Chapter 1

Introduction

We are concerned with the use of computational and statistical methods for the analysis of gene expression data. In this chapter, we describe some basic concepts in gene expression and biotechnological methods used to measure gene expression in large-scale experiments. Additionally, we give a brief overview of the main tasks and challenges in the analysis of the resulting data. Finally, we outline the main scientific contributions of this thesis and summarize its contents.

1.1 Gene Expression: Transcription, Translation and Control

First, we briefly review the process of gene expression. A detailed description can be found in many textbooks, see for example [4]. The genetic information of organisms is stored in deoxyribonucleic acid (DNA) molecules. These molecules are composed of two polynucleotide chains (or strands) forming the double helix structure (Figure 1.1). The nucleotides, which are the building blocks of a DNA molecule, are characterized by the base attached to a sugar phosphate. The classical four types are: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). One particularity of the double stranded DNA is the complementary base pairing, i.e., a particular base on a strand only binds to a complementary base on the opposite strand. More precisely, “A” binds only to “T”, and “C” to “G” (Figure 1.1). The reaction in which a single stranded DNA molecule binds to a complementary strand is called hybridization, a reaction exploited by many molecular biology techniques.

In eukaryotes, i.e., organisms which have cellular nucleus, several linear DNA molecules, called chromosomes, are present in the cell nucleus. Each of these chromosomes is formed by billions of base pairs. Genes are regions of the chromosome that code one or more proteins[1]. They represent the basic units responsible for storing and passing on hereditary characteristics.

Gene expression is the process by which the genetic information contained in the genes

[1] Some genes will code functional RNA structures, which are not translated into proteins.

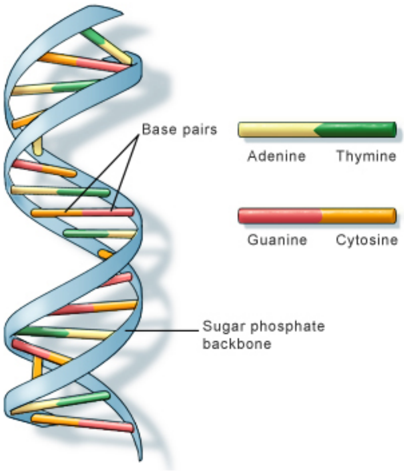

Figure 1.1: *Example of a double stranded DNA molecule. Figure reproduced from the US National Library of Medicine.*

is translated into ribonucleic acid (RNA) molecules and, later, into protein molecules [4]. This process is divided into two main steps: transcription and translation. In the transcription step, regions of DNA, which code genes, are transformed into RNA molecules by the RNA polymerase (Figure 1.2, Step 1). RNA molecules are different from DNA in several aspects: (1) they are only single stranded; (2) they have Uracil (U) instead of Thymine (T); and (3) they have a quicker degradation time.

Next, in the translation process (Figure 1.2, Step 2), RNA molecules leave the nucleus and are "read" by the ribosome in order to synthesize proteins. Triplets of RNA bases are mapped via the genetic code into one of the twenty amino acids. These amino acids are the building blocks of the proteins. Proteins, the final products of genes, are vital to the cell functioning, since they constitute the structural components of the cells and catalyse biochemical reactions.

While most cells of an eukaryotic organism encode the same genetic information, they express genes at distinct levels. The expression of a particular set of genes is either a response to distinct environmental conditions or is part of the specific repertoire of a given cell type. Understanding the mechanisms controlling gene expression is a central question in molecular biology. This control can happen at several levels of the gene expression process. The first level and the one of main concern in this work is the transcriptional control. At this level, proteins, called transcription factors, bind the upstream (or regulatory) regions of genes. These factors act as initiators (or repressors) of transcription by facilitating (or blocking) the access of the RNA polymerase to initiate transcription.

