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

Wilson Leung  08/2023

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: TTTCACACGTAAGTATAGTGTGTGA
    - Path 1: $\text{E E E E E E E E E E E E E E E E E E E E}$
    - Path 2: $\text{E E E E E E E E E E E E E E E E E E E}$
    - Path 3: $\text{E E E E E E E E E E E E E E E E E E}$
- 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
  - NCBI Prokaryotic Genome Annotation Pipeline (PGAP)
    - Li W., et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation, NAR. (2021) 49(D1), D1020-D1028
    - [https://github.com/ncbi/pgap](https://github.com/ncbi/pgap)
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., mod(mdg4))
  - Pseudogenes (e.g., swaPsi)

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
  - Gnomon, MAKER, EVM, JIGSAW, Ipred, GLEAN

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

Norvig P. How to write a spelling corrector. https://www.norvig.com/spell-correct.html

GEP curriculum on HMM

- Use an HMM to predict a splice donor site
  - Use Excel to experiment with different emission and transition probabilities
- See the Curriculum section of the GEP website
  - Also available on CourseSource
Ways to create training sets to estimate transition and emission parameters

- Manually curated genes for the target species
- Bootstrap with ab initio gene predictions
  - GeneMark-ES, GENSCAN
- Sequence similarity to orthologs in informant species
  - BUSCO, BRAKER2
- Whole genome conservation profiles
  - Augustus-cgp, N-SCAN, SGP2
- RNA-Seq (splice junctions, assembled transcripts)
  - BRAKER1

GENSCAN HMM Model

- GENSCAN considers:
  - Promoter, splice sites and polyadenylation signals
  - Hexamer frequencies and base compositions
  - Probability of coding and non-coding DNA
  - Distributions of gene, exon and intron lengths


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


BRAKER training protocols

Training with genome assembly only

Training with proteins and RNA-Seq alignments


Use multiple HMMs to describe different parts of a gene


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
Intron predictions based on spliced RNA-Seq reads

Processed mRNA

RNA-Seq reads

Contig

Splice junctions

Cufflinks – reference-based transcriptome assembly

1. Build graph of incompatible RNA-Seq fragments

2. Identify minimum path cover (Dilworth’s theorem)

3. Assemble isoforms

Use TransDecoder to identify coding regions within assembled transcripts


Transcriptome assembly of RNA-Seq reads remains an area of active research

Sensitivity (%)

Precision (%)

StringTie

StringTie + SR

Cufflinks

Scripture

IsoLasso

Precision of assembling Illumina RNA-Seq reads into transcripts depends on expression levels

Improve transcript assembly using PacBio IsoSeq and Nanopore RNA-Seq

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

EVIDenceModeler (EVM)


TSEBRA

Gabriel L et al. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics (2021) 22(1), 566
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


Eukaryotic genomes annotated by NCBI

- RefSeq annotations available for more than 1000 species


Drosophila RefSeq 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:
  - https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/

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

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

- Examine the “Comments on Gene Model” and the “Sequence Ontology” sections of the FlyBase Gene Report

Non-canonical start codon:

- Check Gene Model: Gene model reviewed during ED
- Sequence Ontology: Class of Gene: Non-canonical start codon

Stop codon read through:

- Check Gene Model: Gene model reviewed during ED
- Sequence Ontology: Class of Gene: Non-canonical stop codon

Frequency of non-canonical splice sites in FlyBase Release 6.52 (Number of unique introns: 72,063)

<table>
<thead>
<tr>
<th>Donor site</th>
<th>Count</th>
<th>Acceptor site</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC</td>
<td>603</td>
<td>AC</td>
<td>34</td>
</tr>
<tr>
<td>AT</td>
<td>30</td>
<td>TG</td>
<td>28</td>
</tr>
<tr>
<td>GA</td>
<td>15</td>
<td>AT</td>
<td>16</td>
</tr>
</tbody>
</table>
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:
  - Click on the feature, and then click on the **Predicted Protein** link

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?

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