprint arXiv:1111.5028*, 2011.

90. W. Li, S. Zhang, C.-C. Liu, and X. J. Zhou. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, on line, 2012.

91. D. Lin. An Information-Theoretic Definition of Similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

92. R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 4. Wiley New York, 2002.

93. R. D. Luce and A. D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

94. N. M. Luscombe, M. Madan Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312, Sept. 2004.

95. D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences*, 107(14):6286–6291, Apr. 2010.

96. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

97. F. Markowetz and R. Spang. Inferring cellular networks–a review. *BMC bioinformatics*, 8 Suppl 6(Suppl 6):S5, 2007.

98. M. Mason, G. Fan, K. Plath, Q. Zhou, and S. Horvath. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*, 10(1):327, 2009.

99. D. Maxwell Chickering and D. Heckerman. Efficient Approximations for the

Marginal Likelihood of Bayesian Networks with Hidden Variables. *Machine Learning*, 29(2):181–212, Nov. 1997.

100. P. Meyer, F. Lafitte, and G. Bontempi. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics*, 9(1):461, 2008.

101. T. Milenković, J. Lai, and N. Przulj. GraphCrunch: A tool for large network analyses. *BMC Bioinformatics*, 9(1):70, 2008.

102. T. Milenković, V. Memišević, A. Bonato, and N. Pržulj. Dominating biological networks. *PLOS ONE*, 6(8):e23016, 2011.

103. modENCODE Consortium, S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White, and M. Kellis. Identification of functional elements and regulatory

circuits by Drosophila modENCODE. *Science (New York, N.Y.)*, 330(6012): 1787–1797, Dec. 2010.

104. C. Myers, D. Barrett, M. Hibbs, C. Huttenhower, and O. Troyanskaya. Finding function: evaluation methods for functional genomic data. *BMC Genomics*, 7 (1):187, July 2006.

105. R. R. Nayak, M. Kearns, R. S. Spielman, and V. G. Cheung. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome research*, 19(11):1953–1962, Nov. 2009.

106. R. r. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

107. C. J. Needham, J. R. Bradford, A. J. Bulpitt, and D. R. Westhead. A primer on learning in Bayesian networks for computational biology. *PLoS computational biology*, 3(8):e129, Aug. 2007.

108. S. J. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

109. G. Pandey, B. Zhang, A. N. Chang, C. L. Myers, J. Zhu, V. Kumar, and E. E. Schadt. An Integrative Multi-Network and Multi-Classifier Approach to Predict Genetic Interactions. *PLoS Comput Biol*, 6(9):e1000928, Sept. 2010.

110. E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

111. M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.

112. D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17(suppl 1):S215–S224, June 2001.

113. H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, Aug. 2005.

114. P. Pons and M. Latapy. Computing Communities in Large Networks Using Random Walks Computer and Information Sciences - ISCIS 2005. In p. Yolum, T. Güngör, F. Gürgen, and C. Özturan, editors, *Computer and Information Sciences - ISCIS 2005*, volume 3733 of *Lecture Notes in Computer Science*, chapter 31, pages 284–293. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2005.

115. F. Provost, T. Fawcett, R. Kohavi, et al. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the fifteenth international conference on machine learning*, volume 445, 1998.

116. Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3):490–500, 2006.

117. J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32(supp):496, 2002.

118. J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

119. A. Rao, A. Hero 3rd, and J. D. Engel. Using directed information to build biologically relevant influence networks. *Computational systems bioinformatics*

/ Life Sciences Society. Computational Systems Bioinformatics Conference, 6: 145–156, 2007.

120. A. K. Rider, G. H. Siwo, S. J. Emrich, M. T. Ferdig, and N. V. Chawla. A supervised learning approach to the unsupervised clustering of genes. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 322–328. IEEE.

121. G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, 1966.

122. E. E. Schadt, S. H. Friend, and D. A. Shaywitz. A network view of disease and compound screening. *Nature Reviews Drug Discovery*, 8(4):286–295, Apr. 2009.

123. S. Schappert and E. Rechtsteiner. Ambulatory medical care utilization estimates for 2007. *Vital and Health Statistics. Series 13, Data from the National Health Survey*, (169):1, 2011.

124. M. R. Segal, K. D. Dahlquist, and B. R. Conklin. Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10:961–980, 2003.

125. J. Sethuraman. A constructive definition of dirichlet priors. Technical report, DTIC Document, 1991.

126. B. Shahbaba and R. Neal. Nonlinear models using dirichlet process mixtures. *The Journal of Machine Learning Research*, 10:1829–1850, 2009.

