CHAPTER 1

Introduction

1.1 Motivation

Gene Regulation and Transcription Factor Binding Sites

Every living organism is composed of multiple different cells. These cells contain genetic material encoded in the form of DNA molecules, also known as genome. The genome can be represented as a categorical vector $\mathbf{g} = \langle g_1, \cdots, g_n \rangle$, where $g_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ represents the nucleotide at genomic position i. Certain substrings of the genome $\mathbf{g}$, denoted as $\mathbf{g}[u..v]$, from genomic positions u to v for $u < v \leq n$, represent genes. Genes can be read by specialized proteins to produce other proteins. This protein-producing cycle is the key mechanism for maintenance of life.

A couple of years ago, it was believed that, in possession of the complete genomic sequence $\mathbf{g}$ for a given organism, it would be possible to exactly determine its phenotype and disease susceptibility. However, after the analysis of the first genomes, it was clear that the simple determination of an organism's DNA nucleotide sequence is not enough to explain the great diversity of biological processes. Such processes are governed by a complex chain of events called "gene regulation". Gene regulation includes a wide range of mechanisms that happen inside a cell in which genes are turned "on" (i.e. they are expressed) and "off" (i.e. they are not expressed) dynamically. Depending on which genes are "on" or "off", the cell specializes in different functionalities (Alberts et al., 2007).

In the so-called post-genomic era, attention is turning to the understanding of how protein-coding genes (about 25,000 in humans) and their products are regulated (Maston et al., 2006). These regulatory mechanisms drive the correct execution of biological processes and require a set of carefully orchestrated steps that depends on the correct spatial and temporal expression of genes (Maston et al., 2006). Therefore, the deregulation of gene expression is often linked to diseases (ENCODE Project Consortium, 2012).

not sure exactly what deregulation here is, maybe write it something like: errors in regulation?

To understand the molecular mechanisms that dictate the cell's expression patterns, it is important to identify the regulatory elements involved in these activities. One of the most important regulatory features are transcription factors (TFs) – proteins that bind on the DNA enhancing or repressing the expression of genes. These proteins bind to particular genomic regions called transcription factor binding sites (TFBSs) (Maston et al., 2006). TFBSs may be active if they are currently being bound by a TF or inactive, ~~if they are not currently being bound by a TF.~~

I would just write: or inactive otherwise.

Importance of the Identification of All Active TFBSs of a Cell

The identification of all active TFBSs of a cell is a very important task, since they are the key players on regulatory mechanisms. By identifying active TFBSs we can develop regulatory networks, which encode the interplay between different genes to control specific cell functions. Such a task leads to the understanding of cellular mechanisms and the particular derregulatory steps which leads to disease.

it's not the task, rather it's the successful identification that leads to an understanding

There are a great number of successful experimental studies that benefited from the proper identification of active TFBSs. For instance, studies were able to: (1) unravel cellular mechanisms Lin et al. (2015); Tsankov et al. (2015); (2) unravel disease mechanisms Schaub et al. (2012); Vernot et al.

(2012); Charos et al. (2012); (3) understand the function of different regions in the genome Yip et al. (2012); Whitfield et al. (2012); Natarajan et al. (2012) and (4) understand other cellular regulatory elements such as long noncoding RNAs Tilgner et al. (2012); Bánfai et al. (2012).

In summary, the identification of active transcription factor binding sites is important because of its broad impact on many other cellular processes. Given the importance of the proper identification of cell-specific active TFBSs, our research focuses on performing such a task by applying computational methods to biological experimental data.

Computational Detection of Active TFBSs Must Consider the Chromatin Dynamics

Historically, the first computational approach to identify TFBSs was based solely on the DNA sequence (Stormo, 2000). Each TF has a particular DNA sequence affinity, i.e. they tend to bind to specific DNA sequences. The computational sequence-based methods search the genome $\mathbf{g}$ for DNA substrings $\mathbf{g}[u..v]$ that correspond to the affinity sequence of target TFs. However, although computational sequence-based methods are able to detect TFBSs, they are not able to tell whether these sites are active or inactive (Pique-Regi et al., 2011). This happens because such computational approach does not consider the fact that only a few regions in the genome are accessible for TFs to bind. These regions are called "open chromatin regions". The number of open chromatin regions and their location vary between different cell types and ultimately dictates which genes are accessible and being expressed (ENCODE Project Consortium, 2012).