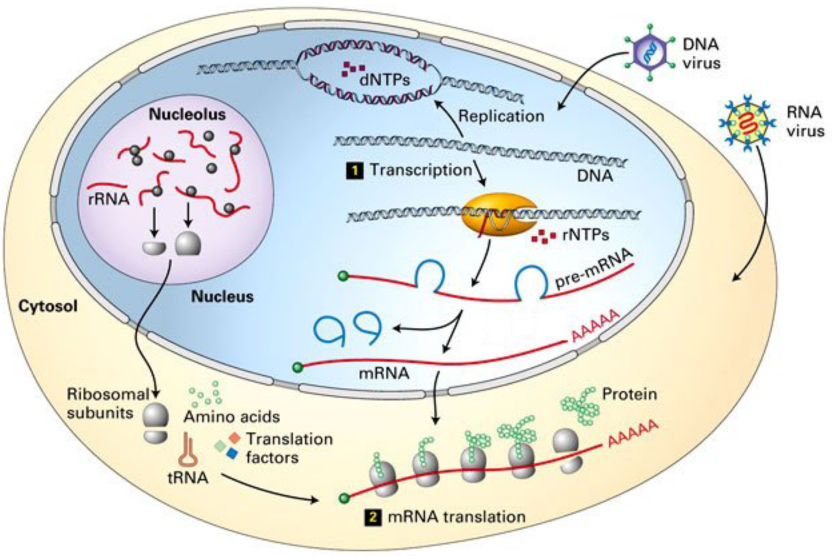

Figure 1.2: *We depict here the main stages of gene expression. Step 1 corresponds to the transcription of DNA to RNA molecules. Step 2 corresponds to the translation of messenger RNA (mRNA) to protein molecules. Figure reproduced from [136].*

1.2 Measuring Gene Expression with Microarrays

Microarray technology allows the simultaneous measurement of the concentrations of RNA molecules (or transcripts). More precisely, this technology allows the measurements of the expression patterns of genes—also known as expression profiles. For example, by comparing the expression profiles of disease and normal cells [5], responses of cells to environmental conditions [82], or during biological process such as cell cycle [201] and development [214], the researchers can explore the dynamics of gene expression, to form hypotheses about regulatory and functional roles of genes, and to obtain molecular signatures of cell types and all this on a genome-wide scale.

Microarray Technology. The main idea behind DNA microarrays is to exploit the fact that two complementary single stranded DNA molecules hybridize [111]. For each gene of interest, a short sequence complementary to its sequence is select. These sequences are called probes and have lengths ranging from 20 to 60 bases. The probes should be selected in such a way that there is a low chance of hybridizing with sequence others than the target gene sequence.

Then, with the aid of robotics or nano manufacture technologies, thousands of copies of a particular probe are placed in a tiny area of a hard surface — the array. Thousands of such probe spots can be placed side by side forming a grid on the array. Each spot contains probes designed to hybridize with RNA from a specific gene. In the end, one can have as many as 10^5 spots arranged in a 2×2 cm array.

In the next step, the RNA molecules of the cell population of interest are separated and transcribed to single stranded complementary DNA (cDNA) molecules. This step is needed as RNA molecules are unstable and would quickly degrade. Afterwards, the cDNA molecules are marked with fluorescent or radioactive labels. The cDNA molecules are, then, poured onto the slide. After some time, the slide is washed, removing the cDNA molecules that did not hybridize with the probes.

Next, the slide is scanned, resulting in an image with all the spots intensities (see Figure 1.3 middle). Such an image is further processed using computational methods. The aim is to calculate the intensity at each spot, which is proportional to the number of transcripts of a gene that a probe is complementary to (the whole process is illustrated in Figure 1.3).

There are several distinct microarrays technologies such as cDNA microarrays [183] and Affymetrix Gene Chips (also known as Oligonucleotide arrays) [134]. They differ mostly by how the chips are manufactured and on methodologies for probe selection. The particular characteristics of such technologies are important for decisions concerning experimental design, experimental costs, measurements reliability and data pre-processing aspects. See for example [119] for a complete description of microarray technologies.

One important aspect of microarrays is the use (or not) of reference RNA samples. In double channel microarrays, such as cDNAs microarrays, two cell samples are poured in the same microarray: cells of interest (e.g., disease cells, treated cells) and reference cells (e.g., healthy cells, untreated cells). Each of these cell populations is dyed with a distinct marker, for instance, a red Cy3 dye versus a green Cy5 dye. Double channel microarrays return a relative quantification of the RNA expression in relation to the reference cell, usually measured by taking the logarithm of the ratio between the red and green signals (see Figure 1.3 (a) for an example of a two-channel microarray).

In single channel arrays, only one RNA sample is poured in the array, and no reference sample is used. Single channel arrays return estimates on the number of copies of a particular transcript in a given sample. With Affymetrix microarrays, an example of single channel array, 20 to 40 distinct probes, which are complementary to the sequence of an unique gene, are placed in distinct spots on the array in order to obtain reliable estimates of RNA quantities. Additionally, a mismatch spot (MM) containing a sequence, where a base in the middle of the original probe (PM) sequence is exchanged, is placed next to each PM spot. These reduce the effect of cross-hybridization, increase the signal to noise ratio and improve the accuracy of the RNA quantification (see Figure 1.3 (b) for an example of an single-channel microarray).

