Key for BioConcept Questions, Chapter 10

1. Why is codon usage a poor predictor of the point where an exon and intron are joined? Why is the 5′ splice site consensus also a poor predictor?

   Codon usage is a good reflection of where there is an intron and where there is an exon, but it is difficult to define an exact boundary in this way, as there is likely not going to be one specific point where the usage shifts drastically. (It would be hard to tell the start of an intron from simply a couple of less-used codons at the end of an exon.)

   There is a recognizable consensus at the points where exons and introns meet, but it is a fairly weak consensus: really, only a GT pair is absolutely conserved at the 5′ site. This means that a lot of potential sites could be found in the DNA that are not actual splice sites, and it is hard to tell just from sequence which are the real ones.

2. How much of a typical human gene is usually coding sequence, versus intron sequences that will be spliced out (you may wish to recall the gene displays you saw in the UCSC Genome Browser in Chapter 1)? How does this pattern affect the difficulty of predicting introns and exons?

   Although there are human genes that have no introns at all, most have many introns and the introns are much longer than the exons. On average, each human gene is divided into about 9 exons, and 80% of these are shorter than 200 bp in length, with introns ranging up to 11,000 bp. Typically, some 80% of an average gene will be introns that will ultimately be spliced out. The small exons and high intron content make it more difficult to locate the fragments of coding sequence.

3. Why are CpG islands considered valuable for gene prediction? Where would you expect to find one with respect to a eukaryotic transcription unit? What other elements might you look for in connection with the CpG island to increase the strength of a gene prediction?

   CpGs (C-G nucleotide pairs) are a site where methylation can occur in eukaryotic cells, and methylation in promoter regions is commonly used as a means of gene regulation. Thus, "islands" where many CpGs occur are equated with potential promoter regions. Generally, they would be upstream of the coding sequence and their predictive value is strengthened if the TATA box, *Inr* element or transcription-factor binding sites are found in the same region.

4. How could alignment of a sequence with orthologous sequences contribute to the prediction of exons and introns? How could expression data (e.g., cDNA sequences) contribute?

   Other than the splice sites themselves, there is little selective pressure to maintain the sequence of an intron: mutations can occur here without affecting how the amino acids of the protein are encoded. The exons, however, are under much stronger selective pressure, so in an alignment, regions of high similarity separated by regions of low similarity are likely to represent exons and introns, respectively.

   Expression data often comes in the form of cDNA sequences, and the cDNA is produced by reverse-transcribing mRNA from the cytoplasm. Since this is spliced, mature mRNA, it should contain only exons, which can then be aligned with DNA sequences.