Abstract

A major goal of molecular biology is to understand the mechanisms behind the transcriptional regulation and expression of genes. Transcription of a gene is controlled by both the binding of transcription factors, and the state and accessibility of nearby chromatin, which in turn determines the accessibility of DNA to transcription factors and RNA polymerase. The state of chromatin is controlled by a suite of factors, including transcription factor binding and covalent post-translational modifications to histone proteins. Transcription factor binding is sequence-specific, with each transcription factor being thermodynamically predisposed to bind particular sequences of DNA. These sites are generally short (5-15 bases) and degenerate, occurring frequently throughout the genome. Chromatin immunoprecipitation and sequencing (ChIP-seq) is the gold-standard method for the genome-wide measure of the sites bound in a given tissue. Unfortunately, however, ChIP-seq suffers from a major drawback: only one transcription factor (TF) in a single tissue, condition, cell type, or state—which I generalise as "tissue-specific"—may be analysed at a time, at a relatively high cost.

In contrast to ChIP-seq, current *in silico* methods of predicting transcription factor binding sites are extremely cheap and fast. Generally, however, they only consider the sequence specific binding preferences of each TF, and do not consider the tissue-specific accessibility of the DNA region where binding may occur. As a result, *in silico* prediction methods of transcription factor binding suffer from high false positive rates, predicting sites where a TF could possibly bind in any tissue, given a favourable and accessible chromatin state. Given both the large number of TFs and tissues, there is a strong need for tissue-specific *in silico* transcription factor binding site (TFBS) scanning methods. Developing tissue-specific *in silico* TF binding prediction methods is a key focus of this thesis.

Histone modification data provides tissue-specific information on chromatin accessibility and the transcriptional state of nearby genes. In this thesis, I integrate histone modification data with *in silico* methods that predict transcription factor binding. I show that by using this data, the accuracy of predictions can be significantly increased across a wide variety of TFs and in several different tissues. Although histone modification data is obtained via ChIP-seq, it has the advantage that for a given histone modification, the ChIP-seq only needs to be performed once per tissue. I show that integrating widely-available histone modification data with *in silico* TFBS predictions allows more accurate, tissue-specific binding site predictions to be made.

Predicted transcription factor binding sites, while informative of gene regulation, shed little light on the specific action of a transcription factor on gene expression. Gene expression modelling allows direct insight into the roles of individual TFs and histone modifications in controlling gene expression. Previously, the most accurate model of gene expression relied upon ChIP-seq data for each TF included in the model. To date, most models that aim to predict gene expression rely solely on transcription factor data to described gene expression. One notable exception is the study performed by Karlić *et al.* [85], which exclusively used histone modifications instead.

I find that models using seven histone modifications and DNase I hyper-sensitivity data

alone are as accurate as ChIP-seq TF-based models. Furthermore, I show in a log-linear model of expression in mouse embryonic stem cells that using tissue-specific *in silico* predictions of TF predictions can be as accurate as using ChIP-seq TF binding data, with the exception of one indirectly binding TF, E2f1. Combining histone modification and DNase data with ChIP-seq TF data improves model performance even further.

In silico methods to understand gene regulation extend beyond identifying binding sites of individual TFs and their role in modulating expression. Motif enrichment analysis (MEA) is the task of searching for sequence motifs that are statistically over-represented in a set of sequences, such as co-factors to a "ChIPped" TF. Most MEA methods that have previously been published, such as AlignACE [82] and Clover [63], rely on a binary-labelled set of genes. Laboratory technologies such as expression microarrays and chromatin immunoprecipitation on chip (ChIP-chip), however, do not produce binary labels of whether a gene is up- or down-regulated or TF-bound, but instead report a continuous score. Typically, a threshold is applied to this score to discretise the experimental results. If this threshold is too strict, it typically will result in mislabelling positives as negatives; if it is too lax, it will cause the inclusion of false negatives.

In this thesis, I compare a number of different existing MEA methods, and I develop a novel, linear-regression-based MEA method—RAMEN—that does not require the use of thresholds. Some of these methods require a specified threshold, while some attempt to identify an optimum threshold in a data-driven way. I evaluate these methods, along with two existing methods (Clover and PASTAA), using yeast ChIP-chip data. MEA methods using data driven threshold determination can perform poorly unless the range of thresholds is limited *a priori*. In contrast, the novel linear-regression-based method—RAMEN—was the most accurate method tested on our validation set.

This thesis focuses on developing *in silico* tools to better understand gene regulation. It has improved the understanding of how to predict TFBSs in a tissue-specific manner, and has investigated how to use these predictions to more accurately model gene expression. I have shown how combining TF ChIP-seq data and histone modifications can improve the accuracy of gene expression models even further. In addition, several tools for identifying tissue-specific bound genes, tissue-specific TFBSs, and enriched motifs have been published, and may be applied to analyse future experiments.

Keywords

Bioinformatics, Transcriptional Regulation, Gene Regulation, Transcription Factor, Histone Modification, Chromatin, Gene Expression

Australian and New Zealand Standard Research Classifications (ANZSRC)

060404	Epigenetics (incl. Genome Methylation and Epigenomics)	20%
060405	Gene Expression (incl. Microarray and other genome-wide approaches)	30%
060407	Genome Structure and Regulation	40%
080110	Simulation and Modelling	10%