Introduction to *ab initio* and evidence-based gene finding

Wilson Leung 01/2019

Outline

- Overview of computational gene predictions
- Different types of eukaryotic gene predictors
- Common types of gene prediction errors

Computational gene predictions

- Identify genes within genomic sequences
  - Protein-coding genes
  - Non-coding RNA genes
  - Regulatory regions (enhancers, promoters)
- Predictions must be confirmed experimentally
  - Eukaryotic gene predictions have *high error rates*
- Two major types of RefSeq records:
  - NM_/NP_ = experimentally confirmed
  - XM_/XP_ = computational predictions

Primary goal of computational gene prediction algorithms

- Label each nucleotide in a genomic sequence
- Identify the *most likely sequence of labels* (i.e. optimal path)

Sequence: TTCACACGTAAAGTAGTGTTGTA

<table>
<thead>
<tr>
<th>Path 1</th>
<th>Path 2</th>
<th>Path 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEEEEEEEEEIIIIIIIIIIIIIII</td>
<td>EEEEEEEEEIIIIIIIIIIIIIII</td>
<td>EEEEEEEEEIIIIIIIIIIIIII</td>
</tr>
</tbody>
</table>

Labels: Exon (E) 5’ Splice Site (S) Intron (I)

Basic properties of gene prediction algorithms

- Model must satisfy *biological constraints*
  - Coding region must begin with a start codon
  - Initial exon must occur before splice sites and introns
  - Coding region must end with a stop codon
- Model rules using a finite state machine (FSM)
- Use *species-specific characteristics* to improve the accuracy of gene predictions
  - Distribution of exon and intron sizes
  - Base frequencies (*e.g.*, GC content, codon bias)
  - Protein sequences from the same or closely related species

Prokaryotic gene predictions

- Prokaryotes have relatively simple gene structure
  - Single open reading frame
  - Alternative start codons: AUG, GUG, UUG
- Gene finders can predict most prokaryotic genes accurately (> 90% sensitivity and specificity)
  - Glimmer
Eukaryotic gene predictions have high error rates

- Gene finders generally do a poor job (<50%) predicting genes in eukaryotes
- More variations in the gene models
  - Alternative splicing (multiple isoforms)
  - Non-canonical splice sites (e.g., toy)
  - Non-canonical start codon (e.g., Fmr1)
  - Stop codon read through (e.g., gish)
  - Nested genes (e.g., ko)
  - Trans-splicing (e.g., mxd/mdg6)
  - Pseudogenes (e.g., swaPse)

Types of eukaryotic gene predictors

- Ab initio
  - GENSCAN, geneid, SNAP, GlimmerHMM
- Evidence-based (extrinsic)
  - Augustus, genBlastG, GeMoMa, Exonerate, GenomeScan
- Comparative genomics
  - Twinscan/N-SCAN, SGP2
- Transcriptome-based (RNA-Seq)
  - Cufflinks, StringTie, Trinity, CodingQuarry
- Combine ab initio and evidence-based approaches
  - GLEAN, Gnomon, JIGSAW, EVM, MAKER, Ipred

Ab initio gene prediction

- Ab initio = from the beginning
- Predict genes using only the genomic DNA sequence
  - Search for signals of protein coding regions
  - Based on a probabilistic model
    - Hidden Markov Models (HMM)
    - Support Vector Machines (SVM)
- GENSCAN

Hidden Markov Models (HMM)

- A type of supervised machine learning algorithm
- Uses Bayesian statistics
- Makes classifications based on characteristics of training data
- Many types of applications
  - Speech and gesture recognition
  - Bioinformatics
    - Gene predictions
    - Sequence alignments
    - Chip-seq analysis
    - Protein folding

Supervised machine learning

Use previous search results to predict search terms and correct spelling errors

GEP curriculum on HMM

- Use a HMM to predict a splice donor site
  - Use Excel to experiment with different emission and transition probabilities
- See the Beyond Annotation section of the GEP web site
  - Also available on CourseSource

**GENSCAN HMM Model**

- GENSCAN uses the following information to construct gene models:
  - Promoter, splice site and polyadenylation signals
  - Hexamer frequencies and base compositions
  - Probability of coding and non-coding DNA
  - Distribution of gene, exon and intron lengths


**Use multiple HMMs to describe different parts of a gene**


**Evidence-based gene predictions**

- Use sequence alignments to improve predictions
- EST, cDNA or protein from closely-related species

Exon sensitivity: Percent of real exons identified
Exon specificity: Percent of predicted exons that are correct


**Predictions using comparative genomics**

- Use whole genome alignments from one or more informant species
- CONTRAST predicts 50% of genes correctly
- Requires high quality whole genome alignments and training data

Flicek P. *Gene prediction: compare and CONTRAST*, Genome Biology (2007), 8, 233

**TopHat junction predictions from spliced RNA-Seq reads**

1. Build graph of incompatible RNA-Seq fragments
2. Identify minimum path cover (Dilworth’s theorem)
3. Assemble isoforms


**Cufflinks – reference-based transcriptome assembly**

- Use TransDecoder to identify coding regions within assembled transcripts

Generate consensus gene models

- Gene predictors have different strengths and weaknesses
- Create **consensus gene models** by combining results from multiple gene finders and sequence alignments
  - GLEAN
    - Eisik CG et al. Creating a honey bee consensus gene set. Genome Biology 2007, 8:R13
  - EvidenceModeler (EVM)

**Automated annotation pipelines**

NCBI Gnomon gene prediction pipeline

- Integrate **biological evidence** into the predicted gene models
  - Examples:
    - NCBI Gnomon
    - Ensembl
    - UCSC Gene Build
    - EGASP results for the Ensembl pipeline:
      - 71.6% gene sensitivity
      - 67.3% gene specificity


**Drosophila Gnomon gene predictions**

- Based on **RNA-Seq** data from either the same or closely-related species
- Predictions include **untranslated regions** and **multiple isoforms**
- Gnomon gene predictions are available through the NCBI RefSeq database:

**Common problems with gene finders**

- Split single gene into multiple predictions
- Fused with neighboring genes
- Missing exons
- Over predict exons or genes
- Missing isoforms

**Non-canonical splice donors and acceptors**

- Many gene predictors strongly prefer models with canonical splice donor (GT) and acceptor (AG) sites
- Check **Gene Record Finder** or FlyBase for genes that use non-canonical splice sites in *D. melanogaster*

<table>
<thead>
<tr>
<th>Introns with non-canonical splice sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcrip. ID</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Caden_D</td>
</tr>
</tbody>
</table>

**Annotate unusual features in gene models using *D. melanogaster* as a reference**

- Examine the “**Comments on Gene Model**” section of the FlyBase Gene Report

**Stop codon read through**

- Examine the “Stop codon read through” section of the FlyBase Gene Report

Non-canonical start codon:
Gene prediction results for the GEP annotation projects

- Gene prediction results are available through the GEP UCSC Genome Browser mirror
  - Under the **Genes and Gene Prediction Tracks** section
  - Access the predicted peptide sequence:
    - Click on the feature, and then click on the **Predicted Protein** link
  - Original gene predictor output available inside the **Genefinder** folder in the annotation package
    - The Genescan folder contains a PDF with a graphical schematic of the gene predictions

Questions?

https://flic.kr/p/6okjA W

Summary

- Gene predictors can quickly identify potentially interesting features within a genomic sequence
- The predictions are hypotheses that must be confirmed experimentally
- Eukaryotic gene predictors generally can accurately identify internal exons
- Much lower sensitivity and specificity when predicting complete gene models