Investigation of Gene Regulation and its Application to Disease Using Machine Learning and Network Models

BY

MATTHEW B. CARSON
B.Sc., University of Oklahoma, 2000

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Defense Committee:
Hui Lu, Chair and Advisor
Yang Dai
Jie Liang
Warren Kibbe, Northwestern University
Caiyan Jia, Beijing Jiaotong University

This work is dedicated to my father, Robert T. Carson, and my grandfathers, Dr. Clarence M. Parker and William J. Camp, for their inspiration and guidance throughout my life, each in their own way.

ACKNOWLEDGMENT

I would first like to thank my advisor, Dr. Hui Lu, for his direction and insight during my time as a graduate student. His encouragement and wisdom was immeasurably helpful, and many interesting avenues were explored because of his enthusiasm for science and willingness to think outside of the box. Professor Lu's uncanny ability to ask the right questions kept my focus on the "big picture" and always provided a unique and valuable perspective.

I would also like to thank my dissertation committee members Yang Dai, Jie Liang, Warren Kibbe, and Caiyan Jia for their guidance and advice concerning my dissertation and its contents. In addition, I would like to express my gratitude to Dr. R. John Solaro for his generous funding toward my research (NIH Training Grant Fellowship # 5 T32 HL 07692-19, "Training in Cellular Signaling in the Cardiovascular System", P.I.: R. John Solaro, Ph.D.).

I gratefully acknowledge the members of my lab past and present for their camaraderie and help with the projects included in my thesis: Dr. Robert Langlois, Dr. Nitin Bhardwaj, Dr. Guijun Zhao, Georgi Genchev, Dr. Morten Källberg, Dr. Gang Feng, Dr. Ognjen Perišić, Xishu Wang, Wenyi Qin, Cong Liu, Gamze Gursoy, Dr. Joe Dundas, Dr. Lei Huang, Gobind Singh, Donna Esbjornson, Anuradha Mittal, Rima Chaudhuri, and Adam Carlson. In particular, I'd like to acknowledge Rob for his collaboration and for sharing his knowledge on the subject of machine learning. I would also like to thank Nitin for his teamwork and for showing me the ropes when I was a fledgling graduate student.

ACKNOWLEDGMENT (Continued)

To my parents, Robert and Judi Carson, I extend my deepest appreciation for their love, encouragement, and belief that I had it in me to pursue this degree. I would also like to express my gratitude to my sister and brother, Linsey and Eric Carson, and my grandparents, Dr. Clarence and Mrs. Peggy Parker, and Bill and Mary Camp for their unwavering support.

Finally, to my wife, Suzanne Pham, without whom I would be lost...thank you for your optimism, for your inspiration, for keeping me going, and for preventing me from going insane. I love you and I couldn't have done this without you.

MBC

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC4.5	Adaboost C4.5
ACC	Accuracy
AdaBoost	Adaptive Boosting
ADTree	Alternating Decision Tree
ASA	Solvent-accessible Surface Area
AUC	Area Under the ROC Curve
C4.5ADA	C4.5 with AdaBoost
C4.5BAG	C4.5 with Bootstrap Aggregation
C4.5BAGCST	C4.5 with Bootstrap Aggregation and Costing
CpG	Cytosine Phosphate Guanine (dinucleotide)
CpGI	CpG Island
CV	Cross-validation
DO	Disease Ontology
DOLite	Disease Ontology Lite
DO	Disease Ontology
DORIF	Disease Ontology + GeneRIFs
DOR	Dense Overlapping Regulons

LIST OF ABBREVIATIONS (Continued)

DNA	Deoxyribonucleic Acid
FPR	False Positive Rate
GeneRIF	Gene Reference Into Function
LOO CV	Leave-one-out Cross-validation
MCC	Matthews Correlation Coefficient
miRNA	microRNA
mRNA	Messenger RNA
MSA	Multiple Sequence Alignment
n-CV	n-fold Cross-validation
NA-binding	Nucleic Acid-binding
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PRE	Precision
PSSM	Position-specific Scoring Matrix
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SEN	Sensitivity
SCM	Side Chain Center of Mass

LIST OF ABBREVIATIONS (Continued)

siRNA	Small Interfering RNA
SNP	Single-nucleotide Polymorphism
SPE	Specificity
SVM	Support Vector Machines
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TG	Target Gene
TPR	True Positive Rate

SUMMARY

Gene regulation is one of the most important functions in the cell. This process is responsible for differentiation within tissues of a single entity as well as diversity between organisms. It allows for adaptation to stimuli in the environment and for more efficient use of environmental resources. Changes in the way a gene is regulated can either increase the versatility and adaptability of an organism or it can be detrimental. Cells have mechanisms to deal with gene deletions and mutations while accepting beneficial changes. In this work we study gene regulation from three perspectives.

Firstly, we focus on nucleic acid-binding prediction, both on the protein and the residue levels. We predict DNA-binding proteins with 88% accuracy. On the residue level, we create an ensemble classifier based on C4.5, bootstrap aggregation, and a cost-sensitive learning algorithm. We demonstrate that by using this method with a number of residue attributes, we obtain balanced sensitivity, specificity, and precision with high accuracy when training and testing on unbalanced data sets.

Secondly, we turn our attention toward DNA-binding sites and two current bioinformatics problems concerning protein binding: transcription factor binding site prediction and DNA methylation prediction. We define a knowledge-based binding potential for genome-wide transcription factor binding site prediction. In an application using this potential to search for 142 experimentally-determined binding sites for the CRP protein in *E. coli*, 59% of the true binding sites are ranked as one of the top 5 sites with the strongest potential. Next, we use an

alternating decision tree to shed light on sequence patterns that are characteristic of methylated CpG islands. We find highly discriminating rules that differentiate between methylated and non-methylated islands in human DNA.

Thirdly, we take a global perspective and study human molecular networks in the context of disease, focusing on transcription factor and protein-protein interaction networks. We examine the number of partnership interactions between transcription factors and how it scales with the number of target genes regulated. In several model organisms, we find that the distribution of the number of partners vs. the number of target genes appears to follow an exponential saturation curve. We also find that our generative transcriptional network model follows a similar distribution in this comparison. We then analyze human disease and its relationship to two molecular networks. We first identify conserved motifs in the human transcription factor network and pinpoint the location of disease- and cancer-related genes within these structures. We find that both cancer and disease genes occupy certain positions more frequently. Next, we examine the human protein-protein interaction (PPI) network as it relates to disease. We are able to predict disease genes with 79% area under the ROC curve (AUC) using ADTree with 10 topological features. Additionally, we find that a combination of several network characteristics including degree centrality and disease neighbor ratio help distinguish between these two classes. Furthermore, an alternating decision tree (ADTree) classifier allows us to see which combinations of strongly predictive attributes contribute most to protein-disease classification. Finally, we create a co-occurrence matrix for 1854 diseases based on shared gene uniqueness and find both previously known and potentially undiscovered disease relationships. This matrix

will be useful for making connections between diseases with very different phenotypes, or for those disease connections which may not be obvious. It could also be helpful in identifying new potential drug targets through drug repositioning.

CHAPTER 1

INTRODUCTION

In 1944 Erwin Schrödinger made the keen observation that all living beings share a common goal: to fight decay into equilibrium (144). This implies that what is alive must make a constant effort to resist destructive forces which would seek to reduce it to its lowest energy state. Life, then, is a struggle, and each living thing must constantly monitor, control, and modify its cellular components in order to fight this tendency to decay. Luckily, organisms are equipped with mechanisms that help to maintain internal stability, or *homeostasis*, and to adapt to changes in the environment. Attempts to understand these processes began with the discovery of a peculiar molecule.

1.1 The Discovery of DNA

The molecule we know today as deoxyribonucleic acid (DNA) was first observed in 1869 by Swiss biologist Friedrich Miescher, who stumbled upon a substance which was resistant to protein digestion. At the time he referred to the molecule as "nuclein" (137). Though Miescher remained in obscurity, Russian biochemist Phoebus Levene continued work with this substance and in 1919 discovered the three major components of a nucleotide: phosphate, sugar, and base. He noted that the sugar component was ribose for RNA and deoxyribose for DNA, and he proposed that nucleotides were made up of a chain of nucleic acids (109). He was largely correct, and in 1950 Erwin Chargaff, after reading a paper by Oswald Avery in which Avery

identified the gene as the unit of hereditary material (13), set out to discover whether the deoxyribonucleic acid molecule differed among species. In contrast to Levene's proposal that nucleotides are always repeated in the same order, he found that nucleotides appear in different orders in different organisms. These molecules did however maintain certain characteristics. This led him to develop a set of rules (known as "Chargaff's Rules") in which he states that the total number of purines (Adenine and Guanine) and the total number of pyrimidines (Cytosine and Thymine) are almost always equal in an organism's genetic material. In 1952 Rosalind Franklin and Maurice Wilkins used X-ray crystallography to capture the first image of the molecule's shape, and in 1953 James Watson and Francis Crick finally proposed the three dimensional model for DNA (178). The four main tenants of their discovery still hold true today: 1) DNA is a double-stranded helix, 2) the majority of these helices are right-handed, 3) the helices are anti-parallel, and 4) the DNA base pairs within the helix are joined by hydrogen bonding, and the bases can hydrogen bond with other molecules such as proteins.

1.2 The Central Dogma of Molecular Biology

The Central Dogma of Molecular Biology, first proposed by Francis Crick (50), describes the directional processes of conversion from DNA to RNA and from RNA to protein. The gene expression process starts with DNA, a double-stranded molecule consisting of base-paired nucleic acids adenine (A), cytosine (C), guanine (G), and thymine (T) on a sugar-phosphate backbone. This genetic material serves as the *information storage* for life, a dictionary of sorts that provides all of the necessary tools for an organism to create the components of itself. During the process of *transcription*, the DNA molecule is used to make messenger RNA

(mRNA), which carries a *specific instance* of the DNA instructions to the machinery that will make protein. Proteins are synthesized during *translation* using the mRNA molecule as a guide. Gene expression is a deterministic process during which each molecule is manufactured using the product of the previous step. The end result is a conversion from the genetic code into a functional unit which can be used to perform the work of the cell. As you can imagine, this process must be controlled by an organism in order to make efficient use of resources, respond to environmental changes, and differentiate cells within the body. Gene regulation, as it is sometimes called, occurs at all stages along the way from DNA to protein.

1.3 Types of Gene Regulation

Regulation falls into four categories: 1) epigenetic (methylation of DNA or protein, acetylation), 2) transcriptional (involves proteins called transcription factors), 3) post-transcriptional (sequestration of RNA, alternative splicing of mRNA, microRNA (miRNA) and small interfering RNA (siRNA)), and 4) post-translational modification (phosphorylation, acetylation, methylation, ubiquitination, etc. of protein products). Epigenetic regulation of DNA involves a reversible, heritable change that does not alter the sequence itself. DNA methylation occurs on the nucleic acid cytosine. Arginine and lysine are the most commonly methylated amino acids. When proteins called *histones* contain certain methylated residues, these proteins can repress or activate gene expression. Often this occurs on the transcriptional level, and thus prevents the cell from manufacturing messenger RNA (mRNA), the precursor to proteins. Proteins are often referred to as the workhorse of the cell and are responsible for everything from catalyzing chemical reactions to providing the building blocks for skeletal muscles. Some proteins, called

transcription factors, help to up- or down-regulate gene expression levels. These proteins can act alone or in conjunction with other transcription factors and bind to DNA bases near gene coding regions.

1.4 Motivation

What can be gained by studying gene regulation? In general, it allows us to understand how an organism evolves and develops, both on a local scale (42; 180), and on a more global network level (15; 162; 16; 72; 84). There are, however, more specific reasons to investigate this process closely. Failure in gene regulation has been shown to be a key factor in disease (157). Additionally, learning how to interrupt gene regulation may lead to the development of drugs to fight bacteria and viruses (124). A clearer understanding of this process in microorganisms may lead to possible solutions to the problem of antimicrobial resistance (49).

There are two major factors that motivate the studies herein. Firstly, the size and quality of biological data sets has increased dramatically in the last several years. This is due to high-throughput experimental techniques and technology, both of which have provided large amounts of interaction data, along with X-ray crystallography and nuclear magnetic resonance (NMR) experiments which have given us the solved three-dimensional structure of proteins. Secondly, machine learning has become an increasingly popular tool in bioinformatics research because it allows for more sound gene and protein annotation without relying solely on sequence similarity. If a collection of attributes which distinguish between two classes of proteins can be assembled, function can be predicted.

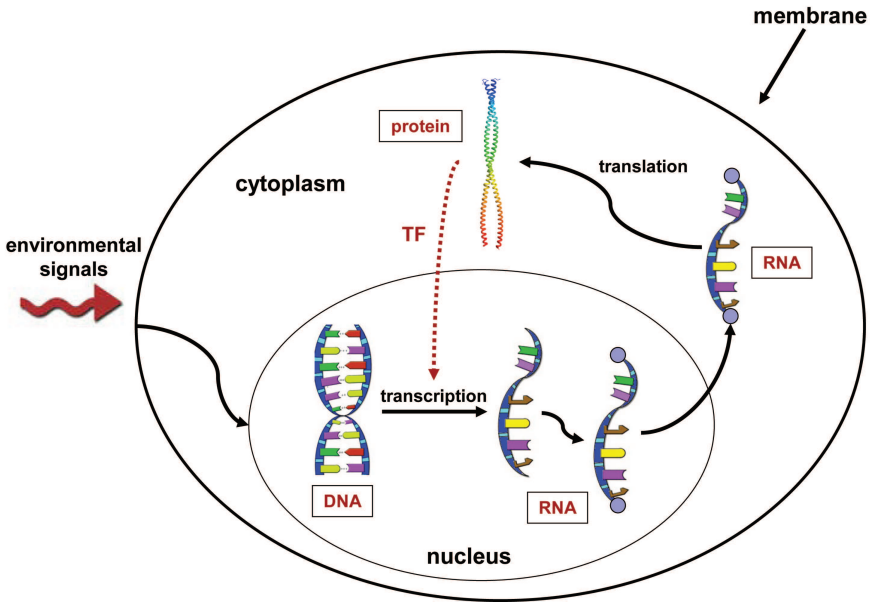


Figure 1: This is a general schema for gene expression. DNA is a double-stranded molecule consisting of base-paired nucleic acids A, C, G, and T on a sugar-phosphate backbone and is used as information storage. mRNA is made during transcription and carries a specific instance of the DNA instructions to the machinery that will make the protein. Proteins are synthesized during translation using the information in mRNA as a template. This is a deterministic process during which each molecule is manufactured using the product of the previous step. mRNA requires a 5' cap and a 3' poly(A) tail in order to be exported from the nucleus. The cap ensures that the mRNA will be recognized by the ribosome and protected from enzymes called RNases that will break down the molecule. The poly(A) tail helps protect it from degradation by enzymes called exonucleases. DNA: `http://www.biologycorner.com/bio1/DNA.html`, protein: `http://ucsdnews.ucsd.edu/graphics/images/2008/03-08M1protein.png`.

In this work we focus mainly on regulation at the transcriptional level and the components which play a commanding role in this operation. So-called nucleic acid-binding (NA-binding) proteins, a group which includes transcription factors, are involved in this and many other cellular processes. Disruption or malfunction of transcriptional regulation may result in disease. We identify these proteins from representative data sets which include many categories of proteins. Additionally, in order to understand the underlying mechanisms, we predict the specific residues involved in nucleic acid binding using machine learning algorithms. Identification of these residues can provide practical assistance in the functional annotation of NA-binding proteins. These predictions can also be used to expedite mutagenesis experiments, guiding researchers to the correct binding residues in these proteins.

1.5 Project Overview

Toward the ultimate goal of attaining a deeper understanding of how nucleic acid-binding proteins facilitate the regulation of gene expression within the cell, the research described here focuses on three particular aspects of this problem. We begin by examining the nucleic acid-binding proteins themselves, both on the protein and residue levels. Next, we turn our attention toward protein binding sites on DNA molecules and a particular type of modification of DNA that can affect protein binding. We then take a global perspective and study human molecular networks in the context of disease, focusing on regulatory and protein-protein interaction networks. We examine the number of partnership interactions between transcription factors and how it scales with the number of target genes regulated. In several model organisms, we find that the distribution of the number of partners vs. the number of target genes appears

to follow an exponential saturation curve. We also find that our generative transcriptional network model follows a similar distribution in this comparison. We show that cancer- and other disease-related genes preferentially occupy particular positions in conserved motifs and find that more ubiquitously expressed disease genes have more disease associations. We also predict disease genes in the protein-protein interaction network with 79% area under the ROC curve (AUC) using ADTree, which identifies important attributes for prediction such as degree and disease neighbor ratio. Finally, we create a co-occurrence matrix for 1854 diseases based on shared gene uniqueness and find both previously known and potentially undiscovered disease relationships.

1.5.1 Chapter 3: DNA-binding Protein Prediction

The goal for this project is to predict nucleic acid-binding on both the protein and residue levels using machine learning. Both sequence- and structure-based features are used to distinguish nucleic acid-binding proteins from non-binding proteins, and nucleic acid-binding residues from non-binding residues. A novel application of a costing algorithm is used for residue-level binding prediction in order to achieve high, balanced accuracy when working with imbalanced data sets.

1.5.1.1 Protein-level Prediction

In order for a protein to perform its function, it must interact with other molecules in the environment. Due to the ever-increasing number of crystallized protein structures available from databases such as the Protein Data Bank (PDB) (`http://www.rcsb.org`), a corresponding structural annotation approach to the identification of these interactions is warranted. Concur-

rently, machine learning has gained popularity in bioinformatics and provides robust annotation for genetic elements without relying solely on sequence homology. In this work we describe a machine learning protocol used to identify DNA-binding proteins. Because there is no general set of rules for choosing the best machine learning algorithm, we run a systematic comparison of several classification algorithms known to perform well over various protein data sets. We find that a tree classifier used in conjunction with a boosting algorithm gives the best performance, achieving 88% accuracy when discriminating between non-homologous DNA-binding proteins and non-binding proteins, significantly outperforming all previously published works.

This section is based on the following publication:

1. Langlois, **Carson**, Bhardwaj, and Lu. *Learning to translate sequence and structure to function: identifying DNA-binding and membrane binding proteins.* Annals of Biomedical Engineering. 2007 June; 35(6):1043-1052.

1.5.1.2 Protein Residue-level Prediction

In this section we present a method for the identification of amino acid residues involved in DNA and RNA binding using sequence-based attributes. This method, which combines the C4.5 algorithm with bootstrap aggregation and cost-sensitive learning, has been implemented in a publicly-available server. Given a protein sequence, the server returns a list of binding and non-binding residues within the protein along with a score, which measures the confidence in the prediction. In our prediction trials, the DNA-binding model achieved 79.1% accuracy, while the RNA-binding model reached an accuracy of 73.2%. In comparison with five other commonly used algorithms for NA-binding residue prediction, we find that our ensemble method performs

better for sequence-based prediction of both DNA and RNA binding. Additionally, we built our classifiers over several DNA- and RNA-binding protein data sets from previous works and found that we are able to improve on the previously published results. The NAPS web server is freely available at `http://bioinformatics.bioengr.uic.edu/NAPS`.

This section is based in part on the following publication:

1. **Carson**, Langlois, and Lu. *NAPS: A Residue-level Nucleic Acid-binding Prediction Server*. Nucleic Acids Research. 2010. July; 38:W431-5.

1.5.2 Chapter 4: DNA Binding Site and Methylation Prediction

This project has two parts. Firstly, we describe a method used to recognize the location on a DNA molecule at which regulating proteins bind. We develop interaction potentials which are able to assess protein-DNA interactions and test them on known binding sites of a transcription factor. Secondly, we predict the location of a particular type of epigenetic modification of DNA called *methylation* which can affect protein binding. We use an alternating decision tree (ADTree) to discover which DNA sequence patterns are characteristic of methylated CpG islands, and search for rules that differentiate between methylated and non-methylated CpG islands in human DNA.

1.5.2.1 DNA-binding Site Prediction Using Interaction Potentials

Gene regulation requires specific protein-DNA interactions. Detecting the short and variable DNA sequences in gene promoter regions to which transcription factors (TFs) bind is a difficult challenge in bioinformatics. Here we have developed two-body and three-body interaction

potentials that are able to assess protein-DNA interaction and achieve a higher level of specificity in the recognition of TF-binding sites. The potentials were calculated using experimentally characterized 3-D structures of protein-DNA complexes. We implemented two approaches in order to evaluate the potentials. Using the first method, we calculated the Z-score of the potential energy of a true TF-binding sequence when compared to 50,000 randomly generated DNA sequences. The second method allowed us to take advantage of the ability of statistical potentials to recognize novel TF-binding sites within the promoter region of genes. We found that the three-body potential, which takes into account the interaction between a DNA base and a protein residue with regard to the effect of a neighboring DNA base, had a better average Z-score than that of the two-body potential. This neighbor effect suggests that the local conformation of DNA does play a critical role in specific residue-base recognition. In all cases, the potentials developed here outperformed published results. The two sets of potentials were tested further by applying them in genome-scale TF-binding site prediction for the CRP protein in *E. coli*. Out of the 142 cases, 28% of the true binding sites ranked first (i.e., had the lowest Z-score), while in 59% of cases the true binding site ranked in the top 5. We show with these results that statistical potentials can be used in genome-scale TF-binding site prediction.

This section is based in part on the following publication:

1. Zhao, **Carson**, and Lu. *Prediction of specific protein-DNA recognition by knowledge-based two-body and three-body interaction potentials.* Engineering in Medicine and Biology Society, Proceedings of the 29th Annual International Conference of the IEEE EMBC. 22-26 August. 2007 Pages:5017-5020.

1.5.2.2 Predicting Methylation of CpG Islands within DNA

CpG island (CpGI) methylation is an epigenetic modification that occurs in eukaryotes and is based on the addition of a methyl group to the number 5 carbon of the pyrimidine ring of cytosine. When methylation of a CpGI occurs, the associated gene (if any) is not expressed (86). Aberrant methylation is thought to be a causative agent in disease (168) and drug sensitivity (158; 148). In this work, we have predicted the methylation status of CpGIs in human chromosome 21 using sequence patterns. These patterns showed a significantly different distribution between methylated and unmethylated islands in a previous work (30). Using C4.5 with bagging and cost-sensitive learning, we achieved 85.6% accuracy, 82.8% sensitivity, and 86.4% specificity. We then constructed 1000 alternating decision trees using a bootstrapping method and analyzed conserved nodes between the trees. This allowed us to find specific combinations of sequence patterns that distinguished between methylated and unmethylated CpGIs. Analysis of these characteristics offers certain insight into the conditions that permit or prevent methylation.

This section is based on the following publication:

1. **Carson**, Langlois, and Lu. *Mining Knowledge for the Methylation Status of CpG Islands Using Alternating Decision Trees.* Engineering in Medicine and Biology Society, Proceedings of the 30th Annual International Conference of the IEEE EMBC. 20-25 August. 2008 Pages:3787-3790.

1.5.3 Chapter 5: Gene Regulation and Network Models

Through combinatorial regulation, transcriptional regulators partner with each other to control common targets. This allows a small number of regulators (e.g., transcription factors) to control a large number of targets (e.g., target genes). How does the number of regulators scale with the number of targets? We answer this question by creating and analyzing co-regulation networks related to transcription and phosphorylation that describe partnerships between regulators controlling common genes. We perform analyses across six diverse species from *Escherichia coli* to *Homo sapiens*. These reveal many properties of partnership networks, such as the absence of a classical power-law degree distribution despite the existence of nodes with many partners. We also find that the number of co-regulatory partnerships follows an exponential saturation curve in relation to the number of targets, except for the prokaryotes *E. coli* and *B. subtilis* for which only the beginning linear part of the curves are evident due to arrangement of genes into operons. In order to gain insight into the saturation process, we relate biological regulation to more commonplace social contexts where a small number of individuals can form an intricate web of connections on the Internet. We find that the size of partnership networks saturates even as the complexity of output increases. Then, we build a more general simulation of network growth and find agreement with a wide range of real networks.

This chapter is based in part on the following publication:

1. Bhardwaj, **Carson**, Abyzov, Yan, Lu, and Gerstein. *Analysis of Combinatorial Regulation: Scaling of Partnerships between Regulators with the Number of Governed Targets.* PLoS Computational Biology. 2010. May 27; 6(5): e1000755.

1.5.4 Chapter 6: Molecular Networks and Human Disease

During the past few decades, the amount of biological data available for analysis has grown exponentially. Along with this vast amount of information comes the challenge to make sense of it all. One subject of immediate concern to us as humans is health and disease. Why do we get sick, and how? Where do our bodies fail on a molecular level in order for this to happen? How are diseases related to each other, and do they have similar modes of action? These questions will require many researchers from multiple disciplines to answer, but where do we start? We take a bioinformatics approach and examine disease genes in a network context. In this chapter we analyze human disease and its relationship to two molecular networks. Firstly, we identify conserved motifs in the human transcription factor network and identify the location of disease- and cancer-related genes within these structures. We find that both cancer and disease genes occupy certain positions more frequently. Next, we examine the human protein-protein interaction (PPI) network as it relates to disease. We find that we are able to predict disease genes with 79% AUC using ADTree with 10 topological features. Additionally, we find that a combination of several network characteristics including degree centrality and disease neighbor ratio help distinguish between these two classes. Furthermore, an alternating decision tree (ADTree) classifier allows us to see which combinations of strongly predictive attributes contribute most to protein-disease classification. Finally, we build a matrix of diseases based on shared genes. Instead of using the raw count of genes, we use a *uniqueness* score for each disease gene that relates to the number of diseases with which a gene is involved. We show several interesting examples of disease relationships for which there is some clinical evidence

and some for which the information is lacking. This matrix will be useful in finding relationships between diseases with very different phenotypes, or for those disease connections which may not be obvious. It could also be helpful in identifying new potential drug targets through drug repositioning.

This chapter is based in part on the following publications:

1. **Carson** and Lu. *Network-based Knowledge Mining For Disease Genes.* SUBMITTED.
2. **Carson**, Kibbe, and Lu. *Discovering Disease Relationships Using a Co-occurrence Matrix.* TO BE SUBMITTED.

CHAPTER 2

BACKGROUND

The processes described in the Central Dogma of Molecular Biology dictate the course of gene expression and the path toward cellular development. The regulation of this process is fundamental for organisms at all levels of complexity. From adaptation to changing environments to the development of new functions, gene regulation provides several advantages to the cell. By choosing which genes to express at a particular time, an organism can increase versatility and adaptability to its environment. Gene regulation allows for more efficient use of resources. For example, if an organism is able to metabolize both lactose and glucose but only glucose is available, the organism can shut down the machinery responsible for the uptake and metabolism of lactose, thereby conserving energy. Gene expression allows for adaptation to stimuli in the environment, such as exposure to extreme heat or cold, and gives the organism protection from these potential hazards. In higher level eukaryotes, gene regulation assists the differentiation process, which produces specific cell types such as liver, kidney, lung, etc. Control over gene expression can be exerted by the cell at the transcriptional, post-transcriptional, translational, and post-translational levels, as well as epigenetically through the addition of specific compounds to nucleic acids. This wide range of regulation types allows cells to construct very complicated and extensive regulation mechanisms.

2.1 Nucleic Acid-binding Proteins

NA-binding proteins are involved in gene regulation at many levels, and understanding how they perform their tasks is crucial to understanding the regulatory process. DNA-binding proteins are an integral part of the gene regulation process and are also responsible for DNA repair. A particular subclass of these proteins, transcription factors, help control both the initiation and the level of transcription, which is currently the most studied and best understood type of regulation. RNA-binding proteins are directly involved with activities such as protein synthesis, regulation of gene expression, and RNA splicing and editing, as well as other post-transcriptional activities. Both DNA- and RNA-binding proteins are essential to the replication of specific types of viruses. The mechanisms underlying the behavior of these proteins require an understanding of the interactions that occur between specific amino acid residues within proteins and the nucleic acids to which they bind. Prediction of NA-binding residues can provide practical assistance in the functional annotation of NA-binding proteins. Predictions can also be used to expedite mutagenesis experiments, guiding researchers to the correct binding residues in these proteins. Identifying these residues is a complex and difficult problem. The characteristic traits of a residue which enable binding are largely unknown. Whether or not certain characteristics of its neighbors affect a residue's binding capability is also poorly understood, further complicating the issue. Because of this, machine learning has often been employed in an attempt to discover precisely which residues confer binding functionality.

2.2 Machine Learning

Machine learning is a subcategory of artificial intelligence and involves "teaching" computers to make informed decisions based on prior data. These data points often have hundreds or even thousands of characteristics, or attributes, associated with them. Finding connections between these characteristics is a highly complex problem, and finding these connections "manually" is next to impossible. Often, the learning problem is one of pattern recognition, in which, by exposing the machine to an increasing amount of data with relevant attributes, we increase its ability to recognize complex patterns that may not at first be obvious. There are several categories of learning including supervised, unsupervised, semi-supervised, multiple instance, reinforcement, and others. In this work we focus on *binary classification* in supervised learning problems (e.g., a protein does/does not bind DNA, a residue does/does not bind RNA, a cytosine is/is not methylated, etc.). The goal for these types of problems is for the algorithm, after having been shown many instances belonging to both classes, to classify a new, unseen example as belonging to one class or the other, preferably with high confidence. The classifier makes a decision as to which class the example belongs by assigning a *confidence score* between 0 and 1 or -1 and 1 for each example in the data set. This score reflects the level of certainty in the assigned class. For instance, if the upper bounds of the score are 0 and 1 respectively, an example with a confidence score between 0 and 0.5 would be assigned to the negative class, and one with a score between 0.5 and 1 would be placed in the positive class (using 0.5 as the threshold). For a positive example, a larger number indicates a higher level of confidence.

Likewise, a lower score indicates a higher confidence that the example belongs to the negative class.

Many algorithms have been developed for different types of learning problems. It is important to note that the appropriate algorithm for a particular prediction problem depends entirely on the nature of the problem as well as the types of attributes associated with the examples. No one algorithm works well in all cases, although some are more widely used than others. In this work we focus mainly on several versions of the decision tree algorithm, as well as an implementation of the support vector machines (SVM) algorithm called libSVM (41). These algorithms are included in the machine learning workbench Malibu, which was developed in-house by Robert Langlois (102).

2.2.1 Supervised Learning

Supervised learning requires that the data points have both an input value (usually a vector of features) and an output value, often referred to as the *class label*, which indicates the desired outcome of the prediction. The data is usually divided into two sets, one for training and one for testing. Once the algorithm observes the training data set, it builds an inferred function called a *classifier* (in the case of discrete output). This classifier should predict the correct output for each of the inputs, and, ideally, would correctly identify the class label for each of the previously unseen data instances in the testing set as well. In order for this process to work correctly, the classifier must be able to generalize and apply what it learned during the training phase to the testing set. Below we describe some specific algorithms for supervised learning that have been used in this work.

2.2.2 Support Vector Machines (SVM)

SVM (47) is binary classifier that can perform both linear and non-linear comparisons. It attempts to separate two classes of data by the widest margin, or *hyperplane*, possible. The training examples that lie on this hyperplane are called *support vectors*, and test examples compared to the resulting training model are classified based only on these support vectors. The wider the margin is, the better the classifier performance. While SVM has been shown to perform very well on high-dimensional data sets (i.e., those with a large number of attributes), it has a significant drawback for our purposes. Given a set of inputs, the classifier returns output without details related to the procedure it followed to arrive at its decisions. Therefore, we will gain no knowledge of which features were important in making these decisions. Often referred to as the *Black Box Effect*, this phenomenon prevents us from identifying the most important characteristics (or combination thereof) separating the two classes. Since our ultimate goal is to identify the importance of the features related to our examples, we choose to focus primarily on decision tree variants in our predictions.

2.2.3 Decision Trees

A decision tree is a general graphical model for displaying an algorithm. The tree-like structure is composed of internal nodes (each is an attribute), branches (each indicating the outcome of a test on an attribute), and leaf nodes (each is a class label). Each internal node represents a decision made on an attribute that successfully separates the data into two groups. At least one of these groups must contain an overrepresentation of one class or the other. The criterion for this split is the *information gain* that results from the choosing of each attribute.

The attribute associated with the largest information gain value is chosen, and subsequent nodes represent recursive splits on subsets of this data. Several variations on the decision tree are described below.

2.2.3.1 The C4.5 Decision Tree Algorithm

The C4.5 decision tree is a classification tree developed by Ross Quinlan (139). An improvement on his earlier ID3 algorithm, C4.5 has several advantages. It handles large data sets efficiently and does not require data preprocessing (unlike other algorithms such as SVM). It can handle both discrete and continuous attributes, and can handle training examples which may be missing values for these attributes. Additionally, the algorithm avoids *overfitting* the data, which causes the generated model to produce artificial, high-quality results and consequently makes it a poor predictor. The C4.5 tree is pruned in order to keep the model as simple as possible and avoid this problem.