Pre-processing and normalization procedures are the initial computational tasks in the analysis of data from microarrays. These procedures are responsible for improving the estimates of the RNA levels measured by microarrays. In the pre-processing step, one tries to correct the probe intensities for errors introduced by experimental artifacts, such as non-specific hybridization, dye efficiency, spatial biases, and so on. See for example [106]

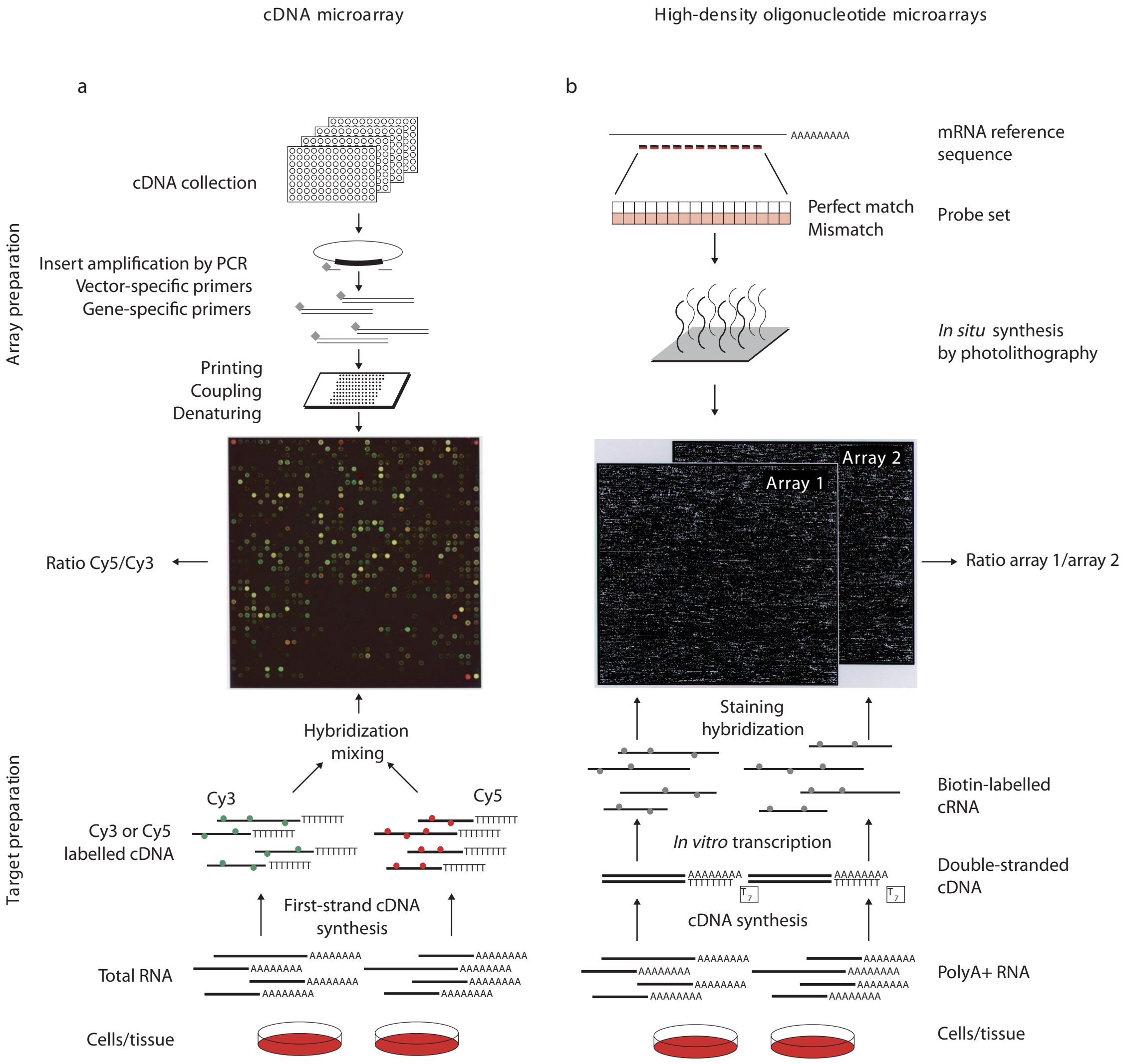

Figure 1.3: *We depict how microarray experiments are performed for cDNA (a) and Oligonucleotide (b) microarrays. In the top, we depict how microarrays are manufactured; and in the bottom, how RNA samples are obtained. In the middle, we can see the images obtained after RNA samples hybridize to the microarrays. For cDNA microarrays (a), each dot represents a probe, and the red (or green) colors are proportional to the counts of RNA hybridized to that probe in the reference (or control) sample. Similarly, the intensity of white dots in Oligonucleotide arrays (b) represents the counts of RNA hybridized to that probe. Figure reproduced from [190].*

for a review of methods on Affymetrix Gene Chips or [230] for protocols for cDNA microarrays. Next, normalization methods are applied with the aim of making expression values obtained in distinct hybridization experiments comparable. See [203] for a review of pre-processing and normalization methods.

