The central dogma of molecular biology

The molecule we know today as deoxyribonucleic acid 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’ (Pray, 2008). 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 (Levene, 1919). 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 (Avery, 1944), set out to discover whether the deoxyribonucleic acid molecule differed among species. He found that although, in contrast to Levene’s proposal that nucleotides are always repeated in the same order, nucleotides appear in different orders in different organisms, these molecules maintained 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 (Watson, 1953). 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.

The Central Dogma of Molecular Biology, first proposed by Francis Crick (Crick, 1958), describes the directional processes of conversion from DNA to RNA and from RNA to protein. This 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.

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.

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 out of the nucleus. The cap is critical for recognition by the ribosome and protection from enzymes called RNases that will break

down the molecule. The poly(A) tail and the protein bound to it aid in protecting mRNA from degradation by other enzymes called exonucleases.

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 (Choe, 2006, Wilson, 2008), and on a more global network level. There are, however, more specific reasons to investigate this process more closely. Failure in gene regulation has been shown to be a key factor in disease (Stranger, 2007). Additionally, learning how to interrupt gene regulation may lead to the development of drugs to fight bacteria and viruses (McCauley, 2008). A clearer understanding of this process in microorganisms may lead to possible solutions to the problem of antimicrobial resistance (Courvalin, 2005).

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.

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, 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.

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.

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.

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. First, we find 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. We believe 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.