Chapter 2

Introns and RNA Splicing: a short review

2.1 An overview of human gene structure

The complexity of gene structure in hgher organisms is one reason that automated annotation of the human and other large genomes has proven difficult (International Human Genome Sequencing Consortium, 2001; Zhang, 1998; Guigo *et al.*, 2000). The average human gene covers nearly 30 kb of genomic sequence and consists of several promoter signals and numerous splice sites as well as at least one transcription start, translation start, translation stop and polyadenylation site. These features determine how mature messenger RNAs (mRNAs) are produced from the genomic DNA sequence and play important roles in the processing of the gene into a final protein product (see Figure 2.1).

In lower eukaryotes, such as *S. cerevisiae*, most genes produce a single mRNA containing a continuous coding sequence flanked by short untranslated regions (UTRs). In contrast, many, if not most, human genes produce multiple messages, typically with long UTRs and interrupted by intervening sequences called introns that are spliced out during mRNA processing. The mRNA sequences that are spliced together when the introns are excised are known as exons and these form the final mRNA transcript. Internal exons tend to be short, with most less than 300 bp in length, whde introns, in contrast, vary greatly in size but are generally longer, with a mean size of nearly 3,400 bp (see Table 2.1). Terminal exons, at the beginning and end of mRNA transcripts, can be significantly longer than internal exons, and these long exons are found quite frequently in the 3′ UTR.

	Median	
Internal exon length	122 bp	145 bp
Number of exons	7	
Intron length	1,023 bp	3,365 bp
Coding sequence length	1,100 bp	1,340 bp
Genomic extent	14 kb	M

Table 2.1 - Characteristics of human genes. The median and mean values for a number of charactensucs of human genes are provided. Data from (International Human Genome Sequencing Consortium, 2001).

Human gene structure varies within the genome as well. For instance, while exon length remains relatively constant across a variety of GC content levels, intron length decreases dramatically in regions of high GC content (International Human Genome Sequencing consortium, 2001).

Exon/intron boundaries are known as splice sites, and as more and more genes

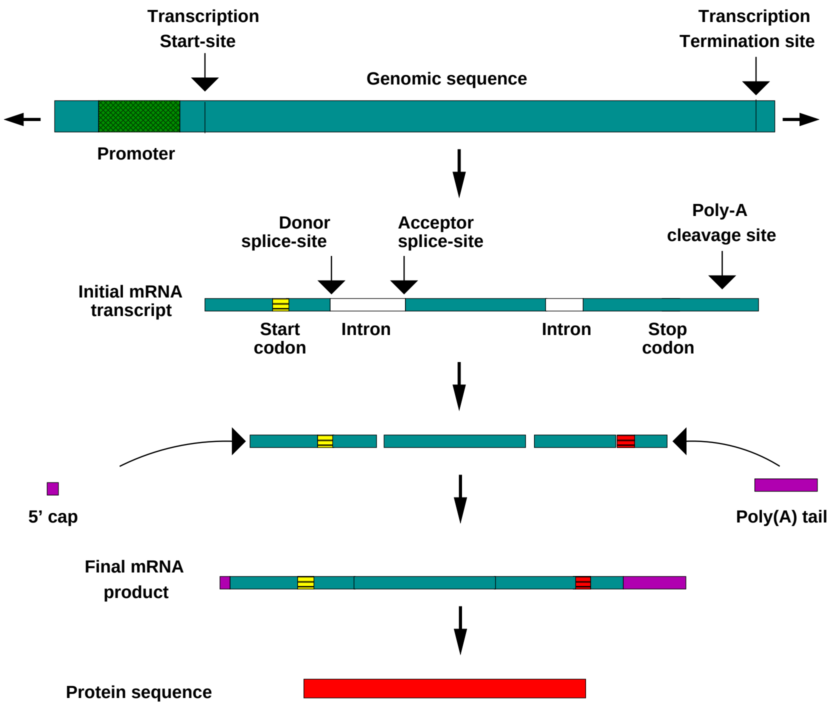

Figure 2.1 - The major components of a typical protein-coding eukaryotic gene and the role these components play in the processing of a nascent transcript into a mature mRNA are shown. (Figure courtesy of K. Howe).

were sequenced, it was quickly noticed that most splice sites had similar sequences (Mount, 1982). In particular, almost all introns started with a GT dinucleotide and ended with an AG dinucleotide. These bases became known as the canonical dinucleotides and formed the basis of the GT-AG rule for introns. Additionally several other bases near splice site junctions show varying degrees of conservation and these form the consensus sequences for donor and acceptor splice sites (see Figure 2.2).

