Advanced eQTL analysis for identification of regulatory factors and their target genes

Qi, Wenjie; Filipovic, David

Final report: CMSE 491 - Computational biology and bioinformatics

Problem

DNA is made of a sequence of nucleotides containing one of the four possible nitrogenous bases (A, C, T and G). All genetic information is precisely encoded in this sequence. Gene, a basic unit of heredity, is the portion of the DNA that encodes for a single protein and is used as an instruction manual in protein synthesis. The central dogma of molecular biology states that protein synthesis occurs in two steps: transcription and translation. In transcription DNA is read by the RNA polymerase and a complementary mRNA (messenger RNA) is constructed and processed. In translation mRNA is read by the ribosome and the final protein product is constructed. A specific mRNA transcript or protein product almost exclusively results from a single gene and the level of that product (either mRNA or protein) can be accurately quantified using biochemical methods. These levels are referred to as gene expression and could be used as targets of machine learning or other statistical algorithms. Gene expression, the process of using gene information for protein synthesis, is also used to refer to the amounts of functional gene product (usually mRNA or protein) resulting from this process. One way in which the gene expression process is regulated is through DNA-protein interactions (which can either stimulate or inhibit transcription). These interactions depend on the structure of the interacting protein, which in turn depends on the gene sequence coding for that protein. The most common way gene sequences exhibit variations in nature is the single nucleotide polymorphism (SNP) where a single nucleotide at a specific position has been replaced with another type of nucleotide. These individual genetic variations have been shown to affect phenotypic traits, most notably gene expression, e.g. in the work of Thomas et al, 2012. Therefore, it should be possible, using SNPs as input features, to construct a prediction framework of gene expression levels (or changes in the levels of their expression), i.e. it should be possible to associate gene expression with genetic variations.

Current Approaches.

Currently, the most commonly used method of associating gene expression with genetic variation is expression quantiative trait loci (eQTL) analysis. This analysis is based on a set of statistical methods that are used to ascertain how much of the genetic variation in gene expression can be attributed (or explained) by variation at a specific locus (a place of SNP occurrence). Xie et al., 2016, have developed a deep learning framework for prediction of gene expression levels, based on genetic variation data (specifically SNP features). They used a well-known microarray yeast dataset, collected by Brem and Krugklyak, 2005, and a deep neural network composed of stacked denoising autoencoders (consisting of 2 denoising autoencoders) with dropout and compared it to the same method without dropout, random forest and Lasso. The problem posed here was interpreted as a multiple regression problem involving supervised learning. However, the dimensionality of the problem is very high. The output contained levels of gene expression, which were quantified for 6111 genes, whereas the input contained genetic variation information for 2956 SNPs. Xie et al., 2016, have noted that this kind of problem was previously approached with simple regression methods, however these methods proved too complicated to be of any use for prediciton Therefore, Xie et al., 2016, have focused only on the SNPs. Additionally, methods like ChIP-seq can reveal DNA binding sites of various proteins, thus enabling us to label some of them as potential transcription factors, however, this kind of data does not tell us what gene targets are affected by this transcription factor-DNA interaction.

Project Goals

The main goal of this project is to come up with an eQTL or a similar analysis method that is more sensitive to variations in the genotype (e.g. by allowing for non-linear dynamics or multiple regression modeling) than currently available methods. The aim is to use this analysis method for identification of regulatory factors and their potential gene targets.

Most of the currently available eQTL models perform a single correlation analysis for every SNP and transcript level combination in order to find statistically significant correlations. This procedure is usually carried out using repeated linear regression analyses (one for every SNP-transcript pair) with special considerations for covariates and heteroscedasticity of the expression data. However, most of these methods disregard any prior knowledge about how regulation actually occurs within the genome. We believe the goal of improving the eQTL models can be achieved by incorporating the following features into the eQTL models - the semantics of codons (silent mutations vs. nonsense and missense mutations); repeated measures of genotype (genes with the exactly or nearly exactly same SNP genotype); non-linear machine learning methods, such as random forests and artificial neural networks; knowledge about the mechanisms related to regulation, such as *cis*- and *trans*- regulation, binding affinity (e.g. a SNP in a gene coding for a transcription factor might change its binding affinity, a SNP in target DNA binding sequence might also change the binding affinity of the associated transcription factor); proteomics data to identify post-transcriptional interactions (these interactions will also influence protein transcript levels (directly) and mRNA transcript levels (indirectly)); incorporating prior information about SNP importance in regards to regulation (e.g. missense SNPs occurring within known DNA binding domains are more likely to cause changes in expression levels of regulated genes, compared to SNPs occurring outside of those domains).

