Chapter 6

Conclusion

The process of annotating the human genome has now begun in earnest with the completion of the draft sequence and will continue for many years as the sequence is finished and as understanding of the genome's complex structure improves. The work described in this thesis aimed to assist this ongoing process by using a variety of computational approaches to help identify introns, and in the process also add to our knowledge about the mechanisms of RNA splicing. Below I outline briefly a few of the many ways these investigations could be taken further.

One step toward this goal was taken with the development of a new model for splice site identification that utilises local GC content to generate improved predictions compared to standard non-stratified models. This model identified differences in intron recognition signals that varied with GC content. These lead to interesting questions regardmg the role of splice site recognition in genome evolution. It is worth exploring whether splice site recognition has played a role in genome evolution, by, for instance, restricting most introns in GC-rich regions to short lengths. Another possibility that merits further examination is that small and large introns may exhibit slightly different splice site signals. If this is the case, my stratified model may capture this signal indirectly as a by-product of most small introns occurring in GC-rich regions of the genome.

Splice site signals are insufficiently informative by themselves and the chance of pinpointing splice sites in genomic sequences based on their signals alone seems slim. However, improvements in splice site identification lead to improvements in gene prediction and are worth exploring. In this regard, it seems that future work will further blur the distinction between splice site prediction and gene prediction with the emphasis falling on programs that take advantage of a variety of signals to accurately identify either introns or exons. An improved understanding of the splicing process *in vivo* will play a key role in this progression as utilising the information provided by intronic and exonic splicing enhancer sites (reviewed in Blencowe, 2000) is likely to be a difficult but important step in this process.

The advent of large expressed sequence libraries is having a dramatic impact on genome annotation. In regard to splicing, alignments of expressed sequences to the genome can identify splice sites with high confidence and are a powerful tool for determining gene structure. Expressed sequence datasets, however, are generally both incomplete and biased toward the start and end of transcripts and, whde a valuable annotation tool, complement rather than replace other annotation efforts.

Due to their scarcity, U12-dependent introns have traditionally been ignored in large-scale annotation efforts, hindering the accurate annotation of several hundred human genes. My analysis of these rare introns will help define these gene structures and work is ongoing to add these introns into Ensembl (International Human Genome Sequencing Consortium, 2001).

I hope that this research will also have implications for understanding the origin of the two spliceosome system found in many eukaryotes. In particular, comparative analysis of the processes of gain and loss of U12-dependent introns and conversion of U12- to U2-dependent introns (and perhaps vice versa) will increase our understanding of both splicing systems, and perhaps offer insights into eukaryotic evolution more generally. Scanning the mouse genome, once its sequencing reaches a suitable stage, would be the logical next step, as many genes and even gene structures should be conserved between mouse and human.

Although my analysis has led to a number of new observations regarding U12-dependent introns, the calculation of a preliminary error rate for U12-dependent splicing is particularly interesting. To the best of my knowledge, no previous estimates for splicing error rates by either spliceosome exist, and whde my estimate is only preliminary, it points the way toward an effective use of expressed sequence data to more accurately estimate this value and provide a comparable estimate for U2 splicing. These estimates could be obtained as by-products of the large-scale EST to genome alignment projects ongoing at the Sanger Centre and elsewhere.

Developing an accurate understanding of the frequency and mechanisms of alternative splicing looms as the next big goal in the ongoing effort to understand eukaryotic gene structure and hopefully a combination of expressed sequence based analysis and *ab initio* predictions can yield significant progress on this front. Although my work has not focused on this problem, I have identified a variety of alternatively spliced U12-dependent introns and hope that my stratified splice predictor will help in the determination of potential alternative U2-dependent splice sites.

Nearly 25 years of research into RNA splicing has yielded enormous progress in terms of understanding, but many key questions remain. Although *in vivo* work is crucial to an eventual understanding of RNA splicing, the wealth of data becoming avadable through genome projects has enabled informatics approaches to also make a significant contribution. As sequence data continues to accumulate, I expect this trend to continue and imagine that computational approaches will play an important role in answering many of the still unresolved questions regardmg RNA splicing.