2.2.3.2 Adaptive Boosting (AdaTree, AdaC4.5, and AdaStump)

A recurring problem in supervised machine learning problems is that of data set bias. As it turns out, a set of "weak learners", i.e., those that make mistakes by misclassifying examples, can be used to create a "strong learner" through a process known as *boosting*. Originally proposed by Freund and Schapire (66), the adaptive boosting (AdaBoost) algorithm constructs a collection of weak learning algorithms over several iterations using various distributions of the data set. First, a weak learning algorithm is trained over a uniform distribution of the data. During the next training iteration, a different distribution of the data set is used, one which places a higher weight (indicating greater importance) on misclassified examples from

the previous cycle, and a lower weight on correctly classified examples. This process is repeated for a predetermined number of iterations. This meta-algorithm has several advantages. It is relatively easy to implement and can be used in concert with other machine learning algorithms. In addition, it is fairly easy to tune compared to algorithms such as SVM and artificial neural networks. For DNA-binding protein prediction we use three algorithms enhanced with adaptive boosting: AdaTree (a custom implementation of the ID3 algorithm developed in-house by Robert Langlois (102)), AdaC4.5 (AdaBoost with the C4.5 algorithm (139)), and AdaStump (AdaBoost with a one-level decision tree serving as the weak learning algorithm).

2.2.3.3 C4.5 with Bagging and Cost-Sensitive Learning

One of the disadvantages of decision tree algorithms is that small changes to the data set may result in a different feature being chosen at a certain node, which will affect the structure of the tree in subsequent nodes and thus lead to instability. In this work we use the C4.5 algorithm with sub-sampling aggregation, which is similar to bagging, or bootstrap aggregating (32). This method attempts to offset this instability by building many different trees from the training data set using random sampling with replacement through a uniform probability distribution. The resulting classifier then uses majority voting to decide to which class the example belongs. When used in combination with an unstable classifier such as a decision tree, bagging can help to improve accuracy by reducing variance. For DNA-binding residue and CpG island methylation prediction, we used 200 trees to build our model, setting aside half of the negative examples as a training set and half as a testing set. The exact number of positive

examples used was determined by a positive class weight, or cost, which was equal to the class distribution.

Cost-sensitive learning (191) is useful for problems in which, as is the case with CpGI methylation and NA-binding residue prediction, one class is more highly represented than the other. The costing algorithm determines the importance of an example by considering its class label and performs an importance-weighted classification by applying weights to each class based on class distribution (191). The final classifier is created from an average of multiple rounds of rejection sampling. For DNA-binding residue and CpG island methylation prediction, we performed 10-fold stratified cross validation over the training set for one run during both the selection and the validation cycles.

2.2.3.4 Alternating Decision Trees

Alternating decision trees, or ADTrees (65), provide the benefits of the decision tree algorithm with the added advantage of the ability to generate an intuitive graphical model. This algorithm builds decision trees over a user-defined number of iterations using confidence-rated boosting, which results in an option tree (37). We used a bootstrapping method with ADTree which allowed us to find those rules that were conserved among multiple trees. In order to find the optimum number of iterations at which to grow the trees, we performed parameter selection on the data set using accuracy as our standard.

2.2.4 Classifier Evaluation

2.2.4.1 Metrics

We use the following metrics to evaluate the binary classifiers developed in our work. Sensitivity (or recall) (Equation 2.1) and specificity (Equation 2.2) measure the correct number of true positive examples and true negative examples, respectively. Accuracy (Equation 2.3) measures the proportion of true positives and true negatives correctly identified by the classifier. Precision, also called *positive predictive value*, measures the fraction of positive results that are truly positive (Equation 2.4). Matthews Correlation Coefficient (MCC) (Equation 2.5) provides a reliable way of measuring accuracy for both balanced (the number of positive and negative examples are equal) and imbalanced (more examples of one class than the other) data sets. A predictor that is always correct would have a MCC of 1.0; a predictor that is always incorrect would have the value -1.0. A random prediction would be 0. The Receiver Operating Characteristic (ROC) curve measures the ability of a classifier to separate positive examples from negative examples, and is generally considered a good measure of overall performance. The curve is created by sorting the continuous-valued outputs (corresponding to the likelihood for an example to belong to the positive class) from the generated model and plotting the false positive rate (FPR, which is equal to $1 - specificity$, Equation 2.6) vs. the true positive rate (TPR, Equation 2.7) for each example in the data set. An area under the ROC curve (AUC) of 0.5 is considered random, and an AUC equal to 1 would be characteristic of a flawless model. The AUC gives an idea of the tradeoff between sensitivity and specificity, and it is a good predictor of how a classifier will perform on future data sets.

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN} \tag{2.1}$$

$$Specificity = \frac{TN}{TN+FP} \tag{2.2}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.3}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.4}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}} \tag{2.5}$$

$$ROC\ plot,\ X\ axis:\ FP = \frac{FP}{FP+TN} \tag{2.6}$$

$$ROC\ plot,\ Y\ axis:\ TP = \frac{TP}{TP+FN} \tag{2.7}$$

2.2.4.2 Cross-validation

In this work we evaluate our classifiers using two types of cross-validation. The n-fold cross-validation (n-CV) method first partitions a data set into n subsets. The classifier is then trained n times, leaving out a subset of the data for each round of training. The withheld subset is used to test the model and to calculate a designated metric (usually accuracy in our work). Each value in the data set makes a contribution to this calculation and an average is taken. n-CV has been found to work well for small data sets (76). For protein-level NA-binding

prediction we also use leave-one-out cross-validation (LOO CV). In this technique, n is equal to the number of examples in the data set, i.e., each example is used for one testing cycle. This method has several drawbacks (92), but does at times perform well. We use it here in order to make comparisons with previous work.

2.2.5 Nucleic Acid-binding Protein Prediction

Previous attempts have been made using machine learning to identify whether or not a protein binds DNA or RNA (38; 2; 24; 67; 79), still others have tried to identify the residues on DNA- (101; 129; 166; 167; 3; 4; 25; 187) and RNA-binding proteins (85; 99; 165; 177; 164) which bind nucleic acids. Each previous attempt made toward the identification of binding residues in NA-binding proteins has used different combinations of binding residue attributes and machine learning classifiers. Some examples of classifier types include SVM (38; 101; 129), neural networks (4; 85), and Naïve Bayesian methods (187; 165). Both sequence- and structure-based attributes have been used including amino acid composition, hydrophobicity, predicted secondary structure, local tertiary structure, both non-position-specific (BLOSUM, (81)) and position-specific (PSSM) measures of evolutionary information, and others.

2.2.6 Using Costing to Overcome Data Set Imbalance

NA-binding proteins almost always contain a smaller number of binding than non-binding residues. This means that any residue-level data set calculated from these proteins will be imbalanced in terms of these two classes, and thus prediction results will be imbalanced in terms of sensitivity and specificity due to *degeneracy*, or the tendency during prediction for a classifier to assign examples to the most common class. For instance, if binding residues

made up the positive class in a training set and non-binding residues comprised the negative class, validation results would reveal much lower sensitivity than specificity, indicating that less true positive examples had been found. To get around this problem, many in the past have removed a number of non-binding residues from the data set in order to achieve a class balance. However, balancing a training set would seem to remove important information about the non-binding class and prevent the classifier from developing a clear distinction between the two types of examples, and in fact has been shown to result in a less robust model and to cause poor test performance (130). Therefore, results using models built with balanced training sets may not actually provide reliable results. Ideally, we would like to provide machine learning classifiers with as much of the data set as possible in order to increase accuracy, while avoiding the problem of degeneracy. One way to do this is to force the developed classifier to apply a weight to the examples in the underrepresented class. This method, called *cost-sensitive learning* or *costing*, has been used in previous research to improve prediction. Sorzano et al. used a Naïve Bayes algorithm to improve recognition of particles in images from cryo-electron microscopy experiments (153). Fan, Lee, Stolfo, and Miller used their own version of costing to improve accuracy for intrusion detection systems (60).

2.3 Protein-DNA Interaction

Protein-DNA interactions play crucial roles in the cell. Many proteins are involved in the processes of DNA replication, repair, recombination, transcription and their regulation. Histones help to condense chromosomal DNA into a compact structure in eukaryotic cells. Restriction enzymes in bacteria aid in host defense by recognizing and cutting foreign DNA. Transcrip-

tion factors recognize and bind specific DNA sequences to activate or repress the transcription of regulated genes depending on environmental factors and cell requirements. Transcription factor binding sequences are usually short and degenerate, which allows one transcription factor to regulate several genes. Identifying targets of a transcription factor may help to discover the number of genes under its regulatory control and how this regulatory process is carried out. Many experimental and *in silico* methods have been developed to detect whether or not a protein binds DNA and at which sites this binding occurs (36). Many experimental methods are complicated, time-consuming, and expensive, which makes a computational approach to the prediction of protein-DNA interaction sites highly desirable. Results from such computational methods may also be used to compliment experimental techniques.

2.4 Development of Protein-DNA Interaction Potentials

Current computer modeling techniques for calculating protein-DNA interaction potentials fall into two categories: sequence- and structure-based. Sequence-based methods commonly use evolutionary information collected from a statistical analysis of the transcription factor binding sites. In general, these methods begin with the collection of DNA sequences known to bind specific transcription factors, followed by statistical analysis of these sequences. One common analysis protocol involves the creation of a multiple sequence alignment (MSA), which can then be used to build a position-specific scoring matrix (PSSM). Genomic sequence is subsequently scanned with the PSSM to identify additional binding sequences for this transcription factor. The main drawback of these methods is that they require many examples in order to gain useful information, and a sufficient number of examples are often unavailable.

In contrast, structure-based methods use three-dimensional structures obtained by X-ray crystallography and NMR to describe protein-DNA interactions at the atomic level and can clearly explain binding recognition and binding affinity (86). An ever-increasing number of experimentally-solved three-dimensional structures provide new models for protein-DNA interaction. These structures allow for a much clearer explanation of binding recognition and binding affinity. Much effort has been focused on the development of a statistical potential for protein structure and protein-protein interaction prediction (115; 197; 116; 114).

Although there is not a one-to-one recognition code between amino acids and DNA bases, protein-DNA interaction preferences have been found and knowledge-based potentials imply that such interaction propensities can be used for protein-DNA binding prediction. A statistical potential, based on hydrogen bonding and hydrophobic contacts, has been used to describe interaction preferences between 20 amino acids and 4 bases. This potential was subsequently used to evaluate the binding affinity of a protein-DNA complex (120). Only short-distance contacts were included in that computation. A distance-dependent statistical potential, which takes into account both long-range interactions up to 15Å and multi-body interactions, has also been built (113). A grid-based potential, which has a different spatial partition than a distance-dependent potential, has been shown to perform quite well in evaluating binding affinity, predicting cooperative binding, and predicting binding specificity of protein-DNA complexes (142; 95). In these works, Cα atoms were used to represent the position of the amino acid and two-body interactions (those involving one amino acid and one DNA base) were considered.

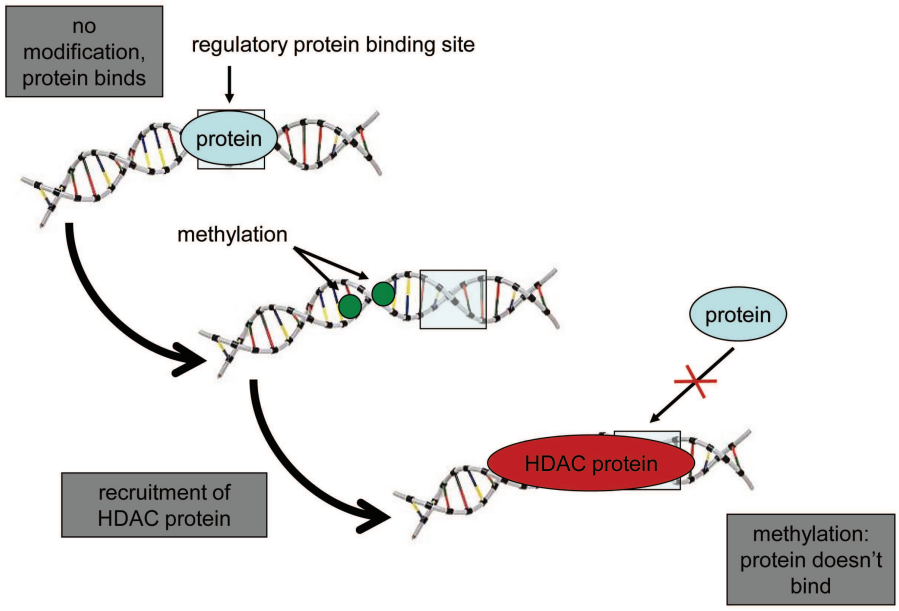


Figure 2: The methylation of DNA, which usually occurs in GC-rich regions called CpG islands (CpGIs), provides signals for recruitment of a histone deacetylase complex (HDAC). The HDAC deacetylates the lysine residues of histones, which increases the protein's affinity for the negatively charged backbone of DNA (not shown). This initiates the formation of heterochromatin, which results in the suppression of transcription due to lack of access by transcription factors. DNA: `http://www.dk.co.uk/static/clipart/uk/dk/future/image_future007.jpg`.

2.5 Methylation of DNA in CpG Islands

2.5.1 Biological Function

CpG island (CpGI) methylation is a type of epigenetic DNA modification. It occurs in eukaryotes and, in human DNA, is based on the addition of a methyl group to the nucleic acid cytosine. The methyl group is added to the number 5 carbon of the pyrimidine ring of a cytosine which is followed by a guanine (CpG). This action is performed by enzymes called DNA methyltransferases, of which there are two types. DNA methyltransferase 1 (Dnmt1) performs maintenance methylation during which methyl groups are added to cytosines of the daughter strand in exactly the same pattern as the parent strand. DNA methyltransferases 3a and 3b (Dnmt3a, Dnmt3b) are involved in *de novo* methylation during which cytosines at new positions in the DNA strand are methylated. This changes the methylation pattern in a localized region of the DNA, and the new methylation pattern is a reversible, heritable change that does not alter the DNA sequence itself or the genotype. This type of modification is called *epigenetic*.

The generally accepted definition of CpGIs involves three criteria (70; 160). Firstly, the GC content of a CpGI must be ≈60% or greater. Secondly, the observed vs. expected number of CpG dinucleotides should be above ≈0.65. Thirdly, the length of the island itself should fall within the range of 200 to 3000 nucleotides. This last characteristic is a matter of some debate and varies within the literature. CpGIs account for ≈1% of the human genome and were initially thought to be located primarily in the 5' region of expressed genes in higher eukaryotes. While CpGIs overlap with the promoter region of 50-60% of human genes, including most

housekeeping genes, recent high-throughput and genome-wide studies indicate as many as 50% of CpGIs occur inter- or intragenically in human DNA (83). Conventional wisdom says that in the human genome CpGs within promoters regions are usually unmethylated, while most CpG dinucleotides in non-coding regions are methylated (29) (there are several exceptions, including X-chromosome inactivation (52)). It has been shown recently, however, that more than one third of CpGs within transcribed regions of *Arabidopsis* DNA may be methylated (194).

Methylation of CpGIs can affect gene expression in the following manner (35): Methylation of an island upstream (5') of a gene by DNA methyltransferase signals the methyl-CpG-binding protein (MeCP) components of a histone deacetylase complex (HDAC). The action of deacetylation by HDAC of the lysine residues of histone restores their positive charge, which increases the affinity of histone for the negatively charged backbone of DNA. HDACs are associated with the formation of heterochromatin through this process. The formation of heterochromatin generally down-regulates DNA transcription by blocking access for transcription factors to the promoter region of genes.

2.5.2 Previous Work In CpGI Methylation Prediction

A number of previous attempts have been made to predict CpGI methylation. Feltus et al. (64) used hierarchical clustering and seven DNA sequence patterns to distinguish between methylation-prone and methylation-resistant CpGIs with an accuracy of 82%. They later used a linear discriminate method and achieved 87% accuracy (63). Fang, Fan, Zhang, and Zhang (61) used SVM with DNA sequence properties and transcription factor binding site (TFBS) information to reach a result of 85% accuracy, 77% sensitivity, and 86% specificity. Bhasin,

Zhang, Reinherz, and Reche (27) attempted to predict specific cytosines within CpGIs that were methylated. This group used an SVM classifier and sequence composition attributes and reported 75% accuracy, 72% sensitivity, and 77% specificity as their best results. The most comprehensive paper written on this subject is that of Bock et al. (30), in which they use 1184 DNA attributes to classify 132 CpGIs from human chromosome 21 as methylation-prone or methylation-resistant. In their study they find that specific sequence patterns, DNA repeats, and DNA structure have a high correlation to methylation. They report that ≈66% of the attributes that are differentially distributed between methylated and unmethylated CpGIs are 4-mer DNA sequence patterns, both strand- and non-strand-specific. Additionally, other work has shown that certain DNA motifs may affect susceptibility to aberrant methylation in human fibroblast clones (64). Further evidence indicates that repetitive sequences can affect the methylation status of CpGIs (117; 189). These studies highlight the importance of sequence patterns and characteristics in CpGI methylation.

2.6 Gene Regulation in a Network Context

Each of the abovementioned areas of study focuses on a particular component of the nucleic acid binding process. While important, none of these individual components gives us a broad understanding of how the system of transcriptional regulation operates as a whole. To achieve this, much focus has been placed in recent years on the study of gene regulation or transcriptional networks, with the goal of understanding how these networks evolve and function. These works include the creation of logical models, continuous models, and single-molecule methods (for a review see (90)). Statistical models of protein-protein interaction networks have also

been created in order to study their evolution (22). This type of global approach is necessary to understand how the individual components function in concert to regulate gene expression.

Recently, the topics of network organization and robustness within biological networks have come into the spotlight. How do gene regulation and protein-protein interaction networks manage to stay robust to genetic changes, some of which are deleterious or mutational in nature? How does an organism maintain fitness? Parallels between gene regulation networks and communication networks have been drawn, focusing on failure and attack tolerance in biological networks (5). In 2005, Wagner discussed two main hypotheses on the mechanistic causes of robustness: redundancy and distributed robustness (172). He pointed out that while there is evidence that duplicate genes play an important role in an organism's tolerance to change, many systems, including metabolic and gene regulation networks, show no gene redundancy but are still able to tolerate removal of highly-connected nodes. Subsequently, the transcription factor co-regulation network in yeast was shown to possess a distributed node degree distribution (18), which is thought to lend a level of robustness to the scale-free gene regulation network. However, the same co-regulation network architecture was not found in *E. coli* (17), which highlights the possibility that there are multiple pathways for achieving fitness. In 2007, Wagner and Wright observed that many regulator-target gene pairs in more than a dozen biological networks had intermediate regulators between them. These 'alternative routes' could be a possible cause of robustness (173).

Many in the past have postulated that the connectivity distribution $P(k)$ of gene regulation networks is scale-free, that is, the probability that a node is connected to k other nodes is

inhomogeneous, with a few nodes having a large number of edges (20; 5; 18). It has also been shown that the removal of nodes with large numbers of connections (hubs) can cause the network to be fatally fragmented (5). Interestingly, some scale-free networks are not vulnerable to the removal of hubs (54; 59). Balaji, Iyer, Aravind, and Babu highlighted an interesting paradox in the yeast genome. They found that although $\approx 1/2$ of the regulatory interactions in yeast are controlled by hubs in the gene regulation network, only 5/33 were essential for survival. Furthermore, only 9/157 transcription factors examined were found to be essential under the conditions provided in the experiments. In addition, they identified a 'co-regulation network' of transcription factors underlying the gene regulation network which appears to possess a distributed architecture (18). This leads to an interesting question: How did the gene regulation network form, and could the method of formation provide for an emergent property such as the formation of this underlying partnership network? Analysis of regulation networks has led to the hypothesis that the evolution of these networks has been incremental rather than being formed by the copying or transfer of motifs or circuits.

Several evolutionary models have been created to study robustness. Ciliberti, Martin, and Wagner showed that robustness is an evolvable trait (43). Crombach and Hogeweg showed that the evolution of gene regulation dynamics can lead to increased efficiency of creating well-adapted offspring, while maintaining a robustness to most mutations (51). Krishnan, Tomita, and Giuliani showed that robustness evolves along with networks and is an emergent property even without selective pressure (97).

2.7 Gene Regulation and Disease

In recent years, researchers have produced a body of work that has given us a clearer (albeit more complicated) picture of how a disease such as cancer comes to be, how it develops, and how it can be treated. The roles of genetics in the form of single-nucleotide polymorphisms (SNPs) (57), epigenetics (147), miRNA (171), copy number variation (98), chromatin structure (33), and protein biomarkers (69) in cancer have been shown. While great scientific advances have been made in the understanding and treatment of this disease in the last 50 years, we still do not have clear knowledge of the 'how' and 'why'. Given a set of initial conditions in the body defined by genetics, lifestyle, environmental exposure, etc., cancer begins and proceeds to develop through an evolutionary process. This results in all cancers having unique characteristics (77). Clearly, cancer is a multi-dimensional problem for which we have an enormous amount of data. Gaining knowledge from the existing data, however, is a non-trivial task.

In the last several years, bioinformatics and computational biology have made a variety of contributions to disease analysis using existing data in an attempt to increase our understanding. Popular topics include the discovery, prediction, and analysis of genes related to disease (176), statistical analysis of SNPs and disease (110), the prediction and discovery of new drug targets (159), the development of the disease ontology (DO) and its application to the human genome (132; 133), the analysis of protein-protein interaction networks as they relate to disease (82), and many others. Of particular interest is the development of 'disease networks' (74; 192), which are in most cases bipartite graphs describing disease-disease as well as disease-

gene relationships (see Figure 3). These edges may signify one or more shared genes, metabolic pathways, miRNAs, or a number of other data types.

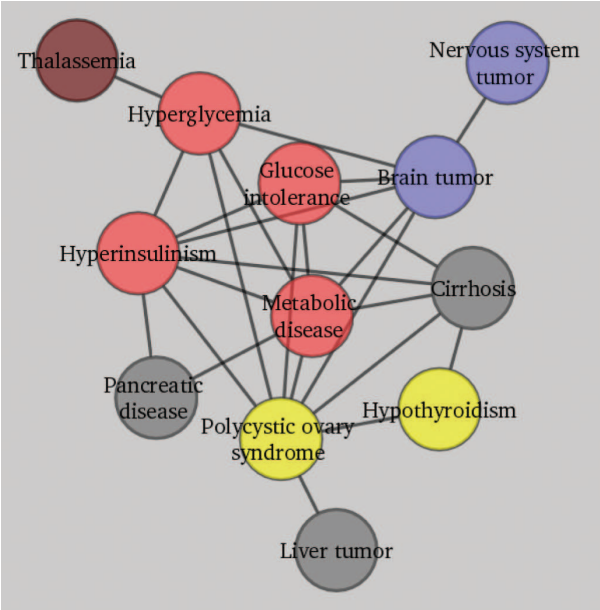


Figure 3: An example subgraph of the disease-gene bipartite graph. This projection of the network describes disease-disease relationships. Nodes indicate diseases; edges between nodes represent disease relationships. Edges may signify one or more genes, metabolic pathways, miRNAs, or a number of other data types. This network was visualized using Cytoscape (151).

The disease network reveals the interconnected nature of various diseases, which begs the question; can we gain new knowledge of a disease such as cancer by studying 'connected', non-cancer diseases? Many diseases including obesity (100; 161), various infections (9), diabetes

(174), and possibly even psychological stress (71) have been reported to have some relationship to cancer. Often the relationship type is unknown or partially known, which indicates that a deeper understanding of these relationships is needed. Furthermore, these relationships have not yet been explored as a whole, but rather as individual links.

2.8 Disease Gene Prediction

There have been several previous attempts at predicting gene-disease association. Özgür, Vu, Erkan, and Radev used SVM, text mining, and several network metrics to rank potential disease genes (134). Gonzalez et al. predicted atherosclerosis-related genes by creating a PPI and adding weights to certain proteins based on text mining of PubMed abstracts (75). Xu and Li used a KNN classifier to predict hereditary disease genes from OMIM over the human PPI network with an overall accuracy of 76%. They found that these hereditary disease proteins tended to have a larger number of interactions and tended to have more shared neighbors than non-disease proteins (185). Wu, Jiang, Zhang, and Li developed CIPHER, a software tool that prioritizes disease genes (182).

Due to the complicated nature of many diseases, which may involve the failure of multiple levels of biological function including DNA repair, gene regulation, epigenetic and histone modifications, metabolic pathways, etc., elucidation of disease relationships requires a systematic and computational solution. Though there may be a plethora of data available to quantify this problem, the data itself does nothing for us unless we can turn that data into knowledge (a similar problem arose after the sequencing of the human genome). Merely combining sources of data is not sufficient. We must identify patterns within the data, which is manually in-

feasible when the number of data points and characteristics to be compared is large. Clearer understanding could be gained by finding, among all attributes of a relationship, those that characterize it most accurately. Several existing machine learning algorithms can help achieve this including multiple instance learning (53), positive/unlabeled (PU) learning (112), Bayesian inference (28), ADTree (65), and others. We have used the ADTree algorithm to analyze methylation patterns on DNA (39) and to predict DNA-binding proteins (103). In both cases, this algorithm helped us to understand what characteristics have the most influence on determining the class to which the examples belonged. A similar method of rule discovery is needed in the case of the disease network.

CHAPTER 3

DNA-BINDING PROTEIN PREDICTION

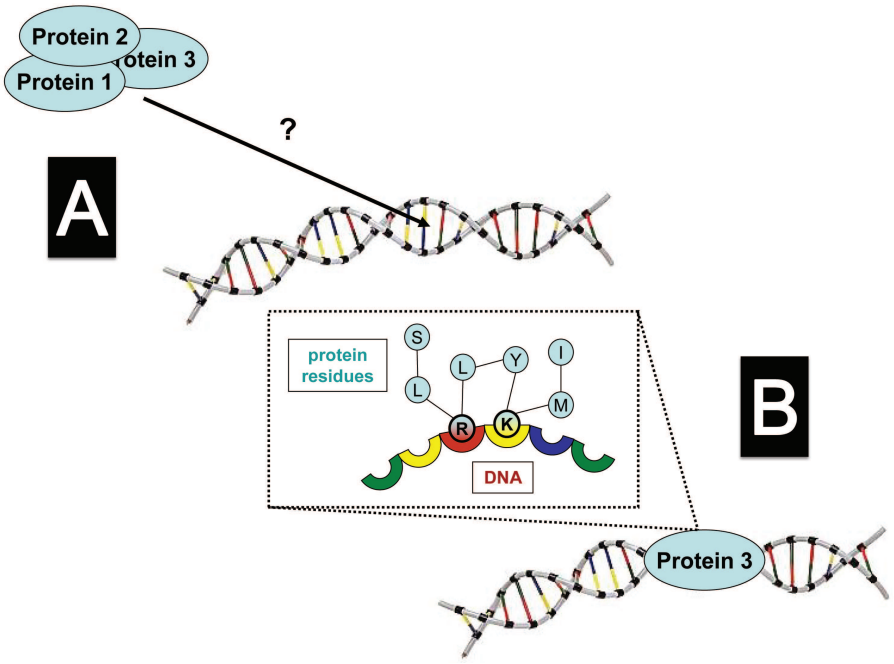


Figure 4: A) Which of these proteins binds DNA? B) Given that a particular protein binds DNA, which residues are involved in binding? DNA: `http://www.dk.co.uk/static/clipart/uk/dk/future/image_future007.jpg`.

3.1 Project Goals

We approach the DNA-binding prediction project in two ways, illustrated in Figure 4. We asked two questions: 1) Given a set of proteins, which of the proteins in this set binds DNA (104)? 2) Given that a particular protein binds DNA, which residues are involved in binding (40)? Though these are two very different questions, the prediction strategy used during the implementation of these ideas is very similar.

3.2 Prediction Strategy

Our prediction strategy (Figure 5) starts with either a protein structure or protein sequence. We then gather features for either the entire protein or each residue depending on what type of prediction we are making. Next, we follow one of two paths: 1) calculate structure-based features, which are attributes of three-dimensional structures acquired using X-ray crystallography or NMR, or 2) calculate sequence-based features. These features are collected and possibly normalized or converted in some way depending on the algorithm. We then train models using this attribute information and subsequently test our models with new data (for which we calculate the same features) in order to predict either binding or non-binding.

3.3 Protein-level Prediction

3.3.1 Data Set

In Langlois, Carson, Bhardwaj, and Lu (104), our data set for DNA-binding prediction included two classes of proteins; those that bind DNA (positive class) and those not known to bind DNA (negative class). It consisted of 75 DNA-binding proteins and 214 others not known to bind DNA including membrane-binding proteins, chaperones, and enzymes. This negative

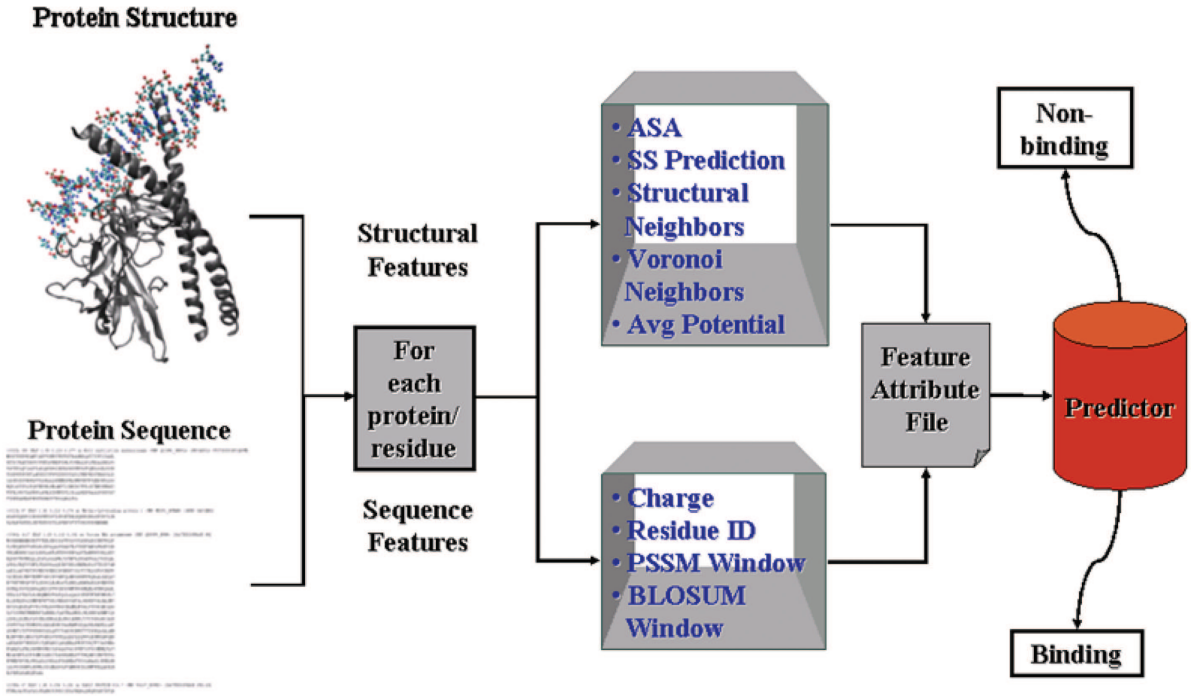


Figure 5: NA-binding prediction strategy

set is a subset of one used by Stawiski, Gregoret, and Mandel-Gutfreund (154). These sets were culled using the PISCES web server (174) and only structures with a sequence identity of $\leq$ 20% and a resolution of $\leq$ 3Å were included in our experiments.

3.3.2 Calculated Attributes

For protein-level binding prediction, we used a set of 42 features related to protein structure and sequence. The sequence-based features included the amino acid composition (20 features) and the net charge calculated using the CHARMM (34) force field (1 feature). Two structure-based features were included. Firstly, we calculated the surface amino acid composition (20 features) as follows. Residues were identified as 'surface' if the solvent accessible surface area (ASA) was $\geq$ 40%, otherwise they were considered 'internal'. This distinction was made by first adding hydrogen atoms to the resolved structures using REDUCE (181), and then calculating the ASA for these modified structures using DSSP (88). The second structural feature we used was the size of the largest positively charged patch (24; 26) (1 feature).

3.3.3 Classifier Evaluation

Refer to the section "Classifier Evaluation Metrics" in **Chapter 2** for details on our evaluation methods.

3.3.4 Results

We performed DNA-binding protein classification using sequence- and structure-based features with several algorithms (Table I). We varied the size of the data sets via 2- and 5-fold cross validation. These results demonstrate our ability to distinguish between DNA-binding and non-DNA-binding proteins. AdaTree performed the best with 88.5% accuracy, 66.7% sensitivity,

96.3% specificity, and an AUC of 88.7%. These results can be interpreted in the following way. Given a random data set of proteins with the same class distribution, AdaTree should correctly assign ≈88% of these proteins to the correct class. If we know that a protein binds DNA, our classifier will correctly categorize it ≈66% of the time. Similarly, if we have prior knowledge that a protein that does not bind DNA, we can correctly predict this ≈96% of the time. All of these metrics are dependent on class distribution with the exception of the AUC, and because our data set was imbalanced, we placed more confidence in the AUC. As mentioned earlier, the AUC gives an idea of the tradeoff between sensitivity and specificity and is a good predictor of how a classifier will perform on future data sets. One interesting finding is that the results for SVM (a very slow algorithm to train) and the second fastest (AdaStump) were fairly close. This tells us that we can use a fast tree algorithm such as this and not sacrifice much accuracy. This is useful because the SVM algorithm ran at a rate ≈25 times slower than AdaStump.

3.4 Residue-level Prediction

After identifying DNA-binding proteins, we calculated features for each amino acid residue in a set of proteins and predicted whether that residue binds either DNA or RNA (40). We follow the same protocol for prediction as we did for protein-level prediction (Figure 5). We then created an ensemble classifier using C4.5 with bootstrap aggregation (bagging) and cost-sensitive learning (costing).

