Chapter 6. Conclusions

In this project, I developed a number of methods which give information about the location of promoter regions and transcription start sites in mammalian genomes. The EponineTSS method from chapter 3, based on a novel machine learning approach, offers state-of-the-art performance in predicting promoters, including information about the transcription start site, and is now used as part of the Ensembl genome annotation pipeline. Making use of a rather different source of information, in chapter 5 I have shown that simple methods of comparative genomics between the human and mouse genomes can reveal highly conserved blocks of sequence – including blocks of 100 bases or more with 95% nucleotide identity – which seem likely to correspond to functional promoter regions. These two methods are strongly complementary: EponineTSS can be used to predict transcription start sites, which give an indication of the 3' end of the promoter, and then information from comparative genomics can be used in conjunction with this to indicate how far upstream the functional region is likely to stretch. This combined approach meets my objective of providing promoter detection methods which can be used to support more detailed future analyses of promoter signals.

A third strand of this project, covered in chapter 4, was also related to human-mouse comparative genomics but took a very different approach, considering the complete set of non-coding similarities between the two species. Interestingly, but not entirely surprisingly, this also led me to a promoter signal, and provided a second predictive method which could be used for scanning genomes, albeit with somewhat lower accuracy than EponineTSS, and no direct predictions of the transcription start sites. While the advantages of EponineTSS mean that the models trained from homologies will probably not be useful in themselves as predictive tools, the principle that it is possible to learn such a predictor from raw comparative data is intriguing, and might prove helpful in "bootstrapping" genomic knowledge if sequencing is carried out in

clades of organisms which don't have a prior tradition of genetic or molecular biology research, and consequently no body of knowledge to draw on when annotating their genomes. In addition, I still believe that with improvements – perhaps including automatic choice of windows rather than using a fixed window of arbitrary size – the sensitivity of this method might be increased sufficiently that, once all the promoter regions have been stripped out of the training set, it might identify other types of functional non-coding region in the genome.

When evaluating the EponineTSS method, I found that it detected primarily CpG-enriched promoters. Moreover, the set of promoters detected overlapped strongly with the set detected by another computational method, PromoterInspector. This could be explained away by the fact that both EponineTSS and PromoterInspector were trained on promoters from the EPD database, and there might be some bias in the entries of this database. However, it was subsequently found that the EponineHomol 1 and 2 predictors, which were not trained using any information from EPD, also preferentially detected the same subset of promoters. This suggests that there is one group of promoters which are based around a common set of core signals (captured in the EponineTSS model). There also appear to be other types of promoter which do not follow this pattern. Attempts were made with both the EAS and EWS model families to train models on a dataset which had these core promoter sequences removed, in the hope that this might reveal a second type of core promoter. Neither of these attempts produced a model which could be usefully applied to detecting a distinct set of promoters, which suggests that there may not be an single alternative core promoter, but that each of these atypical promoters is different. As discussed in section 3.5.2, the distribution of the core promoters is not entirely uniform: metabolic enzymes, protein kinases, and transcription factors are all likely to have common core promoters, while the majority of receptors and immune system components are apparently transcribed from atypical promoters. This distinction illustrates one more level of complexity in the regulatory story.

To support the sequence analysis requirements of this project, I investigated the recently developed field of Sparse Bayesian Learning algorithms, and developed a practical implementation of a Sparse Bayesian binary classification system. This can be applied either as a

one-step method similar to the previously described Relevance Vector Machine [Tipping 2000], or using a novel incremental approach which offers a pragmatic method of sparse learning from very large sets of basis functions and even, given suitably correlated basis function, a possibly infinite family of basis functions. This library code proved to be a good basis for the learning applications described in chapters 3 and 4, but it is a general-purpose implementation and has been applied to other tasks, including analysis of microarray gene expression data. In addition, the Eponine Anchored Sequence system from chapter 3 has been applied as-is to other problems such as prediction of transcription termination sites [A. Ramadass, personal communication], and I hope that it will find many further applications in the future.

Finally, during the course of this project I had many opportunities to work on the infrastructure of genomics and sequence analysis. Throughout the time, I was involved in the development of the BioJava toolkit (section 1.5.1), which provided a foundation for all the applications and experimental programs which I developed here. I was also involved in the early development of the DAS protocol [Dowell *et al.* 2001], and wrote a BioJava-based server for this protocol [Down and Pocock 2001]. I used DAS on a number of occasions to view results from my analysis methods in a genomic context. For some parts of this work, I made direct use of Ensembl's relational databases, either by direct SQL queries or using the bj-ensembl library.

At the conclusion this project, I remain interested in the mechanisms of transcriptional regulation, and how regulatory sequences can be decoded *in silico*. Understanding transcriptional regulation would be relatively easy if we could firstly accurately determine the structure of every protein produced by the genome, and then predict interactions between those proteins and the nucleic acids. Realistically, though, the current state of the art in structure prediction still leaves much to be desired [Moult *et al.* 2001], and while Richard Lavery's group have had some successes in physical modeling of DNA-protein complexes (see, for example, [Harvey *et al* 2003]), their methods currently require a high-resolution structure of the full complex before predictions of binding specificities can be made. Therefore, to further our understanding of promoters and gene regulation, it seems important to continue direct study of

the sequences themselves, and also take advantage of any experimental techniques which can give better information about gene expression patterns and the factors which influence them.

One such experimental approach is direct study of which proteins bind to which regions of DNA sequence. Ideally, the results from high throughput experiments of this kind could give the same kind of information as simulation of DNA-protein interactions, and have the advantage that they can be performed with real chromatin, rather than the idealized situation of a single protein interacting with a "naked" DNA helix. The results of my comparative promoter analysis are currently being considered in planning a new experimental study, using a combination of chromatin immunoprecipitation and microarray techniques [Ren *et al.* 2000] to localize the *in vivo* binding locations of proteins to DNA. A comparison approach based on that described here will be used as an indicator of how much sequence upstream of each gene under consideration should be tiled onto the array. Since each probe on the array requires an individual PCR reaction with custom primers [D. Vetrie, personal communication], optimizing the choice of probes could help to maximize the number of genes which can be studied for a given budget.

Returning to computational methods, this project has introduced some powerful tools for classifying sequence data. Both the EWS and EAS models are flexible sequence analysis techniques which have the advantage that the motifs learned by the model are immediately visible for user examination. It may prove possible to apply these, or similar, methods to classification of promoters based on expression patterns, thus learning the signals which confer those particular patterns. An obvious extension for this purpose is to use a multi-class variant of the training algorithm, so that a number of different patterns can be considered in a single training run. A more radical variant of this approach, and one which I am interested in developing, is to extend this approach to unsupervised machine learning, where patterns are discovered in previously unlabeled data. In this case, the aim is to learn both the "labeling" (which groups of genes share either complete or – more likely – partial expression patterns) and the sequence-based signals which regulate this. This is a demanding problem, with no off-the-shelf solution. However, if promoter sequences are treated as a "mixture" of sequence

motifs or regulatory modules, identifying the set of modules present in a large set of promoters shows some similarities to the well-known problem of Independent Component Analysis (see, for example, [Miskin 2000]). I believe it will be possible to adapt an approach analogous to ICA to conceptually de-mix promoter regions.