Chapter 7

Conclusion

Since the work within this dissertation was started in January 2000, the initial sequencing of the human genome has been completed and the sequencing of the related mouse genome has been started and is now close to completion. Other pairs of evolutionarily related genomes such as the nematodes C. elegans and C. *briggsae* or the fruit fly ***Drosophila*** *melanogaster* and the mosquito ***Anopheles gambiae*** are emerging. The availability of these data opens the new opportunity to understand these genomes by comparing evolutionarily related genomes to each other. Comparative studies will greatly increase our understanding of genomes as both, regions which are conserved between the genomes and those which are not conserved will teach us something important.

Methods for comparative ab initio gene prediction have only started to emerge in 2000. The two novel methods presented in this dissertation are among the first to solve the gene prediction and sequence alignment problem simultaneously. They make use of the different types of conservation within two related **DNA** sequences in order to predict protein coding genes.

One method, **DOUBLESCAN,** simultaneously predicts the genes as well as the alignment of two related input **DNA** sequences by only knowing their sequence of A, C, G, T letters. This approach from first principles makes use of only a few and very basic assumptions on the general structure of eukaryotic genes and the ways in which two similar genes are related. It is capable of predicting partial, single and multiple genes as well as pairs of genes which are related by events of exon-fusion or exon-splitting. The underlying probabilistic pair hidden Markov model is parametrised in a simple way, and all parameters have a clear interpretation. The results presented in this dissertation show that **DOUBLESCAN** can be successfully used

to predict mouse and human genes simultaneously and that its parameters can be easily adapted to analyse other pairs of genomes, as demonstrated in the analysis of C. *elegans* and C. *briggsae* sequence pairs. **DOUBLESCAN** has a high sensitivity for predicting entire known genes correctly and captures the long range constraints imposed by the similar exon-intron structures of related genes well in its comparative model. This is reflected in the performance of **DOUBLESCAN** relative to that of one of the reference non-comparative *ab* initio gene prediction methods, **GENSCAN,** which increases progressively when going from nucleotide (fine scale) to gene level (large scale).

The second method, **PROJECTOR,** can be used to find the gene structures of one **DNA** sequence when those of a related **DNA** sequence are already known. Similarly to **DOUBLESCAN, PROJECTOR** is capable of dealing with partial, single and multiple genes as well as genes which are related by events of exon-fusion or exon-splitting. It was presented here for the comparative prediction of mouse and human as well as C. elegans and C. *briggsae* genes. It is the first gene prediction method which makes use of homology directly at **DNA** level and also simultaneously predicts genes and an alignment. As it makes use of gene structure information, it should have a superior sensitivity especially for detecting remotely related genes with respect to gene prediction methods which employ protein homology information.

Both methods not only detect genes, but also comparatively predict conserved subsequences within their genomic context. This should highlight novel regulatory elements which cannot be reliably predicted by non-comparative methods which have a very low specificity for detecting these typically short subsequences. For example, **PROJECTOR** can be used with one **DNA** sequence containing known genes to find both, the related genes and conserved subsequences in another related **DNA** sequence of yet unknown annotation. Both, **DOUBLESCAN** and **PROJECTOR** are the first methods to comparatively predict conserved subsequences in their genomic context.

The two above methods not only introduce new theoretical concepts for comparatively predicting protein coding genes, but have also been implemented into efficient computer programs so that they *can* be applied to realistic large scale problems. The latter was achieved by introducing a new algorithm, the Stepping Stone algorithm, whose memory and time requirements both scale essentially linearly with the length of the input sequence. The predictions generated by the Stepping Stone algorithm were compared to those of the exact Hirschberg algorithm and shown to provide a very good practical solution for the analysis of long sequences.

As **DOUBLESCAN** and **PROJECTOR** both require that the pairs of input sequences exhibit similarities in collinearity, the application of **DOUBLESCAN** and **PROJECTOR** to entire genomes requires more care in the preparation of the input sequences than non-comparative methods which can essentially be given any genomic sequence **as** input. The genomes to be analysed with **DOUBLESCAN** and **PROJECTOR** first have to be partitioned into pairs of sequences in which sequence similarities appear in collinearity using simple alignment programs like **BLASTN** [AGM+90] or **DOTTER** [SD96]. The maximal length of these sequence pairs will vary not only between different pairs of genomes, but also within one pair of genomes and will depend on the local level of divergence. Concerning the performance, it is a **priori** not clear how the performance of **DOUBLESCAN** and **PROJECTOR** on multi gene sequences will compare to that on single gene sequences. The results in [GAA+00b] show that the specificity of **GENSCAN** on semi-artificial long genomic sequences is significantly lower than on single gene sequences while its sensitivity remains essentially unchanged, whereas the specificity of similarity based programs like GENEWISE and PROCRUSTES is not significantly altered and depends mainly on the strength of the similarity to a homologous protein. The change in performance of comparative ab **initio** gene prediction methods when analysing multi gene instead of single gene sequences has so far only been investigated in [WGJMOG01]. The authors evaluate **GENSCAN** and their comparative ab **initio** gene prediction program SGP-1 on one single gene set and several multi gene sets which are derived from different regions of two genomes. Whereas GENSCAN's specificity generally decreases when analysing the multi gene sets, that of **SGP-1** increases on some of the multi gene sets. And whereas GENSCAN's sensitivity only slightly decreases when analysing the different multi gene sets, that of **SGP-1** shows both positive and negative changes with a higher amplitude than **DOUBLESCAN.** The authors conclude that the performance of their comparative ab **initio** method depends more on the level of conservation between the regions of the two genomes from which the sequences are derived than the single or multi gene nature of the sequences. Although the behavior of DOUBLESCAN's or **PROJECTOR's** performance on multi gene sequences remains to be investigated, we expect them to behave similarly.