3.4.1 Data Sets

The proteins comprising our data sets were extracted from the PDB database (`http://www.rcsb.org`) and culled using the PISCES web server (174) with a sequence identity of ≤

CV	Metric	AdaTree	AdaC4.5	AdaStump	SVM	C4.5
2-fold	ACC	86.5	86.3	83.6	84.5	79.4
	SEN	61.4	61.5	60.5	57.5	59.8
	SPE	95.3	95.0	91.7	94.0	86.3
	AUC	88.0	88.8	81.6	84.4	61.3
5-fold	ACC	87.2	86.6	84.1	85.7	79.4
	SEN	63.8	61.4	61.2	59.1	59.9
	SPE	95.4	95.4	92.2	95.0	86.3
	AUC	88.4	89.6	82.7	85.9	54.1
LOO	ACC	88.5	86.5	85.1	86.3	80.0
	SEN	66.7	61.3	62.7	62.7	65.3
	SPE	96.3	95.3	93.0	93.9	85.0
	AUC	88.7	89.8	84.6	86.3	74.0

TABLE I: Comparison of four metrics and three different cross validation techniques using five different classifiers for protein-level prediction over the protein-DNA data set.

25%. All structures were determined by X-ray diffraction and had a resolution of ≤ 3.0Å. Our sequence-based DNA-binding residue data set consisted of 54 proteins and 14,780 residues, 2,083 of which were identified as DNA-binding and 12,697 considered non-binding based on distance from the DNA molecule in the bound structure (class ratio of $\approx 1/6$). The RNA-binding residue data set used for sequence-based prediction contained 84 proteins and 60,016 residues, 5,934 classified as RNA-binding and 54,082 as non-binding (class ratio of $\approx 1/9$).

In order to compare our strategy with previous works, we collected two DNA-binding sets (PDNA-62 (146) and 274 proteins from Ofran, Mysore, and Rost (129)) and one RNA-binding set (109 protein chains used by Terribilini et al. (165)). We calculated both structure- and sequence-based features for these sets (described below).

3.4.2 Definition of Binding Residues

Because we have formulated residue prediction in this case as a binary classification problem, each residue in the data set must be defined as DNA-binding or non-DNA-binding. As with previous studies (3; 4; 101), we based this class distinction on a residue's distance from the DNA molecule in the crystallized protein-DNA complex. A residue was defined as binding if any heavy atom (carbon, nitrogen, oxygen, or sulfur) belonging to the residue fell within a distance of 4.5Å of any atom in the DNA molecule. In agreement with Kuznetsov, Gou, Li, and Hwang (101), we found that this distance provided the best accuracy for predictions. Any residues without atomic coordinates in the PDB file were not included in the data set.

3.4.3 Definition of Surface Residues

For structure-based data sets, residues were identified as 'surface' if the solvent accessible surface area (ASA) was $\geq 40\%$, otherwise they were considered 'internal'. This distinction was made by first adding hydrogen atoms to the resolved structures using REDUCE (181), then calculating the ASA for these modified structures using DSSP (88). Only surface residues were included when calculating structural features over the data set ($\approx$38% for DNA-binding, $\approx$34% for RNA-binding).

3.4.4 Structure-based Attributes

In this work we used 24 structure-based attributes, which we define as any residue feature that requires a protein crystal structure in order for it to be calculated. These features describe a residue environment in the natural three-dimensional space. Further details of these features are described below.

3.4.4.1 Solvent-accessible Surface Area (ASA)

For each residue, the ASA was calculated using DSSP (88). This attribute is a real number between 0 and 1 and represents the percentage of the total surface area of the residue that is exposed on the surface of the protein (1 feature).

3.4.4.2 Predicted Secondary Structure

The predicted local secondary structure to which each amino acid belongs was also predicted using DSSP (88). The secondary structure was identified as either 'helix', 'beta', or 'extended' (3 features).

3.4.4.3 Structural Neighbors

In order to represent the local tertiary structure within the protein, we incorporated distance-based neighborhood information into our prediction. This feature identifies the neighbors of each residue. Any amino acid y lying within 12Å of another amino acid x was identified as a neighbor of x (as used by Kuznetsov, Gou, Li, and Hwang (101)). The number of neighbors of each type were recorded and represented as a 20-dimensional array for each residue (20 features).

3.4.5 Sequence-based Attributes

We calculated a total of 301 sequence-based attributes for each residue in our data sets. We considered a 'sequence-based attribute' any residue feature that can be calculated without the use of a crystal structure (i.e., only protein sequence). The individual descriptors are described below.

3.4.5.1 Residue ID

A 20-dimensional feature vector representing the 20 common amino acids was used to identify each residue, where a single non-zero entry indicates the current residue (20 features).

3.4.5.2 Residue Charge

Since DNA molecules are negatively charged, basic amino acid residues with a positive charge can play an important role in nucleic acid binding. Accordingly, we included a charge attribute for each residue. Arginine and lysine residues were assigned a charge of +1, histidines +0.5, and all others 0 (1 feature).

3.4.5.3 Measures of Evolutionary Conservation

In order to consider the level of evolutionary conservation of each residue and its sequence neighbors, we created a position-specific scoring matrix (PSSM) for each residue in the test protein. Along with the NCBI-NR90 database (4), which contained $\leq 90\%$ sequence identity between any two proteins, PSI-BLAST (8) was used to create a matrix representing the distribution of all 20 amino acids at each position in the protein sequence. A 7-residue sliding window, which represented the distribution of amino acid residues at the positions occupied by three sequence neighbors on either side of the central residue, was subsequently created. This resulted in a 140-element feature vector for each residue. A similar 7-residue window was created using a BLOcks SUbstitution Matrix (81). We chose BLOSUM62, which was built using sequences with $\leq 62\%$ sequence identity, in order to capture non-position-specific evolutionary conservation information for the sequence neighborhood of each residue. This feature contributed another 140-element feature vector (280 features total).

3.4.6 Classifier Evaluation

Refer to the section "Classifier Evaluation Metrics" in **Chapter 2** for details on our evaluation methods.

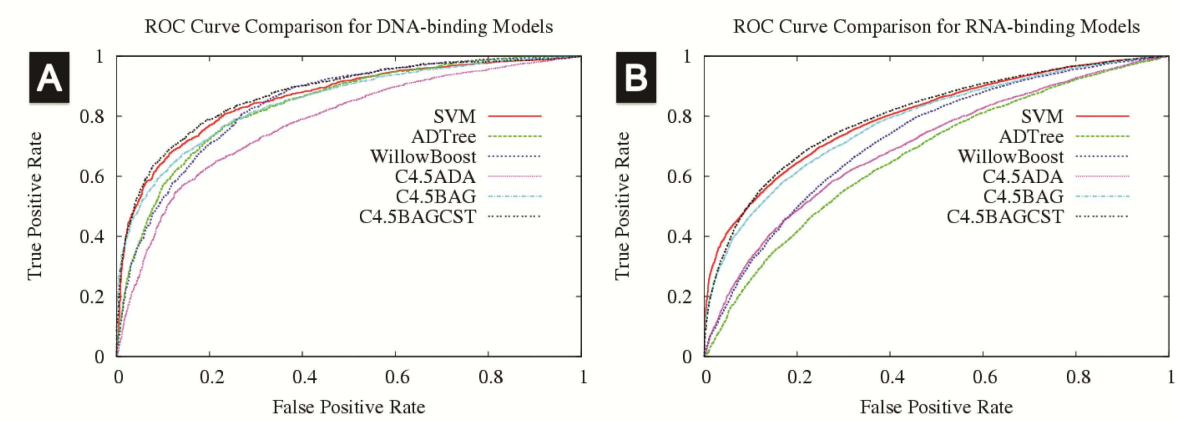
3.4.7 Results

3.4.7.1 Classifier Comparisons Using Sequence-based Features

We evaluated the performance of our C4.5BAGCST models against five other classification algorithms (SVM, ADTree, WillowBoost, C4.5 with AdaBoost, and C4.5 with bootstrap aggregation). We built two models for each using sequence-based features: one for DNA-binding proteins and one for RNA-binding proteins. Figure 6 describes the results for this comparison and shows the performance of each algorithm in terms of accuracy, sensitivity, specificity, precision, Matthews Correlation Coefficient (MCC), and the area under the Receiver Operating Characteristic Curve (AUC). The AUC provides a measure of a model's ability to separate positive and negative examples and is generated from a plot of the true positive rate versus the false positive rate for each example in the data set (Figure 6). A perfect model would have an AUC of 1, while a random model would have an AUC of 0.5.

3.4.7.2 Evaluation Using Previous Data Sets

In order to demonstrate the stability of our classifiers, we built models using previously compiled data sets for both DNA- and RNA-binding residue predictions. Figure 7 shows the comparisons between the original classifier and ours using two previously compiled DNA-binding protein data sets and one RNA-binding protein data set used in seven publications (4; 99; 101; 129; 165; 175; 177). The classifiers were created using 10-fold cross-validation for both selection



C

	DNA						RNA					
ALGORITHM	ACC	SEN	SPE	PRE	AUC	MCC	ACC	SEN	SPE	PRE	AUC	MCC
SVM	0.781	0.682	0.881	0.851	0.861	0.574	0.723	0.644	0.802	0.764	0.799	0.451
ADTree	0.834	0.470	0.927	0.622	0.838	0.444	0.626	0.621	0.630	0.627	0.669	0.251
WillowBoost	0.758	0.762	0.752	0.755	0.842	0.515	0.671	0.668	0.674	0.672	0.729	0.342
C4.5ADA	0.710	0.730	0.690	0.702	0.776	0.421	0.641	0.648	0.633	0.638	0.700	0.281
C4.5BAG	0.769	0.768	0.770	0.769	0.848	0.537	0.708	0.678	0.737	0.721	0.782	0.416
C4.5BAGCST	0.791	0.805	0.777	0.783	0.873	0.583	0.732	0.707	0.757	0.744	0.802	0.464

Figure 6: Receiver Operating Characteristic (ROC) curves comparing six classifiers for A) DNA-binding residue prediction models and B) RNA-binding residue prediction models. C) Results of a 10-fold cross validation over the binding residue data sets. Six metrics describe the performance of each of the 12 classifiers: ACC = accuracy, SEN = sensitivity, SPE = specificity, PRE = precision, AUC = area under the ROC curve, MCC = Matthews Correlation Coefficient. Algorithms: SVM (Support Vector Machines), ADTree (alternating decision tree), WillowBoost (in-house-developed tree algorithm w/ boosting), C4.5ADA (C4.5 decision tree w/ AdaBoost), C4.5BAG (C4.5 decision tree w/ bootstrap aggregation), C4.5BAGCST (C4.5 decision tree w/ bootstrap aggregation and costing). The highlighted classifiers (C4.5BAGCST) are those used for the NAPS web server. These figures appear in Carson, Langlois, and Lu (40).

NA TYPE	DATA SET	PUBLICATIONS	ACC	SEN	SPE	AUC	MCC
DNA	56 DNAbp		81.61	83.86	79.60	90.01	0.634
		Jones 2003	68				
	63 DNAbp		82.20	84.53	79.88	90.09	0.645
		Tsuchiya 2004					avg=0.42
	PDNA-62	w/ WillowBoost (str, b)	83.30	85.37	81.20	91.38	0.666
		Kuznetsov 2006 (str, b)	82.3	79.2	85.4		0.66
	PDNA-62	seq, str	83.06	85.09	80.49	89.98	0.656
		Ahmad 2004 (seq, str)	79.1	40.3	81.8		
		Kuznetsov 2006 (seq, str, ub)	78.1	79.2	77.2	84	0.49
	PDNA-62	seq	78.47	79.74	77.20	85.68	0.570
		Ahmad 2005 (seq)	66.4	68.2	66.0		
		Kuznetsov 2006 (seq, ub)	76.0	76.9	74.7	83.6	0.45
		Wang 2006 (seq)	70.31	69.40	70.47	75.24	
	274 DNAbp	seq, str	87.93	90.20	85.66	95.06	0.759
	274 DNAbp	seq	86.37	84.62	87.87	93.13	0.725
		Ofran 2007 (seq)	89				
	50 DNA-bp	seq, str	82.09	83.92	80.27	90.16	0.642
		Bhardwaj 2007 (seq, str)	74.94	75.54	74.77		
RNA	109 RNA-bp chains	seq, str	81.78	78.20	82.84	88.09	0.611
		seq	76.17	75.39	76.94	83.20	0.523
		seq (LOO)	75.31	84.14	66.49	83.94	0.514
		Terribilini 2006 (seq)	84.8	51	38		0.35
		Kumar 2008 (seq)	81.16	53.05	89.55		0.45
		Wang 2008 (seq)	87.4	48.2			0.457

Figure 7: The results shown are from 10-fold CV using a C4.5BAGCST classifier (unless otherwise indicated) built using each of the previous data sets for DNA- and RNA-binding residue prediction. White rows indicate the performance of our classifier over a particular data set; gray rows indicate the performance of the classifiers reported in the original publication. Gaps identify metrics which were not reported. str: sequence- + structure-based predictor; seq: sequence-based predictor; b: balanced data set; ub: imbalanced data set.

and validation. For the costing algorithm, the weight assigned to each class was equal to the class distribution and 200 costing iterations were run. Net accuracy was used to find the best model. The prediction metrics from previous works shown are either those reported as the best results from the publications, or if the author's intended best result is unclear, the results with the best accuracy or MCC.

Overall we found that, based on the metrics reported in these previous publications, we were able to improve on those results over each of three previously compiled data sets. We first built our own classifier with the PDNA-62 data set, which was originally compiled by Selvaraj, Kono, and Sarai (146) and used for binding residue prediction in three subsequent publications (4; 101; 175). Our C4.5BAGCST model achieved ≈78% accuracy, ≈80% sensitivity, ≈77% specificity, ≈86% AUC, and an MCC of 0.57, which is an improvement of +0.12 in the MCC for the best previous result (101). The second data set we tested was compiled and used by Ofran, Mysore, and Rost (129) and consisted of 274 proteins. Our classifier reached ≈86% accuracy, ≈85% sensitivity, ≈88% specificity, ≈93% AUC, and an MCC of 0.725. The only directly comparable metric reported in this previous work is accuracy. While our accuracy is slightly lower than that reported by Ofran (129), we believe that our model actually offers a more reliable result. In their work, they used sequence to derive evolutionary profiles, sequence neighborhood, and predicted structural features. Their SVM classifier gave its best performance at 89% accuracy. However, their 'positive accuracy' (precision) and 'positive coverage' (sensitivity) were imbalanced. For example, at a sensitivity rate of ≈80% (the number of true positive examples correctly classified), the precision rate is quite low (≈55%), which indicates that the

classifier has low confidence that the predicted positive examples are actually positive. Finally, we tested 109 RNA-binding protein chains originally collected by Terribilini et al. (165) and used in three works (99; 165; 177). Our model achieved ≈76% accuracy, ≈75% sensitivity, ≈77% specificity, ≈83% AUC, and an MCC of 0.523 over this set, which is an improvement of +0.07 in the MCC over the best result (177).

The feature sets used in the previous publications varied between works, as did the type of classifier used for prediction and the type of validation performed. Unequal comparisons of this type are not ideal. However, each of the classifiers we have built using C4.5 with bagging and costing provided consistent results in terms of overall accuracy when trained over various data sets, thus increasing our confidence in this ensemble method.

3.4.7.3 A Prediction Test Case: GP16

In order to validate our method, we predicted the binding residues of the gene-16 protein (GP16), a DNA-packing motor protein in Bacillus phage phi29, for one of our collaborators. This protein contains an ABC transporter nucleotide-binding domain and is known to bind ds-DNA. However, the DNA-binding residues for GP-16 are unknown, and there are no highly-related crystallized protein structures available. Our collaborator had some prior evidence that pointed toward two particular regions of interest in the protein. Using our methods, we were able to focus the costly experimental validation of nucleic acid binding residues on a few key locations in the sequence. We predicted the binding residues of this protein using our sequence-based DNA-binding classifier based on a Platt-calibrated version of the cost-sensitive method described above, which was built using residue charge, identity, and sequence homology

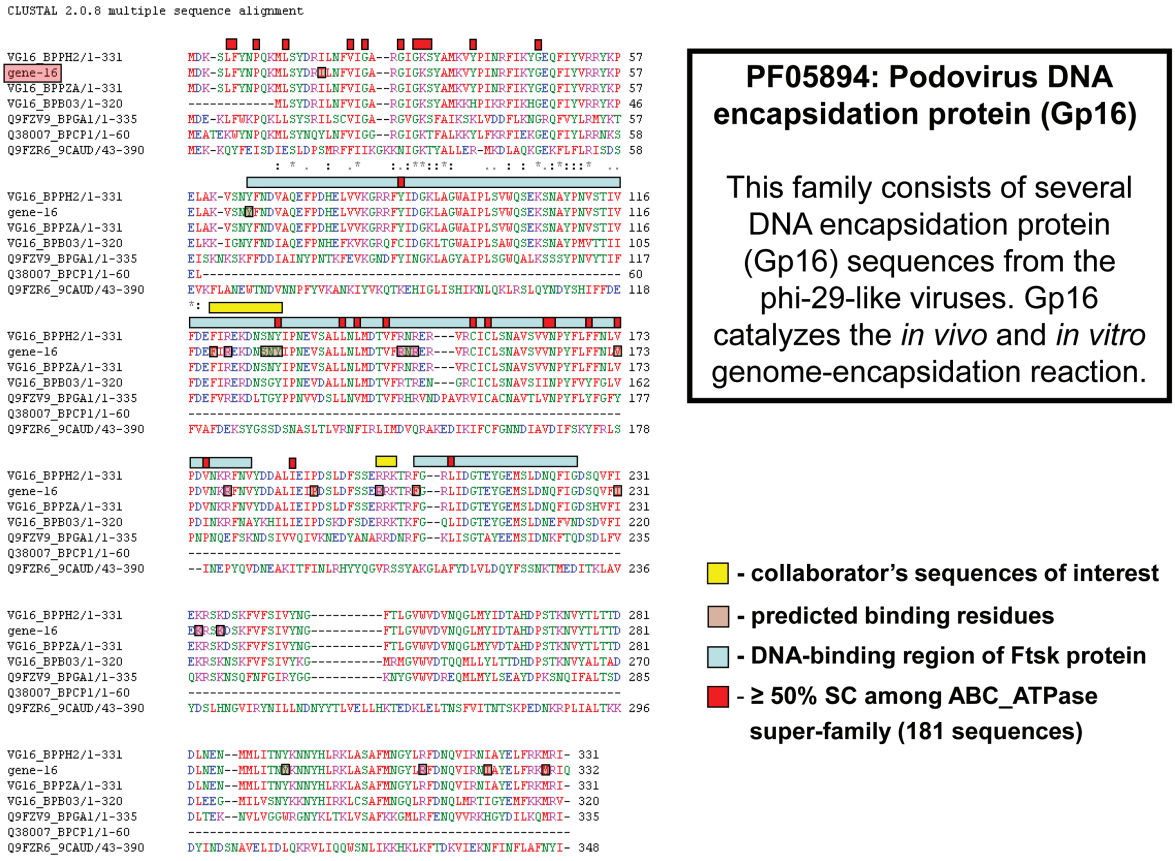


Figure 8: Predicted binding residues of GP16. Yellow boxes indicate regions with potential DNA-binding residues as identified by our collaborator. Pink boxes indicate binding residues predicted by our classifier. Light blue boxes indicate regions which align to the DNA-binding region of the FtsK protein (PDB ID: 2IUS, (122)) in *E. coli*. At the time of this work, FtsK was the most closely-related protein to GP16 having an experimentally-determined 3-D structure in the Protein Data Bank (PDB). FtsK is a multifunctional protein that acts in cell division and chromosome segregation. ≈55% of the predicted binding residues lie within the DNA-binding region of FtsK. Red boxes indicate regions of GP16 having $\geq$ 50% sequence conservation among the 181 sequences in the ABC_ATPase superfamily. This multiple sequence alignment was created using ClustalW (105).

information. Figure 8 shows that our predicted binding residues overlap significantly with the collaborator's residues of interest.

3.4.8 The Nucleic Acid Prediction Server: NAPS

The Nucleic Acid Prediction Server (NAPS, `http://bioinformatics.bioengr.uic.edu/NAPS`) is a publicly-available web server that performs NA-binding residue prediction using sequence-based attributes. NAPS takes a DNA- or RNA-binding protein sequence as input and calculates a set of 301 sequence-based attributes (as described above) for each residue in the test protein. It then returns a list of residues, the predicted class (binding or non-binding), and a score indicating the classifier's confidence in the decision (Figure 9). The model classifier assigns a confidence score between 0 and 1 for each residue in the test protein. This score reflects the level of certainty in the assigned class with a threshold of 0.5. Residues with a confidence score between 0 and 0.5 are classified as non-binding residues; those with a score between 0.5 and 1 are classified as binding residues (Figure 9C). A table of calculated statistics, including the total number of residues binned by confidence score, the number of binding and non-binding residues in the protein, the percentage of each class, and the mean confidence value, is also returned.

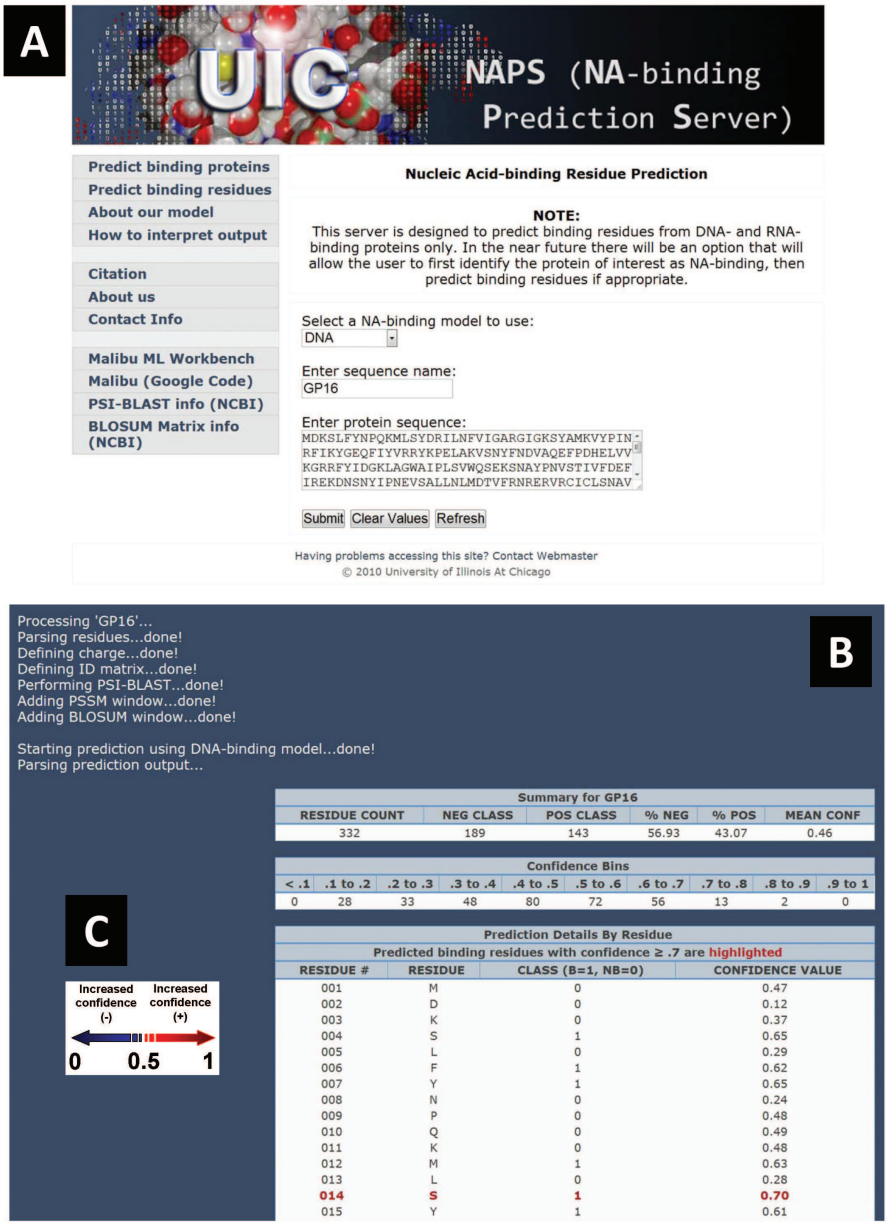


Figure 9: NAPS (`http://bioinformatics.bioengr.uic.edu/NAPS`) screen shots. A) The NAPS home page; B) example output from the DNA-binding prediction of GP16; C) an illustration of the confidence score used by NAPS to determine residue class.

CHAPTER 4

DNA BINDING SITE AND METHYLATION PREDICTION

4.1 DNA Binding Site Prediction Using Interaction Potentials

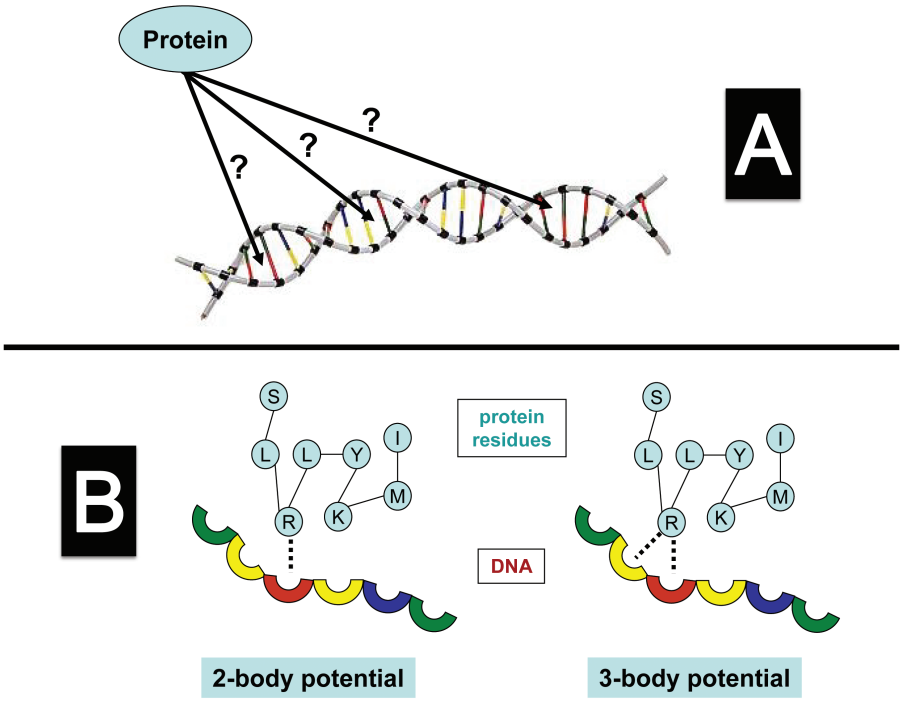


Figure 10: A) At which sites on DNA does this protein bind? B) The two-body potential considers one AA and one DNA base, while the three-body potential considers one AA and two DNA bases. DNA: `http://www.dk.co.uk/static/clipart/uk/dk/future/image_future007.jpg`.

4.1.1 Project Goals

After pinpointing the binding residues within NA-binding proteins, our next step was to identify the sites on DNA to which these residues bind (196). Our goal in this work was to develop a statistical potential based on hydrogen bonding and hydrophobic contacts in order to improve the recognition of protein-DNA interactions. By creating and applying these potentials, we were able to locate the sites on DNA where proteins bind (Figure 10 part A). We adopted a grid-based potential strategy similar to previous works (142; 95). However, we also tested other amino acid representation methods such as Cβ, side chain center of mass (SCM), and the nearest atom to DNA. We also calculated potentials for both two-body interactions (one amino acid, one base) and three-body interactions (one amino acid, two bases) (Figure 10 part B).

4.1.2 Data Set

We selected protein-DNA complexes with a resolution of ≤ 3 Å from the PDB database and removed single-stranded binding proteins and those complexes with bound DNA sections less than 5 base pairs. BLAST (7) was used to keep protein sequence redundancy below 35%. We also checked each complex and discarded all repeated units. This selection process resulted in a final data set of 111 protein-DNA complexes.

4.1.3 Local Coordinate System

Two coordinate systems were used in this work: distance-based interactions and spatial grid-based interactions. While distance-based interactions can be calculated in a straight-forward manner, spatial grid-based interactions require the superposition of DNA bases. To calculate the spatial relationship between a base and an amino acid, one must attach a local coordinate

system to each base. The origin of this local system is placed on the N9 atom for purines and the N1 atom for pyrimidines. The X-axis of this local coordinate system runs from N9 towards C4 for purines and from N1 towards C2 for pyrimidines. The X-Y plane is formed by the N9-C4-N3 atoms for purines and the N1-C2-N3 atoms for pyrimidines. The Y-axis points from N9/N1 to the side where N3 is located. The Z-axis is chosen to complete a right-handed, orthogonal coordinate system. Within this coordinated local environment for each DNA base, we divided the space into cubic boxes of the same size.

4.1.4 Two-body and Three-body Potentials

In our method, we built two-body (one amino acid, one DNA base) and three-body (one amino acid, two DNA bases) knowledge-based potentials from 111 protein-DNA complexes. First, we placed these complexes in a spatial grid-based coordinate system. Then, for the two-body potential we used the expression

$$P_{(i,j,g)} = -RTln(\frac{N_{obs(i,j,g)}}{N_{exp(i,j,g)}}), \tag{4.1}$$

where $N_{obs(i,j,g)}$ is the observed number of residue i in grid g interacting with base j at the origin, and $N_{exp(i,j,g)}$ is the expected number of residue i in grid g interacting with base j at the origin in the reference states. R and T are the Boltzmann constant and the absolute temperature, respectively. The expected number of interacting pairs can be calculated from the quasi-chemical approximation

$$N_{exp(i,j,g)} = x_i x_j N_{obs(g)}, \tag{4.2}$$

where x_i is the percentage of residue type i, x_j is the percentage of residue type j, and $N_{obs(g)}$ is the total number of interaction pairs in grid g. The three-body potential, designed to explore the effects of neighboring DNA bases on contact residue-based recognition, was calculated in a similar manner with the addition of a second base.

The two-body propensity was collected by examining the spatial distribution of each amino acid around each of the four DNA bases. Each base was superimposed on a three-dimensional axis, and amino acid residues were collected based on their relative position to that DNA base. Four kinds of centroids were used to represent the position of the amino acid: Cα, Cβ, side chain center of mass (SCM), and the nearest atom to DNA. The three-body interaction was analyzed through the spatial distribution of each amino acid with respect to a DNA base using the neighboring DNA base as an extra index. The neighboring base could be 5' or 3' with respect to the current base in the DNA sequence, depending on where the amino acid centroid was located. Three types of statistical potentials were calculated: 1) distance-dependent two-body potentials, 2) grid-based two-body potentials, and 3) grid-based three-body potentials. These potentials were evaluated by Z-score and TF-binding site recognition.

4.1.5 Potential Validation

We validated our results in two ways. In the first method, we calculated a Z-score for the native structure complex compared to randomly generated sequences. These random sequences were created using uniform base composition throughout. Fourteen protein-DNA structures from the PDB and 50,000 DNA sequences were used to evaluate the current potentials in order

to compare our results with previously published data (95). The Z-score was calculated using the equation

$$\frac{(E_{native} - < E >)}{(\sigma)}, \tag{4.3}$$

where $< E >$ is the average of the potentials for all random sequences, E_{native} is the potential of the complex, and σ is the standard deviation. A lower Z-score for the native binding sequence indicates a more successful potential (Table II). The second validation method, which we consider a more significant test, involved the recognition of true binding sites among gene regulation sequences using one-dimensional threading. This method utilizes a sliding window that moves along the upstream and downstream regions of a gene. At each position a statistical potential is calculated for each possible binding sequence. The ranking of the potential for a true binding sequence was used to evaluate our potential. Ideally, the true binding site should have the strongest binding affinity in the regulatory region and its associated developed potential should rank first. We used two data sets for this method. The first, 17 known binding sites for the CRP protein in *E. coli*, was used for comparison with previous results (119). The second set included 142 experimentally-determined transcription factor binding sites for the same protein (198).

Potential	14-2-6-2	13.5-3-6-3	9-3-6-3
two-body Cα	–3.91	–3.63	–3.5
two-body Cβ	–3.97	–3.91	–3.74
two-body SCM	–4.11	–3.9	–3.74
two-body nearest	–4.11	–3.9	–3.85
three-body Cα	–5.01	–4.76	–4.2
three-body Cβ	–4.82	–5.64	–4.54
three-body SCM	–5.24	–4.89	–4.26
three-body nearest	–5.18	–5.05	–4.52

TABLE II: The average Z-scores for 14 protein-DNA complexes with different categories of reaction centers and different types of spatial partitions are shown. Four types of centers were used to represent an amino acid at the residue level: Cα coordinate, Cβ coordinate, side chain center of mass (SCM), and the coordinates of the nearest atom to DNA. Three types of grid systems were used to partition space: 14-2-6-2 (cubic size of 2x2x2 Å^3), 13.5-3-6-3 (cubic size of 3x3x3 Å^3), and 9-3-6-3 (cubic size of 3x3x3 Å^3).

4.1.6 Results

4.1.6.1 Z-score Evaluation

The Z-score of the complex was calculated from the distribution of the potential generated from 50,000 random sequences. Table II lists the average Z-scores for the 14 complexes with different categories of reaction centers and different types of spatial partitions. Four types of centers were used to represent an amino acid at the residue level: 1) Cα coordinate, 2) Cβ coordinate, 3) side chain center of mass (SCM), and 4) coordinate of the nearest atom to DNA. In the spatial partition, three types of grid systems are compared. In general, the three-body potentials have a significantly better Z-score than that of the two-body potentials.