Recent advances in biological techniques (Shendure and Ji, 2008) have enabled the creation of experimental methods to identify these open chromatin regions (ENCODE Project Consortium, 2012). We will explore two of these so-called "open chromatin next-generation sequencing (NGS) techniques": the chromatin immunoprecipitation followed by NGS – termed ChIP-seq (Johnson et al., 2007); and the DNase I cleavage followed by NGS – termed DNase-seq (Crawford et al., 2004; Sabo et al., 2004b). These techniques generate signals which span the entire genome and indicates open chromatin regions. These signals can be viewed as a numeric vector $\mathbf{x} = \langle x_1, \cdots, x_n \rangle$ where high $x_i \in \mathbb{N}^0$ indicates open chromatin regions. Moreover, certain patterns in the signals generated by DNase-seq and ChIP-seq are indicative of active TFBSs. Therefore, we can apply computational methods to process the DNase-seq and ChIP-seq signals and to identify these patterns. By doing so, we can detect active TFBSs considering the open chromatin information.

what is signal here, and I don't mean the numeric vector below?

Computational Detection of Active TFBSs Using DNase-seq and ChIP-seq

The DNA is found wrapped in proteins called histones. There are a number of post-translational modifications on these histones which are indicative of open chromatin regions, such as the ~~so-called~~ H3K4me1 and H3K4me3. By performing a histone modification ChIP-seq experiment we are able to identify cell-specific open chromatin regions. Furthermore, the DNase-seq data also provides a robust map of open chromatin regions with a very high spatial resolution. By combining these two experimental data, we observe very characteristic patterns indicating the active binding of TFs in the genome (see Figure 1.1). This pattern is commonly referred to as TF "footprints". A TF footprint is defined as a region likely to be associated to an active TFBS (Boyle et al., 2011; Gusmão et al., 2012).

The experiments presented in this thesis focus on the computational treatment of DNase-seq and histone modification ChIP-seq data to perform computational predictions of active transcription binding sites. Such prediction is performed by searching the distinctive patterns that the DNase-seq and histone modification ChIP-seq signals exhibit around active TFBSs. These distinctive patterns are termed footprints. We use the traditional term "computational footprinting methods" for computational methods that searches for footprints using open chromatin data, such as DNase-seq and histone modification ChIP-seq. The computational footprinting framework presented in this thesis can be used in multiple different biological experiments to understand the regulation of genes.

redefinition / same def. for footprint again? it's already there in the previous paragraph

