Chapter 6

Conclusions

In the previous chapters, I have proposed a semi-formal framework for the integration of gene component evidence from arbitrary sources into complete predictions of gene structure. I have shown that this concept is valid by implementing it in a program called GAZE, and applying it to the prediction of gene structures in worm and vertebrate genomic sequences. I envision GAZE to be useful both as a research tool for investigating signal and content recognition methods, and as the final stage in a genome annotation pipeline, drawing together the results of previous localised analyses.

One of the key aspects of the system is that it does not rely upon a fixed underlying model of gene structure. This allowed for example the modelling of *trans*-spliced genes in *C.elegans* with little effort (see chapter 3). GAZE therefore provides a platform for the modelling of structural properties that are not currently accounted for in existing programs. One example would be nested genes, i.e. genes situated in the introns of other genes. The constructs currently offered by GAZE do allow the definition of a model that allows a gene to occur in the intron of another gene. However, it is necessary to define an identical gene model for each of the six different introns that can occur (for each phase, on each strand). Extension of the system to allow sub-models to be defined once and referred to elsewhere in the configuration would therefore be a natural avenue for future work.

As well as providing an integrated tool for genome annotation and research, there are ideas in this thesis that might be applicable to to other integrated gene prediction approaches. One of these is the idea of *dominance* in the context of gene prediction, introduced in chapter 2. This was used as the basis for a search-space pruning strategy that allows GAZE to run in time that grows effectively linearly with the input sequence length. The primary advantage of the technique is that it is applicable under a wider range of conditions than is typically the case with other pruning methods. The linear run-time growth of GENEID for example is offset by the fact that it neither allows arbitrary length penalty functions nor reports posterior probabilities. GENSCAN on the other hand does compute posterior probabilities, but pseudo-linear run-time is achieved by restricting the use of arbitrary length probability distributions to alternating (in practice protein-coding) states. The GAZE pruning strategy is robust under arbitrary (although eventually monotonically increasing) length penalties for all types of gene region. It would therefore be interesting to see whether the improvements in accuracy resulting from explicit modelling of exon lengths (see [20] and chapter 5) are matched by explicit length modelling of introns and intergenic regions.

The other substantial new idea described in this thesis is the Maximal Feature Discrimination method for determining optimal weights for the scores of the different types of evidence employed in an integrated gene prediction system. This was shown to have better correspondence with the standard gene prediction accuracy assessment metrics than the classical maximum likelihood method. Another advantage that it offers is the ability to parameterise sophisticated models of gene structure when the location of certain features, for example the site of transcription initiation, are not known for the training sequences. Although presented in the context of GAZE, MFD can be applied to other probabilistic gene prediction systems. For Hidden Markov Models for example, the approach can be viewed as the maximisation of the posterior probabilities of the correct state transitions at specific positions, although this is only possible if such transitions can be unambigu-

ously defined. Because the posterior probabilities can be computed by the standard forward-backward algorithm, the method is likely to be directly applicable in many situations.

Given the current trend of systems that make use of genomic sequences from more than one organism, it is natural consider the applicability of GAZE to comparative gene prediction. One simple application would be to use TBLASTX matches of a genome of interest to a variety of other genomes to support coding regions. This is the approach essentially taken by an extension of GENEID called SGP-2 [83]. It has the apparent advantage that it is not limited to hits from a single genome. However, by treating matches to different genomes equally, we ignore the fact that they will share different levels of background genomic conservation to the target genome. This means that a relatively poorly scoring match to a distantly related organism can be more meaningful than a high scoring match to a closely related one. By supplementing genomic similarity with a phylogenetic tree, we can start to discriminate between those conserved regions that are functional, and those that can be explained by evolutionary distance alone. In this example, we could therefore adjust the scores of the conserved regions according to phylogeny, up-weighting matches to distantly related genomes and down-weighting matches to closely related ones. This suggests a less principled strategy that is immediately applicable, which involves attaching a separate weight to the matches to each genome and using MFD to optimise their values.

The true power of GAZE in the use of comparative information in gene prediction comes from its clean separation of feature detection and gene prediction. Multiple alignments of corresponding regions of several genomes can be used as the basis for models of the evolution of splice sites and coding regions (for example). Supplementing standard signal recognition methods (such as those described in chapter 1) with such models will improve their accuracy. In this sense, GAZE can therefore be viewed as an efficient and convenient way for sophisticated evolutionary models to be employed in the prediction of complete gene structures.

To end, it is interesting to consider the future of research into computational gene prediction. The recent appearance of comparative and similarity-based methods, and the accompanying lack of new *ab initio* single-sequence methods, is suggestive of a feeling that there is a ceiling to what is achievable with the latter, and that we are pretty close to it with current methods. However, the fact that a living cell is able to determine the precise gene structure of its genome with high fidelity, without the use of external sequences, suggests that continued research into single-sequence techniques is not in vain.