The improvement of the Z-score for three-body potentials ranges from 0.7 to 1.7, depending on which grid and which amino acid reaction centers were used.

Next, we determined which reaction center/spatial partition combination for an amino acid residue gave the best performance. Three types of spatial partitions were used: 1) 14-2-6-2, where the grid cubic size was 2x2x2$Å^3$, 2) 13.5-3-6-3, where the grid cubic size is 3x3x3$Å^3$, and 3) 9-3-6-3, where the grid cubic size is 3x3x3$Å^3$. For the two body potential, when the same centroid was used, the grid 14-2-6-2 consistently had the best performance. The grid 13.5-3-6-3 showed median performance among the three. The grid 9-3-6-3 consistently had the worst performance among the three. These results show that in a two-body potential, the grid size of 2Å performed better than 3Å, which indicates that the data distribution varied in the scale below 3Å. The long-range interactions also played an important role, since the box with 14Å outperformed the box with 9Å. Among the same spatial partition, using the nearest atom as the interaction site yielded the best performance, and using center of mass as the centroid gave very similar performance. Using Cβ was preferable to using Cα. The results showed that the nearest atom and center of mass best represent the interaction between a residue and a DNA base at the residue level, and Cα does not represent the interactions adequately.

For the three-body potentials, the results were more complicated. When using Cα, SCM, or nearest atom as the centroid, the 14-2-6-2 grid generated the best performance, followed by the 13.5-3-6-3 grid and the 9-3-6-3 grid. However, when the Cβ atom was used as the reaction center, the 13.5-3-6-3 grid ranked first.

Using Table II, we can also compare the performance of the two-body and three-body potentials under the same conditions of grid size and centroid used. In every comparison, the three-body potential gives a better Z-score than that calculated from the two-body potential. The largest increase came from the 13.5-3-6-3 grid using the Cβ atom as the centroid. The three-body potential outperformed the two-body potential by a Z-score value of 1.73. The smallest increase resulted from the 9-3-6-3 grid using SCM as the centroid, where the three-body potential was better by a Z-score of 0.52. We also calculated the Z-scores for individual protein-DNA complexes using both two-body and three-body potentials with grid 13.5-3-6-3 and the nearest atom as the centroid, then compared them to the Z-scores from published results in which two-body potentials were used for calculation (95). The average Z-score of our two-body potential was 3.9, which is an improvement over the published result of 2.86. In nine of the cases, our potential performed better, and in two cases, our potential performed worse. The performance was the same for the remaining five cases. Our three-body potential had an average Z-score of 5.05 among the proteins in our data set. The three-body potential outperformed the previously published two-body potential in 12 cases, while falling behind in one case and performing equally in another.

4.1.6.2 Recognition of TF Binding Sites

The most important goal in building a statistical potential for protein-DNA binding is to recognize the TF binding site in the promoter region of a regulated gene. We evaluated our potential using the *E. coli* cyclic AMP receptor protein (CRP) as a validation case. CRP is an important transcription factor that regulates more than 100 genes primarily involved in

Ranking	Two-body Number	Two-body %	Three-body Number	Three-body %
1	38	27%	40	28%
2–5	45	58%	43	59%
6–10	26	76%	17	71%
11–20	14	86%	10	78%
21–50	18	99%	19	91%
51–100	1	100%	9	97%
>100	0	100%	4	100%

TABLE III: The statistical potential for 142 CRP binding sites from *E. coli* is shown. The ranking of the potential for a true binding sequence was used to evaluate our potential. 'Two-body Number' and 'Three-body Number' indicate the raw counts of true binding sites ranked at a given level. 'Cumulative %' indicates the percentage of the total number of binding sites.

the glucose-depleted pathway. The contact map was calculated from the protein-DNA complex structure (PDB ID: 1J59, (135)). To test our potential, we scanned upstream sequence from the gene promoter region. Each frame was evaluated with a statistical potential, and the ranking of the known binding site was recorded. The results of the ranking of 17 known CRP-binding sites using the current grid potential are as follows. In seven cases using the two-body potential, the true binding site ranked first, and in 14 others the true site ranked in the top five. Using the three-body potential, in nine of the cases the true binding site ranked at the top, and in 14 cases the true site ranked in the top five. For comparison, the previously published distance-dependent potential (119) has three and nine cases ranked in top one and top five, respectively. Both of our two-body and three-body potentials outperformed this previous result. In addition, we have tested our potential on a larger data set (Table III). We predicted 142 experimentally-

identified binding sites of CRP in *E. coli* (198). When using the two-body potential, in 38 cases the true binding site ranked number one in the promoter region, and in 83 cases the true site ranked in the top five. When using the three-body potential, in 40 cases we were able to rank the true site at the top, and in 83 cases the site fell within the top five. This shows the advantage of our structure-based potential; we do not need to rely on a large number of known binding sites of this transcription factor in order to identify other potential binding sites. Rather, once we have a small number of experimentally-determined structures of protein-DNA complexes, we can glean information such as general binding propensity and recognition of TF-binding DNA sequences from these complexes to predict novel TF-binding sites. Also demonstrated here is that the neighboring effect, represented here as the three-body potential, can be used to improve TF-binding site prediction.

4.2 Prediction of CpG Island Methylation In Human DNA

4.2.1 Project Goals

In Carson, Langlois, and Lu (39) we classified CpGIs as being methylated or unmethylated by analyzing DNA sequence patterns. This was done by training machine learning classifiers from the Malibu machine learning workbench (102) using characteristics from experimentally studied CpGIs. Once the classifiers were trained on known examples, we performed cross-validation over the model. We then identified which combinations of attributes were most important for predicting the methylation status of the CpGIs by analyzing an ADTree built from these attributes. By finding specific combinations of attributes of these islands that determine the state of methylation we may be better able to understand aberrant methylation

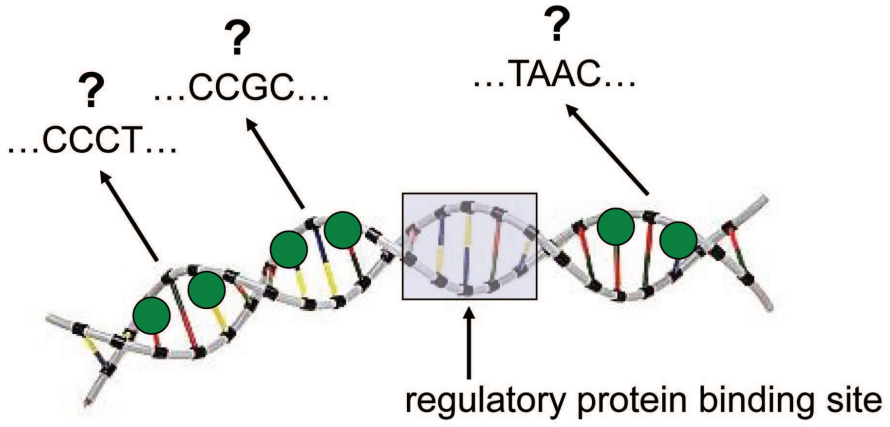


Figure 11: Which DNA sequences will be methylated? DNA: `http://www.dk.co.uk/static/clipart/uk/dk/future/image_future007.jpg`.

of CpGIs. We predicted the methylation status of CpGIs in human chromosome 21 and analyzed specific sequence patterns found to be significantly differentially distributed between methylated and unmethylated islands. These patterns showed a significantly different distribution between methylated and unmethylated islands in a previous work by Bock et al. (30). Using C4.5 with bagging and cost-sensitive learning, we achieved 85.6% accuracy, 82.8% sensitivity, and 86.4% specificity. We then constructed 1000 alternating decision trees using a bootstrapping method and analyzed the nodes that were conserved between the trees.

4.2.2 Data Set and Attributes

The data set for this work was calculated using the DNA sequence of the CpGIs from human chromosome 21, which were originally identified by Yamada et al. (186) and used by Bock et al. (30). We used 27 4-mer sequence patterns that were both strand- and non-strand-specific.

The frequency of occurrence of each sequence pattern within the CpGIs was used as the actual descriptor value.

4.2.3 Machine Learning Algorithm Parameters

For C4.5 with bagging and costing (C4.5BAGCST), we used a 10-fold, stratified cross validation was performed over the training set for one run during both the selection and the validation cycles. For the ADTree classifier, we performed parameter selection on the data set using accuracy as our standard in order to find the optimum number of iterations at which to grow the trees. We found that the best parameters were reached between 20 and 26 iterations. For the purpose of displaying the ADTree, we chose 20 iterations for readability. A total of 1000 trees were created during the bootstrap process, and each tree was run for 20 iterations.

4.2.4 Classifier Evaluation

Refer to the section "Classifier Evaluation Metrics" in **Chapter 2** for details on our evaluation methods.

4.2.5 Results

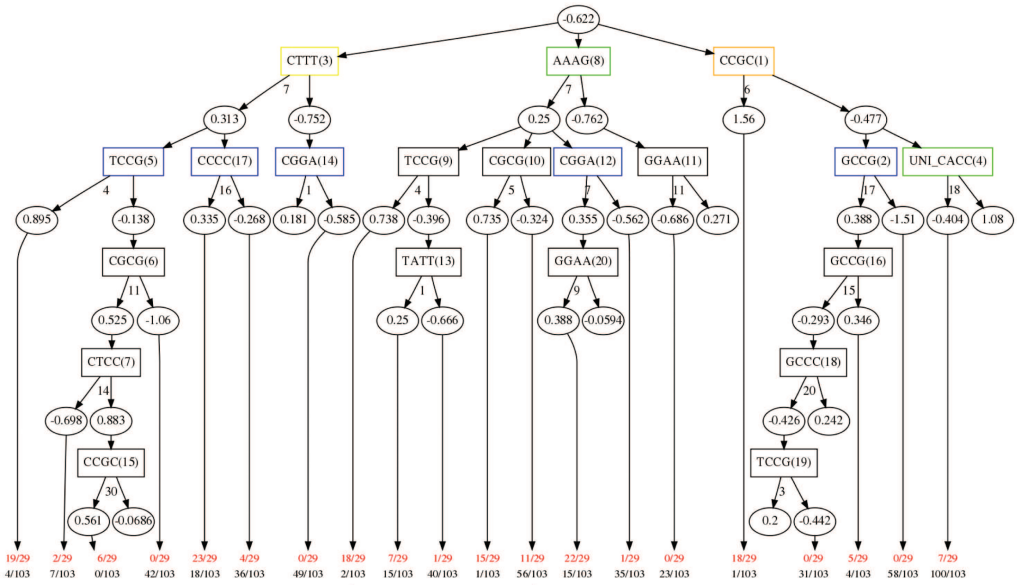
The results from our two classifiers are shown in Table IV, along with results from Bock et al. (30) for comparison. Our models achieved a similar accuracy to those of Bock with more balanced sensitivity and specificity. C4.5BAGCST was able to find many more true positive examples, outperforming the other models by up to 20%. The bootstrapped ADTree achieved a significantly higher MCC than that of the three models reported by Bock. The AUC (area under the ROC curve) for both C4.5cst and ADTree is significantly large. However, no comparison can be made because the AUC was not reported for the Bock models. Though the

metrics allowed us to evaluate the performance of these particular models, they provided little in terms of knowledge about why a particular CpGI is predicted methylated or unmethylated. This was the primary motivation for using a decision tree. The ADTree was chosen for its high accuracy in predicting the methylation status of CpGIs. The intuitive, graphical nature of the ADTree allowed us to find possible (anti-) correlations between sequence features and to elucidate the interdependencies between 4-mer patterns present in the CpGIs. It also allowed us to discover whether repeated patterns of these 4-mers were important in distinguishing between the classes, and if so, how many of the repeated patterns were required to make a decision. Using the ADTree bootstrapping method, we built a tree which shows conserved attribute nodes, decision thresholds, and the number of examples from each class that followed a particular decision path (Figure 12(a)). The most conserved node, the pattern CCGC, was present in $\geq$ 70% of the trees. This agrees with the results from Bock et al., wherein they find that CCGC is the most differentially distributed pattern between methylated and unmethylated CpGIs, with a higher significance for the unmethylated islands (30). However, we can see from the ADTree that if a CpGI contains 6 or more of these patterns, it is highly likely to be methylated ($confidence = 1.56$). This is underscored by the fact that ≈62% (18/29) of the positive examples fit this criterion, in comparison to only 0.1% (1/103) of the negative examples. Furthermore, if a CpGI has less than 6 CCGC patterns and 36 or more strand-independent CACC patterns (UNI_CACC, as per the convention used by Bock et al. (30)), it is highly likely that the island is unmethylated, since ≈97% (100/103) negative examples followed these criteria. One possible reason for both the large number of the UNI_CACC patterns in CpGIs

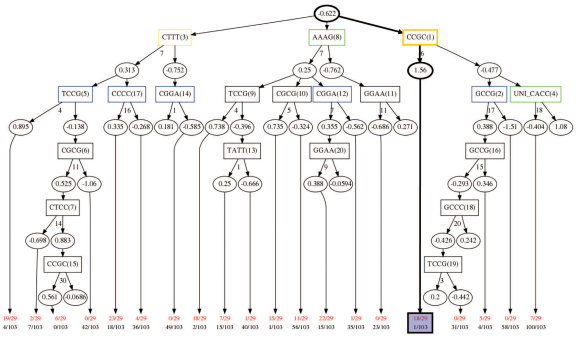
Classifier	ACC	SEN	SPE	AUC	MCC
ADTree[a]	90.63	75.00	95.83	89.06	0.742
C4.5BAGCST[a]	85.60	82.76	86.41	93.54	0.672
libSVM[b]	89.80	61.03	97.96	-	0.684
AdaBoost[b]	89.20	59.83	97.52	-	0.664
C4.5[b]	85.20	65.52	90.73	-	0.566

TABLE IV: The performance of five classifiers over the CpGI methylation data set is shown. [a]The results from our work. [b]The results from Bock et al., reported as being obtained from classifiers built using sequence patterns and the frequency distribution of cytosines, guanines, CpGs, and of the observed/expected ratio of these (426 attributes). ACC = accuracy; SEN = sensitivity; SPE = specificity; AUC = area under the ROC curve. These four metrics are reported in percent form. MCC = Matthews Correlation Coefficient.

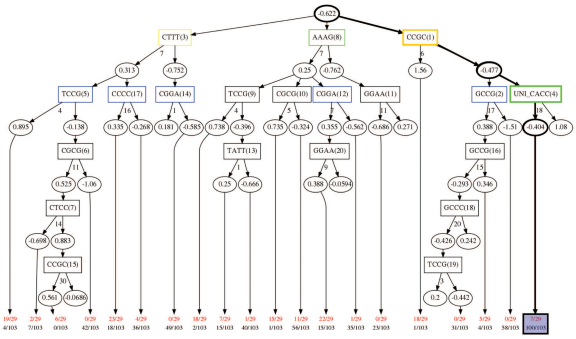
and the conservation of this particular rule is that methylated cytosine has a propensity to mutate to thymine. This leads to a tendency for CpG dinucleotides to become TpG or CpA (87).



(a) An ADTree describing rules for CPG Island methylation



(b) A highly-discerning positive rule



(c) A highly-discerning negative rule

Figure 12: (A) An ADTree created using bootstrap validation is shown. The root node indicates the bias in the data set, i.e., the ratio of positive to negative class examples (in this case methylated vs. unmethylated CpGIs). The number in parentheses within each node indicates the order in which the rules were found. The amount of node conservation between each of the 1000 trees generated is indicated by the color of the box (orange: $\geq 70\%$; yellow: $\geq 50\%$; green: $\geq 30\%$; blue: $\geq 10\%$; black: $\leq 10\%$). The ratios at the bottom of the figure indicate the number of examples from each class that followed that particular decision path (red = methylated, black = unmethylated). Those patterns occurring on either strand are designated as UNI_[pattern] in accordance with Bock et al. (30). An example of a highly-discerning positive rule (B) and negative rule (C) is also shown. A variation of these figures appears in Carson, Langlois, and Lu (39).

CHAPTER 5

GENE REGULATION AND NETWORK MODELS

Most previous work with gene regulation network models attempts to capture snapshots of a gene regulation network in action through (usually stochastic) simulation, with the focus being on the explicit relationship between a transcription factor and a target gene. However, transcription factors often share regulation responsibilities with other transcription factors. This process, called combinatorial regulation, allows relatively few regulators to control the entirety of the transcriptional program. A different type of network can be derived from a gene regulation network, one which describes the relationship between transcription factors that regulate common target genes. This partnership network could be analyzed to gain insight into how transcription factor partners co-regulate their target genes.

We examined how the number of partnership interactions between transcription factors scaled with the number of target genes (23). We analyzed the transcription factor partnership networks of six organisms (*E. coli*, *B. subtilis*, yeast, mouse, rat, and human) and the phosphorylation network for yeast (Figure 14). We noticed two interesting properties in particular. Firstly, the partnership networks lacked the often-observed power law-like degree distribution despite the existence of hubs. Secondly, we found that the number of transcription factor partners varied with the number of target genes and appears to follow an exponential saturation curve of the form $f(x) = a(1 - e^{-bx})$, with the exception of *E. coli* and *B. subtilis*, for which the

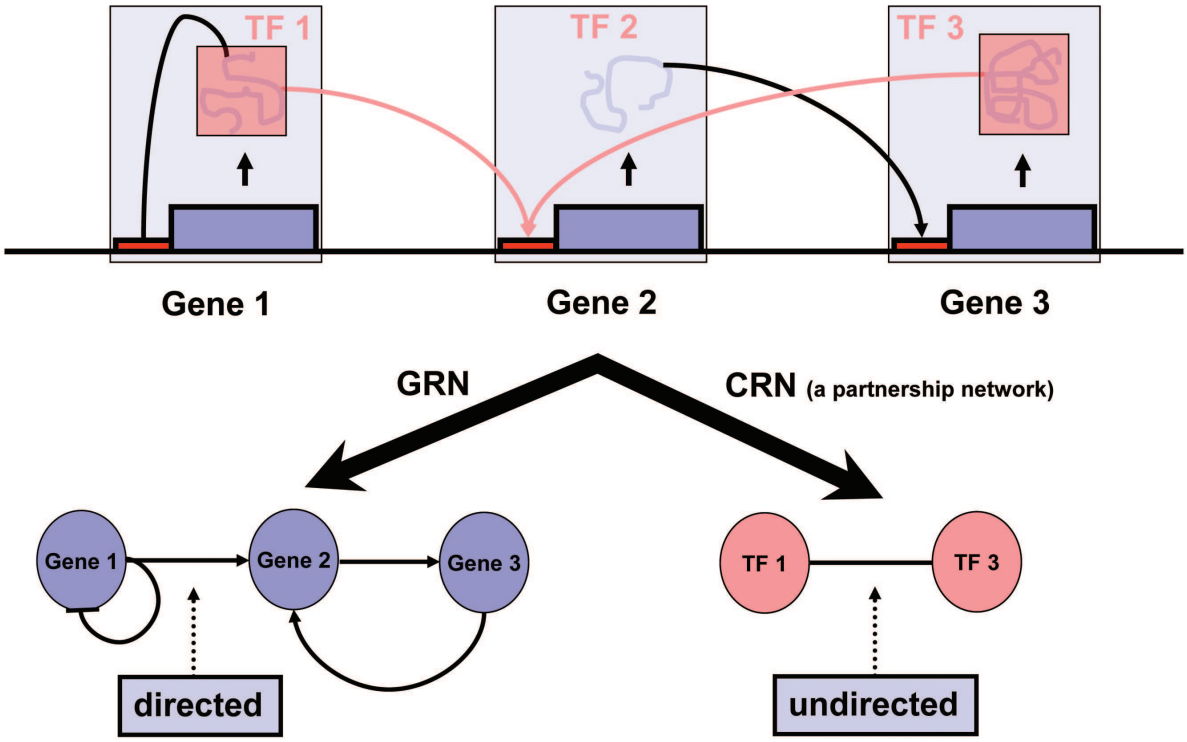


Figure 13: In our work we represent transcriptional regulation in a network context as follows. Given a protein-coding region, a regulatory region, and a protein product from a transcription factor, these three elements are grouped and represented as a single node. From here, two types of graphs can be constructed. A gene regulation network (GRN) describes a set of interactions involving the protein product of one transcription factor binding to the regulatory region of another (or its own) gene and performing some type of regulatory role. This is a directed network. A transcription factor partnership or co-regulation network (CRN) describes the relationship between transcription factors, where an edge between two nodes indicates that these two transcription factors have one or more regulated genes in common. This type of network is undirected. Inspired by a figure from the presentation "Understanding gene regulation from a network perspective" by Gabor Balazsi, PhD, 7/02/2008.

relationship is linear (we show that this could possibly be explained by the operon structure of prokaryotes).

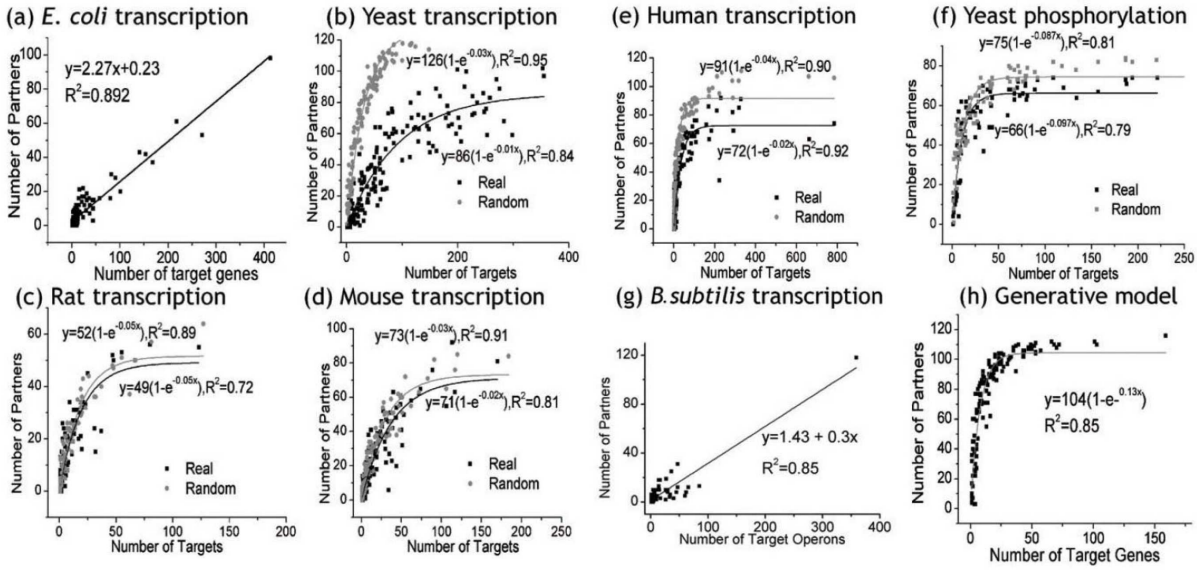


Figure 14: The number of partners vs. the number of target genes for each regulator is shown. (a-e) The transcription network for five species; (f) the phosphorylation network in yeast; (g) the transcription network of *B. subtilis* and (h) the generative model. Black and gray lines correspond to real and random networks respectively. Random networks were generated by shuffling the edges in real networks while maintaining the in- and out-degree of each node. The best fit line and corresponding R^2 value is indicated for each sub-graph. This figure appears in Bhardwaj et al. (23).

In order to find possible explanations of this saturation relationship, we created a generative network model to see if the same co-regulation characteristics appear in a simulated evolutionary environment. Starting with one regulator and one target, generative networks were grown

over 10,000 iterations using a probability-based move set. Possible moves included the addition, deletion, or duplication of regulators and targets as well as the addition or deletion of interactions between a regulator and a target. The resulting co-regulatory network (Figure 14h) showed a similar trend to that in the model organisms. At a certain point, the number of regulatory partners levels off even as the number of regulated targets increases, leading to the characteristic saturation curve. This indicates that the saturation curve seen in these co-regulatory networks could be a product of evolutionary development, during which regulators gain and lose interactions with targets over time. With our model we simulated the evolution of a transcription factor/target gene network in order to observe its development over time. Our model simulates conditions similar to those used to study the development of social networks, which have been studied previously (126; 107), by building the network incrementally.

5.1 Model Development

We describe the generative transcription factor network model we created below.

5.1.1 Overview

Our generative model creates networks with three node types: transcription factors (TFs), target genes (TGs), and transcription factor/target genes (TF-TGs). Each model begins with one TF, one TG, and one edge between the two nodes. Models are allowed to develop over a predetermined number of cycles. One move from a move set derived from prior biological knowledge is applied during each cycle on the basis of a predetermined probability. Essentiality will be defined in terms of network architecture, e.g., 'target gene X must be regulated by at least two transcription factors' or 'genes X, Y, and Z must always be regulated by the same

transcription factor(s)'. If the essential condition for that particular model is not met after removal of the transcription factor with the highest degree in the current network (i.e., the largest hub), the network fails. If the condition continues to be met after failure, we proceed with development. In this way we hope to gain some insight into co-regulation behavior. Each move in our move set and the reasoning behind it is described in more detail below.

5.1.1.1 Addition of a New TF/TG

In order to grow the network, nodes must be added. There are two options in the move set: 'Add TF' and 'Add TG'. Both of these moves are constructors; they merely add one node to the set of current nodes of that type. No ingoing or outgoing edges are added during this move. Besides being necessary for growing the network, these moves also help our model account for known biological phenomena. The addition of new nodes allows each network to develop uniquely. Balaji, Iyer, Aravind, and Babu showed that evolutionary conservation and node connectivity showed no correlation in several yeast species, meaning that regulatory hubs are not essential in these organisms. Each organism has developed its own set of highly connected regulators, allowing it to survive in different types of environments (18). This interpretation is supported by independent studies that show that different lineages of eukaryotes have evolved their own set of transcription factors and regulating hubs (14; 108; 48). Additionally, *E. coli* and yeast only have two DNA-binding domain families in common (162), further indicating that each organism has its own model for gene regulation. Addition of nodes also mimics an important phenomenon in prokaryotes; the horizontal transfer of genes between organisms. This phenomenon causes bacteria to acquire new functions from other bacteria in close proximity.

Gelfand points out that horizontal transfer of bacterial enzymes outnumber duplications by as much as 10 times (72).

5.1.1.2 Deletion of TF/TG

Our model also accounts for the possibility that a gene could become inactive due to mutation(s). If some type of deleterious DNA mutation occurs within a gene, the gene may become a pseudogene (i.e., an inactive gene). As in the case of the addition moves, a separate probability is assigned to the deletion of a TF and the deletion of a TG This move simulates random failure within the network as defined by Balaji, Iyer, Aravind, and Babu as well as Albert, Jeong, and Barabási (18; 5).

5.1.1.3 Duplication of TF/TG with Partial Edge Inheritance

Two main theories exist for the formation of gene regulation networks (44). The first involves the duplication of a gene and subsequent diversification of interaction from the parent gene. This phenomenon, which is referred to as the duplication and divergence model, is thought to be ubiquitous in nature. Both yeast and *E. coli* show extensive gene duplication, which implies that this phenomenon is not biased by prokaryotic horizontal transfer or operon structure. Additionally, the duplications appear to be gene-by-gene instead of by genetic circuit or module. Most duplicated genes have interactions in common with their parent (77% *E. coli*, 69% yeast) (162). Our model allows both TFs and TGs to be duplicated with separate probability. The duplicate copy inherits a certain percentage of the parent interaction (currently 30% for both TFs and TGs).

Transcription factor/target gene interactions are represented in a network model by directed edges. The second theory for the development of the gene regulation network hypothesizes that regulation motifs arose independently through recruitment of unrelated genes. Conant and Wagner observed that regulatory motifs do not have a common ancestry and that duplicate regulatory genes have a random distribution in relation to different motif types, which implies that these regulators can easily gain/lose interactions with target genes (44). It has also been observed that up to 1/2 of the interactions between TFs and TGs have been gained after a gene has been duplicated, and that roughly 10% of regulatory interactions are innovations, meaning they appear to have been gained randomly (162).

5.1.1.4 Transformation from TF to TF-TG

In order to allow for the multiple levels of transcriptional control seen in model organisms, we allow TFs to become target genes themselves without losing their current regulatory interactions. During this move, a TF is picked at random to become a target gene, and a second TF is chosen to become the first transcription factor's regulator.

5.1.1.5 Self Interactions

This type of regulation involves the protein product of a gene performing a regulatory action on its own gene. These auto-regulatory loops are the only example of feedback loops that have been observed in the *E. coli* regulatory network (149). We do not include a separate move in our model for adding a self-regulating interaction. Self-interactions can still be acquired during the transformation from TF to TF-TG if a transcription factor is chosen to be its own regulator.

5.1.1.6 Attack on Highly-connected Nodes

To simulate a direct attack on the networks, we include a move which removes the most highly-connected node in the network during the current cycle. Networks with an inhomogeneous degree distribution are known to vulnerable to direct attacks such as this (18; 5). By applying this move we hope to 'weed out' those networks which do not provide backup vertices and edges, the idea being that if there is no backup path to our defined essential interactions, the network will be non-viable. This parameter can also be varied to mimic a harsh environment in order to analyze any differences that may exist in either the partnership network between transcription factors or the gene regulation network when attacks are more prevalent.

5.2 Saturation Curves In Other Partnership Networks

5.2.1 Social Networks

The World Wide Web has made available a vast amount of interaction data. The number of social networks of all types, email exchanges, and blogs provide a rich source for comparative analysis with biological networks. To see how the saturation curve fit social data, we examined partnership in two directed social networks. We studied a blog linkage network that consisted of inter-linked blog entries where the nodes are blogs. An edge exists between two nodes if the two blogs link to a common blog. We also studied an email network obtained using a set of emails exchanged among users. Nodes are email users, and a link exists between them if they have sent an email to a common user. We found that both these networks displayed the same kind of exponential saturation relationship between the output (the number of out-going links or email recipients) and the number of partners (co-linked blogs or co-senders of email) (Figure 15). This

suggests that in social networks as well, the size of partnership network saturates at a certain value even as the output of the group gets exponentially complex, highlighting the similarities between the organizational structure of social and biological networks.

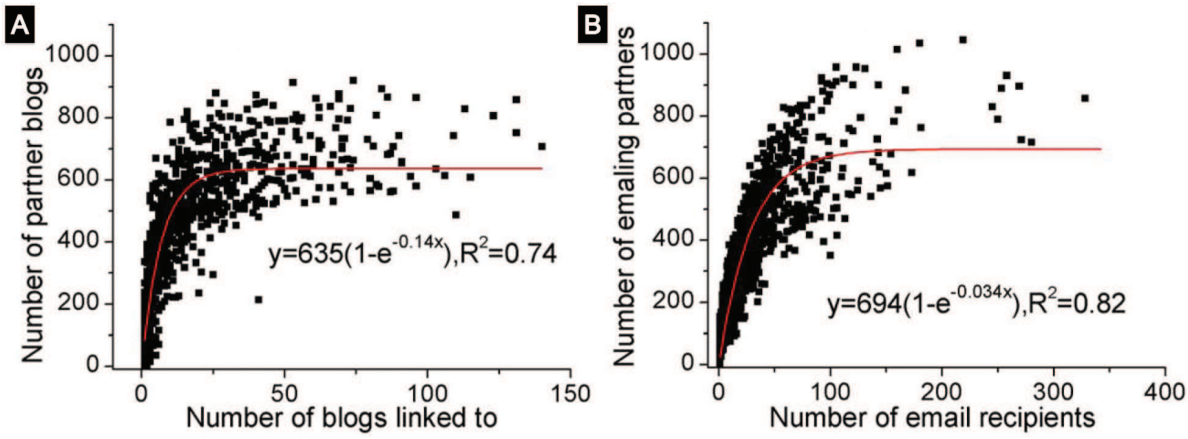


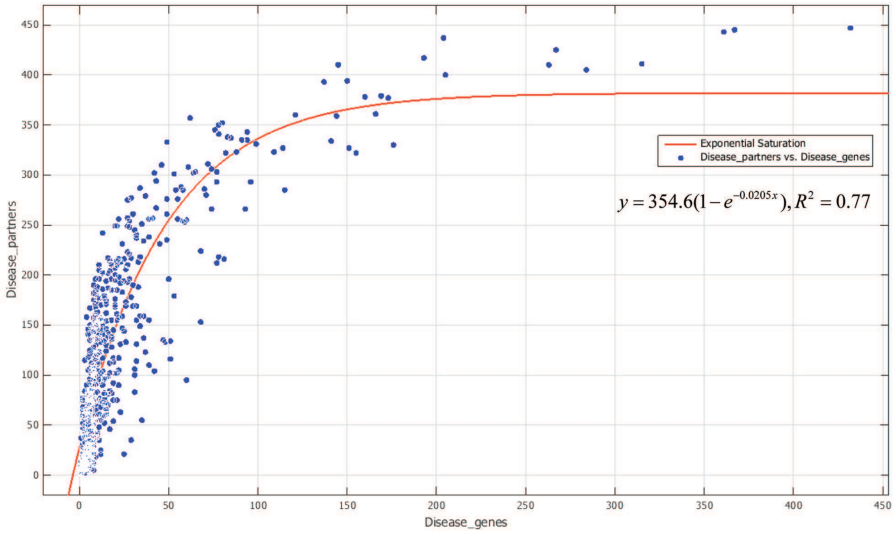
Figure 15: (A) A plot of the number of blogs linked to a user's blog (X-axis) versus the number of blogs which point to the same blogs (Y-axis). Each data point corresponds to a blog in the blogs network. (B) A plot of the number of recipients of a user's email versus the number of other users who email the same recipients. Each data point corresponds to a user who sends an email in the email network. This figure appears in Bhardwaj et al. (23).

