Part III – Annotation

**Slide 1:** After learning how to access sequence information in the Databases slide deck (part II), let's move on to Annotations.

**Slide 2:** In this PowerPoint, we will learn about annotations and what it means, introduce the concept of Hidden Markov Models, and talk a bit about a prokaryotic genome annotation pipeline. We’ll follow this slide deck with an Annotation Exercise to get some practice.

**Slide 3:** The first question we need to ask is what is annotation? And second, why is it important? Let’s look at an example. In this slide, we see a map of the United States with certain cities marked on it and certain areas that are different shades of purple. If we had to guess what this map represents, it is difficult without a map legend.

**Slide 4:** Here’s the same picture, but now with the annotation revealed: a map of airline passengers per capita with the darker shaded areas having airports that carry larger number of passengers. So just like this map legend helped us make sense of this information, genome annotation provides genome information for raw sequences. This information could be in the form of location of genes, open reading frames etc.

**Slide 5:** Annotations can be divided into two types: Structural and Functional annotations. First, we’ll focus on structural annotations. Some of the properties of structural annotations are locating the ORFs (open-reading frames), genes, regulatory elements, and repeat elements.

**Slide 6:** Before we start gene prediction we have to locate the reading frames in the data.

**Slide 7:** Here we see a double strand of DNA. Recall that for most pieces of DNA, the protein-coding region will be on one strand or the other, but not both (i.e., there will not be overlapping genes). So, the first challenge is to determine which strand.

**Slide 8:** The next challenge is to determine where the protein starts. Recall that on a double-stranded piece of DNA it is possible to have six reading frames, three on the forward strand and three on the reverse strand.

**Slide 9:** The next step in locating the reading frame is to look for the stop codons. The sequence between two stop codons in the same reading frame constitutes an open reading frame (ORF).

If we look at our piece of DNA with its six reading frames again, we can identify the ones with the stop codons (highlighted in red) and eliminate those which do not have them (greyed out).

Finally, we know that the ORF is the longest sequence of nucleotides that can potentially translate into a polypeptide chain. Thus, identifying the start codon among the three remaining reading frames can tell us which one is the ORF.

**Slide 10:** To summarize: we need to find the potential coding regions in a stretch of nucleotides. We can search for potential stop and start codons within each of the 6 reading frames for a given piece of DNA. The longest sequence of nucleotides that can potentially code for a protein is defined as the ORF.

Certain other features of regions with ORFs is that stop codons are located in AT rich regions and GC rich regions tend to have longer ORFs. The location of certain structural elements like the Shine-Dalgarno sequence that is located 8 bps upstream of the initiation codon also provides valuable hints to the location of these ORFs. In other words, these additional features can be used to support the choice of one reading frame as the protein coding region.

Simple rules like these can be used to develop software programs that can automate the structural annotation process.

**Slide 11:** The challenge still remains that not every ORF has coding potential and although the ORF gee begins with a start codon, it may not be the actual translational start site. So even though programs are getting really good at automating this kind of annotation, there remains the possibility for incorrect results. In this simple example, there is an ATG (codes for Met) but there is also a TTG (codes for Leu). Although most genes in bacteria start with Met, there is a reasonably high chance that the ORF begins with Leu or Val instead.

**Slide 12:** Certain other features can help make the “rules” for annotation. For one, stop codons are located in AT-rich regions and so GC-rich regions tend to have longer ORFs. The location of certain structural elements like the Shine-Dalgarno sequence located 8 bps upstream of the initiation codon also provides valuable hints to the location of ORFs. Etc.

**Slide 13**: A further summary – the point here is that the rules are relatively simple, but not always straightforward. It’s for this reason that where possible, biological data is useful for confirming what the computer algorithms determine.

**Slide 14:** There are a number of different software programs that are useful for automated gene prediction - this is not meant to be an exhaustive list, but merely to illustrate some of the tools available for different kinds of genes.

**Slide 15:** Now let’s move on to functional annotation. Often we’re more interested in the biological functions of these structural elements.

**Slide 16:** How do we infer function from sequence? One important tool, called Hidden Markov Models or HMMs, looks at patterns of sequence. These are statistical tools that can predict the function of an unknown proteins. By creating a multiple sequence alignment of proteins (in this example) and using the inference that structural similarity will lead to functional similarity, your query sequence can be placed in a group whose function has been previously defined. So by using these statistical models, the query protein can be assigned a function based on its similarity to others

**Slide 17:** Another important concept in Functional Annotation is Gene Ontology. Ontologies are used to describe and represent an area of knowledge. As data from bioinformatics analyses started to be made available there was no common system of classifying genes based on functionality. Thus, creating a gene ontology was an attempt to unify gene and gene product attributes from across species. Most importantly, gene ontology creates a controlled vocabulary to define gene functions so that it can be easily used for comparison between and among species.

**Slide 18:** As an example, think back to the old days of Netflix, when they sent you DVDs in the mail based on your queue. We can label these by genre (action, drama, comedy, etc.).

**Slide 19:** We can also label movies by other features: title (name), year made, actor, director. In IMdB, you can, for example, search for all the movies starring Brad Pitt or movies directed by the Coen Brothers.

**Slide 20**: Similarly, if we’re thinking about genes and gene products (proteins), Gene Ontology gives us a way to compare across species.

**Slide 21:** If we take the example of the *dnaA* gene, we would be interested to know the biological function of this gene product (it’s molecular function), the biological process it is involved in, and the place in the cell where it acts (the cellular component). These three areas – molecular function, cellular component, and biological process – can be assigned to every gene product as Gene Ontology or GO terms.

**Slide 22:** Based on the functional annotations associated with this gene we know that it acts in the cytosol (it is a bacterial protein) and the plasma membrane. It is involved in ATP and DNA binding but most importantly the biological process it controls is DNA replication. This information its extremely useful to compare the gene and gene product across species.

**Slide 23**: There are different “strengths” of evidence for GO term assignment. The strongest ones are ones where there is experimental data to back up the assignment (shown in orange here). Less strong are ones where there is no experimental evidence, but humans have looked at the data and evaluated (yellow). The weakest evidence comes from simply a match to something else in the database (purple).

**Slide 24:** In practice, how do you annotate a genome? For prokaryotic genomes, one way is illustrated on this slide. A pipeline of different computer algorithms is linked together. Starting with the raw sequence (the A’s, C’s, G’s, and T’s), the sequence is analyzed for tRNA and rRNA genes, then for potential protein-coding genes. These are then examined for close relatives in existing databases by phylogeny. From this comes two classes of genes – determined ones and putative ones. Determined ones are the ones where there is good evidence to suggest that he have identified the *DnaA* homolog for example. Putative ones are ones that pass the rules for being an open-reading frame, but we don’t have much evidence based on comparisons to other genomes currently available. This information goes into making functional assignments (i.e. the GO terms). The results of such analysis can then be viewed on a browser.

One such pipeline is called RAST (Rapid Annotation using Subsystem Technology). It is available at <http://rast.nmpdr.org>.

**Slide 25: F**or this slide deck we learned about genome annotation, including the two types of annotation – structural and functional. We also introduced the HMMs and the Gene Ontology classification.

Please do the exercise on annotations using the tool on the IMG database.

Following this exercise please turn to the next slide deck – Part IV – Comparative Genomics.