Chapter XXX
Data Integration for Regulatory Gene Module Discovery

Alok Mishra
Imperial College London, UK

Duncan Gillies
Imperial College London, UK

ABSTRACT

This chapter introduces the techniques that have been used to identify the genetic regulatory modules by integrating data from various sources. Data relating to the functioning of individual genes can be drawn from many different and diverse experimental techniques. Each piece of data provides information on a specific aspect of the cell regulation process. The chapter argues that integration of these diverse types of data is essential in order to identify biologically relevant regulatory modules. A concise review of the different integration techniques is presented, together with a critical discussion of their pros and cons. A very large number of research papers have been published on this topic, and the authors hope that this chapter will present the reader with a high-level view of the area, elucidating the research issues and underlining the importance of data integration in modern bioinformatics.

INTRODUCTION

A network of transcription factors regulating transcription factors or other proteins is called a transcriptional regulatory network or gene regulatory network. The understanding and reconstruction of this regulation process at a global level is one of the major challenges for the nascent field of bio-informatics (Schlkopf et al., 2004).

Considerable work has been done by molecular biologists over the last few years in identifying the functions of specific genes. In an ideal world it would be desirable to apply these results in order to build

detailed models of regulation where the precise action of each gene is understood. However, large number of genes and the complexity of the regulation process means that this approach has not been feasible. Research into discovering causal models based on the actions of individual genes has encountered a major difficulty in estimating a large number of parameters from a paucity of experimental data. Fortunately however, biological organisation opens up the possibility of modelling at a less detailed level. In nature, complex functions of living cells are carried out through the concerted activities of many genes and gene products which are organized into co-regulated sets also known as regulatory modules (Segal et al., 2003). Understanding the organization of these sets of genes will provide insights into the cellular response mechanism under various conditions. Recently a considerable volume of data on gene activity, measured using several diverse techniques, has become widely available. By fusing this data using an integrative approach, we can try to unravel the regulation process at a more global level. Although an integrated model could never be as precise as one built from a small number of genes in controlled conditions, such global modelling can provide insights into higher processes where many genes are working together to achieve a task. Various techniques from statistics, machine learning and computer science have been employed by researchers for the analysis and combination of the different types of data in an attempt to identify and understand the function of regulatory modules.

There are two underlying problems resulting from the nature of the available data. Firstly, each of the different data types (microarray, dna-binding, protein-protein interaction and sequence data) provides a partial and noisy picture of the whole process. They need to be integrated in order to obtain an improved and reliable picture of the whole underlying process. Secondly, the amount of data that is available from each of these techniques is severely limited. To learn good models we need lots of data, yet data is only available for few experiments of each type. To alleviate this problem many researchers have taken the path of merging all available datasets before carrying out an analysis. Thus there can be some confusion regarding the term integrative because it has been used to describe both of these two very different approaches to data integration: one among datasets of the same type, for example microarrays, but from different experiments, and the other among different types of data, for example microarray and DNA binding data.

In the rest of the chapter we will describe various techniques proposed to carry out both of these types of integration and will discuss their pros and cons. We will review some of the prominent research following the former approach by Ihmels et al. (2002) and Segal et al. (2005), and work following the latter approach by Bar-Jospeh et al.(2003), Tanay at al. (2004, 2005) and Lemmens et al. (2006).

BACKGROUND

Biological Background

Higher organisms are made up of various different cell types each of which performs a specific role that contributes to its overall functioning. The fascinating fact is that each of these cells contains exactly the same set of genes. The cells of higher organisms, known as eukaryotes, differ from those of the less evolved prokaryotes in having a well-defined nucleus that carries the genetic material. The remarkable diversity among the cells is a result of a precisely controlled mechanism of expression and regulation of a subset of genes in each cell type. The expression of genes into their complements, called m-RNAs or transcripts, is known as transcription while the next step of the process, which leads to creation of a