Introduction to \textit{ab initio} and evidence-based gene finding

Wilson Leung  
08/2018

Outline

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

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)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Path 1</th>
<th>Path 2</th>
<th>Path 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTTCACACGTAAGTATAGTGTA</td>
<td>EEEEEEEE S II II II II II II II</td>
<td>EEEEEEEEE S I I I I I I I I I</td>
<td>EEEEEEEEE EEEEEEEE S I I I I</td>
</tr>
</tbody>
</table>

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

Rosetta Stone

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)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Path 1</th>
<th>Path 2</th>
<th>Path 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTTCACACGTAAGTATAGTGTA</td>
<td>EEEEEEEE S II II II II II II II</td>
<td>EEEEEEEEE S I I I I I I I I I</td>
<td>EEEEEEEEE EEEEEEEE S I I I I</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., giib)
  - Nested genes (e.g., ko)
  - Trans-splicing (e.g., mxd/mdg64)
  - 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
  - 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

- Processed mRNA 5’ cap M * Poly-A tail
- RNA-Seq reads
- Contig/Intron
- TopHat 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

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

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

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>Donor site</th>
<th>Count</th>
<th>Acceptor site</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC</td>
<td>601</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>14</td>
<td>AT</td>
<td>18</td>
</tr>
</tbody>
</table>

Frequency of non-canonical splice sites in FlyBase Release 6.22 (Number of unique introns: 71,668)
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:

<table>
<thead>
<tr>
<th>Comment on Gene Model</th>
<th>Gene model released during 2011</th>
<th>Gene model released during 2014</th>
<th>Low-frequency RNA splicing potentially re-annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gene model released during 2015</td>
<td>Gene model released during 2016</td>
<td>Gene model released during 2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene model released during 2018</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Genes with stop codon read through are annotated</td>
</tr>
</tbody>
</table>

Stop codon read through:

<table>
<thead>
<tr>
<th>Comment on Gene Model</th>
<th>Gene model released during 2011</th>
<th>Gene model released during 2014</th>
<th>Low-frequency RNA splicing potentially re-annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gene model released during 2015</td>
<td>Gene model released during 2016</td>
<td>Gene model released during 2017</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Genes with stop codon read through are annotated</td>
</tr>
</tbody>
</table>

Trans-spliced gene 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
- Original gene predictor output available inside the *Genefinder* folder in the annotation package
- The Genscan 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?

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