6 Conclusion

In this thesis we have established a framework for comparative analysis and annotation of vertebrate promoters. Promoters are key players in gene regulation. They receive signals from various sources and control the level of transcription initiation, which largely determines gene expression. We have given an introduction to the molecular biology of gene regulation. From that it became apparent that traditional approaches of pattern finding and promoter prediction are error-prone in vertebrate genomes. The search space (intergenic regions) is extremely large and thus the number of false predictions is prohibitively high. We have shown how a comparative approach can mitigate this problem. This has been presented as case studies where prior knowledge on promoter location and components was available. For automated annotation of promoter regions, the database of Comparative Regulatory Genomics and its software components integrate information of two kinds:

Cross-species comparison. Cross-species conservation within upstream regions of orthologous genes. Pairwise as well as multiple sequence comparisons are taken into account.

Experimental evidence. Binding site descriptions (strings and PWMs) are employed to predict regulatory elements. Assembled EST sequences and verified transcription factor start sites are incorporated to pinpoint the first exon and narrow the extent of upstream regions down.

The CORG analyses pipeline imposes no positional constraints on local similarity detection and assesses the significance of local similarities. Applications for this resource include studying evolution of DNA binding sites and promoter constitution, discovery of new sequence elements (i.e. microRNAs and binding sites), systemic studies on transcriptional regulation of biological processes and hypotheses generation for experimental research.

The CORG database is accessible via a web site. This site features an interactive viewer based on JAVA technology, which is tailored to detailed promoter analysis. Large-scale studies make direct use of the MySQL implementation of CORG in conjunction with an application interface.

The benefits of having an extensible promoter annotation framework were demonstrated in two case studies. 1) Functional inference from non-random binding site distributions across expression clusters. Here, we showed that certain binding sites

occur preferentially upstream of genes that are associated with particular cell cycle phases in HeLa cells. 2) Promoter analysis of SRF induced genes. In this setting, we dissected direct from indirect SRF targets *in silico*. Our predictions gained support from follow-up ChIP experiments.

Our promotor annotation framework is flexible enough to meet future demands. Experimental efforts will soon deliver a wealth of information in the area of binding site mapping, chromatin modifications and location of actively transcribed sequence.

To sum it up, gene regulation is far from being understood. The CORG platform is my contribution to foster progress in this area.