Eukaryotic Gene Prediction

- *De novo* gene prediction programs use solely sequence information for their predictions.
- Expression-based gene prediction:
    - Align complementary DNA/protein sequences to genomic sequences to define CDS. Gene str. Prediction by expression very accurate, but limited by lack of cDNA or ESTs (~20-40% of genome codes for expressed genes).
- Single *de novo* gene prediction is used where expression based gene prediction fails
    - GENSCAN(problem with false positives), HMMGene, FGENESH, and Augustus are all programs that analyze DNA to predict exon location.
- Content Sensors: prediction algorithms that depend on features in coding regions (e.g. NT composition).
    - 3rd base program for bacteria uses assumption that the 3rd base in a codon is more random in non-cDNA than it is in cDNA, but the wobble hypothesis exists for the 3rd codon in cDNA.
    - Hexamer frequency most reliable by combining codon bias and dicodon combos. Analyzed using HMM, based on training sets/probability, results used to find similar patterns in unknown sequence, and probability used to predict those coding regions.
- Signal Sensors: clues to find genes based on particular sequence that's assoc with gene.
    - Promoters, splice sites, stop codons, poly(A) are all content, but signals found using "weighted matrix", or PSSM, that gives weight to most conserved potitions(like introns starting with GT and ending AG, those positions given more attention).
- Output of Eukaryotic Gene Prediction: 5' and 3' UTRs usually ignored, so would normally just start at start codon and end at stop codon, but could include:
    - First mRNA exon: TSS and splice donor site (5')
    - Internal exon: splice acceptor site (3') to splice donor (5')
    - Terminal mRNA exon: splice acceptor(3') to transcription termination site.
- Comparative Genomics: one of the best ways to find euk genes. Compare genes that are moderately distant. Too close, and the conserved regions of noncoding sequences could be marked as genes. Too far, and there are not going to be any conserved regions. Can be used small scale for exon locations, large scale to ID potentially undiscovered genes, and generally to improve *ab intio* prediction by combining it with alignment to a similar genome.

Expression-Based Prediction:

- cDNA can align to the genome and be used to find gene locations and exon pattern, but getting that CDS with UTRs requires isolating & sequencing cDNA.
- **SPLIGN**: NCBI program that aligns cDNA to genomic DNA to ID splice-junction loci, potential frame shifts, and alt gene models. Supports cross species aligns and will output a complete gene model for input cDNA sequences.

- **Genomic BLAST:** if you have ESTs, can tell where the gene MIGHT be located, but not the full genomic structure. Algorithm is faster than conventional BLAST and works well with large input sequences.
- **BLAT:** analogous to Splign used to align sequences to the whole genome, and can align mRNA to determine the gene model but doesn't include visuals.

NCBI Genomes, Map Viewer, and Variation Viewer:

- Entrez Genomes database: complete genomes from various organisms: 306.892 prokaryotes, 42.119 viruses, and 15.544 eukaryotes
    - Genomic Data Viewer (map viewer) best way to examine eukaryotic RefSeq genomes.
- Variation Viewer: another option from NCBI that mainly focuses on human genome

Intro to BED/WIG files:

- BED(Browser Extensible Data) Format SHORT: designed to store info on genomic intervals. Uses 3 col table with each representing different information point/ entity. 1st col is chr name, 2nd is start position of that entity, and 3rd is stop position of that entity.
    - Zero-based starting position("Fence-posting"): important to know whether zero-based or one-based, as the starting codon in the 2nd column will indicate the nucleotide before the start of the sequence(ZERO-BASED) or the nucleotide that IS the start of the sequence(ONE-BASED).
- BED Format LONG: can include more columns than the three listed above, and can have 6(chromosome, start, end, name, score, strand), and 12(← + thickStart, thickEnd, itemRgb, blockCount, blockSizes, blockStarts).
- WIG("wiggle") Format: used to store continuous data across large genomic regions, scored displayed for every N nts where N an integer > 1.
- For both formats, header required for metadata to be interpreted by browser for visualization.
- BIG WIG: compressed WIG files that do not include info not used in visualization, fast.
- BED file from scratch easy, all needed is 3 columns of info and save with .bed file extension.