Key for BioConcept Questions, Chapter 9

1. Why are long ORFs sometimes considered to be the same as genes? In what ways is this definition insufficient?

   A DNA sequence that represents a genuine protein-coding gene should start with a start codon, end with a stop codon and have sufficient codons between the two to encode a reasonable-length protein: this would be a long ORF. But: (1) Some genuine proteins may be very short; (2) Sometimes an alternative start codon is used; and (3) In eukaryotes, most coding regions are interrupted by introns.

2. How does RNA polymerase find the transcriptional start site of a gene in prokaryotes? How can we use this information in a gene prediction algorithm?

   In prokaryotes, the sigma subunit of RNA polymerase binds to two recognizable sequences, the -10 and -35 sequences. Because these promoter sequences are present for every gene and well-conserved, it is possible to develop a "consensus" sequence that can be searched by a pattern-matching algorithm.

3. How does RNA polymerase find the transcriptional start site of a gene in eukaryotes? Why is it more difficult to develop an algorithm to find a eukaryotic promoter than a prokaryotic promoter?

   In eukaryotes, RNA polymerase doesn't bind DNA but binds to transcription factors and other proteins bound to DNA. Most genes share a few of these protein-binding sites, but in the regulatory promoter and enhancer regions, the number and types of such sites can be extremely variable. Thus, it is impossible to clearly define one set of promoter sequences that applies to all genes.

5. How does a prokaryotic ribosome find the correct start codon within an mRNA? How can we use this information in distinguishing which ORFs are genes?

   The small subunit of the prokaryotic ribosome binds to a well-conserved Shine-Dalgarno sequence (ribosome binding site) which immediately precedes the start codon. A consensus sequence can be developed for this site and searched via a pattern-matching algorithm.

6. Why can't we use a similar strategy to distinguish which ORFs are genes in eukaryotes?

   In eukaryotes, the start codon is almost always the first AUG from the 5′ end of the mRNA. Where a core promoter can be clearly defined and particularly when a clear *Inr* sequence is present, a good guess can be made about where transcription begins and thus where the first AUG is. But often the promoter region is much more ambiguous and it can be difficult to determine just where to start looking for that first AUG.

7. A simple ORF finding program would do a very poor job of predicting the amino-acid sequences of the proteins encoded in the human genome. Discuss why this is the case.

   An ORF-finding protein is incapable of finding the boundaries between the exons that together make up the coding sequence and the introns that separate them.

8. How might you identify a gene encoding a functional RNA (that does not encode a protein)? How does the discovery of key functions for very small RNA molecules complicate the issue?

Functional RNAs such as rRNAs and tRNAs also have recognizable promoter sequences; in fact, they have distinct promoters recognized by RNA polymerase I and RNA polymerase III, respectively, whereas mRNA promoters are recognized by RNA polymerase II. These promoter sequences can be used to help identify RNA genes that lack any coding sequence. The problem becomes much more difficult when the functional RNA may be 10s of nucleotides long or less, as they become very hard to distinguish from mere random sequence.