A limit on the size of the social network an individual can develop has been reported previously as well. It has been suggested before that a human brain allows a stable network of about 150 contacts (known as the *Dunbar number*) (56). Similarly, the average number of "friends" on social networking sites like Facebook (`https://www.facebook.com/`) has been

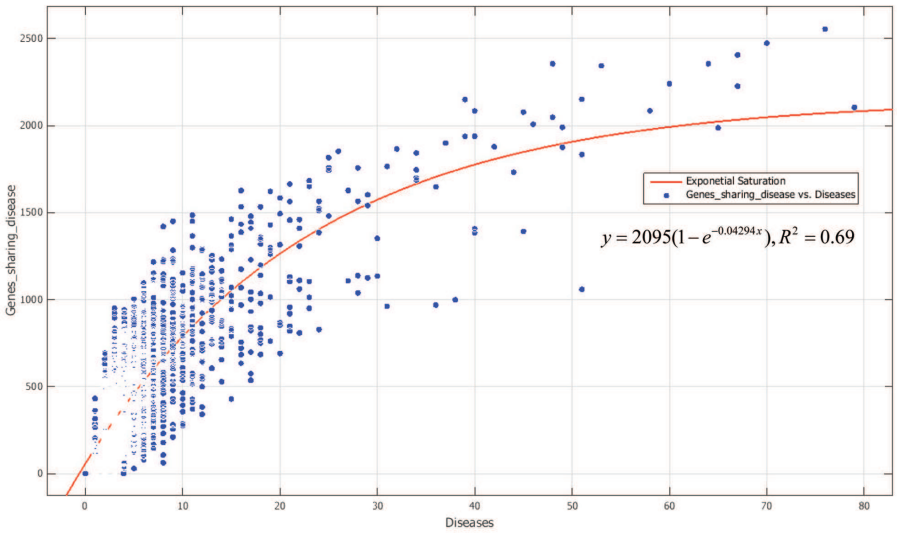
observed to be 120 (10). These observations and our results above are indirectly related: setting a cap on the number of individuals one interacts with loosely limits the number of other individuals (or partners) that interact with the same group.

5.2.2 Human Disease Network

We examined another type of partnership network related to human disease. Disease-gene associations have been found and documented using such constructs as the Online Mendelian Inheritance In Man (OMIM) database (78), the Disease Ontology (143), and NCBIs GeneRIF (`http://www.ncbi.nlm.nih.gov/projects/GeneRIF/`). Using a consolidated version of the Disease Ontology termed DOLite (55), we created a map for the human "diseasome" similar to that from Goh et al. (74). From this bipartite network we extracted two projections: 1) a disease partnership network and 2) a disease-gene partnership network. We then performed curve fitting analysis on the two networks using MATALB (123). For the disease partnership network, which compares the number of diseases that share a gene (disease partners) to the number of genes associated with that disease, the saturation curve $f(x) = a(1 - e^{-bx})$ fits the data with an adjusted R^2 value of 0.77 (16(a)). The disease-gene partnership graph, which compares the number of genes sharing a disease relationship with gene x to the number of disease associated with gene x, fits with an adjusted R^2 value of 0.69 (16(b)). The similarity of these graphs to those of the social networks above suggests that there may be a similar phenomenon underlying disease relationships. We will look more closely at the disease partnership and disease-gene relationships in the next chapter.



(a) Disease Partnership: the number of genes associated with disease x vs. other diseases sharing gene(s) with disease x.



(b) Gene Partnership in Disease: the number of diseases associated with gene x vs. other genes sharing disease(s) with gene x.

Figure 16: Curve fitting analysis for two projections of the human disease network. (a) The exponential saturation equation $f(x) = a(1 - e^{-bx})$ fit the disease partnership network with $R^2 = 0.77$. (b) The disease-gene partnership relationship follows this distribution with $R^2 = 0.69$. This figure was created using MATLAB (123).

CHAPTER 6

MOLECULAR NETWORKS AND HUMAN DISEASE

6.1 A Data Warehouse for Human Molecular Networks

In order to facilitate our work with human disease networks, we created a human molecular data warehouse (Figure 17). We collected human molecular and disease data from several categories including gene annotation, protein sequence, structure, and interaction, TF targets, functional and metabolic data, miRNA targets, drug target information, disease, and data sets from papers of interest. Sources for this data warehouse include NCBI's Entrez Gene (118) and Reference Sequence (RefSeq) (138) databases, the HUGO Gene Nomenclature Committee data set (HGNC) (145), the Human Protein Reference Database (HPRD) (93), EBI's Universal Protein Resource Knowledgebase (UniProtKB) (46), Gene Ontology (GO) (11), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (89), DrugBank (94), the Transcriptional Regulatory Element Database (TRED) (195), the Database of Interacting Proteins (DIP) (184), the OMIM MorbidMap database (78), Disease Ontology and Gene Reference Into Function (DORIF) (133), Disease Ontology Lite (DOLite) (55), and others. In addition, we created several ID conversion tables in order to relate our data sets to one another and perform complex queries. This data warehouse allows us to integrate public databases with data sets from papers and collaborators, facilitating mining for potential connections between data (e.g., genes involved in the same disease and metabolic pathway).

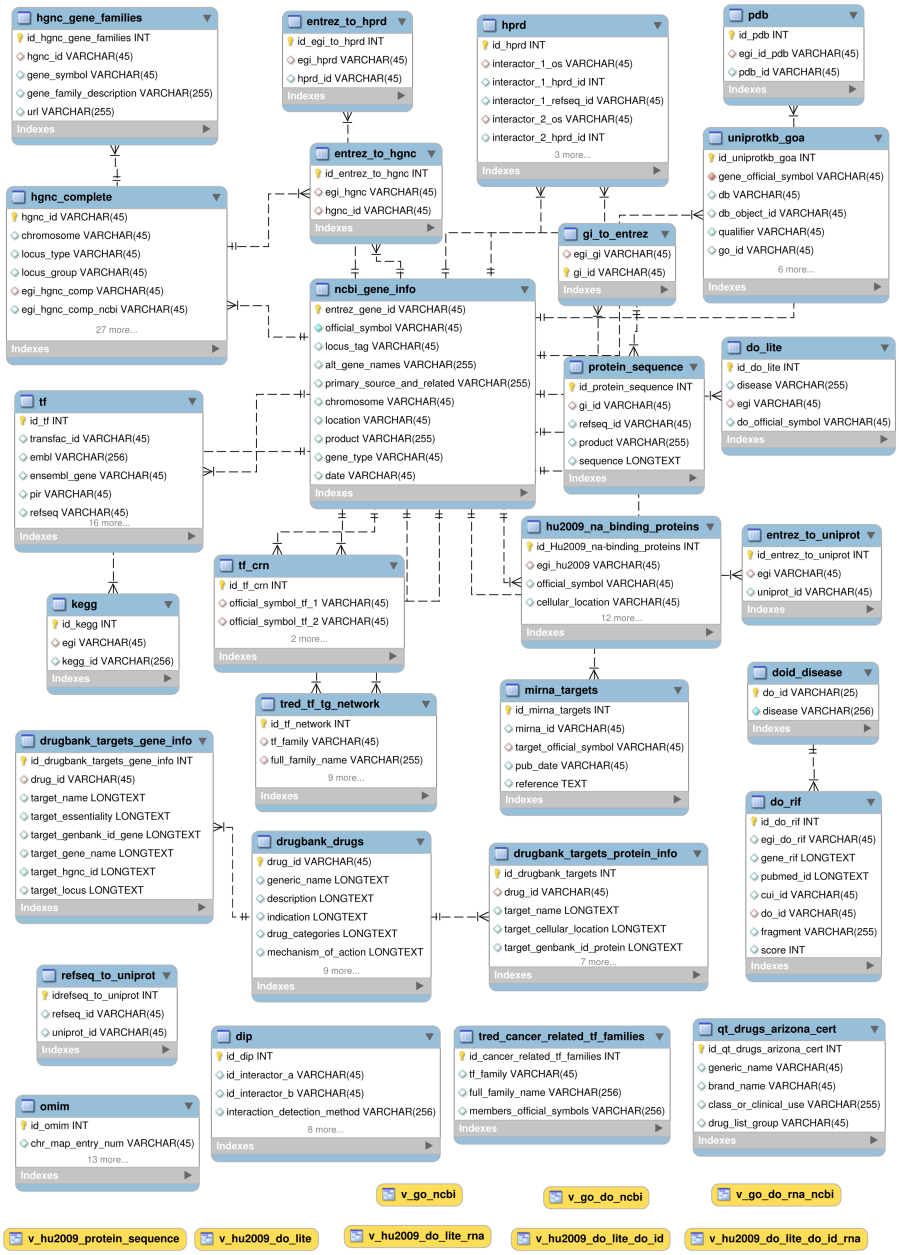


Figure 17: The Human Data Warehouse is currently composed of twenty eight tables from several categories including gene annotation, protein sequence, structure, and interaction, TF targets, functional and metabolic data, miRNA targets, drug target information, disease, and data sets from papers of interest. Also included are various ID conversion tables which allow us to relate data sets to one another to facilitate complex queries. Stored queries, or views, are displayed at the bottom of the figure in yellow. This figure was created using the MySQL Workbench (`http://dev.mysql.com/doc/workbench/en/`).

6.2 Disease-related Genes in Conserved Human TF Network Motifs

A transcription factor network provides a map of interactions between regulating proteins (transcription factors) and regulated genes (target genes). Within this network lie patterns of connections between these two. These patterns, or motifs, give transcription factors (TFs) a variety of tools for regulation depending on how much or how little of a protein product is needed at any given time or within a particular tissue. Having looked into the relationship between co-regulating partners and their target genes (TGs), we now examine more closely how these TFs regulate their targets. We are also curious to know how disease-related genes relate to the transcriptional regulatory network. Past analysis has shown that certain changes in transcriptional regulation can lead to disease phenotypes such as infertility (58), ocular diseases such as glaucoma (1), several developmental diseases, (45), and cancer (68). One way to examine this problem is to identify the presence and location of disease genes in conserved motifs within the human transcription factor network. Are disease genes, specifically those that are cancer-related, overrepresented in statistically significant motifs? Are they regulated by common mechanisms in the TF network? Do cancer genes occupy a particular position within these motifs more often than other disease genes or non-disease genes? What about for specific diseases such as breast, colon, and lung cancer?

6.2.1 Construction of the Network

Following the procedure in Figure 18, we first obtained 154 TFs and 2948 TGs with a total of 6883 interactions from the Transcriptional Regulatory Element Database (TRED, `http://rulai.cshl.edu/TRED/`) (195). TRED is a collection of cis- and trans-regulatory elements

for mouse, rat, and human. It also contains a curated list of 36 families of transcription factors and experimental evidence linking them to their target genes. Although this network does not represent a comprehensive account of human transcription factor/target gene interactions, the proportion of TFs to TGs (≈5.2%) is very close to the observed value of 6% across 32 human tissues as shown by Vaquerizas, Kummerfeld, Teichmann, and Luscombe (170).

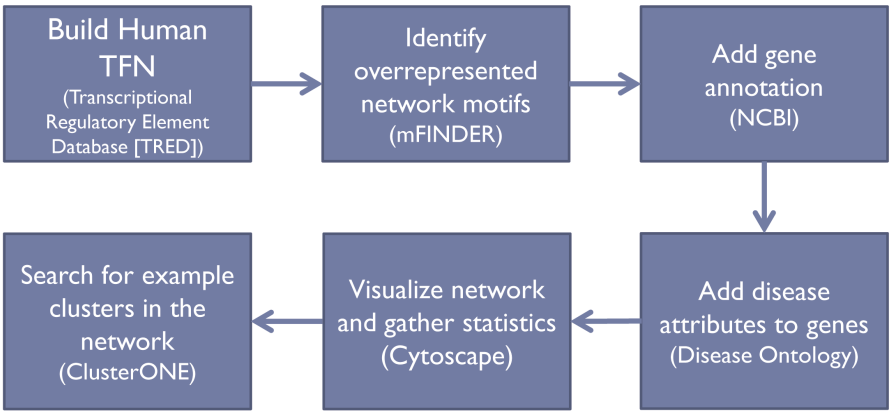


Figure 18: Our protocol for the identification of disease genes in transcription factor network motifs is shown. We began by building a human transcription factor network using the TRED database. The resulting network included 154 TFs, 2948 TGs, and 6883 interactions. We then identified overrepresented network motifs using mFINDER by comparing the human TF network to 1000 random networks. Next, we added gene annotation from NCBI's Entrez Gene database along with disease annotation from the Disease Ontology (560 disease categories). We subsequently calculated network statistics using Cytoscape and Network Analyzer. Finally, we searched for example clusters in the network using ClusterONE.

Our second step was to identify common motifs in the network. We used Motif Finder (mFINDER) (91), a network-centric algorithm capable of creating exhaustive sub-graph lists. mFINDER uses concentration, the ratio of the number of occurrences of a motif versus other motifs of the same size in the real network, as a significance metric. We identified overrepresented patterns in the real network compared to 1000 randomly generated networks. The random networks were created using the same number of incoming, outgoing, and mutual edges as the real network. Each begins as an identical copy of the original TF network. The source and targets of the edges are then randomly switched between nodes, resulting in a randomly-connected set of nodes. The number of times this switching occurs is an arbitrary number between 100 and 200 times the number of edges in the real network.

We then obtained gene annotation from NCBI's RefSeq database (138) and disease information from DOLite (55). DOLite contains 560 disease categories which are a collection of aggregated disease terms from the Disease Ontology (`http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page`). Next, we visualized the network using Cytoscape (151) and subsequently identified example network clusters using the Cytoscape plug-in ClusterONE (127) in order to find examples of the overrepresented motifs in the real network.

6.2.2 Results

We found that 4029 human genes have an associated disease according to DOLite. We identified 1217 genes (≈39% of the human transcriptome) which have a known association with at least 1 of these 560 disease categories (see Figure 19). 622 genes (≈20%) were associated with one or more types of cancer, 176 genes (≈5%) with breast cancer, 106 genes (≈3%) are

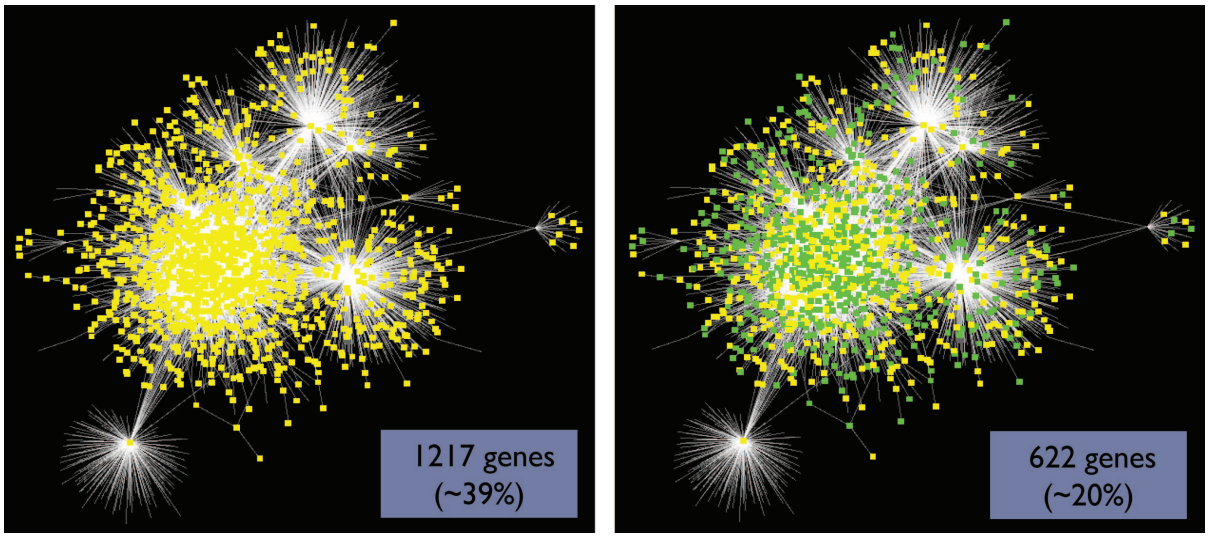


Figure 19: A global view of the human transcription factor network is shown. 154 TFs and 2948 TGs with a total of 6883 regulatory interactions acquired from the TRED database are shown. Left: 1217 genes (≈39% of the human transcriptome) have a known association with at least 1 of 560 disease categories from DOLite (yellow). Right: 622 genes (≈20%) were associated with one or more types of cancer (green), 176 genes (≈5%) with breast cancer, 106 genes (≈3%) with colon cancer, and 97 genes (≈3%) with lung cancer. 48 transcription factors (≈31% of TFs in the database) were associated with cancer (not shown). These networks were visualized using Cytoscape (151).

with colon cancer, and 97 genes (≈3%) with lung cancer. 48 transcription factors (≈31% of TFs in the database) were associated with some type of cancer.

According to our analysis, one 3-node motif and seven 4-node motifs were overrepresented in the human TF network (Table V). Three examples, one 3-node and two 4-node motifs, are shown in Figure 20. Motif 46 (A) is a regulating feedback motif with a feed-forward loop (FFL). This motif is common in developmental and signaling transcription networks (6) and allows for a rapid response to signals (e.g., ON to OFF) while providing a delayed reaction to move in the opposite direction (OFF to ON). This is important for filtering noisy input and ignoring small fluctuations while allowing for a rapid response to stimuli. The most common variations of this pattern are the coherent regulating double positive feedback motif (all interactions enhance transcription) and the coherent double negative feedback motif (all interactions repress transcription). Motif 222 (B) is a combination feedback/bi-fan. This pattern allows for combinatorial control depending on the input function of each gene. The bi-fan is a pattern of joint regulation which usually generalizes to dense overlapping regulons (DORs) in the larger network. Genes in DORs share a global function such as nutrient metabolism and biosynthesis (6). Motif 2206 (C) is a combination feedback/multi-FFL, which is useful for sign-sensitive delay/acceleration and pulse generation. This allows the genes to be expressed in a particular order, and can act as a persistence detector for each output (6).

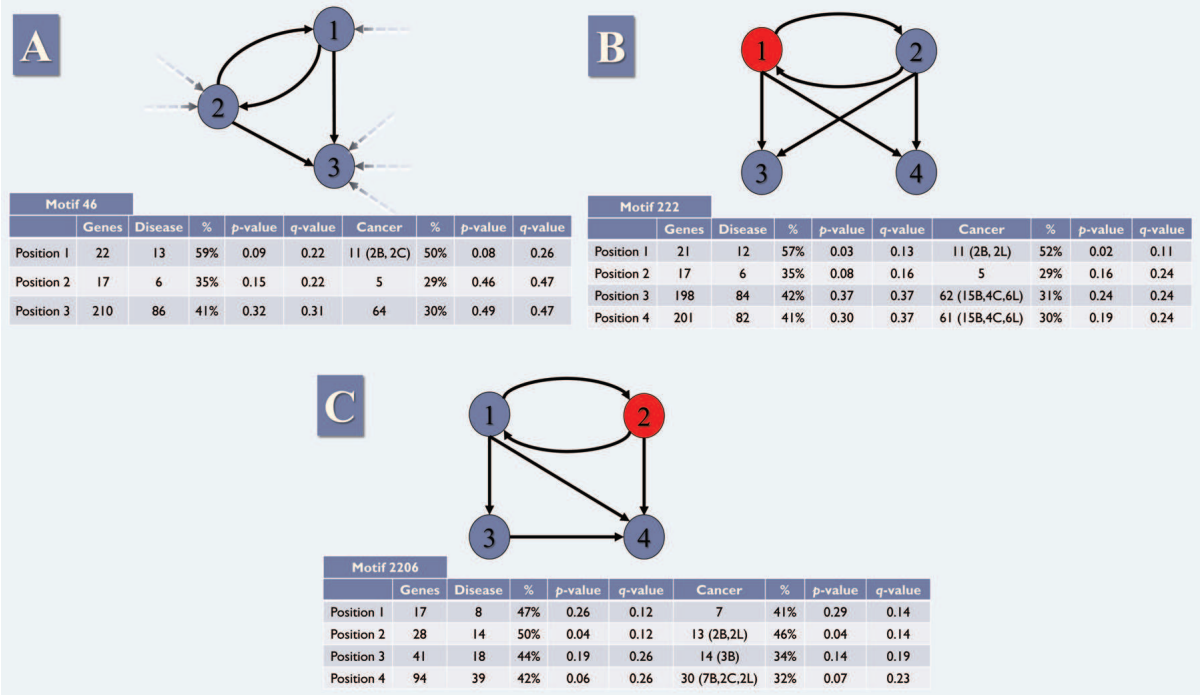
One interesting observation is the high percentage of disease- and cancer-related genes in the motif unit as a whole. In all three motifs, a large percentage of the disease genes are cancer-related. For motif 46, 42% of genes are disease-related and 32% of genes are cancer-related,

Motif ID	**Motif Type**	**Occur (Real)**	**Occur (Rand)**	***p*-value**	**UV**	[ ]
46	3-node	470	332.305+-26.5104	0	9	0.4832
222	4-node	4913	2968	0	8	0.029
904	4-node	1800	1446	0	13	0.0106
908	4-node	950	626	0	7	0.0056
922	4-node	298	216.1667	0	5	0.0018
972	4-node	703	361	0	4	0.0042
2206	4-node	486	300	0	5	0.0029
2462	4-node	264	155	0	4	0.0016

TABLE V: Significant motifs in the human transcription factor network are shown. We identified 8 types of overrepresented 3- and 4-node motifs. 'Occur (Real)' indicates the number of times the motif occurred in the real network, 'Occur (Rand)' is the mean number of occurrences in random networks, 'UV' is the unique value, which is the number of times a motif appears in the real network with a completely different (disjoint) set of genes, '[ ]' is the concentration, which is the ratio of the number of occurrences of a motif versus other motifs of the same size in the TF network.

as compared with the percentage of these categories over the entire network (39% and 20%, respectively). Notably, 76% of disease genes in this motif are involved in at least one type of cancer. The same scenario applies to the other two motifs, with motif 222 having 42% disease-related genes, 32% cancer-related genes, and 76% cancer-related disease genes, and motif 2206 having 44% disease-related genes, 36% cancer-related genes, and 81% cancer-related disease genes.

If we look at specific positions within the motifs, several characteristics apply to each example. Firstly, the number of transcription factors in the first and second positions of motifs 46 and 222 and the first, second, and third positions of motif 2206 are larger than that of the



Motif 46

	Genes	Disease	%	p-value	q-value	Cancer	%	p-value	q-value
Position 1	22	13	59%	0.09	0.22	11 (2B, 2C)	50%	0.08	0.26
Position 2	17	6	35%	0.15	0.22	5	29%	0.46	0.47
Position 3	210	86	41%	0.32	0.31	64	30%	0.49	0.47

Motif 222

	Genes	Disease	%	p-value	q-value	Cancer	%	p-value	q-value
Position 1	21	12	57%	0.03	0.13	11 (2B, 2L)	52%	0.02	0.11
Position 2	17	6	35%	0.08	0.16	5	29%	0.16	0.24
Position 3	198	84	42%	0.37	0.37	62 (15B,4C,6L)	31%	0.24	0.24
Position 4	201	82	41%	0.30	0.37	61 (15B,4C,6L)	30%	0.19	0.24

Motif 2206

	Genes	Disease	%	p-value	q-value	Cancer	%	p-value	q-value
Position 1	17	8	47%	0.26	0.12	7	41%	0.29	0.14
Position 2	28	14	50%	0.04	0.12	13 (2B,2L)	46%	0.04	0.14
Position 3	41	18	44%	0.19	0.26	14 (3B)	34%	0.14	0.19
Position 4	94	39	42%	0.06	0.26	30 (7B,2C,2L)	32%	0.07	0.23

Figure 20: Three examples of significant motifs in the human transcription factor network. Each circle represents a node, and the number inside the circle represents the position of a gene within the motif. Being that the transcription network is directed, the position number is determined by the in-degree of the node in the real network (dashed arrows) so that a lower position number indicates a lower in-degree and thus a higher-level regulatory position. This applies to the two 4-node examples as well; the arrows are left out for simplicity. The total number of genes, the number of disease-related and cancer-related genes, and their percentage is shown, along with the statistical significance at each position. Nodes for which there is a statistically significant number of disease- and cancer-related genes (i.e., $p \leq 0.05$) are shown in red. p-values were calculated using a standard one-tail t-test. q-values were calculated using the false discovery rate (FDR) method of Pike (136).

downstream positions (values not shown). This is expected and is due to the regulatory activity occurring in these locations (indicated by out-going edges). Some TFs occupy the third position of motif 46, the third and fourth positions of motif 222, and the fourth position of motif 2206. This is also expected, since these patterns do not exist in isolation but are part of the larger graph, and, in the real network, these TFs most likely regulate other proteins in motifs located downstream from these examples. For motif 2206, positions 1, 2, and 3 are each regulatory in nature and follow a top-down hierarchy. Positions 2 and 3 are both the second node in a feed-forward loop, and we would expect each of these positions to have similar number of TFs, which they do. Secondly, position 3 in motif 46 and positions 3 and 4 in motifs 222 and 2206 are associated with a larger overall number of genes. This too is anticipated, since one transcription factor may regulate many genes. Thirdly, there is one particular position in each of these motifs that is associated with a larger percentage of disease- and cancer-related genes relative to the other positions. In motif 46, 13/22 (59%) of genes in position 1 are associated with at least one disease compared to 35% and 41% for positions 2 and 3, respectively. Cancer-related genes are also more common in position 1, with 11/22 (50%) genes having a cancer association while the equivalent value for positions 2 and 3 are 29% and 30%, respectively. This pattern holds for motif 222 as well, and to a lesser extent for motif 2206. For this last motif, the more equal distribution of disease- and cancer-related genes may be a result of the nature of the motif itself as mentioned above.

Figure 21 illustrates a comparison between the number of disease- and cancer-related genes at motif positions versus the percentage over the entire network. Position 1 in motif 46 is

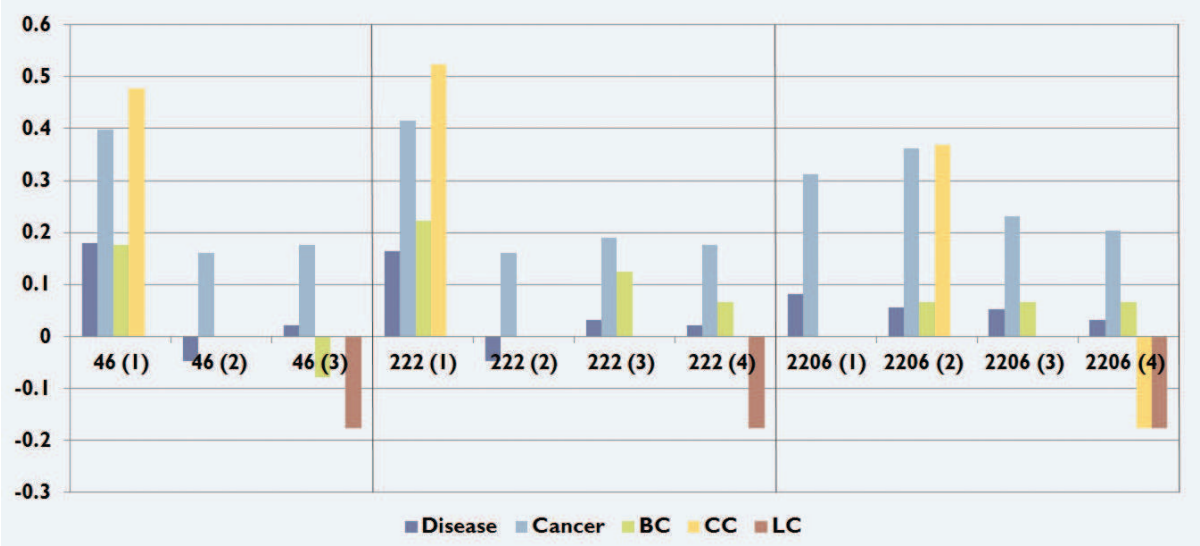
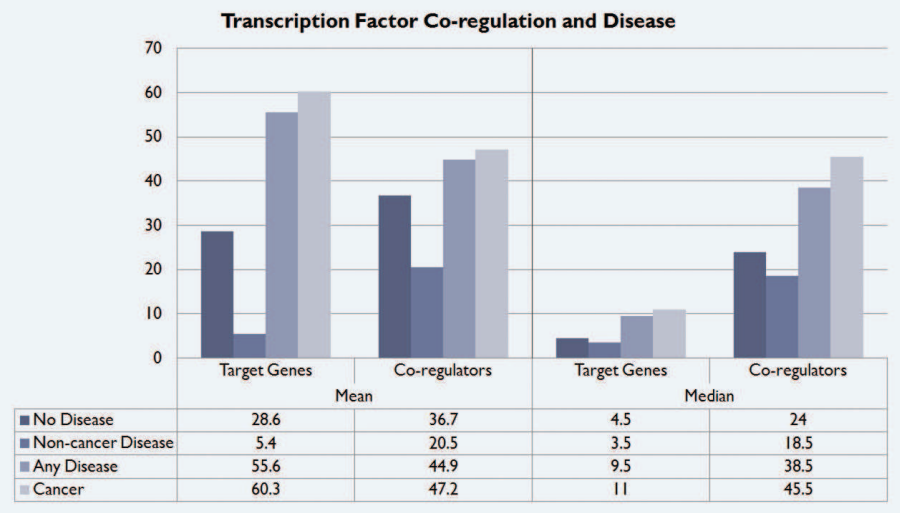


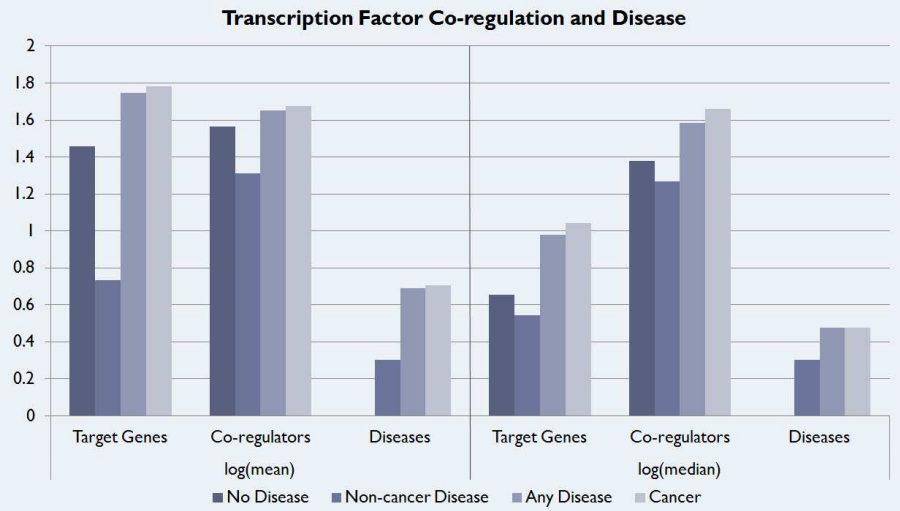
Figure 21: A log-scale graph of the number of genes in three motifs by the position these genes occupy (in parentheses). Five categories are shown: disease, cancer, breast cancer (BC), colon cancer (CC), and lung cancer (LC). Each column is normalized by the percentage of each category in the entire TF network.

occupied by more disease- and cancer-related genes, as well as genes involved in breast and colon cancer. A similar pattern holds at position 1 of motif 222 and position 2 of motif 2206. Interestingly, cancer-related genes are more common in all motif positions relative to both other disease genes at that particular position and the network as a whole. This could be due to the effects of signaling cascades in cancer pathways, with down-stream genes being affected by aberrant signaling upstream, whereas other diseases may be caused by mutations or other malfunctions which remain isolated.