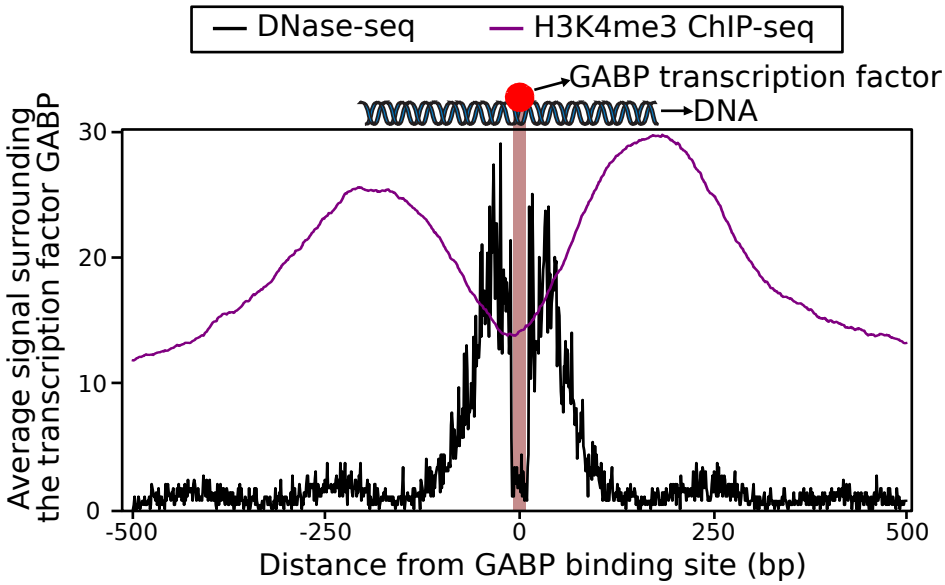

Figure 1.1: Distinctive pattern (footprint) of DNase-seq and ChIP-seq on active TFBSs. Average DNase-seq and histone modification H3K4me3 ChIP-seq signals surrounding the known (biologically verified) TF GABP active binding sites. Active TFBSs happen at depletions between two peaks of the DNase-seq signal (marked in red). Furthermore, these DNase-seq peaks, which determines an open chromatin region, happen at the depletion between two peaks of active histone modification marks. This distinctive pattern of signal depletion between two peaks is called a TF footprint. *Source: Gusmão et al. (2012)* (modified to fit thesis format and/or clarify key points).

1.2 Thesis Overview

In this thesis, we: (1) present a novel computational framework that uses DNase-seq and histone modification ChIP-seq data to detect active TFBSs, (2) evaluate the predictions made by our method using experimentally verified active TFBSs and (3) use our predictions in real biological scenarios to make inferences about the regulatory circuitry of particular cells. ~~The~~ Figure 1.2 presents an overview of this thesis. In the following paragraphs we describe the Figure 1.2 in more detail.

Computational Footprinting Framework

For a particular cell type A we obtain DNase-seq and histone modification ChIP-seq data available in repositories such as the ENCODE Project Consortium (2012) (Figure 1.2a). We process these data using computational methods (Figure 1.2b) to generate a normalized DNase-seq signal $\mathbf{x}_A$ and normalized histone modification ChIP-seq signal $\mathbf{y}_A$ (see Figure 1.2c). Then, we apply a computational footprinting method Θ on $\mathbf{x}_A$ and $\mathbf{y}_A$ (Figure 1.2d) generating a set of genomic regions (intervals) $S_A = \{s_{A,1}, \cdots, s_{A,m}\}$, where each genomic interval $s_{A,i} = [u, v]$ represent a predicted footprint, which is likely to be associated ~~to~~ with an active TFBS (Figure 1.2e).

Evaluation of Predicted Footprints

The predicted footprints S_A are compared to experimentally verified active TFBSs (Figure 1.2e) to create statistics which evaluates how close our predictions are to the true active TFBSs (Figure 1.2f-g). These evaluation statistics (Figure 1.2g) are also used to compare our computational footprinting framework to competing methodologies.

Application to Real Biological Scenarios

Furthermore, the predicted footprints S_A are used, in combination with downstream computational methods (Figure 1.2i) to generate real-scenario biological knowledge about the regulatory circuitry of the cell type A (Figure 1.2j).

