



Decoding Gene Expression Regulation Through Motif Discovery and Classification

Yuan Yuan

Dissertation Advisor: Professor Jun S. Liu

Biological systems are complex machineries with numerous components interacting with each other. Through the regulation of gene expression, the systems work differently at different conditions. The regulatory rules are by and large determined by DNA, as it is the most important inheritable substance. Thus, it is interesting to infer these rules by building connections between DNA sequences and gene expression. Modern high-throughput technologies are able to provide us with massive amounts of data related to sequence features and gene expression. However, the scale of the data also brings the challenges of variable selection and computation efficiency.

This dissertation is mainly composed of three parts. We first discuss the motif discovery problem, including its current status and challenges, with a real data application of motif finding for protein *abrB* in *Bacillus subtilis*, using a specially designed protein binding microarray data. Then we present the problem of predicting gene expression using DNA sequences. Sequence features such as motif matching scores are used as predictors. A Bayesian variable selection scheme is designed to select motifs which are most related to the expression of target genes, and also discover the interaction or synergic effect among them. This method is further extended into a general classifier, called selective partially augmented naïve Bayes (SPAN). We compare this classifier and its variant to several state-of-the-art classifiers, with applications in several real and simulated datasets. SPAN is a very fast classifier, and is shown to have an intrinsic connection with logistic regression. It is able to fit a logistic regression model with large number of covariates, achieving both variable selection and interaction detection at the same time.