Previous work in network analysis

Increasing big data from diverse sources provide much information in a more complex way than before. The formalism of networks provides a common language for understanding a wide range of complex systems. Network science has progressed rapidly in the past decade as large network datasets have become available in a variety of fields. Our knowledge of biological networks, neural networks, social networks, and the Internet has expanded greatly because of technological advances that have made collecting network data easier. The important task of understanding and interpreting this new kind of big data can be advanced using the common language of networks to compare and contrast different classes of complex systems.

In our early work, we developed methods for predicting networks from individual genome features[1,23] Later, we combined different biological datasets to increase the power of our network prediction algorithms[4-67-9] and developed new machine learning techniques.[10] In 2008, this work placed first in the Dialogue for Reverse Engineering Assessments and Methods (DREAM, www.the-dream-project.org) competition for the in silico network prediction challenge. In addition, we have participated in many experimental network determination projects (e.g., (Borneman et al., 2006; Li et al., 2004), in a continual effort to refine our methodologies and keep them at the cutting edge. We have also participated in many experimental network determination projects.[12-15] We have constructed many web tools for network analysis including Topnet,[16] tYNA,[17] and PubNet.[18] We have developed several methods to construct networks based upon genome features (Jansen et al., 2002). Based on this work, we combined several heterogeneous biological datasets to increase the power of prediction (Edwards et al., 2002; Lu et al., 2005; Xia et al., 2006) and we developed new machine-learning techniques to support these research goals (Yip et al., 2009). In addition, we developed a method to study network rewiring on all currently available biological networks. We noted that biological networks show a decreased rate of change at large time intervals. However, different types of biological networks consistently rewire at different rates (Shou et al., 2011).

Cellular networks are also organized in the form of interacting modules, whereby nodes in a module tend to have a larger density of edges connecting them. Biologically, the genes within a module of a genetic regulatory network are co-regulated. We developed various methods to identify the functional modules of various networks. For example, by mapping gene-expression data onto the regulatory network of yeast, we identified different sub-networks that are active in different conditions (Luscombe et al., 2004). We developed a method to extract metabolic modules from metagenomic data, enabling us to identify pathways that are expressed under different environmental conditions (Gianoulis et al., 2009), and a computational framework using spectral analysis to identify connection patterns across three datasets, and applied it in a variety of genomic contexts including chemogenomics data (Gianoulis et al. 2011). We have also developed a way to identify nearly complete, fully connected modules (cliques) present in network interactions (Yu et al., 2006) and we have been using networks to map various kinds of functional genomics data (Gerstein et al., 2010 and 2012). We also defined a module as all accessible nodes downstream of a top regulator and investigated the overlap (share of regulators) between modules. We have found that the modules in E. coli are more independent compared to those within the call-graph of the Linux kernel (Yan et al., PNAS 2010).

Previous work on modeling gene regulatory networks

Various regulator factors control gene expression in a network way. Transcription factors and histone modifications are two interrelated components that regulate the transcriptional output of a gene. To quantify the relationship between TF binding and gene expression, we have constructed linear and non-linear models (see Fig. X) that utilize the binding signals of multiple TFs in the transcription start site (TSS) proximal to genes as the input to "predict" expression levels of protein

coding genes (Cheng and Gerstein, 2012). Similarly, we have also constructed models to predict gene-expression levels based on histone modification signals at different positions proximal to the TSS of different genes (Dong et al., 2012; Gerstein et al., 2010). We applied these models in multiple organisms ranging from yeast to human. Using the machine-learning approaches we developed for identifying individual proximal and distal edges together with miRNA target prediction (and other) algorithms, we have completed the highly ambitious goal of constructing highly integrated regulatory networks for humans and model organisms based on the ENCODE (Gerstein et al., 2012) and modENCODE datasets (Negre et al., 2011).

Proposed work

We aim to develop predictive models to identify gene regulatory networks and mechanisms underlying network connections. For example, to identify regulations of TF/histone modification (HM) from proximal and distal DNA regions, we will apply our machine-learning methods to model gene expression and changes in gene expression based on the amount of TF binding and/or histone modification close to the transcription start site of a gene. By including data from ChIA-PET on distant chromatin interactions, we will define a list of distal regulatory interactions between enhancers and putative target genes. We will build an enhancer model that utilizes the TF binding and/or histone modification signals in candidate enhancers as the input to predict the expression levels of their putative targets.

Beyond studying topologies of regulatory networks, to take into account whether gene expression has an activating or a repressive influence, we will try to integrate logical circuits from Very-large-scale integration design in electronic chips into our analyses. In many cases, the regulatory factors controlling gene expression can be modeled using Boolean models and logical circuits (Arnosti and Ay, 2012; Bonnet et al., 2013; Buchler et al., 2003; Mangan and Alon, 2003; Tu et al., 2013). We plan to utilize the Boolean network approach to define the cooperation between regulators (i.e., TFs, miRNAs and lncRNAs). Also, to incorporate the interactions between various genomic features, we plan to build prediction models based on Dynamic Bayesian network (DBN), which is a probabilistic graphical model that describes dependencies and conditional independencies for a set of random variables over adjacent time points (Zou and Conzen, 2005). It provides a coherent framework for predicting random variable behavior through the integration of prior knowledge and new data. We consider DBN to be an extension of our previous TF/HM modeling due to its ability to incorporate the prior multiplex regulatory network from above, and to simultaneously model local regulatory behavior (e.g., direct regulation) and global regulatory behavior (e.g., indirect regulation and/or conditional independence).

We will further use system modeling such as Ordinary Differential Equations (ODEs) in control theory using time-series data to model the regulatory network comprising multiple types of biomolecules, including genes, mRNAs, regulatory proteins (i.e., TFs) and DNA-protein complexes. The network of biomolecules will be expressed as a set of elementary reactions, such as reactant→ product, with biomolecules serving as reactants or products. The dynamics in and levels of each biomolecule in the elementary reaction can be modeled using the ODEs.

Summary

Our research will integrate variety of biological datasets and systematically predict gene regulatory networks. Our predictive models allow us to identify uncommon regulatory mechanisms in human disease such as the effects of genomic mutations that we found recently (Khurana et al., 2013ab), and make significant progresses for human health.