Applied Bioinformatics
Gene Prediction

Assistant Professor
Department of Biomedical Informatics
Center for Human Genetics Research
Terminology

• What is a gene?
• What is the gene for cystic fibrosis?
• What are the two major databases of gene information?
• Do they use the same definition of the word “gene”?
How do we predict genes?

• For most model organisms and for humans, this is part of a pipeline for genome assembly
• Prokaryotic gene prediction is far easier
  – Genes fall within a single ORF (no splicing)
  – Examination of ribosome binding sites, nucleotide biases and codon sequences
• Methods fall into two general categories
  – \textit{Ab initio} | Comparative
Ab Initio Gene Prediction

• Different organisms often have different patterns of gene organization
  – Intron/exon structure, splicing motifs, gene size, etc

• Using the collection of known genes for an organism, a set of rules for identifying genes is derived

• These approaches use algorithms like HMM, or machine learning tools like neural networks or rule mining
It all starts with the sequence

• GenBank
  – NCBI’s data warehouse for sequence data
  – Observed DNA, RNA, and protein sequences isolated from organisms are submitted online
  – May contain multiple sequences of the same region
    • Think different samples of the same sequence location
    • Can be used to identify genetic variation
An International Effort

International Nucleotide Sequence Database Collaboration

Sites Synchronize Daily
The Reference Sequence Database

• RefSeq
  – Curated by NCBI
  – Contains only one example of each natural biological molecule (transcript and protein)
  – GenBank is like primary literature, RefSeq is like a review (curated consensus)
    • Some entries are copies from GenBank
    • Some are curated annotations from multiple GenBank Records
    • Some are imported from other sources (FlyBase, etc)
The NCBI Pipeline

• Produces new genomic builds
  – Current build is “GRCh37” and “hg19”
  – Coordinates CHANGE between builds
• Genomic contigs are aligned into an assembly
• mRNA and EST sequences from RefSeq are aligned to the assembly to produce gene annotations
• Genes are also predicted ab initio
  – Predictions are inserted into RefSeq
RefSeq Status

• Gene predictions are usually easy to spot (LOC388210)

• If there is doubt, you can check the RefSeq status
  – Inferred | Model | Predicted generally indicate they are a result of prediction algorithms
  – Provisional | Reviewed | Validated generally indicate they are based on a submitted sequence

Entrez-Gene

• Entrez-Gene is a curated collection of genes resulting from the NCBI annotation pipeline
• Entries link back to the RefSeq they are based on
• At this point, most Entrez-Gene entries are from established transcripts
• Entrez-Gene entries have a unique stable numeric ID
• Formerly called LocusLink
Flow of Information

- EBI/BMBL
- DNA Databank of Japan
- GenBank
- RefSeq

NCBI Annotation Pipeline
- RefSeq Alignments
- Ab Initio Predictions

Entrez-Gene
Ensembl

- Ensembl is a joint venture between EMBL-EBI and the Wellcome Trust Sanger Institute
- Ensembl uses a very similar process to annotate genes, but the outcome is slightly different
- *Genebuild* is the name of their annotation pipeline
Genebuild

• Load assembly for the organism
• Targeted stage
  – *Species-specific* proteins are aligned to the genome
  – Transcript structures are generated for each protein
• Similarity stage
  – Proteins from *closely related species* are used to build transcript structures that are missed in the targeted stage
• Next, species specific cDNAs are aligned to the genome, and any non-translated regions are spliced into the transcript structures as a
Ensembl Genes

• Each predicted gene is assigned an Ensembl ID
• Ensembl genes can be thought of as a collection of protein sequences that have been aligned to the same genomic location
• Ensembl provides the supporting evidence for a gene/transcript
  – EMBL | UniProt | RefSeq
• Ensembl doesn’t list *ab initio* predictions