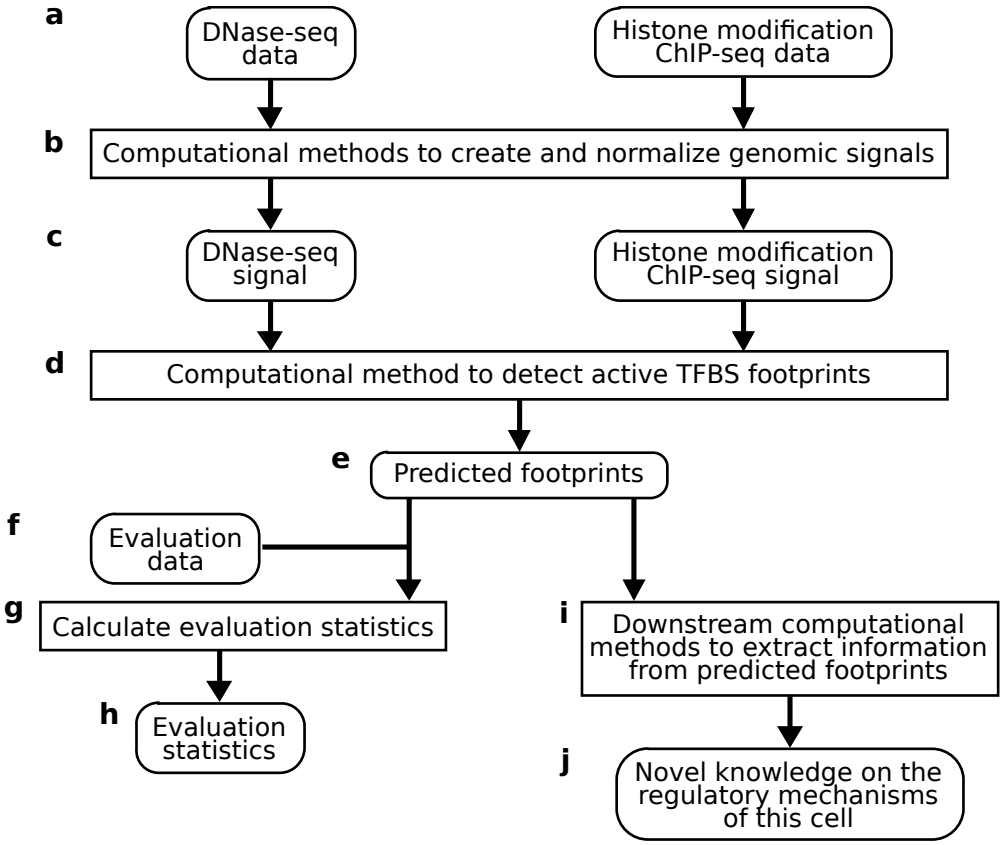

Figure 1.2: Thesis overview. This figure depicts the proposed thesis' workflow. Boxes with round-shaped edges represent data and square-shaped edges represent computational methods.

1.3 Contributions

The main contribution of this work is the development of a novel computational framework to treat data generated with the DNase-seq and ChIP-seq technologies and detect footprints (i.e. putative active TFBSs) based on these data. Our contributions are summarized as follows.

- **Novel signal treatment strategy:** Novel DNase-seq and histone modification ChIP-seq signal treatment approaches were developed and formalized. Such treatment framework has shown to be robust and applicable to a wide range of different datasets.
- **DNase-seq experimental bias correction:** We created an approach to correct for known artifacts on DNase-seq data. Our experiments have shown the efficiency of such correction on bias mitigation.
- **Novel computational footprinting method:** We devised a novel computational footprinting method based on hidden Markov models (HMMs). It was shown to provide robust active TFBS predictions on the basis of an extensive evaluation process.

- **Novel evaluation approach of computational footprinting methods:** Until now, computational footprinting methods have been evaluated using the "TF ChIP-seq approach". However, biases were pointed in such evaluation scheme (Yardımcı et al., 2014). Therefore, we develop a novel computational footprinting method evaluation approach based on gene expression.
- **Comprehensive computational footprinting method comparison:** We performed a comprehensive comparison including: (1) our novel HMM-based approach; (2) nine state-of-the-art computational footprinting methods and (3) four baseline approaches. Our comparative experiment is the most complete so far, with a total of 14 computational footprinting methods and 233 TFs evaluated. most complete with regarding to what? in literature?
- **Analysis of relevant features on computational footprinting:** A number of empirical analyses were performed. These analyses evaluated relevant features for the computational prediction of active TFBSs such as: method's parameter selection, experimental bias correction, optimal footprint scoring strategy and TF binding residence time.
- **Case studies:** We successfully applied our computational footprinting method in two different studies to identify regulatory elements involved in specific biological conditions.

1.4 Document Structure

In Chapter 2 we introduce all the concepts needed for the understanding of our work. We define the current challenges on computational identification of active TFBSs and provide a comprehensive literature review on computational footprinting methods.

In Chapter 3 we formalize our approach to address the detection of active TFBSs. We describe the treatment of the input DNase-seq and ChIP-seq data and the novel approach to detect active TFBSs based on HMMs. Furthermore, in Chapter 4 we describe the full experiment design of this project. We present: the data used in our work, the execution of our computational footprinting approach and the method evaluation strategies.

In Chapter 5 we present the results of our experiments, which encompasses: the analyses on relevant computational footprinting features, a comprehensive comparison of computational footprinting methods and case studies in which our methodology was successfully applied to real biological scenarios. In Chapter 6 we discuss all results presented in this thesis, highlighting all the key findings. Furthermore, we discuss future research opportunities. Further supplementary information and results can be found in the Appendix A.

in the places where you're listing several items, e.g.:

== We present: the data used in our work, the execution of our computational footprinting approach and the method evaluation strategies.

I would mostly write whole sentences, unless the list is very long, like:

==> We present the data used in our work, as well as the execution of our computational footprinting approach and the method evaluation strategies.

it reads a bit nicer :-)