CHAPTER 1

Motivation

The elucidation of transcriptional regulatory networks is crucial for understanding how cells integrate internal as well as external signals, eventually controlling processes like progression through the cell cycle, appropriate response to cellular stress or differentiation of stem cells into adult tissues. For the regulation of many of these processes transcription factor – DNA interactions play a fundamental role. Oftentimes a single transcription factor (TF) is able to bind to the promoters of an entire cohort of genes thereby controlling their expression and ultimately the corresponding cellular process.

For individual genes, TF binding to DNA has been studied experimentally for a long time using a variety of techniques, such as DNAse footprinting and gel-shift assays. Recently, functional genomics technology has opened up the way towards unraveling protein – DNA interactions also on a global scale. In particular, ChIP-chip (Harbison et al., 2004) and protein binding microarrays (PBM, Mukherjee et al., 2004) allow to simultaneously assess the binding strength of a given transcription factor to a large number of diverse DNA sequences.

While such large scale binding data is of great value for discovering regulatory networks the obtained data is oftentimes noisy, containing a considerable fraction of false positive and negative binding predictions. For the validation of experimental findings and in cases where experimental data is not available, theoretical approaches have to be employed. To this end, statistical methods are usually invoked that compute a similarity scores between the known DNA binding motif of a TF and all the individual sites in a given sequence. Such approaches subsequently define a score cutoff, which separates the sites into binding sites and non-binding sites for the TF (Rahmann et al., 2003). An alternative to such techniques is provided by the work of Berg and von Hippel (1987), which allows to compute the binding energy of a TF to a given DNA site. Such energy values can then be converted into continuous binding probabilities. This latter approach pertains presumably much better to the continuous values of TF binding affinity identified in genome wide ChIP-chip and PBM experiments as compared to the statistical hit-based methods, which cement a binary separation between binding sites and non-binding sites. However, available bioinformatics tools for predicting TF target

genes either utilize cutoff based techniques for identifying individual binding sites or apply greatly simplified models to convert predicted binding energies into binding probabilities.

Once binding sites have been discovered either experimentally or based on computational predictions the question arises whether these sites convey a regulatory activity of the corresponding TF. As ChIP-chip and PBM studies suggest, a considerable number of predicted binding sites in the genome might not be functional because the factor is precluded from binding e.g. due to chromatin modifications. Additionally, many bound sites still do not convey regulation because, for instance, the distance to the core promoter is such that the bound factor cannot contact the transcriptional machinery. The functionality of a binding site thus needs to be investigated using for instance reporter gene constructs in conjunction with promoter bashing techniques.

Pertaining to the search for functional TF targets the focus of this thesis is twofold. First, this work aims at deriving a biophysical model that allows to predict the binding strength of TFs measured in large scale experiments without invoking an artificial separation between binding sites and unbound sites. Secondly, by embedding this model into a statistical framework and combining sequence data with gene expression data the approach is used to identify TFs regulating groups of genes and to locate functional high affinity regions within promoters of groups of tissue specific genes from human and mouse. The success and superiority of the model over hit-based as well as other biophysically motivated methods is thereby demonstrated.

1.1 Thesis outline

Chapter 2 will describe the principles of transcriptional gene regulation in eukaryotes and will illustrate several key technologies used for measuring gene expression and TF binding to DNA.

Chapter 3 will discuss bioinformatics approaches frequently used to predict TF binding to DNA. To this end first the most common statistical techniques are outlined, which predict discrete binding sites for TFs. Then the alternative, a biophysical model for predicting binding energies will be explained. Several models exist that convert such predicted binding energies into actual binding probabilities. The most common of these models,

which assumes that the binding probability follows a simple Boltzmann distribution, will be outlined. This model also forms the basis of a common assumption that the statistical and biophysical approaches are equivalent. I will show under what special circumstances this assumption actually holds.

Chapter 4 will show the derivation of the biophysical binding model used in this thesis for predicting the binding strength of a given TF to a DNA sequence. This model contains two parameters that need to be calibrated in order to calculate meaningful binding probabilities. By optimizing the correlation between the model predictions and experimental binding data from yeast a general prescription of how to set these parameters also in the absence of experimental binding data will be derived. The calibration process thereby also reveals fundamental aspects about the biophysical mechanisms guiding TF-DNA binding. The resulting TRAP model for predicting binding probabilities will be shown to outperform classical hit-based methods as well as the models assuming a simplified Boltzmann distribution for binding energy distribution.

Chapter 5 demonstrates the applicability of TRAP to higher eukaryotes by evaluating the binding predictions made for several experimentally well characterized genes and TFs.

In Chapter 6 TRAP is embedded into a statistical framework that allows to predict TFs that regulate groups of genes. The resulting method referred to as PASTAA will first be validated on yeast binding data before it is applied to human and mouse promoters for a detailed analysis of tissue specific genes. Using *in silico* promoter bashing revealed that tissue specific binding signals tend to reside in proximal promoters, preferentially upstream of the TSS, and showed that tissue specific genes often possess a TATA box. It further shows that TFs are preferentially expressed in the tissue whose expressed genes have the corresponding binding signals enriched in their promoters. Results obtained from PASTAA are robust against sequence noise and choice of expression data. Furthermore, PASTAA performs favorably in comparison to a number of state of the art alternative approaches.

Finally, Chapter 7 will feature a broader discussion about the TRAP and PASTAA approaches and outlines future perspectives.

1.2 Publications

Much of the work presented in Chapter 4 has been published in the paper: “Predicting transcription factor affinities to DNA from a biophysical model”, Bioinformatics 2007 and in the chapter “Sequence Annotation” from the book “Modern Genome Annotation: The Biosapiens Network”. The work presented in Chapter 6 is currently submitted in two papers titled: “PASTAA: identifying transcription factors associated with sets of coregulated genes” and “Tissue specific transcription factor binding signals cluster in proximal promoters”.

1.3 Acknowledgements

Foremost I want to thank Martin Vingron, not only for giving me the opportunity to work in his group and funding my research but also for many great ideas and discussions regarding both the TRAP and PASTAA projects. I would like to thank Thomas Manke and Aditi Kanhere for their immense help with writing the TRAP manuscript and Thomas also for his help with the PASTAA paper. The work, particularly on tissue specific genes, would not have been possible without the great support by Stefan Haas and his assistance in manipulating the vertebrate sequences and EST data. In addition, I am very grateful for his help with writing the PASTAA manuscript as well as proofreading this thesis and providing very many helpful comments. Thanks go also to Szymon Kielbasa for supplying most of the standard TF binding site annotations used in this study. Special merits go to Sean O’Keeffe for his help with implementing the PASTAA webpage and to Ho-Ryun Chung for many valuable discussions regarding the biophysical model. Finally, I want to thank all people of the department of Computational Molecular Biology and especially Birgit Löhmer for their help and support throughout the years.