Although more than 99 percent of introns obey the GT-AG rule, both GC-AG and AT-AC are known to be valid intron boundaries as well (Jackson, 1991, Burset *et al.*, 2000). In addition to the splice site consensus sequences, a typical intron has two other semi-conserved sequences, a polypyrimidine tract upstream of the acceptor signal and a branch site region upstream of the polypyrimidine tract.

2.2 The RNA splicing process

Introns are excised from pre-mRNA by a large complex consisting of five snRNAs and 50-100 proteins. A comprehensive review of RNA splicing and spliceosome formation is beyond the scope of this chapter, but I will provide a brief overview of the splicing process and make note of some interesting recent results. For a readable introduction to splicing, I direct the reader to chapter 22 of Lewin's *Genes VII* (2000), whilst for more detailed coverage, I refer the reader to a number of recent reviews below.

In brief, the excision of a single intron from a nascent pre-mRNA transcript is a two-step process requiring two distinct transesterification reactions. Initially cleavage occurs at the 5' splice site and the first base of the intron forms a lariat by binding to an adenosine nucleotide at the branch site, upstream of the **3'** splice site. Next a new phosphodiester bond is formed between the last base of the upstream exon and the first base of the downstream exon and the intron is released as a product of this reaction (reviewed in Burge *et al.,* 1999).

The reactions described above occur within the spliceosome complex, whch is responsible for the crucial tasks of recognising the appropriate splice sites and catalysing the splicing reactions. The spliceosome consists largely of five RNA-protein complexes known as small nuclear ribonucleoprotein particles (snRNPs).

The first step in splicing is typically the ATP-independent recognition of the 5' splice site by the U1 snRNA and the association of the U1 snRNP with this region, which results in the formation of the commitment (E) complex. This interaction, whde thought to occur at the vast majority of introns, is not strictly required as some introns have been identified in whch splicing proceeds efficiently in the absence of the U1 snRNP *in vitro* (Crispino *et al.*, 1994, 1996; Tarn and Steitz, 1994).

A key role of the U1 snRNP is to promote the association of the U2 snRNP with the branch point region of the intron. U2 snRNP association depends on two key interactions: reception of the polypyrimidine tract region by the protein U2AF^{65} and the recently discovered interaction between the protein U2AF^{35} and the intron's terminal AG dinucleotide (Zorio and Blumenthal, 1999; Wu *et al.,*1999; reviewed in Reed, 2000; Moore, 2000). The association of the U2 snRNP with the branch point region is an ATP-dependent process in which at least *six* proteins, components of the essential splicing factors SF3a and SF3b, bind either upstream or downstream of the branch point region (Kramer *et al.,* 1999; reviewed in Reed, 2000). 'fhe association of both the U1 and U2 snRNPs defines complex A (the pre-spliceosome).

Association of the tri-snRNP complex containing the **U4,** U5 and U6 snRNPs with the pre-spliceosome is required to form the B complex. Recently, the splicing factor SPF30 has been shown to play a key role in the integration of the tri-snRNP complex into the pre-spliceosome although this transition remains poorly defined (Rappsilber *et*

