GUIDING PRINCIPLES

OF BACTERIOPHAGE GENOME ANNOTATION

1. In any segment of DNA, typically only one frame in one strand is used for a protein-coding gene. That is, each double-stranded segment of DNA is generally part of only one gene.

2. Genes do not often overlap by more than a few bp, although up to about 30 bp is legitimate.

3. The gene density in phage genomes is very high, so genes tend to be tightly packed. Thus, there are typically not large non-coding gaps between genes.

4. Protein-coding genes should have coding potential predicted by Glimmer, GeneMark, or GeneMark Smeg. Start sites are chosen to include all coding potential. These are, by far, the strongest pieces of data for predicting genes.

5. If there are two genes transcribed in opposite directions whose start sites are near one another, there typically has to be space between them for transcription promoters in both directions. This usually requires at least a 50 bp gap.

6. Protein-coding genes are generally at least 120 bp (40 codons) long. There are a small number of exceptions. Genes below about 200 bp require careful examination.

7. Switches in gene orientation (from forward to reverse, or vice versa) are relatively rare. In other words, it is common to find groups of genes transcribed in the same direction.

8. Each protein-coding gene ends with a stop codon (TAG, TGA, or TAA).

9. Each protein-coding gene starts with an initiation codon, ATG, GTG, or TTG. But note that TTG is used rarely (about 7% of all genes). ATG and GTG are used at almost equivalent frequencies.

CONTINUED…

GUIDING PRINCIPLES

10. An important task is choosing between different possible translation initiation (i.e., start) codons. The best choice of start site is gene-specific, and gene function and synteny must be carefully considered. As phage genes are frequently co-transcribed and co-translated, less weight may be given to optimal ribosome binding site sequences in start site selection. Identifying the correct start site is not always easy and is predicated on the following sub-principles:

    a. The relationship to the closest upstream gene is important. Usually, there is neither a large gap nor a large overlap (i.e., more than about 7 bp). If the genes are part of an operon, a 4bp overlap (ATGA), where a start codon overlaps the stop codon of the upstream gene, is preferred by the ribosome. Therefore RBS scores may have little bearing in this type of gene arrangement.

    b. The position of the start site is often conserved among homologues of genes. Therefore, the start site of a gene in your phage is likely to be in the same position as those in related genes in other genomes. But be aware that one or more previously annotated and published genes could be suboptimal, and you may have the opportunity to help change it to a more optimal one. Homologues in more distantly related genomes (those of a different cluster) may prove more informative because alternate incorrect start sites are less likely to be conserved. Use Starterator!

    c. The preferred start site usually has a favorable RBS score within all the potential start codons, but not necessarily the best. A notable exception is the integrase in many genomes, which has a very low RBS score. Our experimental data suggests that some genes do not have an SD sequence.

    d. Manual inspection can be helpful to distinguish between possible start sites. The consensus is as follows: **AAGGAGG – 3-12 bp – start codon**.

    e. Your final start-site selection will likely represent a compromise of these sub-principles.

11. tRNA genes are not called precisely in the program embedded in DNA Master, and require extra attention. (Please refer to **Section 9.5**.)