1 Motivation

Deciphering gene regulation networks is a major challenge of research in molecular biology and bioinformatics. Our present understanding of gene regulation is largely derived from studies of single gene promoters. Recently, large-scale approaches like shotgun sequencing and gene expression monitoring of whole genomes started to add valuable data on the systemic level. Luckily, bioinformatics supplies a handle to analyze this increasing amount of biological data and, more importantly, to broaden our knowledge on gene regulation. However, traditional machine learning approaches perform poorly on detecting promoter regions and regulatory elements therein. This is mainly because regulatory elements are tiny and many have a low information content. Thus, the corresponding sequence patterns often occur by chance alone. The signal-to-noise ratio in pinpointing active regulatory elements can be improved by changing the search strategy. Considering cross-species sequence conservation, searching for combinations of binding site motifs and integrating experimental evidence cuts down on the false positives, which hamper the usefulness of single sequence approaches. This thesis work concentrates on detecting and describing proximal regulatory elements. In this context, the comparison of upstream regions of orthologous genes is particularly valuable. This approach is commonly referred to as *phylogenetic footprinting*. This new term was coined from the experimental procedure of *DNA footprinting* where DNA – protein complexes are exposed to nucleases ("DNA eating enzymes"). DNA sequence that is not covered by a bound protein is degraded. DNA sites that interact with proteins are left over and can be mapped onto the original sequence. By analogy, *phylogenetic footprinting* across multiple species should unveil local sequence similarities that are likely to originate from functional regulatory elements. These elements stand out of the background of neutrally diverging sequence as they are under selective pressure.

Concomitantly, we would like to see significant associations of gene groups defined by co-expression or co-function to the conserved predicted regulatory elements. If gene regulation is substantially determined by proximal promoter elements, we would undoubtly discover significant associations in the form of preferential enrichment or depletion of specific binding site populations in preset gene groups.

It is the scope of this thesis to provide an extensive framework for promoter detection and analysis. The benefits of having such a resource are presented in the form of detailed studies on biological questions.

1.1 Thesis structure

This thesis starts with a general introduction to genome organization and gene regulation in eukaryotes (Chapter 2). Attention is especially drawn to events at the promoter level and conservation of components therein (see Section 2.2 and 2.3).

What follows is a more formal description of comparative sequence analysis in Chapter 3. In Section 3.1, the concept of pairwise sequence alignment is presented, which occupies a central position in this thesis. Different alignment types and scoring schemes are reviewed and our method of choice (suboptimal local alignments) is explained in greater detail. A related issue is to assess the statistical significance of an alignment result. Section 3.2 provides a comprehensive insight into our random model of alignment scores and the corresponding significance computation. An extension to multiple alignments is discussed in Section 3.3.

Chapter 4 presents the software components and work-flows to build the infrastructure of CORG, our promoter annotation framework. Firstly, we motivate our definition of an upstream region encompassing the start of transcription (Section 4.1). Secondly, in Section 4.2 we show ways to elucidate the phylogenetic relationships of the corresponding genes, and discuss pros and cons of the footprinting approach. We cover our approach of detecting and annotating local sequence similarities in multiple species in Section 4.3. The annotation step brings in experimental evidence as diverse as ESTs, binding site representations and mapped start sites of transcription. Design issues with respect to database structure and user interface are subsequently presented in Section 4.4 and 4.5.

Two example applications of our system are shown in chapter 5. In Section 5.1, a large-scale study on binding site distributions in upstream regions of co-expressed genes, unravels putative regulators of cell cycle progression. Furthermore, a detailed study on predicting conserved binding sites with subsequent experimental evaluation stresses the quality of the CORG system (Section 5.2). Putative binding sites of the Serum Response Factor were evaluated in a collaborative experimental effort.

This thesis closes with a summary on the progress that has been made and an outlook on forthcoming improvements.

1.2 Acknowledgments

I exactly remember the day when I was traveling from York to Berlin to apply for a PhD student position in the newly formed Department of Computational Molecular Biology. The city of Berlin did not receive me well since it was incredible cold and snowy. Nevertheless, I felt most welcome when I was meeting up with the people at the CMB. Luckily, I got the job.

First of all, I am most grateful to my supervisor, Martin Vingron, who offered me the opportunity to work in the exciting field of Comparative Genomics. I didn't want to miss his constant support and broad experience. Secondly, I deeply appreciated all enjoyable discussion with past and present members of the Gene regulation group: Special thanks go to Brian Cusack for getting the comparative business started and proofreading the manuscript. I also cherish the help of Haiyan Wang in designing the CORG web-site. Thanks and credits are given to Steffen Grossmann, Holger Klein, Thomas Manke and Stefan Röpcke. All of them contributed to the success of my thesis work by theoretical input and/or applied measures.

I also owe a lot to Sven Rahmann who helped me out on software issues and illuminated the world of statistics. Many thanks go to Peter Arndt for discussing DNA scoring schemes, Stefan Haas for his help on making use of the GeneNest database and Antje Krause for computing gene phylogenies.

Nothing could have been attained without our systems administrator Wilhelm Rüsing. He relentlessly cracked down on problems with the infrastructure and thus kept the project going.

Additionally, I am thankful to Philip Verdemato a good friend and critical proofreader.

Above all, I deeply admire the everlasting patience and love of my wife Silke. I couldn't have endured these three years without her.