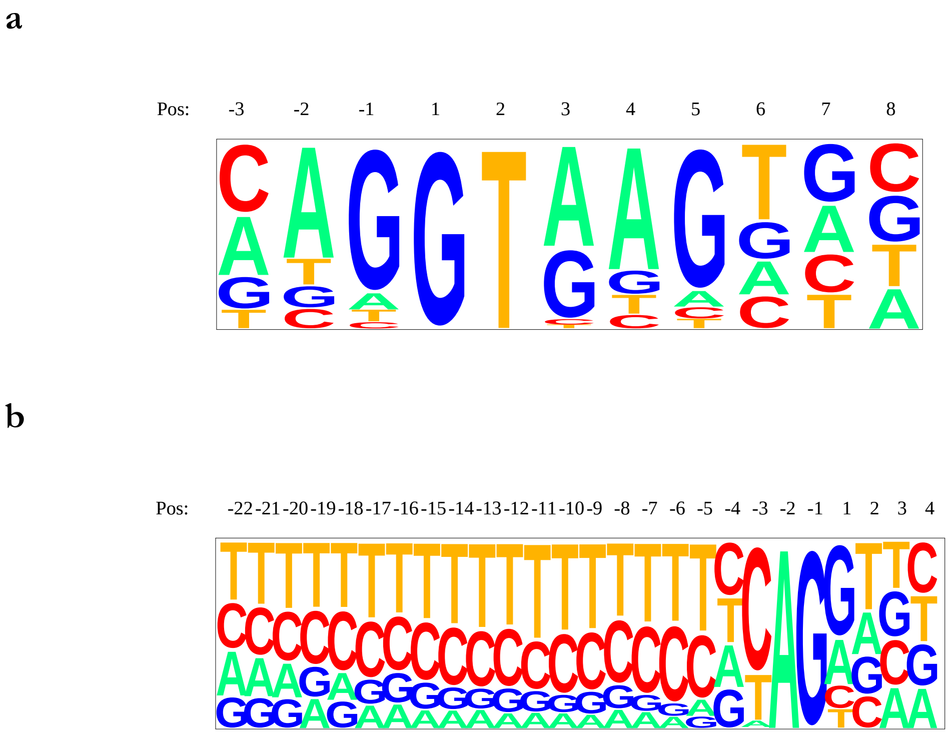

Figure 2.2 - Consensus sequences at canonical donor (a) and acceptor (b) splice sites. The size of each base represents the relative frequency at the specified position. Position one represents the first base of the intron for donor sites and the first base of the exon for acceptor sites. Graphical representations of frequencies were generated using Pictogram (Burge,C., unpublished, available at http://genes.mit.edu/pictogram.html).

al., 2001; Meister *et al.*, 2001). Although the full role of the tri-snRNP complex is not understood, it interacts with both the 5' and 3' ends of the intron and contains the highly conserved splicing factor Prp8, which is thought to contribute to the catalytic core of the spliceosome (reviewed in Collins and Guthrie, 2000). The exact nature of the catalytic core remains an open question.

Tri-snRNP addition initiates a series of RNA-RNA rearrangements, most notably the displacement of the U1 snRNA from the 5' splice site by the U6 snRNA, which create the catalytically competent C complex (reviewed in Burge *et al.*, 1999; Konarska, 1998). Recent results have found that many of these rearrangements are directed by RNA helicases in the DExD/H box family (Staley and Guthrie, 1999; reviewed in Schwer, 2001).

In yeast, the organism in which most of the interactions detailed above were worked out, introns are generally short and often interrupt larger exons. In this situation, the spliceosome is thought to form directly around the intron, in a process known as intron recognition (Talerico and Berget, 1994). In contrast, most human genes have short exons interrupted by long introns, potentially causing problems for an intron recognition based splicing system. However, an alternate model, known as exon recognition is thought to function in the splicing of longer introns; in this model, the spliceosome assembles initially around the shorter exon sequence as opposed to around the intron (Berget, 1995). Recognition in both models involves splicing-associated SR proteins, which are believed to play an important role in bridging the sequence between neighbouring splice sites and bringing spliceosome components together (reviewed in Gravely, 2000).

2.3 U12-dependent introns and the U12 spliceosome

Careful analysis of splice junctions in the early 1990's revealed a small number of introns with highly unusual donor and acceptor sites containing AT and AC in place of the typical GT and AG (Jackson, 1991; Hall and Padgett, 1994). Experimental work quickly verified suggestions that this subset of introns was excised by a novel spliceosome and characterisation of the so-called U12 spliceosome, whch contains the U11, U12, U4atac, U6atac and U5 snRNPs, began (Tarn and Steitz, 1996a, 1996b).

Many genes contain both U2- and U12-dependent introns but little is known about how the two spliceosomes cooperate to identify and splice the correct introns *in vivo.* Distinct differences are observed, however, between the splice site signals associated with the two types of introns. U12-dependent introns exhibit strongly conserved and informative donor and branch signals, whereas U2-dependent introns exhibit only moderately informative signals at the donor and acceptor sites and a highly degenerate branch site signal. Additionally the polypyrimidine tract seen between the branch site and acceptor site of U2-dependent introns is lacking, or is at least significantly weaker (see Chapter 4), in U12-dependent introns.

The evolutionary history of these two classes of introns and their respective spliceosomes remains unclear. Burge *et al.* (1998) have reported both intron subtype switching (e.g. conversion from AT-AC to GT-AG termini among U12-dependent introns) and U12- to U2-dependent intron conversion and concluded that U12-dependent introns tend to convert to U2-dependent over evolutionary time. They also reported a biased distribution of U12-dependent introns within a variety of genomes, a result they found suggestive of a fission-fusion model of spliceosome evolution in whch the U2 and U12 systems diverged in separate lineages and were later united through a merging of genetic material in a progenitor of hgher eukaryotes (Burge *et al.,* 1998).

