Introduction to Genome Annotation

AGCGTGGTAGCGCGAGTTTGCGAGCTAGCTAGGCTCCGGATGCGA
CCAGCTTTTGTAGATGAATATAGTGTGCGCGACTAGCTGTGTGTTG
GAATATATAGTGTTGTCTCTCCTGATATGTTAGTTCTGGATCTAGTGTTGGTGTAGATGGAGATCGCGTGCTTGAGTTCGTTCGTTTTTTTATGCTGATGATATAAATATATAGTGTTGGTGGGGGGTACTCTACTCTCTCTAGAGAGAGCCTCTCAAAAAAAAAGC
CGGGGATCGGGTTCGAAGAAGTGAGATGTACGCGCTAGCTAGTATATCTCTTTCTCTGTCGTGCTGCTTGAGATCGTTCGTTTTTTTATGGATGATATAAATATATAGTGTTGGTGGGGGGTACTCTACTCTCTCAGAGAGAGCCTCTCAAAAAAAAAGCTCGGGGATCGGGTTCGAAGAAGTGAGATGTACGCGCTAGXTAGTATATCTCTTTCTCTGTCGTGC
What is Annotation?

• **dictionary definition of “to annotate”:**
  – “to make or furnish critical or explanatory notes or comment”

• **some of what this includes for genomics**
  – gene product names
  – functional characteristics of gene products
  – physical characteristics of gene/protein/genome
  – overall metabolic profile of the organism

• **elements of the annotation process**
  – gene finding
  – homology searches
  – functional assignment
  – ORF management
  – data availability

• **manual vs. automatic**
  – automatic = computer makes the decisions
    • good on easy ones
    • bad on hard ones
  – manual = human makes the decisions
    • highest quality

**Due to the VOLUMES of genome data today, most genome projects are annotated primarily using automated methods with limited manual annotation**
Annotation pipeline

Generation of Open Reading Frames
Homology Searches
Putative ID
Frameshift Detection
Ambiguity Report
Role Assignment
Metabolic Pathways
Gene Families
DNA Motifs
Regulatory Elements
Repetitive Sequences
Comparative Genomics

Completed Sequence Region (1-10 Kb)
Analysis (ex. Locating ORFs)
Genome Structure

Prokaryote

Intergenic Region  Monocistronic Gene  Intergenic Region  Polycistronic Genes in an Operon

Eukaryote

Gene  Intergenic Region  Gene  Intergenic Region  Gene
Prokaryotic Gene Structure and Transcript Processing

A 'typical' bacterial operon

- DNA
  -35
  -10
  promoter
  transcription starts here
  terminator

- RNA
  gene a
  gene b
  gene c
Eukaryotic Gene Structure and Transcript Processing
Structural Annotation: Finding the Genes in Genomic DNA

Two main types of data used in defining gene structure:

Prediction based: algorithms designed to find genes/gene structures based on nucleotide sequence and composition

Sequence similarity (DNA and protein): alignment to mRNA sequences (ESTs) and proteins from the same species or related species; identification of domains and motifs
Finding Genes (ORFs)

Gene finders are programs that can identify genes computationally.

Running a Gene-finder is a two-part process

1) Train Gene finder for the organism you have sequenced.

2) Run the trained Gene finder on the completed sequence.
Candidate Genes

6-frame ORF map

Stop codons (TAA, TAG, TGA) (long hash marks)
Start codons (ATG, GTG, TTG) (short hash marks)

Minimum ORF Length

ORFs over minimum length highlighted
Annotating ORFs

Possible translations represented by arrows, moving from start to stop, the dotted line represents an ORF with no start site.

Glimmer chooses the set of likely genes.

ORF00001  ORF00002  ORF00003  ORF00004
Eukaryotic Gene Finding

Identifying the protein coding region of genes

AAAGCATGCATTTAACGAGTGCATCAGGACTCCATACGTAATGCCG

Gene finder (many different programs)

AAAGC ATG CAT TTA ACG A GT GCATC AG GA CTC CAT ACG TAA TGCCG

*This is a eukaryotic gene as evidenced by the intron
Signals Within DNA

- Splice sites to identify intron/exon junctions
- Transcription start and stop codons
- Promoter regions
- PolyA signals
Experimental Evidence

**DNA sequence evidence:** Transcript sequence (EST, full length cDNA, other expression types); more restrictive in evolutionary terms

**Protein Evidence:** alignment to protein that suggests structural similarity at the amino acid level; can be more distant evolutionarily
Experimental Evidence

Transcript evidence:

-Demonstrates gene is transcribed

-Delineates exon boundaries

-Defines splice sites and alternative transcripts

-If EST based, indicates expression patterns
Functional Assignments

**Name**
Descriptive common name for the protein, with as much specificity as the evidence supports; gene symbol.

**Role**
Describe what the protein is doing in the cell and why.

**Associated information:**
Supporting evidence: Domain and motifs
EC number if protein is an enzyme.
Paralogous family membership.
Evidence for Gene Function

• PROSITE Motifs
  – collection of protein motifs associated with active sites, binding sites, etc.
  – help in classifying genes into functional families when HMMs for that family have not been built

