Chapter 37
Prioritize Transcription Factor Binding Sites for Multiple Co-Expressed Gene Sets Based on Lasso Multinomial Regression Models

Hong Hu
University of Illinois – Chicago, USA

Yang Dai
University of Illinois – Chicago, USA

ABSTRACT

Computational prediction of cis-regulatory elements for a set of co-expressed genes based on sequence analysis provides an overwhelming volume of potential transcription factor binding sites. It presents a challenge to prioritize a set of functional transcription factors and their binding sites on the regulatory regions of the target genes that are relevant to the gene expression study. A novel approach based on the use of lasso multinomial regression models is proposed to address this problem. We examine the ability of the lasso models using a time-course microarray data obtained from a comprehensive study of gene expression profiles in skin and mucosal in mouse over all stages of wound healing.

INTRODUCTION

Transcription is a key component in the flow of the genetic information within the living organisms by replicating the information into messenger RNA (mRNA) from a section of the DNA (Levine & Tjian, 2003). Transcription factors (TFs) are a class of proteins that bind to transcription factor binding sites (TFBSs) on DNA and regulate the gene expression (Latchman, 1997) in a context-specific manner. The transcription regulation enables cells to respond to intra- and extra-cellular signals to maintain the cellular activities. Previous studies have estimated that in human genome there are about 20,000~25,000

DOI: 10.4018/978-1-5225-8903-7.ch037

genes, which cover only~5% of the DNA sequences. The rest of the genome associates with non-coding RNA molecules, regulatory DNA sequences, short or long interspersed elements, introns, and sequences currently with unknown functions (Lander et al., 2001). Due to the high complexity of regulatory relationship within living organisms, the identification of TFs and their TFBSs in a context-specific manner is an important but unsolved problem in genomic study (Chai et al., 2014; Collins, Green, Guttmacher, Guyer, & Institute, 2003; W. P. Lee & Tzou, 2009; Spellman et al., 1998).

EXPERIMENTAL IDENTIFICATION OF TRANSCRIPTION FACTOR BINDING SITES

Large scale experimental methods have been proposed for the identification of TFBSs on human genome for a given TF. The recently developed technologies of chromatin-immunoprecipitation followed by massively sequencing (ChIP-seq) or microarray (ChIP-array) allow researchers to capture the genome-wide binding profile of a TF under a given experimental condition (Iyer et al., 2001; Johnson, Mortazavi, Myers, & Wold, 2007; Landt et al., 2012; Robertson et al., 2007). The Encyclopedia of DNA Elements (ENCODE) Consortium, which aims to build a comprehensive list of annotation for functional elements in human genome, has performed a large number of ChIP-seq experiments (Dunham et al., 2012). However, this technology is limited by the availability of the TF-specific antibodies and the high experimental cost. Even with the consortium effort, the ChIP-seq data generated in the ENCODE project only covers a limited number of TFs for a few cell types (Dunham et al., 2012). Although the cost for ChIP-seq study is reducing with the rapid development of the next generation sequencing technology, it is not always possible to use this experimental approach. This is because there are situations where the key set of TFs that regulate the underlying process are unknown *a prior*.

Sequence-Based Computational Identification of Transcription Factor Binding Sites From Co-Expressed Genes

Studies have shown that TFs bind to the DNA sequences in a sequence-specific manner. That is, each TF binds to the specific DNA regions to control the rate of transcription (Latchman, 1997). Therefore, using the prior knowledge of sequence features as well as the TF binding preference, the TF binding sites can be predicted from the sequence information. This in turn provides clue on potential involving TFs.

Measuring temporal gene expression profiles upon certain stimulus to a specific type of cells or tissues have been used to interrogate the transcription regulatory mechanism (Hirose et al., 2008; Y. Liu, Jiang, & Zhang, 2009; Schena, Shalon, Davis, & Brown, 1995; Spellman et al., 1998). Gene expression time-course data capture the dynamics of transcription regulation over time, thus providing valuable information for the inference of the transcriptional regulatory relationship between TFs and their target genes (L. Chen et al., 2010; D'Haeseleer, Liang, & Somogyi, 2000; Ljung, 1987; Wahde & Hertz, 2000; Yeung, Tegner, & Collins, 2002). From the analysis of a temporal gene expression profile, genes can be divided into groups based on their co-expression patterns (Eisen, Spellman, Brown, & Botstein, 1998). Some groups comprise of genes that respond immediately upon the stimulus, representing either early up-regulated or early down-regulated expression changes. These genes are likely to be the direct targets of some key TFs that regulate the cellular activities responding to the stimulation. It is extremely important to know what TFs are involved, what genes are being targeted and where are the binding sites of the TFs