Part 1: Prokaryotic Gene Annotation Pipeline

- Prokaryotic Gene Prediction: bacterial mRNA is polyscistronic, making one promoter responsible for the regulation of multiple genes
    - This known as an Operon: where one promoter for more than one coding region (CDR), and is where the RNA pol binds the promoter and transcribes all of the genes (with only 1-2 NTs between each gene).
- Early Annotations: Open Reading Frames (ORF) was a good method for finding CDS in prokaryotic DNA as those sequences lack introns.
    - Choosing the right start codon complicates choosing frame start locations, also limited by RNA sequences having hard to predict start/stop sequences.
    - Gene predictions built off of ORF finding, but utilized comparative genomics.

Part 2: Bacterial Gene Prediction and Comparison

- ORF Finder: shows all/unusually the longest ORFs that have no overlaps with real genes
- FGENESB: automatic annotation of bacterial genomes including their GenBank annotations.
    - Based on Markov Chain Models(MCM) to find CDRs, promoters, operons, termination sites.
- GLIMMER: Unix-based program available on BFX server. Use long ORFs as a training set to analyze the rest of the genome
- Also available: GeneMark & Easy Gene 1.2b

Part 3: Bacteria and Archaea

- Classifying Prokaryotic Genomes:
    - Genome size and geometry: prok genomes can range from 0.5Mb to 10mb, with circular or linear structure
    - Lifestyle: extracellular, extremophilic, mesophilic, epicellular, symbiotic, parasitic
    - Disease relevance: primarily to humans
    - Ribosomal RNA sequences: the first molecular phylogeny studies
    - Other molecular sequences: including proteins, genes, gene order
- *Escherichia coli*: heavily studied genome over the past 70 years due to its small size.
- GC content: a good measure in individual differences between organisms, expressed as a percentage. GC content indicative of codon usage, as multiple codons can create one amino acid, with each organism having some preference of one over the other.
- seqinr package (R): allows simple analysis of FASTA-formatted sequences.
- Pathosystems Resource Integration Center(PATRIC): database with 10,000+ microbial genomes, allows people to use private data with available tools.
- HaloWeb: database limited to haloarchaea/archaeal species that grow in high salt environments.
- UCSC Archaeal Genome Browser: database for archaea and some bacteria, linked to Galaxy
- Greengenes: Uses 16S rRNA to compare organisms.

Part 4: Bacterial Promoters

- Prokaryotic Gene Structure: RNA pol binds regions near to TSS. Seq's at -10 and -35 are likely sigma factor binding sites, and the alpha-subunit of the RNA pol binds to the UP-element at -40 to -60.
- Sigma Factor: a subunit of RNApol that binds to the promoter region, σ^{70} is the most commonly used, with σ^{32} being used in heat-shock conditions.
- Strategies to finding bacterial promoters
    - Comparative Genomics: compares intergenic regions of related species with assumptions that promoter regions must stay similar to maintain binding with sigma factor. Challenge to find genomes that have appropriate evolutionary distance, can't be too close or too far.

- Finding Sequences Unique to promoters: because promoters have consensus, they can be found using statistically over represented short sequences in intergenic regions of the genome, or look at areas of weak DNA stability consistent with that of promoter regions.
- Incorporate microarray/RNA seq data: Use transcriptomic data with "guilt-by-association" method to look for co-expressed genes which could share common regulators (CTSP or BioProspector).