Guiding Principles of Bacteriophage Genome Annotation

OBJECTIVE

This protocol contains the guiding principles of bacteriophage genome annotation. Printing these pages, reading them carefully, and having them close at hand during the process of gene calling are highly recommended.

BACKGROUND

Though the automated annotation created using DNA Master will usually identify more than 80% of genes correctly, some genes will need to be manually added, modified, or deleted. Therefore, all gene calls must be reviewed to identify those that must be changed. In this section, we provide a set of principles that should serve as a guide throughout the process of evaluating and improving the draft annotation.

It is helpful to think of the process of evaluating draft annotation's gene calls as an application of these principles: together they will help make the best possible gene predictions. It is essential to understand that any annotation consists of making a **prediction** as to how the genetic information is organized and used. In the absence of experimental evidence to support a given gene call, there is no right or wrong answer; there are, however, well- supported or ill-supported predictions.

As with any set of principles, the ones presented here will conflict with one another at times. It's important to weigh one against another and make the best gene calls possible. Because of the importance of these principles, this section is dedicated wholly to presenting them. Read them carefully before beginning an annotation, and keep them nearby throughout the process.

HELPFUL TIPS

- The Complete DNA Master Guide is available on www.phagesdb.org. This protocol outlines the guiding principles of the biology of predicting genes.

PROCEDURES

The following two pages present the guiding principles themselves.

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. 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.

5. 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.

6. Protein-coding genes should have coding potential predicted by *either* Glimmer, GeneMark, or GeneMark TB. Start sites are chosen to include areas of strong coding potential.

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

OF BACTERIOPHAGE GENOME ANNOTATION

…CONTINUED

10. An important task is choosing between different possible translation initiation (i.e., start) codons. The correct start site can often be distinguished by association with a credible ribosome-binding site (RBS; Shine-Dalgarno (SD) sequence). Identifying the correct start site, however, is not always easy and is predicated on the following sub-principles:

    a. The preferred start site usually has one of the higher SD scores of all the potential start codons, but not necessarily the highest.

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

    c. 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 4 bp). A short overlap of 1-4 bp—where the start codon overlaps the stop codon of the upstream gene—is very common.

    d. 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.

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

        i. A start codon that overlaps the stop codon of a previous gene trumps a somewhat lower score.

        ii. A higher SD score or canonical RBS trumps a more extended gene overlap. iii. If choosing between several starts with similar SD scores, it is usually best to choose the one that gives the longest open reading frame.

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