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

Wilson Leung  08/2015

Outline

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

Computational gene predictions

- Identify genes in genomic sequence
  - 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: `TTTCACACGTAAGTATAGTGTGTGA`

<table>
<thead>
<tr>
<th>Path 1</th>
<th>Path 2</th>
<th>Path 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon (E)</td>
<td>Splice Site (S)</td>
<td>Intron (I)</td>
</tr>
</tbody>
</table>

Basic properties of gene prediction algorithms

- Predictions must satisfy a set of *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 these 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)

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
  - Alternate start codon (e.g., Fmr1)
  - Stop codon read through (e.g., gish)
  - Nested genes (e.g., b7)
  - Trans-splicing (e.g., mod(mdg4))
  - Pseudogenes

Types of eukaryotic gene predictors

- Ab initio
  - GENSCAN, geneid, SNAP, GlimmerHMM
- Evidence-based (extrinsic)
  - Augustus, genBlastG, Exonerate, GenomeScan
- Comparative genomics
  - Twinscan/N-SCAN, SGP2
- Combine ab initio and evidence-based approaches
  - EVM, GLEAN, Gnomon, JIGSAW, MAKER

Ab initio gene prediction

- Ab initio = from the beginning
- Predict genes using only the genomic DNA sequence
  - Search for signals of protein coding regions
  - Typically 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 aggregated data of previous search results to predict the search term and the correct spelling

GEP curriculum on HMM

- Developed by Anton Weisstein (Truman State University) and Zane Goodwin (TA for Bio 4342)
- 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
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

Yeh RF, et al., Computational Inference of Homologous Gene Structures in the Human Genome, Genome Res. (2001) 11, H5-H16

Augustus gene prediction service

http://bioinf.uni-greifswald.de/augustus/

Augustus [gene prediction]


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

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
  - GLEAN-R (reconciled) reference gene sets for 11 Drosophila species available at FlyBase
GLEAN-R prediction for the *ey* ortholog in *D. erecta*

- Single GLEAN-R prediction per genomic location
- Models have not been confirmed experimentally
- GLEAN-R RefSeq records have XM and XP prefixes

Automated annotation pipelines

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


New Gnomon predictions for eight *Drosophila* species

- Based on RNA-Seq data from either the same or closely-related species
  - *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. willistoni*, *D. virilis*, and *D. mojavensis*
- Predictions include intronless regions and multiple isoforms
- Records not yet 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 only predict genes 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*

Frequency of non-canonical splice sites in FlyBase Release 6.06 (Number of unique introns: 71,476)

<table>
<thead>
<tr>
<th>Intron type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC</td>
<td>594</td>
</tr>
<tr>
<td>AT</td>
<td>27</td>
</tr>
<tr>
<td>GA</td>
<td>44</td>
</tr>
</tbody>
</table>

Stop codon read-through: non-canonical start codon:

- GC

Stop codon suppression (UGA) predicted: FK5066884

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

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

Non-canonical start codon:

- GC stop codon

Stop codon suppression (UGA) predicted: FK5066884
Nested genes in *Drosophila*

A special type of RNA processing where exons from two primary transcripts are ligated together

**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 by clicking on the feature, then click on the *Predicted Protein* link

- Original gene predictor output available inside the folder *Genefinder* in the annotation package
  - The Censcan folder contains a PDF with a graphical schematic of the gene predictions

**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

**Questions?**

[Image: http://www.flickr.com/photos/cristinacosta/430498451/sizes/m/in/photostream/]