Key

BIO/CMPSC 300
Chapter 10 Genome Annotation II
Fall 2017

1. Why is codon usage a poor predictor of the exact point where an exon and intron are joined? Why is the 5′ splice site consensus (GT) 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. Suppose you use the sliding-window algorithm described to analyze codon bias. At several points in a DNA sequence, you notice that you see a high score in your first window and a low score in your second window. But, when you slide the window by one or two nucleotides, you get low scores in both windows. How would you explain this pattern? How might you want to account for it in deciding where your exon-intron boundaries are? (Hint: think reading frames!)

This pattern actually makes sense, because the codon usage depends on the reading frame. So, where you see the high score in the first window and a low score in the second, the first window is likely in-frame with the coding sequence and the boundary between the windows is near an intron-exon boundary. But when you shift one or two nucleotides, the first window is then out-of-frame and it’s no surprise that the codon usage pattern no longer matches expectation.

A way to use this to your advantage in an exon-prediction program is to deliberately look for cases where you get the high-low pattern at multiple points three nucleotides apart, good evidence your upstream window is in a 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 (using BLAST) contribute to the prediction of exons and introns? How could expression data (e.g., mRNA 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.