The developed method would be partially validated by using it to identify genomic targets that are already known for well established transcription factors, such as p53 or UBC.

Division of Work

Downloading and preprocessing of the genotype and phenotype data

Data downloading will be performed by Wenjie Qi and data preprocessing jointly by David Filipovic and Wenjie Qi.

Finding additional datasets, especially datasets of gene annotations

The task of finding additional relevant and useful datasets will be performed by both Wenjie Qi and David Filipovic.

Developing a basic eQTL model and potentially a fast model such as Matrix eQTL detailed by Shabalin (2012)

The task of developing a basic eQTL model will be performed by Wenjie Qi.

Developing a machine learning based nonlinear, multiple eQTL model

The task of developing a machine learning based eQTL model will be performed by David Filipovic.

Modifying the basic models with meta information mentioned in Project Goals.

The task of modifying the basic models will be performed by both Wenjie Qi and David Filipovic, with each implementing at least two modifications for their respective underlying model.

Testing and validating the developed models

Wenji Qi will test and validate the classic eQTL modified model and David Filipovic will test and validate the machine learning eQTL modified model.

Milestones

Downloading and preprocessing of the genotype and phenotype data

We plan on completing the data downloading and preprocessing by March 7th.

Finding additional datasets, especially datasets of gene annotations

We plan on completing the identification of additional useful datasets by March 9th.

Developing a basic eQTL model and potentially a fast model such as Matrix eQTL detailed by Shabalin (2012).

We plan on developing the basic eQTL model by March 31th.

Developing a machine learning based nonlinear, multiple eQTL model.

We plan on developing a machine learning based model by March 31st.

Modifying the basic models with meta information mentioned in Project Goals.

We plan on completing the modifications to the basic models by April 15th.

Testing and validating the developed models

We plan of testing and validating the developed models by April 23nd.

Datasets

Initially we plan on using a well-known microarray yeast dataset, collected by Brem and Krugklyak, 2005. This dataset includes data about 2956 SNPs and microarray expression data for 6111 genes, over 211 samples, in the *S. cerevisiae* yeast model organism. Additionally, we plan on using the typical model eQTL dataset, the cystic fibrosis (CF) dataset from Wright *et al.* (2011). This dataset contains genotype information for 573337 SNPs and the gene expression measurements for 22011 genes over 840 samples, in humans.

Challenges

One of the main challenges of eQTL analysis is the high number of tests that have to be performed, presenting a significant computational problem. Additionally, epistatic effects of gene interaction represent a significant challenge, as described by Huang *et al.* (2013), namely the fact that the effect of a gene on a phenotype (e.g. gene expression) can be modified by one or more other genes.

References

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

2. Huang, Yang, Wuchty, Stefan & Przytycka, Teresa M (2013). eQTL epistasis--challenges and computational approaches. Frontiers in genetics, 4, 51.

3. Shabalin, Andrey A (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics, 28*, 1353-1358.

4. L. F. Thomas and P. Strom. Single nucleotide polymorphisms can create alternative polyadenylation signals and affect gene expression through loss of microrna-regulation. PLOS Computational Biology, 8(8):1–12, 08 2012.

5. R. Xie, A. Quitadamo, J. Cheng, and X. Shi. A predictive model of gene expression using a deep learning framework. In Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on, pages 676–681. IEEE, 2016.

6. Wright, Fred A, Strug, Lisa J, Doshi, Vishal K, Commander, Clayton W, Blackman, Scott M, Sun, Lei, Berthiaume, Yves, Cutler, David, Cojocaru, Andreea, Collaco, J Michael & others (2011). Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13. 2. *Nature genetics, 43*, 539.