127. R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(1), Mar. 2007.

128. G. C. Y.-D. Z. Sheng Tang, Yan-Tao Zheng and J.-T. Li. Ensemble learning with lda topic models for visual concept detection, multimedia - a multidisciplinary approach to complex issues. ISBN 978-953-51-0216-8. doi: 10.5772/37716.

129. T. Shimamura, S. Imoto, R. Yamaguchi, and S. Miyano. Weighted lasso in graphical gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics*, 19:142–153, 2007.

130. Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

131. E. N. Smith and L. Kruglyak. Gene-environment interaction in yeast gene expression. *PLoS biology*, 6(4):e83, Apr. 2008.

132. P. H. Sneath and R. R. Sokal. Numerical taxonomy. *Nature*, 193:855–860, Mar. 1962.

133. R. W. Solava, R. P. Michaels, and T. Milenković. Graphlet-based edge clustering reveals pathogen-interacting proteins. *Bioinformatics*, 18(28):i480–i486, 2012. Also, in Proceedings of the *11th European Conference on Computational Biology* (ECCB), Basel, Switzerland, September 9-12, 2012 (acceptance rate: 14%).

134. N. Srebro. Maximum likelihood bounded tree-width markov networks. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 504–511. Morgan Kaufmann Publishers Inc., 2001.

135. C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1):D535–D539, Jan. 2006.

136. R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2):S231–S240, Oct. 2002.

137. G. Stolovitzky, D. O. N. Monroe, and A. Califano. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences*, 1115(1):1–22, Dec. 2007.

138. D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(Database issue):D561–D568, Nov. 2010.

139. Y. Tamada, S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara, and S. Miyano. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 19 Suppl 2, Oct. 2003.

140. A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2981–2986, Mar. 2004.

141. S. Taner, K. Andrzej, and J. Robert. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics*, 7(1):355, July 2006.

142. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

143. Z. Tu, L. Wang, M. N. Arbeitman, T. Chen, and F. Sun. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*, 22(14):e489–e496, July 2006.

144. B. A. Turlach. Bandwidth Selection in Kernel Density Estimation: A Review.

145. D. Ucar, I. Neuhaus, P. Ross-MacDonald, C. Tilford, S. Parthasarathy, N. Siemers, and R.-R. Ji. Construction of a Reference Gene Association Network from Multiple Profiling Data: Application to Data Analysis. *Bioinformatics*, page btm423, Sept. 2007.

146. D. Ucar, A. Beyer, S. Parthasarathy, and C. T. Workman. Predicting functionality of protein-DNA interactions by integrating diverse evidence. *Bioinformatics*, 25(12):i137–144, June 2009.

147. M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *The New England journal of medicine*, 347(25):1999–2009, Dec. 2002.

148. V. van Noort, B. Snel, and M. A. Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports*, 5(3):280–284, Mar. 2004.

149. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan. 2002.

150. M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using l˜ 1-regularized logistic regression. *Advances in neural information processing systems*, 19:1465, 2007.

151. H. Wang, H. Shan, and A. Banerjee. Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, 4(1):54–70, 2011.

152. J. H. Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, Mar. 1963.

153. E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. E. Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Véronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, and R. W. Davis. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science (New York, N.Y.)*, 285(5429):901–906, Aug. 1999.

154. T. Wittkop, J. Baumbach, F. P. Lobo, and S. Rahmann. Large scale clustering of protein sequences with force-a layout based heuristic for weighted cluster editing. *BMC bioinformatics*, 8(1):396, 2007.

155. L. Yang, Q. Mei, K. Zheng, and D. A. Hanauer. Query log analysis of an electronic health record search engine. In *AMIA Annual Symposium Proceedings*, volume 2011, page 915. American Medical Informatics Association, 2011.

156. A. Yip and S. Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8(1):22, Jan. 2007.

157. S. Zhang, C. Zhang, and Q. Yang. Data preparation for data mining. *Applied Artificial Intelligence*, 17(5-6):375–381, 2003.

158. X. Zhou, M.-C. C. Kao, and W. Hung. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A*, 99 (20):12783–12788, Oct. 2002.

159. J. Zhu, B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner, and E. E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40(7): 854–861, June 2008.

This document was prepared & typeset with LaTeX 2$_\varepsilon$, and formatted with NDdiss2$_\varepsilon$ *classfile (v3.2013β[2013/01/31]) provided by Sameer Vijay and updated by Megan Patnott.*