To assess the significance of disease- and cancer-related genes occupying specific positions within these motifs, we calculated the p-values for each. Using a p-value threshold of 0.05,

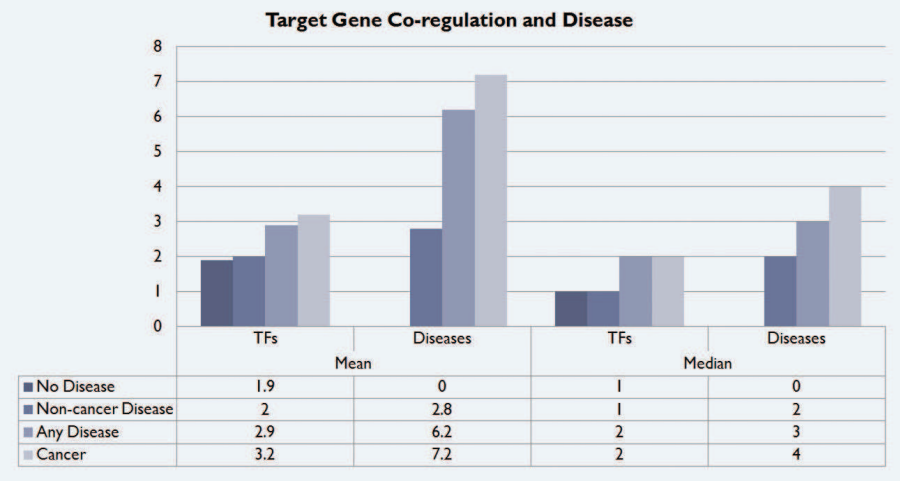


(a) Transcription Factor Co-regulation and Disease

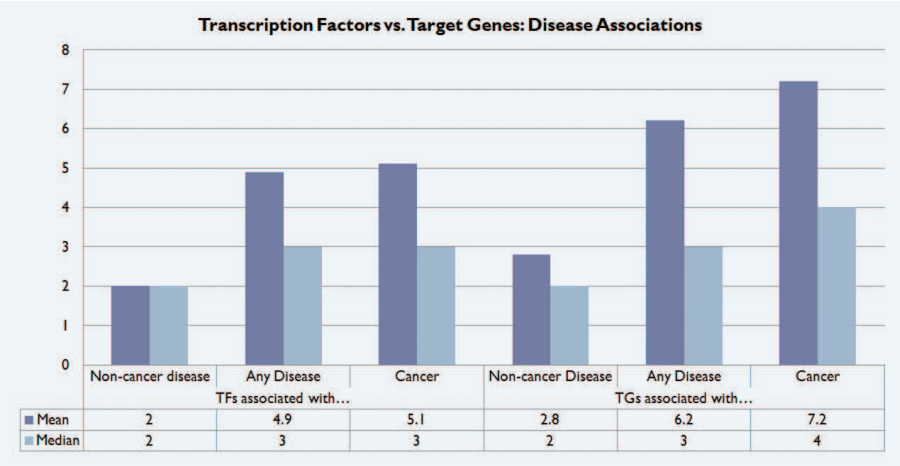


(b) Transcription Factors and Disease Number

Figure 22: A) The mean and median values for the target genes and co-regulators of transcription factors as they relate to disease. The number of both target genes and co-regulators is greatest for cancer-related TFs. B) A similar comparison on a log scale with the number of associated diseases shown. Cancer genes are related to a larger number of total diseases compared with non-cancer-related disease genes.



(a) Target Gene Co-regulation and Disease



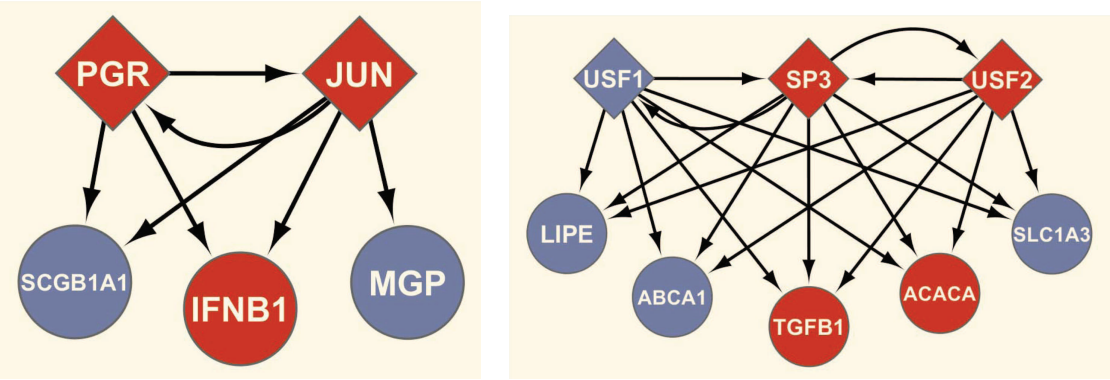
(b) Transcription Factors versus Target Genes: Disease Association

Figure 23: A) The mean and median values for the transcription factors and diseases as they relate to target genes. The number of both transcription factors and diseases is greater for cancer-related TGs. B) A comparison of TFs and TGs and the number of associated diseases of different types. Cancer genes are related to a larger number of total diseases compared with non-cancer-related disease genes.

we found that none of the positions in motif 46 were occupied by a statistically significant number of disease- or cancer-related genes. For motif 222, position 1 has a p-value = 0.03 for disease-related genes and a p-value = 0.02 for those that are cancer-related (Figure 20, in red). Similarly, the second position of motif 2206 has a p-value = 0.04 for both disease- and cancer-related genes. We subsequently performed error correction using the false discovery rate method (136; 125) and found that these significant p-values do not translate to significant q-values. A test for significance of specific cancer-related genes (breast, colon, and lung) were similarly high q-values. Therefore, the results on whether disease- and cancer-related genes occupy specific positions within these regulatory motifs are inconclusive.

Example clusters from the human TF network are shown in Figures 24(a) and 24(b). These were identified using ClusterONE (127) in combination with Cytoscape (151). For each cluster, a combination of the identified significant motifs in Figure 20 is apparent. For the cluster in Figure 24(a), the combination of multi-output FFLs and bi-fan motifs provide for complex and precise regulation of target genes. In Figure 24(b), multi-output FFLs are noticeable, but this time in conjunction with multiple bi-fan motifs, resulting in the DOR pattern mentioned previously. It is important to emphasize that the individual motifs identified here overlap in real networks, and that analyzing these regulatory interactions in the context of small sub-graphs is done in order to reduce complexity and to try to understand the transcription factor's modes of action.

Next, we identified the tissues in which the genes in our data set were expressed using the UniProt database (46). A total of 97 tissue types were represented. The number of genes



(a) PGR/Jun cluster

(b) USF1,2/SP3 cluster

Figure 24: Key: TFs: diamonds, TGs: ellipses, cancer-related: red, non-cancer-related: blue. Left: A portion of the signaling pathway of the progesterone receptor PGR (breast cancer), which involves direct DNA-binding and regulation of target genes. These include the oncogene JUN as well as IFNB1 (colon and prostate cancer). A combination of multi-output FFL and bi-fan motifs can be seen. Right: Upstream stimulating factors (USF1 and USF2) are evolutionarily conserved and ubiquitously expressed. These TFs are major players in the transcriptional regulation of chromatin remodeling enzymes. USF2 and TGFB1 have been linked to tumor growth, while SP3 and ACACA have been linked to breast cancer. SP3 can act as an activator or repressor and is involved in cell-cycle regulation, hormone-induction, and housekeeping. Clusters were identified using the ClusterONE (127) plug-in with Cytoscape (151).

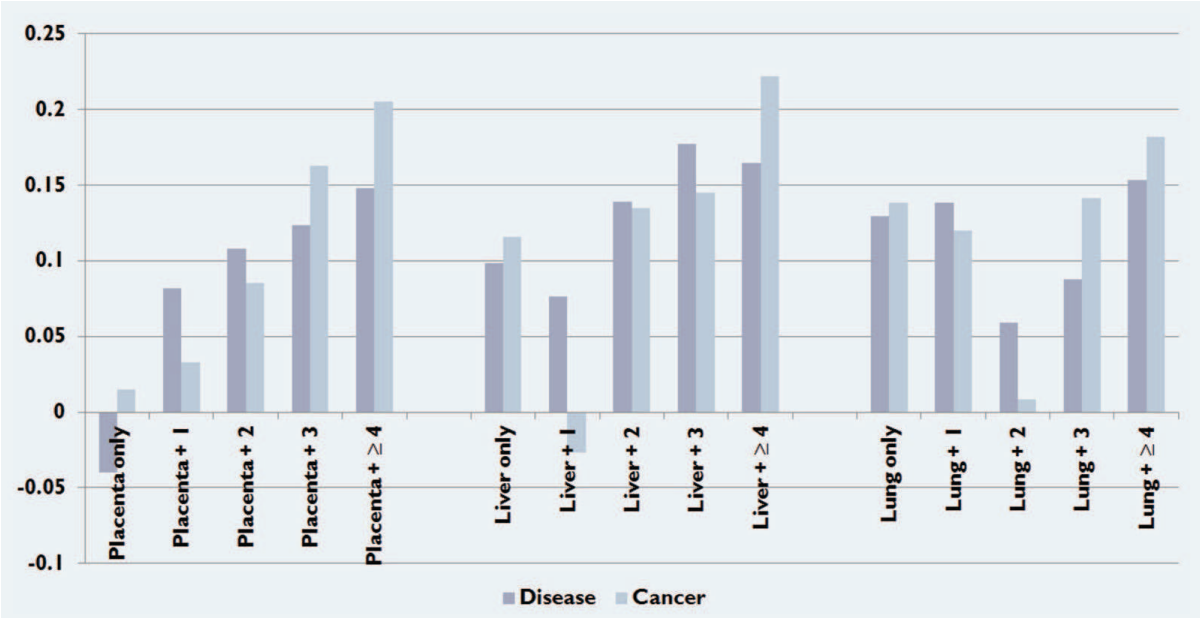


Figure 25: A log-scale graph of the number of disease- and cancer-related genes expressed in the three most common tissue sources for our data set. '$+n$' indicates that the genes are also expressed in n additional tissues.

originating in each tissue varied from 3 to 630 (mean: 63.6, median: 316). Figure 25 shows the three most common tissues in our data set: placenta, 630 genes (20% of data set, 1.8% of the genome), liver, 488 genes (16% of data set, 1.4% of the genome), and lung, 482 genes (16% of data set, 1.4% of the genome). We found that genes were more likely to be associated with both general disease and cancer as the number of associated tissues increased. Interestingly, the number of genes related to cancer surpassed the level for other disease-related genes as they become more ubiquitously expressed, especially for those found in 3 or more tissues. This is in line with the observation from Goh et al. that cancer genes which acquire somatic mutations

are more ubiquitously expressed, while inherited diseases do not show the same expression pattern (74). Many cancer genes also often have house-keeping functions or are involved in cell signaling, either by direct DNA-binding or through a signaling pathway activated by kinases.

We then looked at the transcription factor co-regulation network and its relationship to disease (Figure 22). We found that cancer-related TFs regulated a higher number of target genes than non-disease-related TFs, and that they also had more co-regulating partners. This seems logical given what we know about the cascading effects of cancer development, and that cancer-related genes often have a larger number of interactions with other genes than non-disease genes (74). We also found that cancer-related TFs were associated with more diseases of any type than non-cancer-related TFs. This highlights the connection between cancer and other diseases. The primary disease in a patient is often accompanied by secondary diseases, a phenomenon known as *co-morbidity* (163).

Additionally, we identified the number of TFs associated with each target gene and found that target genes associated with cancer were regulated by a higher average number of TFs than both non-disease and non-cancer disease genes (Figure 23). Furthermore, these cancer-related target genes were involved in a higher than average number of diseases, similar to what we observed for TFs. When comparing the number of disease associations for TFs versus TGs, we found that TGs were related to a higher number of diseases on average than TFs.

6.3 Disease and Protein-protein Interaction Networks

In this project we analyzed the currently-known human protein-protein interaction (PPI) network and its relationship to disease using ADTree. We examined the topological properties

of this network in order to discover differences between proteins which are known to be disease-related and those with no disease association. We also looked for conserved rules over multiple trees in order find the most differentiating characteristics of disease-related proteins in a network context.

In previous work, Gonzalez et al. collected disease genes from NCBI's PubMed database (`http://www.ncbi.nlm.nih.gov/pubmed/`), identified disease-related proteins in the protein-protein interaction network, and weighted protein interactions based on connectivity (75). Wu, Jiang, Zhang, and Li acquired disease-related genes from OMIM (78), identified these in the PPI network using HPRD (93), and then used linear regression and a concordance score to measure the functional relatedness and the phenotypic similarities between genes (182). Özgür Vu, Erkan, and Radev also extracted disease genes from OMIM (78), overlaid the PPI network, and then used an SVM classifier with four centrality measures as features (degree, eigenvector, betweenness and closeness) to predict unknown disease genes (134).

6.3.1 Data Set

We analyzed protein-protein interactions using the Human Protein Reference Database (HPRD) (93) Release 9, which contained 9616 proteins and 39,240 binary interactions. We again used disease information from the DOLite database in order to create two groups for binary classification: those proteins with known/suspected disease association as determined by DOLite (positive class) and proteins with no annotation in the DOLite database (negative class). 3104 of the genes identified in DOLite corresponded to a protein product in the PPI

Data Set	DR Proteins	NDR Proteins	+/- Ratio
HPRD/DO ≥1 disease	3104	6512	≈1/2
HPRD/DO ≥2 disease	2119	6512	≈1/3
HPRD/DO ≥3 disease	1401	6512	≈1/4.6
HPRD/DO ≥4 disease	1026	6512	≈1/6.4
HPRD/DO ≥5 disease	792	6512	≈1/8.2

TABLE VI: We created five versions of the PPI data set. '$\geq n$' in the data set name indicates the number of diseases that must be associated with a protein for it to be a member of the positive class. 'DR' indicates the number of disease-related proteins (+ class), 'NDR' indicates the number of non-disease-related proteins (- class).'+/- Ratio' indicates the ratio of positive and negative class examples for a data set.

network, resulting in ≈32% of HPRD proteins having a disease association. Within this positive class the average number of diseases associated with a given protein was 4.3.

We created 5 versions of the data set, each with a different minimum number of disease associations required for the positive class (Table VI). We analyzed the protein-protein interaction network using Cytoscape (151) and generated the node parameters using Network Analyzer (12). The average number of neighbors for the entire network was ≈7.7, the diameter of the network was 14, and the characteristic path length was ≈4.2.

6.3.2 Prediction of Disease-related Proteins Using ADTree

We used ADTree to predict disease-related proteins in the HPRD database (Figure 26). We found that, using accuracy as our parameter selection standard, 20 iterations provided the best results. Ten-fold cross validation was used for both the parameter selection and validation

steps. Finally, we used a bootstrap sampling method in order to find conserved rules among multiple trees. These conserved rules correspond to the most important features in determining the class to which a sample belongs.

6.3.3 Features

We used ten topological features to build our ADTree model. Each is described below.

6.3.3.1 Degree Centrality

The degree of a vertex is its total number of edges. In a protein-protein interaction network, which is undirected, edges represent molecular interactions between proteins. The degree k of vertex i can be described in this case as

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{6.1}$$

where n is the number of nodes in the network, and A represents an adjacency matrix with elements i and j. This measure has been shown in several works to be a distinguishing characteristic between disease and non-disease genes (21; 62; 131; 111).

6.3.3.2 Closeness Centrality

Closeness centrality measures the average distance between a given node and other nodes in the network. Because it is an inverse measure, a larger value indicates lower level of centrality. It can be thought of as a measure of the rate at which information spreads to neighboring

nodes (128). This measure has been used by Ortutay and Vihinen (131) to identify primary immunodeficiency-related genes. Closeness can be defined as

$$C_c(n) = \frac{1}{avg(L(n,m))} \tag{6.2}$$

where *L(n,m)* indicates the distance of the shortest path between nodes n and m. This metric yields a value between 0 and 1.

6.3.3.3 Betweenness Centrality

This metric indicates the number of shortest paths passing through each vertex. Nodes with high betweenness centrality (often called bottlenecks) have been shown to correspond to essential genes in directed networks (21). The betweenness centrality of a node can be written as

$$C_b(n) = \sum_{s \neq n \neq t} (\frac{\sigma_{st}(n)}{\sigma_{st}}) \tag{6.3}$$

where s and t are vertices other than n, σ_{st} represents the count of shortest paths from s to t, and $\sigma_{st}(n)$ is the count of shortest paths from s to t with which n is involved.

6.3.3.4 Clustering Coefficient

The clustering coefficient (CC) of a node n is the ratio of existing edges between n and its neighbors and the number of possible connections. This is a measure of edge density for the

node's neighborhood (82; 62; 111). For undirected networks, the clustering coefficient of n can be written as

$$CC_n = \frac{2e_n}{(k_n(k_n-1))} \tag{6.4}$$

where k_n is the total number of neighbors of node n and e_n is the count of linked pairs of nodes between the neighbors of n (179; 19).

6.3.3.5 Stress Centrality

The stress centrality (31; 150) for a node n corresponds to the total number of shortest paths passing through it. If the number of shortest paths is large, the stress will be large as well. This metric is described in the following way:

$$C_s(n) = \sum_{s \neq n} \sum_{t \neq n,s} \sigma_{st}(n) \tag{6.5}$$

where s and t are network vertices other than n, σ_{st} indicates the number of shortest paths from vertex s to vertex t, and $\sigma_{st}(n)$ is the count of shortest paths from s to t passing through n.

6.3.3.6 Neighborhood Connectivity

The neighborhood connectivity of a vertex is equal to the average degree of its neighbors (121).

6.3.3.7 Topological Coefficient

This metric describes the average number of shared connections between a node and other nodes. In a social context this would measure the number of mutual friends two people share. The topological coefficient (156) can be represented as

$$T_n = \frac{avg(J(n,m))}{k_n} \tag{6.6}$$

where k_n indicates the neighbors of node n and *J(n,m)* is the count of neighbors that n and m share, plus one if there is an edge between n and m. *J(n,m)* is defined only for all nodes m that have at least one neighbor in common with n.

6.3.3.8 Eccentricity

Eccentricity is the longest path between a node n and another node. The eccentricity value is 0 for isolated nodes. The maximum value for eccentricity is the diameter of the network.

$$C_{ecc}(n) = \frac{1}{max\{dist(v,w) : w \in V\}} \tag{6.7}$$

6.3.3.9 Radiality

Radiality is a measure of centrality (169; 31). It is the average shortest path length of a vertex n minus the diameter of the connected component to which n belongs plus 1. The

resulting value is a number between 0 and 1. A high radiality value indicates that a vertex can easily reach other vertices (96).

$$C_{rad}(n) = \frac{\sum_{w \in V}(\Delta_G + 1 - dist(v,w))}{n-1} \tag{6.8}$$

6.3.3.10 Disease Neighbor Ratio

This metric describes the local environment of a node in terms of its disease-related neighbors. We represent the disease neighbor ratio (DNR) as

$$DNR_i = \frac{n_{disease}}{k_i} \tag{6.9}$$

where $n_{disease}$ is the number of neighbors of node i identified as disease-related proteins and k_i is the degree of i.

6.3.4 Results

Using Receiver Operating Characteristic (ROC) curves, we found that the classifier created using proteins associated with $\geq$5 diseases yielded the best results (Figure 27). We predicted 1622 disease-related proteins using the ADTree classifier. We then created a bootstrapped ADTree for the PPI-disease network (Figure 26) using this model. As indicated by the order in which the rules were found and by the conservation of rules discovered during the bootstrapping process, the attributes which are the most effective for distinguishing disease- from non-disease proteins are degree, disease neighbor ratio, eccentricity, and neighborhood connectivity. The next most conserved feature, present in at least 50% of the trees, is betweenness centrality.

This feature is conserved when used in conjunction with degree and disease neighbor ratio. The remaining rules are conserved in $\leq 50\%$ of the bootstrapped trees. Degree, disease neighbor ratio, and neighborhood connectivity were expected to play an important role, particularly given some recent analysis (80; 193). However, the ADTree illustrates a pathway by which rules work together to discriminate between disease- and non-disease-related proteins.

Figure 28 shows box plots for the first two rules in the bootstrapped tree (created using the R Statistical Environment (140)). Disease-related proteins had a higher degree centrality and disease neighbor ratio compared to non-disease proteins. To test the importance of the most discerning features, we ran the algorithm four more times, each time removing one of these important attributes. Removing the disease neighbor ratio resulted in an 11% decrease in sensitivity (which measures the ratio of true positive examples and those correctly identified as positive). Removal of the degree centrality, neighborhood connectivity, and self interaction features reduced sensitivity by 3%, 3%, and 1% respectively. These results along with the ADTree in Figure 26 make it clear that while individual attributes may contribute more or less to a prediction problem, the combination of these features gives us a multi-dimensional view of how the two classes are separated.

6.3.5 Comparisons with Previous Data Sets

Using our ADTree classifier, we correctly identified 17/17 disease-related proteins which Gonzalez et al. (75) mined from literature. Our classifier also correctly predicted 14/16 known breast cancer genes identified by Wu, Jiang, Zhang, and Li (182). Additionally, we were able to correctly classify 15/16 known disease genes found in literature by Özgür et al. (134).

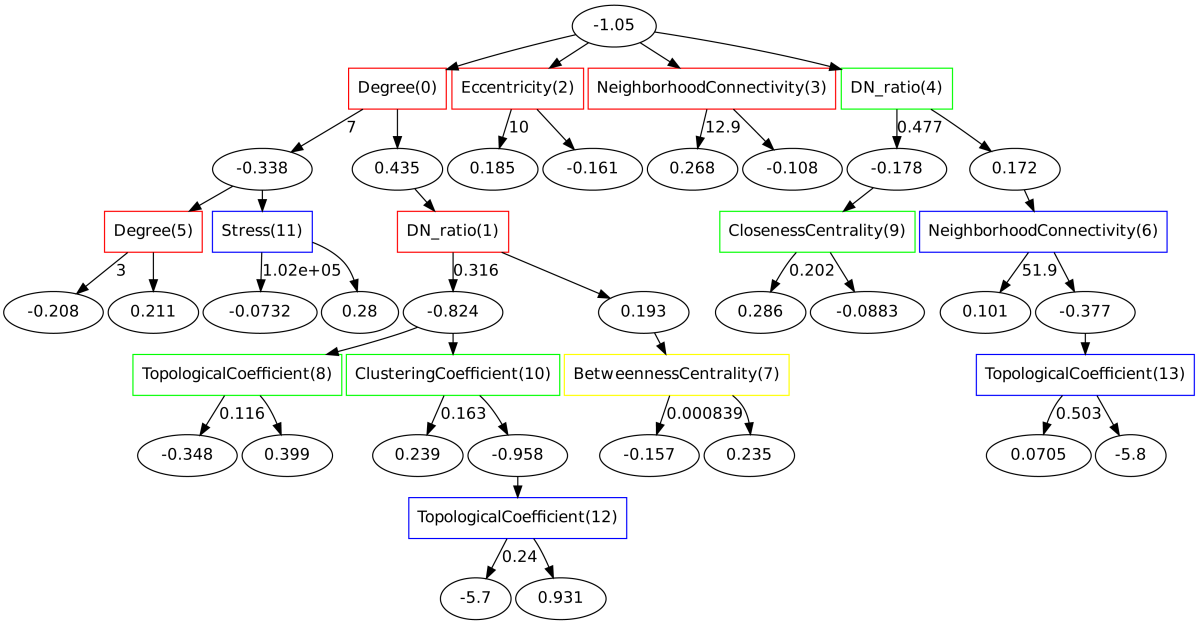


Figure 26: An ADTree created using 10-fold CV and bootstrapping. The root node indicates the bias in the data set, i.e., the ratio of positive to negative class examples (disease-associated proteins versus non-disease-associated proteins). The square nodes contain the feature name and the number in parentheses within each square node indicates the order in which the rule was found. The amount of node conservation between each of the trees generated in the validation step is indicated by the color of the box (red: $\geq$90%, orange: $\geq$70%, yellow: $\geq$50%, green: $\geq$30%, blue: $\geq$10%, black: $\leq$10%). The oval nodes show the value for the weighted vote, where a positive number indicates a prediction for disease-association. The numbers next to the arrows correspond to the threshold for the prediction. If the attribute value exceeds this number, the right path is followed. Otherwise the prediction follows the left path.

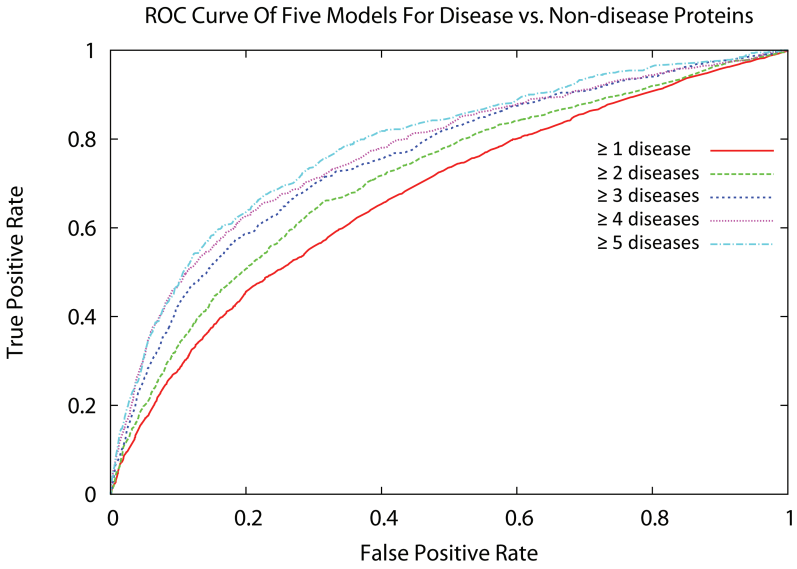


Figure 27: ROC curves for five ADTree classifiers. The performance of the models increases with the removal of proteins associated with few diseases, which affects only the positive class in the prediction. The area under the ROC curve (AUC) is ≈67% for ≥ 1 disease, ≈71% for ≥ 2 diseases, ≈75% for ≥ 3 diseases, ≈76% for ≥ 4 diseases, and ≈79% for ≥ 5 diseases.

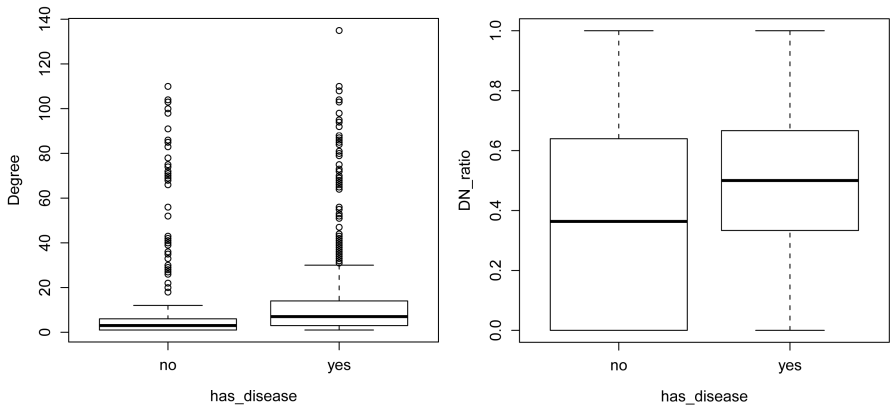


Figure 28: Box plots (created using R (140)) showing two of the most distinguishing features between disease and non-disease proteins: degree centrality and disease neighbor ratio. These features correspond to the first and second rules found using the ADTree model, respectively.

6.3.6 Identification of Potential Disease Genes

Our classifier predicted 98 non-disease-related proteins (i.e., those that lack DORIF annotation) to be members of the positive class with a confidence score ≥ 0.5 (threshold = 0). This indicated that these examples have attribute values which qualify them as potential disease-related proteins. We examined a subset of these examples more carefully and found that there is evidence linking many of these proteins to disease. Table VII shows the top 15 in this group ranked by confidence. Only two proteins (PTCH1 and TCF4) have associated MIM numbers (indicating disease involvement). 9/15 proteins (CDH5, DPP4, GZMB, FGR, FLT1, PECAM1, SREBF2, STAT6, and TOP1) have moderate to strong evidence of disease association, while 4/15 (STAMBPL1, MDH2, GRK5, and CD74) have light evidence. Interestingly, 12/15 are linked to some form of cancer or tumor development.

Conf. Score	OS	DORIF	OMIM	Suspected Disease Relations
6.24096	CDH5	none	none	melanoma, tumor metastasis
5.81860	PTCH1	none	109400, 605462, 610828	basal cell nevus syndrome and carcinoma
5.19721	STAMBPL1	none	none	light evidence, Alzheimer's
5.14813	MDH2	none	none	light evidence, tumor development
1.09972	DPP4	none	none	diabetes, NIDDM (55), colon cancer (3)
1.09972	GRK5	none	none	light evidence, heart failure
0.907016	GZMB	none	none	lymphoma (30), tumors (92)
0.898631	TCF4	none	610954	Pitt-Hopkins syndrome
0.705929	FGR	none	none	breast cancer (3), prostate cancer (1)
0.705929	FLT1	none	none	cancer, various
0.705929	PECAM1	none	none	cancer, various
0.705929	SREBF2	none	none	prostate cancer (2)
0.705929	STAT6	none	none	prostate cancer (3)
0.705929	TOP1	none	none	leukemia, colon cancer, ovarian cancer
0.664823	CD74	none	none	light evidence, lymphoma

TABLE VII: This is a subset of proteins which belonged to the non-disease group (lacking DORIF annotation) but were predicted to be disease-related. These 15 proteins are sorted by the confidence score assigned by the ADTree classifier. Two proteins (PTCH1 and TCF4) have associated OMIM disorders. 9/15 proteins (CDH5, DPP4, GZMB, FGR, FLT1, PECAM1, SREBF2, STAT6, and TOP1) have moderate to strong evidence of disease association, while 4/15 (STAMBPL1, MDH2, GRK5, and CD74) have light evidence linking them to disease. (n) indicates the number of PubMed IDs connecting a protein to a particular disease. 'Conf. Score' is the confidence score assigned by the ADTree classifier, 'OS' is the official symbol of the gene/protein, 'DORIF' is Disease Ontology + Gene Reference Into Function, 'OMIM' is the MIM number associated with the gene/protein, 'light evidence' is defined as having a predicted disease association according to the MalaCards database (141). Disease information for this table was acquired from the GeneCards database (155).

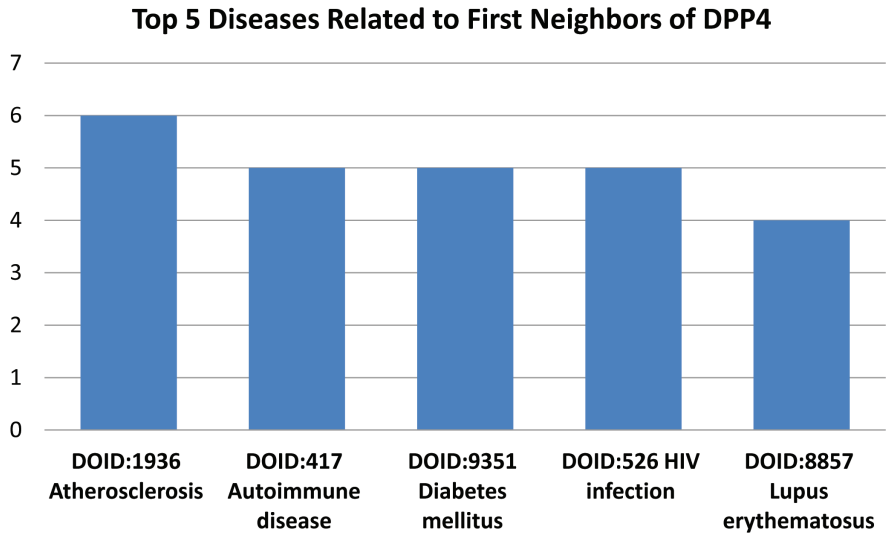


Figure 29: The five most common diseases associated with the first neighbors of DPP4 (those proteins with a direct interaction). DPP4 has 55 PubMed IDs which associated it with non-insulin-dependent diabetes mellitus (NIDDM). Interestingly, NIDDM is often accompanied by beta cell autoimmunity, where the beta cells of the pancreas are destroyed by an autoimmune disorder (190).

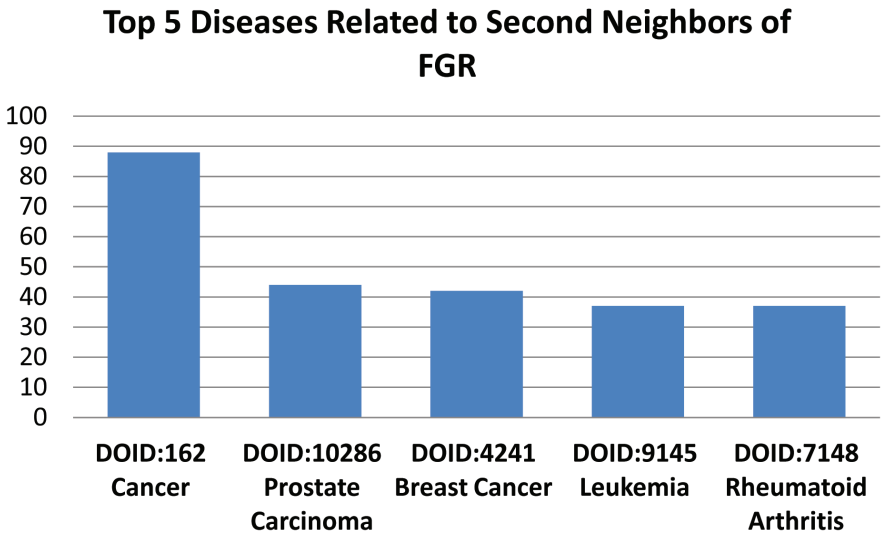


Figure 30: The five most common diseases associated with the second neighbors (i.e., neighbors of neighbors) of FGR. There are 3 PMIDs associating this gene with breast cancer and 1 PMID linking it to prostate cancer.

