Part II – Databases

**Slide 1:** Biological Databases – where to find DNA, RNA, and Protein data for analysis

**Slide 2:** The learning goals for this slide deck are

1) To become familiar with the human microbiome project which has made a major push to sequence genomes of prokaryotes found in humans

2) To learn how to access the biological databases where all the sequences are housed

3) To learn about the utility of accession numbers which serve as unique identifiers for these sequences and

4) To learn how to download sequences of interest once you find them

**Slide 3:** On this slide, time is on the x axis and on the y axis is DNA sequence in kilobases per day per machine. Clearly the rate of sequencing has been growing exponentially while simultaneously the cost of sequencing is going down. So sequence data is made available faster and cheaper than ever before, and this has led to a bottleneck for analysis of these data. Here’s an opportunity for students to get involved in sequence analysis.

**Slide 4:** A few definitions:

The **microbiome** is the aggregate of all microbial genomes in the sample that you are investigating. **Microbiota** refers to individual bacterial species in a specific biome. There are trillions of organisms out there, and out of these 500 species have been identified. In the human body majority of the cells are microbial - so you are made up of more microbial cells than human cells. We know that these microbes have remained in the human body for thousands of years so they must be an integral part of our evolution and must be performing some vital function, which we have not identified yet.

**Slide 5:** We’re now more than 10 years into the first investigations into the human microbiome. We’ve learned a lot, but there’s a lot still to uncover.

**Slide 6:** The Human Microbiome Project (HMP) was started in 2008 as an initiative by the NIH. HMP researchers sequenced the microbes in different regions of the human body and now we can see those results. If you see the figure and look at the pie charts, they are showing the abundance of different phyla of bacteria on different regions of the human body. The next stage of the project is to start identifying the function that these microbes are playing within the human body. There is already evidence that they are involved in certain metabolic functions like production of vitamins and short chain fatty acids. But this research is still in its infancy stage and a there area a lot of things still to be uncovered.

**Slide 7**: There are a variety of other microbiome projects too, some associated with human health, but others are attempted to understand the wealth of microbes in

**Slide 8:** We know now there are trillions of microbes out there; scientists have realized that there is a discrepancy between the number that we can grow in lab conditions and the ones we can see under the microscope in a given sample (and that we can now detect by sequencing). This is termed “The Great Plate Count Anomaly.” Fewer than 1% of known organisms can be grown in the lab. So there are many organisms we are missing.

**Slide 9:** So how do we identify aspects of this hidden diversity? Because of recent changes in technology, the newly emergent field of **metagenomics** has been developed. One way that is used frequently is to amplify a portion of the 16S rRNA genes in an environmental sample and then sequence the individual pieces of DNA. The type and number of sequences provides information about the types of bacteria and their abundance within an population.

**Slide 10:** Now: how do we access these data to begin our analyses? Such data are stored in a variety of sequence databases, data repositories that include sequences of one or more organisms. They include additional details that describe important data features including certain biological properties called **annotations**, which we will talk about more in the next slide deck. These databases sometimes include certain genome browsers, which make it easy to visualize the sequence data and associated features.

**Slide 11:** There are various databases where nucleotide and protein sequences are stored. Genbank is the American initiative by the National Center for Biotechnology Information (NCBI).

**Slide 12:** There is a similar European database called EMBL, an initiative by the European Bioinformatics Institute.

**Slide 13:** Also, there is the DNA Databank of Japan (DDJB). On all these databases, the sequence information is the same, they use different formats, and slightly different interfaces.

**Slide 14:** For our purposes we will be focusing on GenBank. The entries in GenBank are referred to as “flat files” – here’s a screenshot of one entry. This flat file contains information related to the sequence like the date of entry, researchers who made the submission, and most importantly a unique accession number.

**Slide 15:** The accession number is a unique identifier (a string of 4-10 alphanumeric characters) that can be used to search for a sequence of interest. One thing to pay attention to is while each unique accession number points to a unique sequence, a given sequence may have more than one accession number.

**Slide 16:** On this flat file, the accession number is circled in red. We will learn more about its utility when we do the associated exercise at the end of this slide deck.

**Slide 17:** GenBank has both nucleotide and protein databases, but there are some other protein databases that are useful.

One of the best databases for proteins is UniProt. The UniProt consortium consists of three databases: 1) Protein Information Database that is housed in Georgetown 2) Swiss-Prot and 3) Tremble. Swiss-Prot and Tremble are both providing functional information to the protein entries. Swiss-Prot is manually curated whereas Tremble is a computer-annotated supplement to Swiss-Prot.

**Slide 18:** Just like NCBI, data in UniProt are organized in flat files, which give similar information such as the name and origin of the entry, sequence information, and annotations of those sequences.

**Slide 19:** If you are working with prokaryotic data, which we do in Genome Solver, there is a specific database called the Integrated Microbial Genome Database from the Joint Genome Institute, which is very useful. In the next few slides, we’ll talk about some of the features.

**Slide 20:** Conceptual pipeline for accessing a prokaryotic genome.

**Slide 21:** In the IMG database you can search for an organism of interest - click on the “Find Genomes” tab and input the name of the organism you are searching for under “Genome Search.”

**Slide 22:** Alternatively, you can type in the organism of interest under “Genome Browser.” As with most databases and browsers, there is more than one way to get to the information you’re looking for, so spending some time getting familiar with the different types of navigation will help you become more comfortable.

**Slides 23-26: The next several slides show you one way to navigate through the IMG Database to get to sequence information.**

**Slide 27:** This slide shows FASTA format. When downloading sequences in FASTA format, the first line is the identifier and contains useful information (metadata) associated with that sequence (red oval). Most computer programs will “know” that the sequence starts on the second line (after a “return” at the end of the first line). So be careful that you are not manually changing the file in any way as it may lead to loss of sequence information.

**Slide 28:** Similar to the FASTA format is GenBank format. It contains the same nucleotide information but here every line is numbered which makes it easier to look at a particular nucleotide base in the sequence.

**Slide 29:** To recap, in this slide deck we learned about metagenomics and its application in the Human Microbiome Project. We learned about different databases and got a brief introduction to browsers.

This slide deck should be followed by the Accession Number Exercise to get some practice navigating in the different databases.

Next, we’ll turn to Part III – Annotation.