Chapter 1

Introduction

The discovery of the interrupted nature of many eukaryotic genes has to rank as one of the most startling of the era of molecular biology (Berget *et al.,* 1977; Chow *et al.,* 1977). In the last 25 years, remarkable progress has been made understanding the structure of eukaryotic genes and the complex intracellular machine, known as a spliceosome, which processes these interrupted genes to yield final, translatable mRNA products. However, recently the RNA splicing story has been growing more complex rather than less and numerous crucial questions remain unanswered.

For many years, simplicity prevailed and *all* eukaryotic mRNA introns were believed to be processed by the same mechanism. However, this picture of a single spliceosome recognising a simple and conserved set of sequence signals has broken down over the last decade. In its place, we are now aware of two distinct spliceosomes, each processing a disjoint subset of introns, defined not by clear and conserved signals, but by a variety of semi-conserved signals that combine to direct splicing in the cell.

Now as the field of genomics continues to grow, and large-scale analysis becomes an important, if not dominant, research paradigm, the need to improve understanding of RNA splicing becomes even more acute. For the recently completed and ongoing genome projects of higher organisms to reach their full potential, it is critical that researchers be able to accurately identify the protein coding regions of the genome and thus create from the genome sequence a catalogue first of all the genes and then eventually of all the proteins in an organism. And a prerequisite to identifying protein coding regions is an understanding of how the cell demarcates coding and non-coding regions at RNA splice sites.

Additionally, the recent revelations regarding the surprisingly small number of genes in the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.,* 2001) have thrust the intron/exon structure of eukaryotic genes into the limelight as a potential generator of protein diversity (Graveley, 2001). In particular, the process of alternative splicing, due to the potentially combinatorial increase in protein diversity that can result from it, has been hypothesised to account for much of the complexity apparently missing from the relatively gene-poor genome sequence (Modrek *et al.,* 2001).

In short, although RNA splicing has been extensively studied both from an experimental and an informatics perspective over the last 20 or so years, much remains to be learned. Thts thesis addresses a variety of RNA splicing and intron-based analyses,

covering both major and minor class introns and hopefully will prove useful to the ongoing project of annotating the human and other vertebrate genome sequences.

Chapter 2 of this thesis is a brief review, both of the mechanism of RNA splicing and of computational approaches to identifying splice sites. It is intended as an introduction to the non-specialist and should make chapters 3 though 5 more accessible to readers not familiar with either the biology of RNA splicing or basic principles of pattern recognition in DNA sequences.

Chapter 3 discusses the identification of RNA splice sites from genomic DNA sequences. Two new splice site prediction models are introduced, one utilising hgher-order dependencies within the splice site signal and one that considers local GC content in its predictions. This latter model forms the basis of a splice site prediction program called StrataSplice that is discussed at the end of the chapter.

Chapter 4 discusses a scan for members of the rare subclass of U12-dependent introns in the draft human genome sequence. Many new U12-dependent introns were identified and analysis of their properties led to several interesting observations.

Chapter 5 briefly discusses a scan for introns on human chromosome 22. Over three thousand introns were identified of which 20 percent were not part of current gene models and these should prove helpful in the ongoing project to annotate chromosome 22.

Finally chapter 6 concludes the thesis by discussing briefly new directions that could be taken, should this research be continued.