Recent results have found a strikingly high degree of overlap between the proteins and non-coding RNAs involved in U2- and U12-dependent splicing. In addition to the U5 snRNA (Tarn and Steitz, 1996a), all 8 snRNP Sm proteins (Will *et al.,* 1999), the 4 proteins that constitute the splicing factor SF3b (Will *et al.,* 1999), and the splicing-associated protein Prp8 (Luo *et al.,* 1999) have been found in both the U2 and U12 spliceosomes. Recent evidence has indicated that splicing-associated SR proteins, long known to function in the major spliceosome, play functional roles in U12-dependent splicing as well (Hastings and Krainer, 2001). Extensive similarity in secondary structures and interactions between the set of non-coding RNAs U11, U12, U4atac and U6atac involved in the U12 spliceosome and the set U1, U2, U4 and U6 involved in the U2 spliceosome argue for homology of the two systems as well (Burge *et al.*, 1998) as do recent results that have found the stem-loop structures of U6 and U6atac to be functionally analogous (Shukla and Padgett, 2001). Although the evolutionary implications of this high degree of overlap are not entirely clear, these findings may indicate that the U12 spliceosome evolved in the presence of the U2 spliceosome rather than in a different lineage as the fission-fusion model suggests (Will *et al.*, 1999).

As more genomes have been sequenced, U12-dependent introns have been identified in a variety of hgher organisms, including human, mouse, fly and arabidopsis. Interestingly, U12-dependent introns seem to be entirely lacking from the model organisms *S. cerevisiae* and *C. elegans.*

2.4 Computational analysis of RNA splicing

Ever since the recognition of consensus signals for RNA splice sites, research into computational identification of splice sites and, thus, toward the determination of gene structure has been quite active. Originally, splice sites were identified simply by "eyeballing" a DNA sequence and looking for matches to the consensus splice site sequences. However, it quickly became apparent that many functional splice sites shared only a few bases of similarity and more sophisticated computer models were required.

Simple independent weight matrices, or frequency tables, whch yield a probabilistic log-odds score for each base at each position in a sequence, were one of the first methods developed and still prove useful today (see Figure 2.2, Staden, 1984; Harr *et al.,* 1983). Weight matrices and the many derivatives of this method require a training set of true sites to generate the frequency table and then score potential sites by summing the scores of individual bases in a pre-defined window. The incorporation of dependencies between neighbouring bases (first-order dependencies) into the weight matrix framework represents one of the most significant advancements on this simple predictive framework (Zhang and Marr, 1993).

Although several new approaches, such as finite state automata (Kudo *et al.*, 1987) and neural networks (Brunak and Engelbrecht, 1991), were developed to identify splice sites in the 1980's and early 1990's, the next models to gain widespread use were not introduced until 1997 as components of the hghly successful GENSCAN gene prediction system (Burge and Karlin, 1997). Maximal dependence decomposition, which GENSCAN uses to identify donor splice sites, is a tree-based decomposition approach that breaks down donor sites into a set of classes, based on dependencies between bases in the splice site signal, and uses a simple weight matrix to model each class individually (Burge, 1998). The GENSCAN system uses a new model for acceptor sites as well, termed a windowed weight array method, whch models the branch point region using a modification of the first-order dependencies approach that groups sets of neighbouring bases together in order to avoid problems caused by limited data (Burge, 1998).

More recent approaches have tended to integrate multiple signals into the prediction process. For instance, GeneSplicer (Pertea *et al.*, 2001) combines a traditional log-odds score based on a slight variant of maximal dependence decomposition, a measure of local coding potential and a local optimality requirement. Although combining these signals does yield improvements in splice site identification, the utility of this approach for gene prediction is more questionable, as many gene prediction systems already consider the additional signals.

Progress has been made recently as well on the problem of identifying the precise splice site from among a number of nearby, or proximal, false positives, a problem that has significant implications for automated genome analysis. One promising approach uses decision trees to discriminate between true sites and proximal false sites and may prove useful for annotation efforts (Thanaraj and Robinson, 2000). However, the current position is that all these methods generate very large numbers of false positive predictions. Typical behaviour for the analysis of genomic sequences is roughly 12 false positives per kb if thresholds are set to include 99 percent of true sites and 6 false positives per kb if thresholds are set to include 95 percent of true sites.

Finally, the large expressed sequence datasets that have been generated in the last few years have permitted the compilation of EST-confirmed splice sites on a large scale and facilitated analysis of both canonical and non-canonical introns (Burset *et al.*, 2000; International Human Genome Sequencing Consortium, 2001).