We examined two of these proteins more closely to find out which diseases were associated with PPI network neighbors. We identified the first neighbors (i.e., proteins with a direct interaction) of dipeptidyl-peptidase 4 (DPP4, Figure 29). This gene product is a glycoprotein receptor involved in the signaling pathway for T-cell receptor (TCR)-mediated T-cell activation (155). DPP4 has 55 PubMed IDs associating it with non-insulin-dependent diabetes mellitus (NIDDM). Figure 29 shows the five most common diseases of DPP4 first neighbors by Disease Ontology ID (DOID). 'Diabetes mellitus' ranks third, while 'Autoimmune disease' ranks second. Interestingly, NIDDM is often accompanied by beta cell autoimmunity, where the beta cells of the pancreas are destroyed by an autoimmune disorder (190). We used a similar method

for the Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog (FGR), but this time we identified the second neighbors of the protein (i.e., neighbors of neighbors). There are three PMIDs associating this gene with breast cancer and one PMID linking it to prostate cancer. Figure 30 shows the top five diseases related to the second neighbors of FGR. 'Prostate Carcinoma' and 'Breast Cancer' are the second and third most common diseases, respectively, only behind the general category of 'Cancer'.

These two examples illustrate how the PPI network can be used in a post-processing step to examine the network neighborhood of potential disease-related proteins and to identify disease(s) with which these proteins may be associated. This provides a useful filter for experimentalists searching for candidate targets for drug repositioning. This method could also be extended to include other network and biological data types in order to refine these predictions.

6.4 A Disease Co-occurrence Matrix

In order to analyze disease relationships, we built a disease co-occurrence matrix based on shared genes between each pair of diseases. We first calculated the *uniqueness* of each gene i as follows:

$$u_i = \sqrt{\frac{d_i}{d_n}} \tag{6.10}$$

where d_i is the number of diseases associated with each gene i and d_n is the number of diseases in the data set. Next, we created an $N\ X\ N$ matrix. Then, for each pair of diseases we added the uniqueness score of each shared gene:

$$d_{ij} = (u_{s_1} + u_{s_2}...u_{s_n}) \tag{6.11}$$

where d_{ij} is a disease pair and u_{s_n} is the uniqueness value for each gene shared between the two. The diagonal elements of the disease co-occurrence matrix, where $i = j$ for d_{ij}, contain the sum of the uniqueness values for all genes related to disease d_i.

Next, we applied *symmetric approximate minimum degree permutation* to the disease co-occurrence matrix. This algorithm was developed by Stefan I. Larimore and Timothy A. Davis and incorporated into MATLAB (123). This reordering algorithm first creates a permutation vector p from a symmetric positive definite matrix A. This permutation vector, which contains a list of reordered columns from A, is then used to create a new matrix S such that $S = (p,p)$ has a sparser Cholesky factor than the original matrix A. The end result is that the reordered matrix S is less sparse near the lower diagonal and more sparse near the upper diagonal. For our disease co-occurrence matrix, this effectively clusters highly-related diseases in the lower right quadrant around the diagonal.

We first applied this strategy to the OMIM MorbidMap database (78) (`http://omim.org/`). Figure 31 shows the resulting reordered disease co-occurrence matrix for 5224 diseases. While

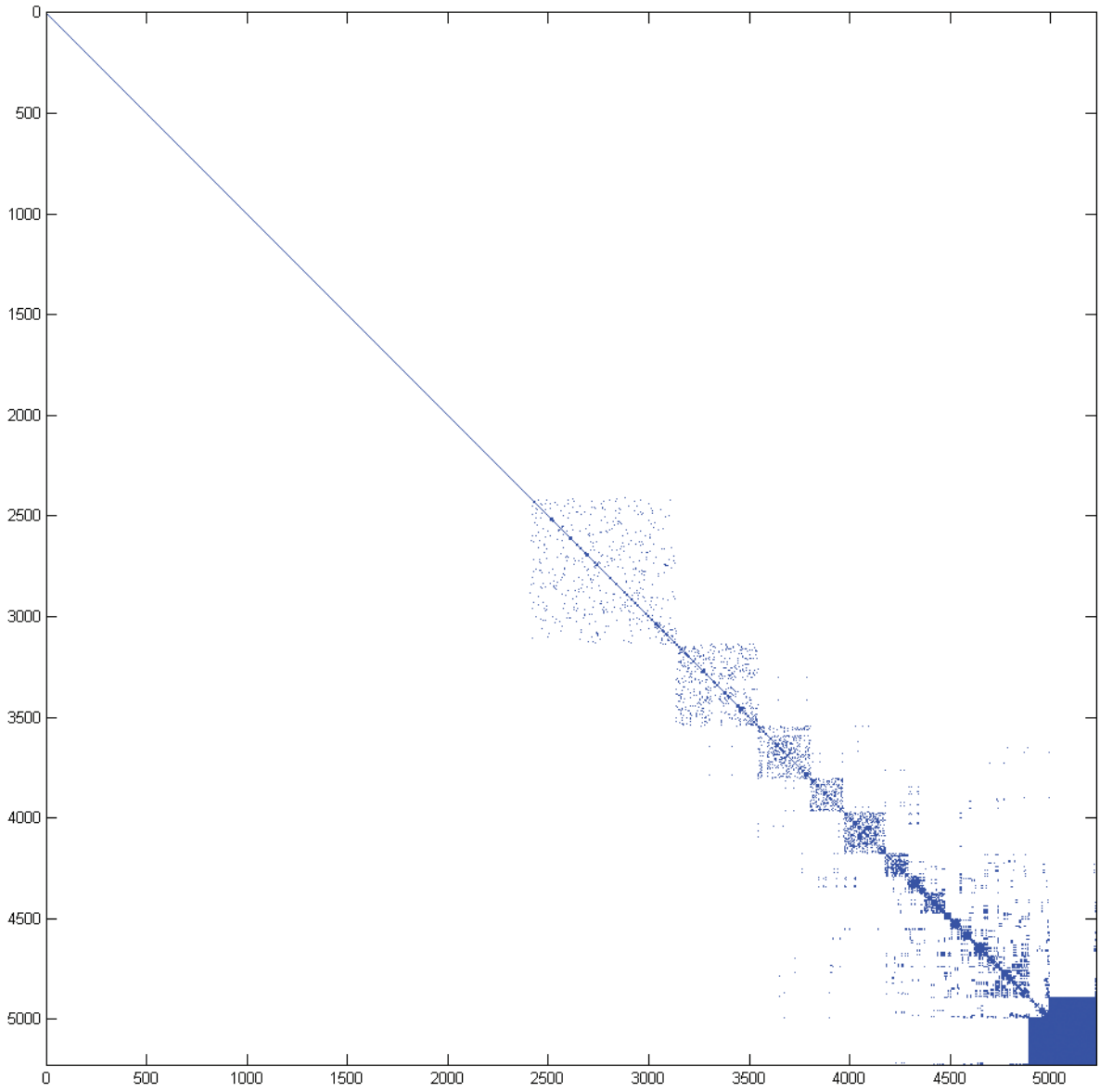


Figure 31: A co-occurrence matrix showing the relationship between 5224 diseases from the OMIM MorbidMap is shown. Matrix elements colored blue indicate a relationship between two diseases, white elements indicate no relationship. Each blue matrix element contains the sum of the uniqueness values for genes related to $disease_i$ and $disease_j$, white elements are equal to 0. Diagonal elements indicate the identity relationship for each disease, i.e., the sum total uniqueness values for all genes associated with $disease_i$. This figure was created using MATLAB (123).

there are well-defined clusters, many of the cluster members are variations of the same disease phenotype or very closely related phenotypes. This is due to the high level of specificity of the OMIM disease categories. For example, the disease "46XY complete gonadal dysgenesis" is listed as two separate disease phenotypes, each with a different MIM identifier. While this distinction is important (the two phenotypes refer to mutations on different chromosomes), it does not serve our purposes in this case. We would like to see more relationships between phenotypically different diseases, and we would like very closely-related phenotypes to be grouped together.

To address this problem, we created another matrix using gene-disease relationships gathered from Disease Ontology (143) and the GeneRIFs (Gene Reference Into Function) database (`http://www.ncbi.nlm.nih.gov/gene/about-generif`). This combination of data sources was used recently by Osborne et al. to annotate the human genome (133) for disease (referred to hereafter as DORIF, (`http://projects.bioinformatics.northwestern.edu/do_rif/`)). They found that DORIF annotation provided a much higher recall rate when compared to OMIM data for validation gene sets. Additionally, the Disease Ontology provides a hierarchical structure in which more specific diseases can be grouped into broader categories, which allows us to compare phenotypically divergent diseases more easily.

The DORIF annotation included 88,343 entries for 5,376 genes. There were 1,854 diseases and 48,436 PubMed references for gene-disease relationships (`http://www.ncbi.nlm.nih.gov/pubmed/`). The DORIF co-occurrence matrix (Figure 32) shows the comparisons between these diseases. There are two notable differences from the OMIM matrix. Firstly, there are noticeably

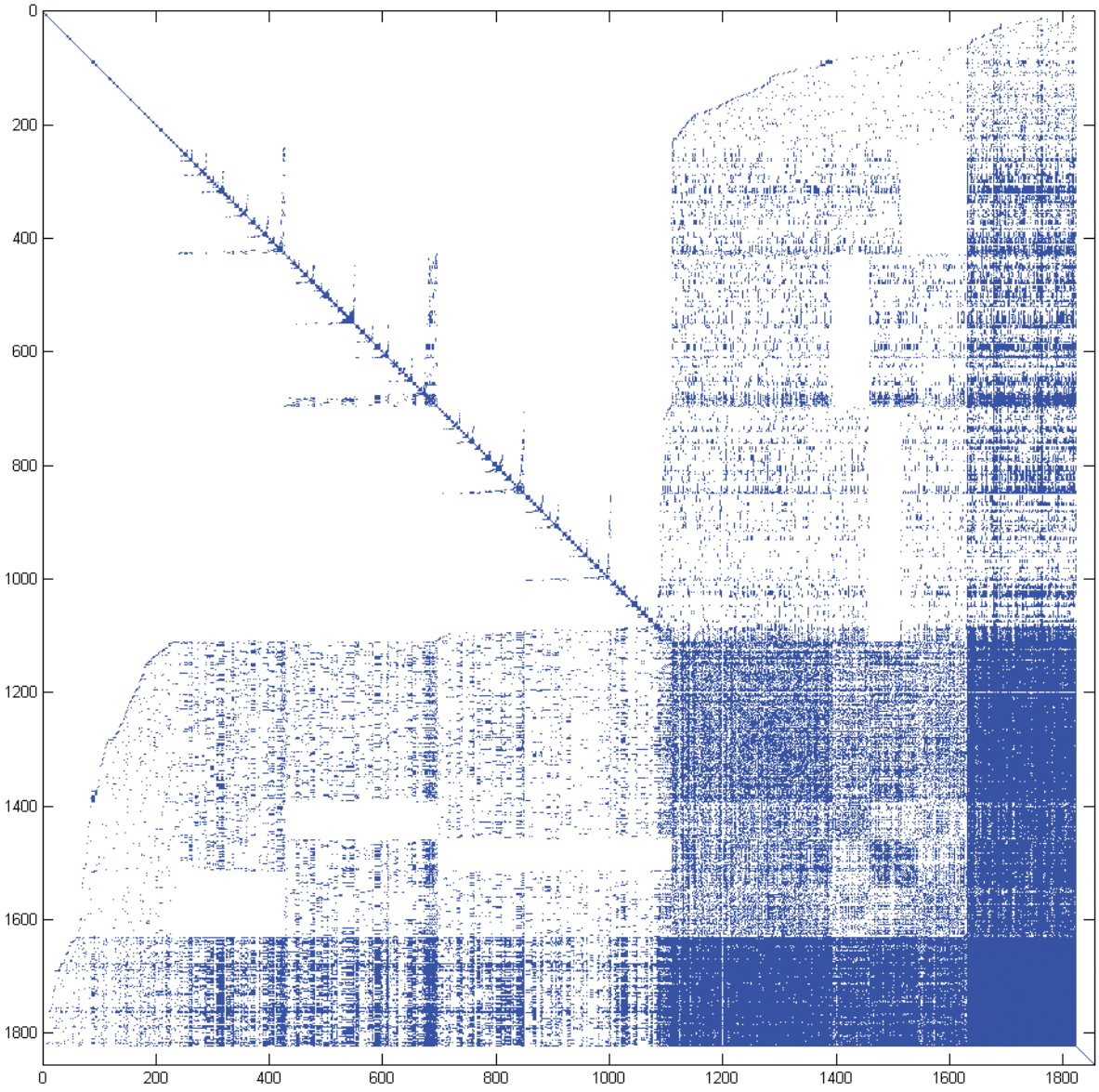


Figure 32: A co-occurrence matrix showing the relationship between 1854 diseases using DORIF data is shown. Matrix elements colored blue indicate a relationship between two diseases, white elements indicate no relationship. Each blue matrix element contains the sum of the uniqueness values for genes related to $disease_i$ and $disease_j$, white elements are equal to 0. Diagonal elements indicate the identity relationship for each disease, i.e., the sum total uniqueness values for all genes associated with $disease_i$. This figure was created using MATLAB (123).

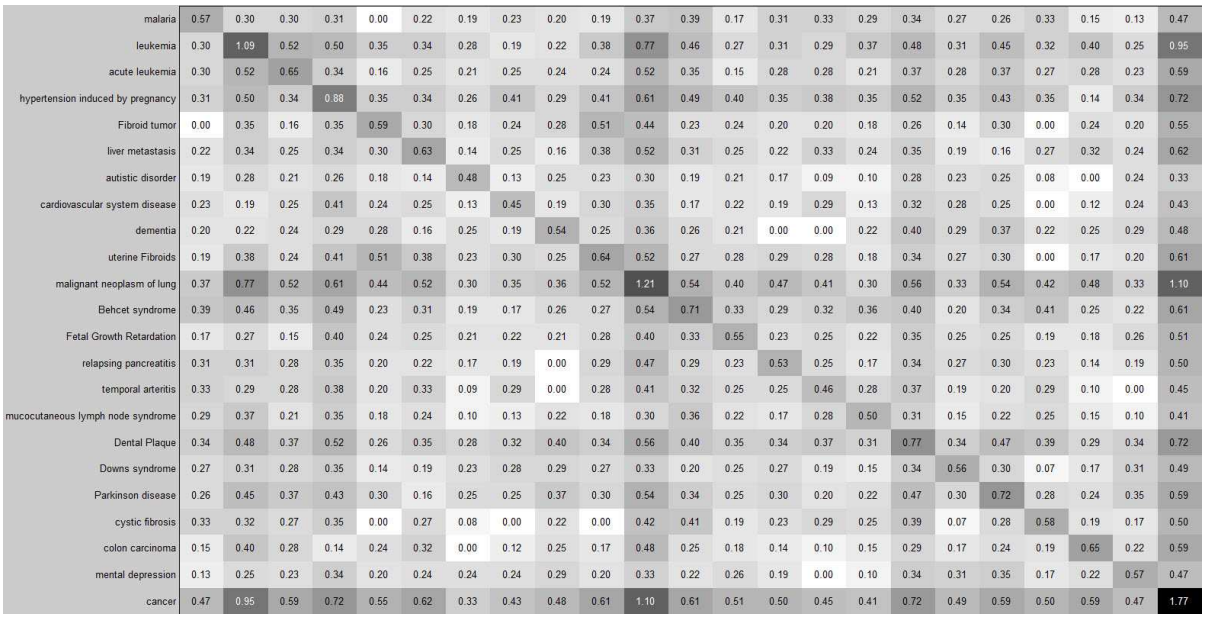


Figure 33: This matrix is a subset of the disease co-occurrence matrix and shows the relationships between 23 diseases beginning with malaria (top) and ending with cancer (bottom). Disease labels for the rows apply to the columns as well. The value of each element *i,j* is the sum of the uniqueness values of each gene related to both $disease_i$ and $disease_j$. Darker squares indicate a higher uniqueness value. This figure was created using MATLAB (123).

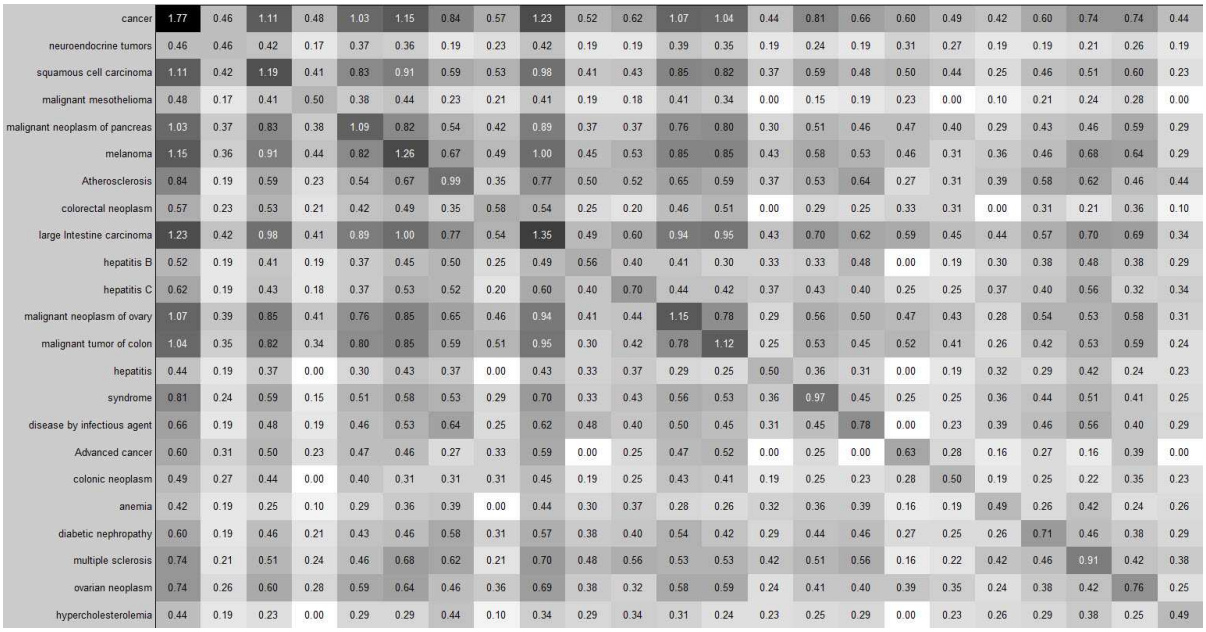


Figure 34: This matrix is a subset of the disease co-occurrence matrix and shows the relationships between 23 diseases beginning with cancer (top) and ending with hypercholesterolemia (bottom). Disease labels for the rows apply to the columns as well. The value of each element *i,j* is the sum of the uniqueness values of each gene related to both $disease_i$ and $disease_j$. Darker squares indicate a higher uniqueness value. This figure was created using MATLAB (123).

more disease relationships. Secondly, the DORIF matrix appears noisier; the relationships are not as tightly clustered as they are in the OMIM matrix. However, these "gray areas" may actually be a benefit if the goal is to find hidden relationships between diseases. A closer look at the individual clusters provides some interesting information.

Figure 33 and Figure 34 show a closer view of two different subsections of a dense cluster. Each of these figures is a 23 X 23 square submatrix of disease relationships from the solid blue cluster in the lower right-hand corner of the matrix in Figure 32. The majority of the diseases in these submatrices are various types of cancers. There are some notable and interesting exceptions, however. For example, in the case of the relationship between malaria and cancer, the uniqueness value is close to that of those genes only related to malaria (0.47 and 0.57, respectively). Recent research provides some interesting findings about these two diseases. A clinical study showed that the mortality rate in patients with any type of cancer was increased after malarial infection (106). Additionally, the malaria drug chloroquine has been shown to reduce tumor size in pancreatic cancer patients (188). Another example is the relationship between hypertension induced by pregnancy and cancer. Recent work has shown that VEGF (Vascular endothelial growth factor) may be the connection. When taking anti-VEGF cancer drugs, patients develop very similar symptoms to pregnancy-induced hypertension. When VEGF expression levels are reduced in solid tumors, growth slows due to the lack of vascular development within. As a side effect, hypertensive symptoms occur (73). Dental plaque and cancer appear to be highly related according to their uniqueness values as well; 0.77 (Dental plaque) and 0.72 (Dental plaque and cancer). Several past studies have made the connection

between oral health and chronic illness. Earlier this year, however, a clinical study spanning the last 24 years was released (152). During this period, researchers followed 1,400 adults. They found that these subjects with high levels of dental plaque were 79% more likely to die prematurely from cancer. This work shows only an association between the two diseases, and thus the true nature of the relationship is yet to be discovered. Other relationships from our matrix share a high uniqueness score, but there is little or no experimental evidence linking them. For example, migraine headaches and large intestine carcinoma have a shared uniqueness score of 0.403, while migraine alone is 0.498 (not shown in figures). Despite this, we could not find research references linking the two. This matrix can be used to identify potential disease relationships and to motivate further study into the elucidation of causal mechanisms in disease.

CHAPTER 7

CONCLUSIONS

7.1 DNA-binding Protein Prediction

We have demonstrated that the boosted decision tree algorithm outperforms other classifiers, achieving 88% accuracy for DNA-binding prediction. This study also provided a rigorous benchmark comparison for five classifiers over a set of four important metrics by varying the size of the training set. The boosted decision trees had the best performance on nearly every metric and for each training set size tested. Additionally, we discussed the area under the ROC curve and how it serves as a better metric for imbalanced data sets due to the fact that it is unaffected by the underlying class distribution. This is important for future work since many types of data sets, including those created for use in DNA-binding and RNA-binding residue prediction, are imbalanced. Furthermore, a larger AUC indicates that the learning algorithm will produce more accurate confidence values and make more robust predictions because it measures the relative ordering of those predictions.

7.2 DNA-binding Residue Prediction

Using imbalanced data sets, we developed classifiers for DNA- and RNA-binding residue prediction based on C4.5 (139), bootstrap aggregation (32), and cost-sensitive learning (191) which produce balanced sensitivity, specificity, and precision with high accuracy. We compared these results to those from five commonly-used algorithms built using balanced training sets

and found that our classifiers achieve higher accuracy than any of the other algorithms. We also compared the results of our classifiers to those from previous works over previously-compiled data sets. We showed that we are able to achieve more accurate results by training on an unbalanced data set. We believe that using this method, as opposed to artificially balancing the data set, provides more realistic results for testing, due to the fact that nearly all NA-binding proteins have many more non-binding than binding residues.

We also created the Nucleic Acid Prediction Server (NAPS, `http://bioinformatics.bioengr.uic.edu/NAPS`), a publicly-available web server based on our C4.5BAGCST classifiers that performs NA-binding residue prediction using sequence-based attributes. NAPS takes a DNA- or RNA-binding protein sequence as input, calculates a set of 301 sequence-based attributes for each residue in the test protein, and returns a list of residues, the predicted class (binding or non-binding), and a confidence score. This web utility provides a convenient tool for researchers to identify potential NA-binding residues from proteins which do not yet have a crystal structure available.

7.3 DNA-binding Site Prediction

We have developed two-body and three-body interaction potentials which are able to assess protein-DNA interaction and achieve a higher level of specificity in the recognition of TF-binding sites. We implemented two approaches in order to evaluate the potentials. We found that the three-body potential, which takes into account the interaction between a DNA base and a protein residue with regard to the effect of a neighboring DNA base, had a better average Z-score than that of the two-body potential. This neighbor effect suggests that the local conformation

of DNA plays a critical role in specific residue-base recognition. In all cases, the potentials developed here outperformed published results. The two sets of potentials were tested further by applying them in genome-scale TF-binding site prediction for the CRP protein in *E. coli*. We showed with these results that statistical potentials can be used in genome-scale TF-binding site prediction.

This work improved protein-DNA statistical potentials by performing two tasks: 1) systematically evaluating the space partition and residue representations, and 2) taking into account the conformational effects of local DNA on the statistical potential. In summary, we found that the grid-based potential outperformed the distance-based potential, that using Cβ as a representative point for an amino acid improved the results, and that three-body potentials provided better specificities than two-body potentials.

7.4 Prediction of CpG Island Methylation In Human DNA

We predicted the methylation status of CpGIs from human chromosome 21 with high, balanced accuracy using an ensemble classifier trained with 27 4-mer sequence patterns that were both strand- and non-strand-specific. In addition, we created an ADTree classifier which allowed us to evaluate the relationships between sequence patterns in CpGIs. One highly-conserved sequence pattern, CCGC, was present in $\geq$ 70% of the bootstrapped trees. This was identified as the most highly-discriminating sequence pattern for distinguishing between methylated and non-methylated CpGIs in a previous work (30). We found that if a CpGI contains six or more of these patterns, it is highly likely to be methylated. However, if a CpGI has less than six CCGC patterns and 36 or more strand-independent CACC patterns, it

is highly likely that the island is unmethylated. We show a possible biological reason for this observation. This highlights the fact that by using ADTree we are able to identify cooperativity between features in classification problems.

7.5 Partnership Networks

We analyzed several transcription and phosphorylation networks and examined the properties of co-acting partners. We found that this partnership follows an exponential saturation curve when related to the number of targets. We also developed a generative transcription factor network model which followed rules consistent with current evolutionary theory. We found that the model created transcription factor networks with very similar characteristics to that of real networks, and that the co-regulatory relationships within followed the exponential saturation curve as well. This suggested that these properties may be inherent in regulatory networks. A comparison with sets of blog and email interactions as well as two projections of the bipartite disease-gene network highlighted similarities in organizational structure. Like the biological networks, these social and disease networks displayed limits on partnership even as the number of interactions grew. It is possible that comparable rules exist in the formation of these networks as well, leading to their architectural similarities.

7.6 Disease-related Genes in Conserved Human TF Network Motifs

We built a human transcription factor network from currently available data and identified eight statistically significant regulatory motifs. We found that both general disease-related genes and cancer-related genes were present more frequently in these motifs when compared to the network as a whole. Disease and cancer genes appeared often in a combination of the

feed-forward loop, regulating feedback, and bi-fan motifs. For two thirds of the motif examples shown, one position within the motif was occupied by a statistically significant number of disease- and cancer-related genes. We identified two clusters from the human transcription factor network that exhibited a combination of statistically significant motifs and disease-related genes. We also made a number of interesting observations. Firstly, we found that more ubiquitously expressed genes were more likely to be associated with both general disease and cancer. Secondly, we found that cancer-related TFs regulated a higher number of target genes than non-disease-related TFs, and that they also had more co-regulating partners. Thirdly, we found that cancer-related TFs were associated with more diseases of any type than non-cancer-related TFs. This highlights the connection between cancer and other diseases. Finally, we identified the number of TFs associated with each target gene and found that target genes associated with cancer were regulated by a higher average number of TFs than both non-disease and non-cancer disease genes. Furthermore, these cancer-related target genes were involved in a higher than average number of diseases, similar to what we observed for TFs. When comparing the number of disease associations for TFs versus TGs, we found that TGs were related to a higher number of diseases on average than TFs.

7.7 Disease and Protein-protein Interaction Networks

We analyzed the human protein-protein interaction network and its relationship to disease and found that both the number of neighbors and whether or not a protein lies in a 'disease neighborhood' contribute greatly to its identity as a disease- or non-disease-related protein. Our classifier predicted 98 non-disease-related proteins to be members of the positive class with

a confidence score ≥ 0.5 (threshold = 0). This indicated that these examples have attribute values which qualify them as potential disease-related proteins. We looked closer at two of these proteins and found that neighbors in the PPI network could be used to identify specific disease associations. These two examples illustrate how the PPI network can be used in a post-processing step after disease-gene prediction. This method could provide a useful filter for experimentalists searching for candidate targets for drug repositioning. This method could also be extended to include other network and data types in order to refine these predictions.

As we have learned from our work, diseases share interactions through molecular networks. One of the next steps in disease-gene analysis could be to study connections between diseases; for instance, various types of cancers as they relate to other illnesses such as diabetes (174), various infections (9), and obesity (100; 161). Though the type or nature of this relationship may be unknown, we may be able to shed light on the subject using these knowledge mining methods along with molecular data such as protein metabolic pathways, regulation networks, and others. We believe that as more complete data sets become available, a higher level of knowledge will be attainable by utilizing this method.

7.8 Disease Relationships in Co-occurrence Matrices

We developed a co-occurrence matrix based on gene uniqueness in order to examine the relationships between diseases from the OMIM and DORIF databases. We found examples of known disease relationships as well as connections with no available evidence. This matrix can be used as a jumping off point for identifying disease-disease associations, providing a map of the connections between diseases and directing focus toward those associations which may not

otherwise be obvious. It could also provide a first step in drug repositioning research, directing focus to new potential protein or DNA targets.

7.9 The Future: Gaining Knowledge from Data

The key to understanding the disease network is to enrich the value of existing edges and to infer new ones based on this enriched value. There is a wealth of information concerning disease, metabolism, gene ontology, drug targets, miRNA, protein-protein interaction, gene regulation, and gene expression. Unfortunately, there are large areas of missing and overlapping data as well as many false positives and even more false negatives. This makes it difficult to assemble the puzzle and gain knowledge. One can use algorithms such as ADTree which can filter through noisy data to find the most informative and conserved characteristics of a disease-disease relationship. Cancer A and non-cancer disease B, though they may not share a causal gene or genes according to OMIM, may be related at some distance through common metabolic pathways, co-regulating transcription factors, or negative regulation by one or more miRNAs. Any of these could be a false positive association. When analyzed together along with other available data, however, a more complete biological picture comes into focus and the noise problem can be mitigated. ADTree allows us to easily visualize which biological processes contribute most to the disease relationship, eliminating the 'black box' effect of many machine learning algorithms.

In conclusion, we believe complex diseases such as cancer are both unique and related to other diseases. By studying all diseases as a system in a network context we believe we can generate many pertinent results. For instance, drugs used for the treatment of related non-

cancer diseases may help to treat the side effects of cancer drugs. Another example lies in the complex relationship between bacteria and cancer: bacteria can be both beneficial and cancer-causing. Disease network research could provide new ideas about cancer and its relationship to infection. Additionally, this research could help clarify the mechanisms and tissue-specificity of non-cancer diseases and how they may prime the cellular environment for metastasis. We expect that in the near future, due to the availability of an enormous amount of genotypic and phenotypic data related to disease, there will be a novel view point for cancer research emerging from disease network study.

CITED LITERATURE

1. Moulinath Acharya, Lijia Huang, Valerie C. Fleisch, W. Ted Allison, and Michael A. Walter. A complex regulatory network of transcription factors critical for ocular development and disease. *Human Molecular Genetics*, 20(8):1610–1624, April 2011.

2. S. Ahmad and A. Sarai. Moment-based prediction of DNA-binding proteins. *Journal of Molecular Biology*, 341(1):65–71, 2004.

3. Shandar Ahmad, M. Michael Gromiha, and Akinori Sarai. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 20(4):477–486, 2004.

4. Shandar Ahmad and Akinori Sarai. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, 6:33, 2005.

5. Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, July 2000.

6. Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, 1st edition, July 2007.

7. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

8. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

9. P. Anand, A. B. Kunnumakkara, C. Sundaram, K. B. Harikumar, S. T. Tharakan, O. S. Lai, B. Sung, and B. B. Aggarwal. Cancer is a preventable disease that requires major lifestyle changes. *Pharm. Res.*, 25(9):2097–116, 2008.

10. Anonymous. Primates on facebook: Even online, the neocortex is the limit. *The Economist*, 2009. Print Edition.

11. 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. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, May 2000.

12. Yassen Assenov, Fidel Ramírez, Sven-Eric Schelhorn, Thomas Lengauer, and Mario Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, January 2008.

13. O. T. Avery, C. M. Macleod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of Experimental Medicine*, 79(2):137–158, February 1944.

14. M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, 14(3):283–291, June 2004.

15. M. Madan Babu and S. A. Teichmann. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Research*, 31(4):1234–1244, February 2003.

16. M. Madan Babu, Sarah A. Teichmann, and L. Aravind. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *Journal of Molecular Biology*, 358(2):614–633, April 2006.

17. S Balaji, M Madan M. Babu, and L Aravind. Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of *E. coli*. *Journal of Molecular Biology*, 372(4):1108–1122, July 2007.

18. S. Balaji, Lakshminarayan M. Iyer, L. Aravind, and Madan M. Babu. Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. *Journal of Molecular Biology*, 360(1):204–212, June 2006.

19. A. L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–13, 2004.

20. Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

21. Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, January 2011.

22. J. Berg, M. Lässig, and A. Wagner. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4(1):1–12, November 2004.

23. Nitin Bhardwaj, Matthew B. Carson, Alexej Abyzov, Koon-Kiu K. Yan, Hui Lu, and Mark B. Gerstein. Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets. *PLoS Computational Biology*, 6(5):e1000755+, May 2010.

24. Nitin Bhardwaj, Robert E. Langlois, Guijun Zhao, and Hui Lu. Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Research*, 33(20):6486–6493, 2005.

25. Nitin Bhardwaj and Hui Lu. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Letters*, 581(5):1058–1066, March 2007.

26. Nitin Bhardwaj, Robert V. Stahelin, Robert E. Langlois, Wonhwa Cho, and Hui Lu. Structural bioinformatics prediction of membrane-binding proteins. *Journal of Molecular Biology*, 359(2):486–495, June 2006.

27. M. Bhasin, H. Zhang, E. L. Reinherz, and P. A. Reche. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Letters*, 579(20):4302–4308, August 2005.

28. Peter J. Bickel and Kjell A. Doksum. *Mathematical statistics : basic ideas and selected topics*. Prentice Hall, Upper Saddle River, N.J., 2nd edition, 2001.

29. Adrian Bird. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21, January 2002.

30. C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genetics*, 2(3):0243–0252, March 2006.

31. Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

32. Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

33. M. V. Brock, J. G. Herman, and S. B. Baylin. Cancer as a manifestation of aberrant chromatin structure. *Cancer Journal*, 13(1):3–8, 2007.

34. Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

35. T.A. Brown. *Genomes*. Wiley-Liss, 2nd edition edition, 2002. Editor: Carlson, S.

36. Martha L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biology*, 5(1):201+, December 2003.

37. Wray Buntine. Learning classification trees. *Statistics and Computing*, 2:63–73, 1992.

38. Y. D. Cai and S. L. Lin. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta*, 1648(1-2):127–133, May 2003.

39. Matthew B. Carson, Robert Langlois, and Hui Lu. Mining knowledge for the methylation status of CpG islands using alternating decision trees. In *Engineering in Medicine and Biology Society, Proceedings of the 30th Annual International Conference of the IEEE EMBC*, pages 3787–3790, 2008.

40. Matthew B. Carson, Robert Langlois, and Hui Lu. NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Research*, 38(suppl 2):W431–W435, July 2010.

41. Chih-Chung Chang and Chih-Jen Lin. libSVM: a library for support vector machines (version 2.31), 2001.

42. Chong Pyo P. Choe, Sherry C C. Miller, and Susan J J. Brown. A pair-rule gene circuit defines segments sequentially in the short-germ insect *Tribolium castaneum*. *Proc. Natl. Acad. Sci. USA*, 103(17):6560–6564, April 2006.

43. Stefano Ciliberti, Olivier C. Martin, and Andreas Wagner. Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Computational Biology*, 3(2):e15+, February 2007.

44. G. C. Conant and A. Wagner. Convergent evolution of gene circuits. *Nature Genetics*, 34(3):264–266, July 2003.

45. Bernard Conrad and Stylianos E. Antonarakis. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annual Review of Genomics and Human Genetics*, 8:17–35, March 2007.

46. The UniProt Consortium. Reorganizing the protein space at the universal protein resource (UniProt). *Nucleic Acids Research*, 40(D1):D71–D75, January 2012.

47. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

48. R. M. Coulson, A. J. Enright, and C. A. Ouzounis. Transcription-associated protein families are primarily taxon-specific. *Bioinformatics*, 17(1):95–97, January 2001.

49. P. Courvalin. Antimicrobial drug resistance: "prediction is very difficult, especially about the future". *Emerging Infectious Diseases*, 11(10):1503–1506, October 2005.

50. F. H. C. Crick. On protein synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

51. Anton Crombach and Paulien Hogeweg. Evolution of evolvability in gene regulatory networks. *PLoS Computational Biology*, 4(7):e1000112+, July 2008.

52. C. D. De Brasi, D. J. Bowen, P. W. Collins, and I. B. Larripa. The CpG island in intron 22 of the factor viii gene is predominantly methylated on the X chromosome of human males. *Journal of Human Genetics*, 47(5):239–242, May 2002.

53. Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez Tom. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

54. John C. Doyle, David L. Alderson, Lun Li, Steven Low, Matthew Roughan, Stanislav Shalunov, Reiko Tanaka, and Walter Willinger. The "robust yet fragile" nature of the internet. *Proc. Natl. Acad. Sci. USA*, 102(41):14497–14502, October 2005.

55. P. Du, G. Feng, J. Flatow, J. Song, M. Holko, W. A. Kibbe, and S. M. Lin. From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*, 25(12):i63–8, 2009.

56. R. Dunbar. Coevolution of neocortex size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–735, 1993.

57. A. Dutt and R. Beroukhim. Single nucleotide polymorphism array analysis of cancer. *Curr. Opin. Oncol.*, 19(1):43–9, 2007.

58. Peter J. Ellis, Robert A. Furlong, Sarah J. Conner, Jackson Kirkman-Brown, Masoud Afnan, Christopher Barratt, Darren K. Griffin, and Nabeel A. Affara. Coordinated transcriptional regulation patterns associated with infertility phenotypes in men. *Journal of Biomedical Informatics*, 44(8):498–508, August 2007.

59. Annette M. Evangelisti and Andreas Wagner. Molecular evolution in the yeast transcriptional regulation network. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, 302(4):392–411, July 2004.

60. Wei Fan, Wenke Lee, Salvatore J. Stolfo, and Matthew Miller. A multiple model cost-sensitive approach for intrusion detection. In *Machine Learning: ECML 2000: Proceedings of the 11th European Conference on Machine Learning, Barcelona, Catalonia, Spain*, volume 1810, pages 3–14. Springer Berlin/Heidelberg, May/June 2000.

61. F. Fang, S. Fan, X. Zhang, and M.Q. Zhang. Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, 22(18):2204–2209, September 2006.

62. I. Feldman, A. Rzhetsky, and D. Vitkup. Network properties of genes harboring inherited disease mutations. *Proc. Natl. Acad. Sci. USA*, 105(11):4323–8, 2008.

63. Alex F. Feltus, Eva K. Lee, Joseph F. Costello, Christoph Plass, and Paula M. Vertino. DNA motifs associated with aberrant CpG island methylation. *Genomics*, 87(5):572–579, May 2006.

64. F. A. Feltus, E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino. Predicting aberrant CpG island methylation. *Proc. Natl. Acad. Sci. USA*, 100(21):12253–12258, October 2003.

65. Yoav Freund and Llew Mason. The alternating decision tree learning algorithm, 1999.

66. Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the 13th Annual International Conference on Machine Learning*, pages 148–156, Bari, Italy, 1996.

67. Kosuke Fujishima, Mizuki Komasa, Sayaka Kitamura, Haruo Suzuki, Masaru Tomita, and Akio Kanai. Proteome-wide prediction of novel DNA/RNA-binding proteins using amino acid composition and periodicity in the hyperthermophilic archaeon *Pyrococcus furiosus*. *DNA Research*, 14(3):91–102, 2007.

68. P. Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, March 2004.

69. A. Gagnon and B. Ye. Discovery and application of protein biomarkers for ovarian cancer. *Curr. Opin. Obstet. Gynecol.*, 20(1):9–13, 2008.

70. M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2):261–282, July 1987.

71. B. Garssen. Psychological factors and cancer development: evidence after 30 years of research. *Clin. Psychol. Rev.*, 24(3):315–38, 2004.

72. M. S. Gelfand. Evolution of transcriptional regulatory networks in microbial genomes. *Curr. Opin. Struct. Biol.*, 16(3):420–429, June 2006.

73. Gh Gluhovschi, A. Gluhovschi, Ligia Petrica, D. Anastasiu, Cristina Gluhovschi, and Silvia Velciov. Pregnancy-induced hypertension–a particular pathogenic model: Similarities with other forms of arterial hypertension. *Romanian Journal of Internal Medicine (Revue roumaine de médecine interne)*, 50(1):71–81, 2012.

74. Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proc. Natl. Acad. Sci. USA*, 104(21):8685–8690, May 2007.

75. Graciela Gonzalez, Juan C. Uribe, Luis Tari, Colleen Brophy, and Chitta Baral. Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. In *Pacific Symposium on Biocomputing*, pages 28–39, 2007.

76. Cyril Goutte. Note on free lunches and Cross-Validation. In *Neural Computation*, pages 1245–1249, 1997.

77. M. Greaves and C. C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–13, 2012.

78. Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–517, January 2005.

79. Lian Yi Han, Cong Zhong Cai, Siew Lin Lo, Maxey C. M. Chung, and Yu Zong Chen. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, 10(3):355–368, 2004.

80. D. Hao and C. Li. The dichotomy in degree correlation of biological networks. *PLoS ONE*, 6(12):e28322, 2011.

81. S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89(22):10915–10919, November 1992.

82. T. Ideker and R. Sharan. Protein networks in disease. *Genome Research*, 18(4):644–52, 2008.

83. R. Illingworth, A. Kerr, D. Desousa, H. Jrgensen, P. Ellis, J. Stalker, D. Jackson, C. Clee, R. Plumb, J. Rogers, S. Humphray, T. Cox, C. Langford, and A. Bird. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biology*, 6(1):0037–51, January 2008.

84. Mark Isalan, Caroline Lemerle, Konstantinos Michalodimitrakis, Carsten Horn, Pedro Beltrao, Emanuele Raineri, Mireia Garriga-Canut, and Luis Serrano. Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, 452(7189):840–845, April 2008.

85. Euna Jeong, I-Fang Chung, and Satoru Miyano. A neural network method for identification of RNA-interacting residues in protein. *Genome Informatics*, 15(1):105–116, 2004.

86. P. A. Jones. The DNA methylation paradox. *Trends In Genetics: TIG*, 15(1):34–37, January 1999.

87. P. A. Jones, W. M. Rideout, J. C. Shen, C. H. Spruck, and Y. C. Tsai. Methylation, mutation and cancer. *BioEssays*, 14(1):33–36, January 1992.

88. Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983.

89. Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(Database issue):D109–D114, January 2012.

90. Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, October 2008.

91. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, July 2004.

92. Michael Kearns and Dana Ron. Algorithmic stability and Sanity-Check bounds for Leave-One-out Cross-Validation. In *Neural Computation*, pages 152–162, 1997.

93. T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database issue):D767–72, 2009.

94. Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi C. Guo, and David S. Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, 39(Database issue):D1035–D1041, January 2011.

95. H. Kono and A. Sarai. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, 35(1):114–131, April 1999.

96. D. Koschutzki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology*, 2:193–201, 2008.

97. Arun Krishnan, Masaru Tomita, and Alessandro Giuliani. Evolution of gene regulatory networks: Robustness as an emergent property of evolution. *Physica A: Statistical Mechanics and its Applications*, 387(8-9):2170–2186, March 2008.

98. R. P. Kuiper, M. J. Ligtenberg, N. Hoogerbrugge, and A. Geurts van Kessel. Germline copy number variation and cancer risk. *Curr. Opin. Genet. Dev.*, 20(3):282–9, 2010.

99. Manish Kumar, Michael M. Gromiha, and G. P. S. Raghava. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins: Struct., Funct., Bioinf.*, 71(1):189–194, 2008.

100. L. H. Kushi, T. Byers, C. Doyle, E. V. Bandera, M. McCullough, A. McTiernan, T. Gansler, K. S. Andrews, and M. J. Thun. American cancer society guidelines on nutrition and physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity. *CA Cancer J. Clin.*, 56(5):254–81; quiz 313–4, 2006.

101. I. B. Kuznetsov, Z. Gou, R. Li, and S. Hwang. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins: Struct., Funct., Bioinf.*, 64(1):19–27, July 2006.

102. R. Langlois and H. Lu. Intelligible machine learning with malibu. In *Proceedings of the 30th Annual International Conference of the IEEE, EMBC*, Vancouver, British Columbia, Canada, August 20-24 2008.

103. R. E. Langlois and H. Lu. Boosting the prediction and understanding of DNA-binding domains from sequence. *Nucleic Acids Research*, 38(10):3149–58, 2010.

104. Robert Langlois, Matthew Carson, Nitin Bhardwaj, and Hui Lu. Learning to translate sequence and structure to function: Identifying DNA binding and membrane binding proteins. *Annals of Biomedical Engineering*, 35(6):1043–1052, 2007.

105. M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and

D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, November 2007.

106. Steven Lehrer. Association between malaria incidence and all cancer mortality in fifty U.S. states and the District of Columbia. *Anticancer Research*, 30(4):1371–1373, April 2010.

107. Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.

108. Olivier Lespinet, Yuri I. Wolf, Eugene V. Koonin, and L. Aravind. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Research*, 12(7):1048–1059, July 2002.

109. P. A. Levene. The structure of yeast nucleic acid: IV. Ammonia hydrolysis. *Journal of Biological Chemistry*, 40:415–424, 1919.

110. H. Li, Y. Lee, J. L. Chen, E. Rebman, J. Li, and Y. A. Lussier. Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory. *J. Am. Med. Inform. Assoc.*, 19(2):295–305, 2012.

111. L. Li, K. Zhang, J. Lee, S. Cordes, D. P. Davis, and Z. Tang. Discovering cancer genes by integrating network and functional properties. *BMC Medical Genomics*, 2:61, 2009.

112. Bing Liu. *Web data mining : exploring hyperlinks, contents, and usage data*. Data-centric systems and applications. Springer, Berlin ; New York, 2007.

113. Zhijie Liu, Fenglou Mao, Jun-tao Guo, Bo Yan, Peng Wang, Youxing Qu, and Ying Xu. Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. *Nucleic Acids Research*, 33(2):546–558, January 2005.

114. H. Lu, L. Lu, and J. Skolnick. Development of unified statistical potentials describing protein-protein interactions. *Biophysical Journal*, 84(3):1895–1901, March 2003.

115. H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44(3):223–232, August 2001.

116. H. Lu and J. Skolnick. Application of statistical potentials to protein structure refinement from low resolution ab initio models. *Biopolymers*, 70(4):575–584, December 2003.

117. A. N. Magewu and P. A. Jones. Ubiquitous and tenacious methylation of the CpG site in codon 248 of the p53 gene may explain its frequent appearance as a mutational hot spot in human cancer. *Molecular and Cellular Biology*, 14(6):4225–4232, June 1994.

118. Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39(suppl 1):D52–D57, January 2011.

119. Y. Mandel-Gutfreund, A. Baron, and H. Margalit. A structure-based approach for prediction of protein binding sites in gene upstream regions. In *Pacific Symposium on Biocomputing*, pages 139–150, 2001.

120. Y. Mandel-Gutfreund and H. Margalit. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Research*, 26(10):2306–2312, May 1998.

121. S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–3, 2002.

122. Thomas H. Massey, Christopher P. Mercogliano, James Yates, David J. Sherratt, and Jan Löwe. Double-stranded DNA translocation: structure and mechanism of hexameric FtsK. *Molecular Cell*, 23(4):457–469, August 2006. PDB ID: 2IUS.

123. MATLAB. *version 7.14.0 (R2012a)*. The MathWorks Inc., Natick, Massachusetts, 2012.

124. M. J. McCauley, L. Shokri, J. Sefcikova, C. Venclovas, P. J. Beuning, and M. C. Williams. Distinct double- and single-stranded DNA binding of *E. coli* replicative DNA polymerase III alpha subunit. *ACS Chemical Biology*, 3(9):577–587, September 2008.

125. Christopher J. Miller, Christopher Genovese, Robert C. Nichol, Larry Wasserman, Andrew Connolly, Daniel Reichart, Andrew Hopkins, Jeff Schneider, and Andrew Moore. Controlling the false discovery rate in astrophysical data analysis. *The Astronomical Journal*, 122(6):3492–3505, July 2001.

126. James Moody. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69:213–238, April 2004.

127. T. Nepusz, H. Yu, and A. Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. In preparation.

128. M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

129. Yanay Ofran, Venkatesh Mysore, and Burkhard Rost. Prediction of DNA-binding residues from sequence. *Bioinformatics*, 23(13):i347–i353, Jul 2007.

130. David L. Olson. Data set balancing. In *Data Mining and Knowledge Management*, volume 3327, pages 71–80. Springer Berlin/Heidelberg, 2005.

131. C. Ortutay and M. Vihinen. Identification of candidate disease genes by integrating gene ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Research*, 37(2):622–628, November 2008.

132. J. D. Osborne, S. Lin, W. A. Kibbe, L. Zhu, M. I. Danila, and R. L. Chisholm. GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM. Associated with the DOLite project, 2007.

133. John Osborne, Jared Flatow, Michelle Holko, Simon Lin, Warren Kibbe, Lihua Zhu, Maria Danila, Gang Feng, and Rex Chisholm. Annotating the human genome with disease ontology. *BMC Genomics*, 10(Suppl 1):S6+, 2009.

134. Arzucan Özgür, Thuy Vu, Güneş Erkan, and Dragomir R. Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. In *Bioinformatics* (183), pages i277–i285. Editor: Edwin Wang.

135. G. Parkinson, C. Wilson, A. Gunasekera, Y. W. Ebright, R. H. Ebright, R. E. Ebright, and H. M. Berman. Structure of the CAP-DNA complex at 2.5 angstroms resolution: a complete picture of the protein-DNA interface. *Journal of Molecular Biology*, 260(3):395–408, July 1996.

136. Nathan Pike. Using false discovery rates for multiple comparisons in ecology and evolution. *Methods in Ecology and Evolution*, 2(3):278–282, June 2011.

137. L. Pray. Discovery of DNA structure and function: Watson and Crick. *Nature Education*, 1:1, 2008.

138. Kim D. Pruitt, Tatiana Tatusova, Garth R. Brown, and Donna R. Maglott. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(Database issue):D130–D135, January 2012.

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

140. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

141. M. Safran, N. Nativ, Y. Golan, I. Dalah, T. Iny Stein, G. Stelzer, and D. Lancet. MalaCards - the integrated human malady compendium. In *20th Annual International Conference on Intelligent Systems for Molecular Biolog (ISMB)*, July 2012.

142. Akinori Sarai and Hidetoshi Kono. Protein-DNA recognition patterns and predictions. *Annual Review of Biophysics and Biomolecular Structure*, 34:379–398, 2005.

143. Lynn Marie M. Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne W. Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden A. Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(Database issue):D940–D946, January 2012.

144. E. Schrödinger. *What Is Life? The Physical Aspect of the Living Cell.* Cambridge University Press, Cambridge, 1944.

145. Ruth L. Seal, Susan M. Gordon, Michael J. Lush, Mathew W. Wright, and Elspeth A. Bruford. genenames.org: the HGNC resources in 2011. *Nucleic Acids Research*, 39(Database issue):D514–D519, January 2011.

146. Samuel Selvaraj, Hidetoshi Kono, and Akinori Sarai. Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. *Journal of Molecular Biology*, 322(5):907–915, October 2002.

147. Amitabh Sharma, Sreenivas Chavali, Rubina Tabassum, Nikhil Tandon, and Dwaipayan Bharadwaj. Gene prioritization in type 2 diabetes using domain interactions and network analysis. *BMC Genomics*, 11(1):84+, February 2010.

148. L. Shen, Y. Kondo, S. Ahmed, Y. Boumber, K. Konishi, Y. Guo, X. Chen, J. N. Vilaythong, and J. P. Issa. Drug sensitivity prediction by CpG island methylation

profile in the NCI-60 cancer cell line panel. *Cancer Research*, 67(23):11335–11343, December 2007.

149. Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1):64–68, May 2002.

150. Alfonso Shimbel. Structural parameters of communication networks. *Bulletin of Mathematical Biology*, 15(4):501–507, 1953.

151. M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–2, 2011.

152. Birgitta Söder, Maha Yakob, Jukka H. Meurman, Leif C. Andersson, and Per-Östen Söder. The association of dental plaque with cancer mortality in Sweden: A longitudinal study. *BMJ Open*, 2(3), January 2012.

153. C. O. S. Sorzano, E. Recarte, M. Alcorlo, J. R. Bilbao-Castro, C. San-Martín, R. Marabini, and J. M. Carazo. Automatic particle selection from electron micrographs using machine learning techniques. *Journal of Structural Biology*, 167(3):252–260, September 2009.

154. Eric W. Stawiski, Lydia M. Gregoret, and Yael Mandel-Gutfreund. Annotating nucleic acid-binding function based on protein structure. *Journal of Molecular Biology*, 326(4):1065–1079, February 2003.

155. Gil Stelzer, Irina Dalah, Tsippi Iny I. Stein, Yigeal Satanower, Naomi Rosen, Noam Nativ, Danit Oz-Levi, Tsviya Olender, Frida Belinky, Iris Bahir, Hagit Krug, Paul Perco, Bernd Mayer, Eugene Kolker, Marilyn Safran, and Doron Lancet. In-silico human genomics with GeneCards. *Human Genomics*, 5(6):709–717, October 2011.

156. U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–68, 2005.

157. Barbara E. Stranger, Alexandra C. Nica, Matthew S. Forrest, Antigone Dimas, Christine P. Bird, Claude Beazley, Catherine E. Ingle, Mark Dunning, Paul Flicek,

Daphne Koller, Stephen Montgomery, Simon Tavaré, Panos Deloukas, and Emmanouil T. Dermitzakis. Population genomics of human gene expression. *Nature Genetics*, September 2007.

158. G. Strathdee, B. R. Davies, J. K. Vass, N. Siddiqui, and R. Brown. Cell type-specific methylation of an intronic CpG island controls expression of the MCJ gene. *Carcinogenesis*, 25(5):693–701, May 2004.

159. S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Computational Biology*, 6(2):e1000662, 2010.

160. D. Takai and P. A. Jones. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA*, 99(6):3740–3745, March 2002.

161. G. Taubes. Cancer research: Unraveling the obesity-cancer connection. *Science*, 335(6064):28, 30–2, 2012.

162. S. A. Teichmann and M. M. Babu. Gene regulatory network growth by duplication. *Nature Genetics*, 36(5):492–496, May 2004.

163. Catherine Terret, Elisabeth Castel-Kremer, Gilles Albrand, and Jean Pierre P. Droz. Effects of comorbidity on screening and early diagnosis of cancer in elderly people. *The Lancet Oncology*, 10(1):80–87, January 2009.

164. M. Terribilini, J. D. Sander, J. H. Lee, P. Zaback, R. L. Jernigan, V. Honavar, and D. Dobbs. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Research*, 35(Web Server issue):1–7, July 2007.

165. Michael Terribilini, Jae-Hyung Lee, Changhui Yan, Robert L. Jernigan, Vasant Honavar, and Drena Dobbs. Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, 12(8):1450–1462, August 2006.

166. Harianto Tjong and Huan-Xiang Zhou. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Research*, 35(5):1465–1477, March 2007.

167. Y. Tsuchiya, K. Kinoshita, and H. Nakamura. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the

shape of molecular surfaces. *Proteins: Struct., Funct., Bioinf.*, 55(4):885–894, June 2004.

168. T. Ueki, K. M. Walter, H. Skinner, E. Jaffee, R. H. Hruban, and M. Goggins. Aberrant CpG island methylation in cancer cell lines arises in the primary cancers from which they were derived. *Oncogene*, 21(13):2114–2117, March 2002.

169. Thomas W. Valente and Robert K. Foreman. Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 20(1):89–105, 1998.

170. Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann, and Nicholas M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, April 2009.

171. R. Visone and C. M. Croce. Mirnas and cancer. *American Journal of Pathology*, 174(4):1131–8, 2009.

172. A. Wagner. Distributed robustness versus redundancy as causes of mutational robustness. *BioEssays*, 27(2):176–188, February 2005.

173. Andreas Wagner and Jeremiah Wright. Alternative routes and mutational robustness in complex regulatory networks. *Biosystems*, 88(1-2):163–172, March 2007.

174. F. Wang, M. Herrington, J. Larsson, and J. Permert. The relationship between diabetes and pancreatic cancer. *Molecular Cancer*, 2:4, 2003.

175. Liangjiang Wang and Susan J. Brown. Prediction of RNA-binding residues in protein sequences using support vector machines. *Engineering in Medicine and Biology Society, Proceedings of the 28th Annual International Conference of the IEEE EMBC.*, 1:5830–5833, 2006.

176. Xiaosheng Wang and Osamu Gotoh. Inference of cancer-specific gene regulatory networks using soft computing rules. *Gene Regulation and Systems Biology*, 4:19–34, 2010.

177. Y. Wang, Z. Xue, G. Shen, and J. Xu. PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*, 35(2):295–302, August 2008.

178. J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

179. D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–2, 1998.

180. Michael D. Wilson, Nuno L. Barbosa-Morais, Dominic Schmidt, Caitlin M. Conboy, Lesley Vanes, Victor L. Tybulewicz, Elizabeth M. Fisher, Simon Tavare, and Duncan T. Odom. Species-specific transcription in mice carrying human chromosome 21. *Science*, 322:329–488, September 2008.

181. J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology*, 285(4):1735–1747, January 1999.

182. Xuebing Wu, Rui Jiang, Michael Q. Zhang, and Shao Li. Network-based global inference of human disease genes. In *Molecular Systems Biology* (183), pages 191–212. Editor: Edwin Wang.

183. Xuebing Wu and Shao Li. *Cancer gene prediction using a network approach - Chapter 11 in Cancer Systems Biology*. Chapman & Hall/CRC Mathematical & Computational Biology. CRC, 2010. Editor: Edwin Wang.

184. Ioannis Xenarios, Danny W. Rice, Lukasz Salwinski, Marisa K. Baron, Edward M. Marcotte, and David Eisenberg. DIP: the database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291, January 2000.

185. Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22(22):2800–2805, November 2006.

186. Yoichi Yamada, Hidemi Watanabe, Fumihito Miura, Hidenobu Soejima, Michiko Uchiyama, Tsuyoshi Iwasaka, Tsunehiro Mukai, Yoshiyuki Sakaki, and Takashi Ito. A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome Research*, 14(2):247–266, February 2004.

187. Changhui Yan, Michael Terribilini, Feihong Wu, Robert Jernigan, Drena Dobbs, and Vasant Honavar. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, 7(1):262–272, 2006.

188. Shenghong Yang, Xiaoxu Wang, Gianmarco Contino, Marc Liesa, Ergun Sahin, Haoqiang Ying, Alexandra Bause, Yinghua Li, Jayne M. Stommel, Giacomo Dell'Antonio, Josef Mautner, Giovanni Tonon, Marcia Haigis, Orian S. Shirihai, Claudio

Doglioni, Nabeel Bardeesy, and Alec C. Kimmelman. Pancreatic cancers require autophagy for tumor growth. *Genes & Development*, 25(7):717–729, April 2011.

189. P. A. Yates, R. W. Burman, P. Mummaneni, S. Krussel, and M. S. Turker. Tandem b1 elements located in a mouse methylation center provide a target for de novo DNA methylation. *The Journal of Biological Chemistry*, 274(51):36357–36361, December 1999.

190. Ji-Won W. Yoon and Hee-Sook S. Jun. Autoimmune destruction of pancreatic beta cells. *American Journal of Therapeutics*, 12(6):580–591, 2005.

191. B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the Third IEEE International Conference on Data Mining, ICDM*, Melbourne, Florida, November 19-22 2003. IEEE.

192. Minlu Zhang, Cheng Zhu, Alexis Jacomy, Long J. Lu, and Anil G. Jegga. The orphan disease networks. *American Journal of Human Genetics*, 88(6):755–766, June 2011.

193. Q. Zhang, F. Y. Wang, D. Zeng, and T. Wang. Understanding crowd-powered search groups: a social network perspective. *PLoS ONE*, 7(6):e39749, 2012.

194. Xiaoyu Zhang, Junshi Yazaki, Ambika Sundaresan, Shawn Cokus, Simon W-L W. Chan, Huaming Chen, Ian R R. Henderson, Paul Shinn, Matteo Pellegrini, Steve E E. Jacobsen, and Joseph R R. Ecker. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell*, 126(6):1025–8, August 2006.

195. Fang Zhao, Zhenyu Xuan, Lihua Liu, and Michael Q. Zhang. TRED: a transcriptional regulatory element database and a platform for in silico gene regulation studies. *Nucleic Acids Research*, 33(Database issue):D103+, January 2005.

196. G. Zhao, M. B. Carson, and H. Lu. Prediction of specific protein-DNA recognition by knowledge-based two-body and three-body interaction potentials. In *Engineering in Medicine and Biology Society, Proceedings of the 29th Annual International Conference of the IEEE EMBC.*, Lyon, France, August 23-26 2007.

197. Guijun Zhao and Hui Lu. Development of a grid-based statistical potential for protein structure prediction. In *Proceedings of the 27th Annual International Conference of the IEEE EMBC*, volume 6, pages 6064–6067, 2005.

198. Dongling Zheng, Chrystala Constantinidou, Jon L. Hobman, and Stephen D. Minchin. Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. *Nucleic Acids Research*, 32(19):5874–5893, 2004.

VITA

Matthew B. Carson

Education

University of Illinois at Chicago, Chicago, IL
Ph.D., Bioinformatics, (June '06 - May '13)
Dissertation: Investigation of Gene Regulation and its Application to Disease Using Machine Learning and Network Models.

University of Oklahoma, Norman, OK
B.Sc., Microbiology, (Aug '96 - May '00)

Grants

1. **NIH Training Grant Fellowship**, Grant # 5 T32 HL 07692-19, 'Training in Cellular Signaling in the Cardiovascular System', Program Director R. John Solaro,(06/30/09 - 06/29/10)

Peer-reviewed Publications

1. **Carson**, Kibbe, and Lu. *Discovering Disease Relationships Using a Co-occurrence Matrix.* TO BE SUBMITTED.
2. **Carson** and Lu. *Network-based Knowledge Mining for Disease Genes.* SUBMITTED.
3. Jia, **Carson**, and Yu. *A Fast Weak Motif-Finding Algorithm Based on Community Detection in Graphs.* SUBMITTED.
4. **Carson**, Liu, and Lu. *Knowledge Mining of Disease Networks can Provide New Insights in Cancer Research through the Analysis of Other Diseases.* J. CARCINOGENE. MUTAGENE. Mar; 3, 1-3. 2012.
5. Siddaramappa, Challacombe, Duncan, Gillaspy, **Carson**, Gipson, Orvis, Zaitshik, Barnes, Bruce, Chertkov, Detter, Han, Tapia, Thompson, Dyer, and Inzana. *Horizontal gene transfer in Histophilus somni and its role in the evolution of pathogenic strain 2336, as determined by comparative genomic analyses.* BMC GENOMICS. Nov 23; 12:570. 2011.
6. **Carson**, Langlois, Lu. *NAPS: A Residue-level Nucleic Acid-binding Prediction Server.* NUCLEIC ACIDS RESEARCH. Jul; 38, W431-5. 2010.
7. Bhardwaj, **Carson**, Abyzov, Yan, Lu, Gerstein. *Analysis of Combinatorial Regulation: Scaling of Partnerships between Regulators with the Number of Governed Targets.* PLOS COMPUTATIONAL BIOLOGY. May 27; 6(5):e1000755. 2010.
8. **Carson**, Langlois, Lu. *Mining Knowledge for the Methylation Status of CpG Islands Using Alternating Decision Trees.* CONF. PROC. IEEE ENG. MED. BIOL. SOC. 378790. 2008.

9. Zhao, **Carson**, Lu. *Prediction of specific protein-DNA recognition by knowledge-based two-body and three-body interaction potentials.* Conf. Proc. IEEE Eng. Med. Biol. Soc. 5017-20. 2008.
10. Langlois, **Carson**, Bhardwaj, Lu. *Learning to translate sequence and structure to function: identifying DNA-binding and membrane binding proteins.* Annals of Biomedical Engineering. Jun; 35(6):1043-1052. 2007.
11. Harrison, Dyer, Gillaspy, Ray, Mungur, **Carson**, Gipson, Gipson, Johnson, Lewis, Bakaletz and Munson. *Genomic sequence of an otitis media isolate of nontypeable Haemophilus influenzae: a comparative study with Haemophilus influenzae serotype d, strain KW20.* J. Bacteriol. Jul; 187(13):4627-36. 2005.
12. Ducey, **Carson**, Orvis, Stintzi, and Dyer. *Identification of the Iron-Responsive Genes of Neisseria gonorrhoeae by Iron-limitation Gene Analysis in Defined Medium.* J. Bacteriol. Jul; 187(14):4865-74. 2005.
13. Munson, Harrison, Gillaspy, Ray, **Carson**, Armbruster, Gipson, Gipson, Johnson, Lewis, Dyer and Bakaletz. *Partial analysis of the genomes of two nontypeable Haemophilus influenzae otitis media isolates.* Infect. Immun. May; 72(5):3002-10. 2004.
14. Ajdic, McShan, McLaughlin, Savic, Chang, **Carson**, Primeaux, Tian, Kenton, Jia, Lin, Qian, Li, Zhu, Najar, Lai, White, Roe, and Ferretti. *Genome sequence of Streptococcus mutans UA159, a cariogenic dental pathogen.* Proc. Natl. Acad. Sci. USA. Oct 29; 99(22):14434-9. 2002.

Selected Non-refereed Publications and Abstracts

1. **Carson** and Lu. *Cancer-related Genes in Conserved Human Transcription Factor Network Motifs.* UIC College of Medicine Research Forum, November 16, 2012.
2. **Carson**, Bhardwaj, and Lu. *Investigating Co-Regulation Networks Using Generative Models.* Biophysical Journal. Volume 98, Issue 3, Supplement 1, January 2010, Page 740a. Biophysical Society Annual Meeting, San Francisco, CA, February, 2010.
3. Källberg, Langlois, **Carson**, and Lu. *Understanding Protein-DNA Interactions through Dynamics.* Biophysical Journal. Volume 98, Issue 3, Supplement 1, January 2010, Pages 470a-471a. Biophysical Society Annual Meeting, San Francisco, CA, February, 2010.
4. **Carson**, Bhardwaj, and Lu. *Investigating Co-Regulation Networks using Generative Models.* Presented at RECOMB Regulatory Genomics Conference, MIT / Broad Institute, Boston, MA, December 2-6, 2009.
5. **Carson**, Langlois, and Lu. *Understanding DNA- and RNA-binding Proteins Using Sequence and Structural Features.* Biophysical Journal. Volume 96, Issue 3, Supplement 1, February 2009, Page 64a. Biophysical Society Annual Meeting, Boston, MA, February, 2009.
6. Zhao, **Carson**, Bhardwaj, Langlois, and Lu. *Discerning Protein Nucleic Acid Interactions Using Machine Learning.* Presented at the 51^{st} Biophysical Society Annual Meeting, Baltimore, MD, February, 2007.
7. Langlois, Bhardwaj, **Carson**, and Lu. *Deciphering Structure from Sequence and Function: Learning to Identify non-homologous DNA-binding and Membrane-binding Proteins.* Presented at the 51^{st} Biophysical Society Annual Meeting, Baltimore, MD, February, 2007.