Computational Challenges in the Analysis of Gene Expression data. Large-scale data produced by microarrays experiments shows how gene expression changes in distinct biological conditions and tissue types. Manual analysis of such amount of data is not feasible. Due to this limitation, statistical and computational methods are vital for analyzing gene expression data. In fact, data arising from microarrays have several particularities that should be taken into consideration by these methods: they should be able to cope with the high dimensionality of the data, be robust to noise and take advantage of the experimental design associated with the biological experiment.

For example, gene expression levels are often measured in few experimental conditions, i.e., tissues types or time points (less the 100) for thousands of genes (more the 10,000). Furthermore, despite improvements of microarrays experiments and protocols, these technologies still suffer from several sources of noise: either by manufacturing failures, problems in the reading procedure, unspecific probes, variability in biological samples, or variations in the environment conditions in which experiments are performed. A recent study [107] showed that at least 10% of expression measurements differ significantly in replication experiments.

Another important aspect is the experimental design procedure used for acquiring the data. For instance, in microarrays experiments measured over time, such as during cell cycle, the cell populations tend to desynchronize with time. This results in deterioration of the expression measurements of later time points [201]. The explicit use of knowledge of the biological process makes computational and statistical methods more robust to this type of inherent noise.

1.3 Thesis Overview

In this thesis, the main focus is on the problem of finding groups of co-expressed genes, or genes that display the same expression behavior through particular biological conditions, such as cell cycle, or developmental processes. The basic rational underlying this approach is the assumption that co-expressed genes should (1) perform a similar functional task, and (2) be regulated by the same transcription regulation program. Thus, exploiting the guilty by association principle, one can deduce the function of an uncharacterized gene by observing the function of co-expressed genes [71]. Also, by including additional data in the analysis, such as regulatory regions, one can explore and uncover regulatory programs controlling the expression of genes [212].

One traditional approach for finding co-expressed genes is the use of clustering methods, also known as unsupervised learning [64]. Clustering methods are usually based on a similarity metric, which defines how close objects (or gene profiles) are in a given multidimensional space, followed by a method that, for example, searches for groups (or clusters) of objects that lie in compact regions and are far apart from other groups. While cluster analysis is a well-developed research area [109], the characteristics of gene expression data impose challenges not previously addressed by classical clustering methods.

This thesis uses mixture models as a statistical formalism for performing clustering of gene expression data [145]. Mixture models are robust to noise, can model uncertainty about cluster assignments, allow the inclusion of prior knowledge, such as intrinsic dependencies of the experimental design, and offer a flexible framework for integration of additional biological data.

In Chapter 2, we introduce the mixture model formalism and the method used for estimating mixture models; the expectation-maximization (EM) algorithm. Then, in Chapter 3, we describe how mixture models can be used to solve the clustering problem, and how questions as choosing the number of clusters and cluster validation can be answered in the context of mixture models. Additionally, in Chapter 3 we propose a novel external index for validating clustering computed by mixtures. With the exception of the proposal of this external index, Chapters 2 and 3 basically review established research on mixture models, and introduce the methodological framework used in the bioinformatic applications described in later chapters.

Mixture models allow, with a proper choice of component models, to make explicit assumptions about the data. This thesis proposes two novel types of components models for analyzing gene expression profiles. The use of hidden Markov models with linear topologies to analyze gene expression time courses will be the focus of Chapter 4. In Chapter 5, we propose a new type of probabilistic model, dependence trees, to model gene expression profiles during a developmental process. This approach assumes that the sequence of changes from a stem cell to a particular mature cell, as described by a developmental tree, are the most important in modeling gene expression from developmental processes. We also explore in Chapter 5 the benefits of using priors of model parameters to obtain maximum-a-posteriori point estimates, and how this improves the robustness of the method.

Once a given component model is defined, it is straightforward to apply any extension of the expectation-maximization (EM) algorithm. We propose, in Chapter 6, the use of an established semi-supervised learning method [123] to integrate additional biological data and improve clusterings of gene expression time-courses. We evaluated the inclusion of Gene Ontology annotations [9] and location analysis of transcription factor biding derived from Chip-on-chip experiments [128]. Additionally, we propose a novel method, which combines gene expression time-courses with location of gene expression in Drosophila embryos [214], for finding groups of syn-expressed genes. Finally, in Chapter 7, we present final remarks and future work with respect to the specific contribution of this thesis.