• InterPro
  – Brings together HMMs (both TIGR and Pfam) Prosite motifs and other forms of motif/domain clustering
  – Results in motif “signatures” for families or functions
  – GO terms have been assigned to many of these
Sequence Alignments

Compare sequence against other databases

<table>
<thead>
<tr>
<th>EGAD</th>
<th>12983</th>
<th>biotin synthase Insert characterized</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMNI</td>
<td>TNL1</td>
<td>EC00751</td>
</tr>
<tr>
<td>GP1</td>
<td>8296</td>
<td>ECO001561</td>
</tr>
<tr>
<td>GP1</td>
<td>79010</td>
<td>3HAC78505</td>
</tr>
<tr>
<td>GP</td>
<td>90019</td>
<td>[1]</td>
</tr>
<tr>
<td>FIJ</td>
<td>J257</td>
<td>SYECBH</td>
</tr>
</tbody>
</table>

Sequence alignments display the comparison of biotin synthase sequences from different species, highlighting regions of similarity and identity. The annotations indicate whether certain insertions or modifications have been characterized in the sequences.
Gene function evidence

sec structure: coil(-), strand(blue), helix(yellow)

TIGR0433: biotin synthase
PF0405: radical SAM domain protein
Characterized match: SP:PI2996

sec structure: coil(-), strand(blue), helix(yellow)

IPR000515/PF00528: Binding-protein-dependent trans
IPR000515/PS00402: Binding-protein-dependent trans
Characterized match: SP:PI6702
Paralogous domain fam_PF00528
Paralogous domain fam_11
Gene Name Assignment: Based on similarity to known proteins in nraa database

Categories:

Known or Putative: Identical or strong similarity to documented gene(s) in Genbank or has high similarity to a Pfam domain; e.g. kinase, Rubisco

Expressed Protein: Only match is to an EST with an unknown function; thus have confirmation that the gene is expressed but still do not know what the gene does

Hypothetical Protein: Predicted solely by gene prediction programs and matches another hypothetical or expressed protein

Hypothetical Protein: Predicted solely by gene prediction programs; no database match
Annotation example

• A good example of this is seen with transporters, what you’ll see:
  – Multiple hits to a specific type of transporter
  – The substrate identified for the proteins your protein matches are not all the same, but fall into a group, for example they are all sugars.

• Give the protein a name with specific function but a more general substrate specificity:
  – “sugar ABC transporter, permease protein”

• Sometimes it will not be possible to identify particular substrate group, in that case:
  – “ABC transporter, permease protein”
Automated Annotation is Not a Solved Problem

What you are getting is output from a series of prediction tools or alignment programs

- Manual curation is often used to assess various types of evidence and improve upon automated gene calls and alignment output

- Ultimately, experimental verification is the only way to be sure that a gene structure is correct
Structural Annotation: Graphic Viewer Annotation Station

Sequence Database Hits
Top: Protein matches
Bottom: EST matches

Not shown graphically: gene name, nucleotide and protein sequence, MW, pI, organellar targeting sequence, membrane spanning regions, other domains.

Gene Predictions
Top: editing panel
Bottom: final curation

Splice site predictions:
red: acceptor sites
blue: donor sites

Screenshot of a component within Neomorphic’s annotation station: www.neomorphic.com
Features Typically Resolved During Manual Annotation

- incorrect exon boundaries
- merged, split, missing genes
- missing untranslated regions (UTRs)
- missing alternative splicing isoform annotations
- degenerate transposons annotated as protein-coding genes
## Increasing Complexity of Genome Annotation

<table>
<thead>
<tr>
<th>Genomic Organism</th>
<th># Genes</th>
<th>bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma pulmonis</td>
<td>780</td>
<td>964,000</td>
</tr>
<tr>
<td><em>Escherichia coli K-12</em></td>
<td>4,300</td>
<td>4,641,000</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>6,300</td>
<td>12,100,000</td>
</tr>
<tr>
<td><em>Plasmodium falciparum</em></td>
<td>5,400</td>
<td>22,850,000</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>19,000</td>
<td>97,000,000</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>16,000</td>
<td>120,000,000</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>25,000</td>
<td>115,400,000</td>
</tr>
<tr>
<td><em>Fugu rubripes</em></td>
<td>35,000</td>
<td>365,000,000</td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>30,000</td>
<td>2,910,000,000</td>
</tr>
</tbody>
</table>

Decrease in gene density and the presence of more, larger introns
Caveats of Genome Annotation

- Greatly impacted by the quality of the sequence; the impact of draft sequencing on whole genome annotation has yet to be seen by Joe/Jane Scientist. There will be disappointment when the research communities realize that they don’t have the “gold” standard of sequence as present in Arabidopsis and rice.

- Annotation is challenging, highly UNDER-estimated in difficulty, highly UNDER-valued until a community goes to use its genome sequence

- Annotation can be done to high accuracy on a single gene level by single investigators with expertise in gene families. The challenge is how to extrapolate this to the whole genome

- Blends of automated, semi-automated, and manual annotation is perhaps the best way to approach genomes in which there are not large communities

- Iterative, never perfect, can always be improved with new evidence and improved algorithms