Gene Prediction
CS795

Huzefa Rangwala
rangwala at cs dot gmu dot edu
Spring 2010
Gene Prediction

- Once the entire sequence of an organism has been obtained, the next major step is the identification of the regions of DNA that code for genes.
  - Gene prediction problem
- This is a sequence annotation problem:
  - Given the entire DNA sequence of an organism, mark the regions that code for genes.
- There is a large difference between the gene structure of prokaryotic and eukaryotic organisms.
  - Specific algorithms for each of them have been developed.
Central Dogma - Gene expression

DNA

transcription

RNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCACAUUUGAUUGAA

PEPTIDE
Background

- **Open Reading Frame – ORF**
  - **Definition**
    - An ORF is a chunk of a DNA sequence that contains a contiguous set of codons, each of which specifies an amino acid. This sequence is terminated by the stop codons.
  
- **Given a DNA sequence there are 6 possible ORFs**

```
ATGACGGATTACG
TACTGCCTAATGC
```
Background

- **Codon Bias**
  - Not all codons are used with the same frequency in the genes of a particular organism.
  - This bias introduces a certain nucleotide use bias that is apparent when viewed at the codon level.
  - Non-coding regions do not have this “codon-use” bias.

<table>
<thead>
<tr>
<th>Species</th>
<th>Codon</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>CCTGT</td>
<td>(Proline, Valine)</td>
</tr>
<tr>
<td>Mouse</td>
<td>CCAGTC</td>
<td>(Proline, Valine)</td>
</tr>
<tr>
<td>Rat</td>
<td>CCAGTC</td>
<td>(Proline, Valine)</td>
</tr>
<tr>
<td>Dog</td>
<td>CCGGTA</td>
<td>(Proline, Valine)</td>
</tr>
<tr>
<td>Chicken</td>
<td>CCCGTG</td>
<td>(Proline, Valine)</td>
</tr>
</tbody>
</table>
Distribution of AA usage

(A) 

(B) 

frequency

I N P A K G F R Q D L C Y S V T H M W E

- red: human
- yellow: fly
- blue: worm
- green: yeast
- purple: E. coli

- blue: gene
- cyan: intergene
Background

- A random DNA sequence should have a stop-codon on the average every 64 bases.
- True genes have a typical length of 450-900 nucleotides (150-300 amino acids).
Prokaryotic Gene Characteristics

- **Gene Structure**
  - A gene starts with ATG (translation start site) and ends with a stop codon.
  - It consists of a contiguous region of the DNA.

- Most of the DNA codes for a gene
  - Greater than 75-90%.
Prokaryotic Gene Prediction

- Overall Methodology
  - Build a model that learns the nucleotide use pattern of the coding and non-coding region.

- However….
  - How do you obtain a training set?

- Key Insight
  - Very long open reading frames are almost certain to be part of genes.
Approach

- **Overall Approach**
  - Scan the genome to identify long ORFs (> 1000 bases)
  - Use these ORFs as positive training and randomly generated sequences as negative training.
  - Build a Markov model-based model
    - Inhomogeneous Markov models
    - Interpolated Markov models
    - HMMs
  - Scan the DNA sequence and classify each ORF using this model.
  - Done!
Approach

- Characteristics
  - High accuracy rates (over 98%).
  - It is a self-contain approach as it does not require any pre-existing information.
  - Performance can be improved by using discriminating models.
- Examples:
  - GeneMark
    - Originally based on 5th order Markov chain and later improved by using HMMs
  - Glimmer
    - Based on interpolated inhomogeneous Markov models
      - 6+1 IIMM are built for classification
Refresher: MC, IMC, SHMM
How about Eukaryotic Genomes?

- It is an entirely different story…
Gene Structure – Eukaryotes

5’ | Exon 1 | Intron 1 | Exon 2 | Intron 2 | Exon 3 | Intron 3 | Exon 4 | 3’

DNA → pre-mRNA → mRNA → protein sequence

ATG - $X_1...X_n$ - STOP

TRANSCRIPTION

SPLICING

TRANSLATION
ORF Scanning Fails!

- Drosophila:
  - 3.4 introns per gene on average
  - mean intron length 475, mean exon length 397

- Human:
  - 8.8 introns per gene on average
  - mean intron length 4400, mean exon length 165

- ORF scanning is defeated
Eukaryotic genes - Signals

- Look at the structure - model
- Lengths of exon/introns
- Is there a particular composition in nucleotides - CpG Islands?
- Start Codons/Stop Codons/ Translation Specific Sites
- Look for more patterns - will make your algorithm better. Can we somehow use the intragenic region? Non-coding?
Eukaryotic Gene Characteristics

Gene Structure

- It is made of a number of exons separated by introns.
  - For human (~averages):
    - exons/gene = 8
    - exon length = 165
    - intron length = 5000
  - Average exon size of 150bp
  - Average intron size of 30Kbp
- Initial and final exons have different characteristics when compared to each other and the internal exons.
  - Initial/Final exons often contain UTR sequences and can be much shorter.
  - A gene starts with ATG and ends with a stop codon.
  - Average length of 1Kbp nucleotides.
Eukaryotic Gene Characteristics

- Most of the DNA does not code for a gene—less than 30%.
  - In human chromosome 6, there are about 1600 genes in 165 megabases.

- The gene density varies at different regions.
Eukaryotic Gene Prediction

- An “unsupervised” approach used for prokaryotic organisms does not work.
  - Why?

- Approaches used for gene prediction
  - Comparative genomics or Homology Based
    - Search against the gene structure of previously characterized organisms
    - Search against previously identified proteins
    - Multi-genome predictions
    - Take advantage of EST information

- Ab initio gene prediction
  - The rely on learning models based on a training set that was determined previously
    - either experimentally (ESTs) or using comparative approaches.
  - A number of approaches have been developed
    - dynamic programming, HMM, and discriminative approaches.

- Hybrid Approaches
  - Combination of the above two set of methods.
Ab initio Gene Prediction

- **Signals**
  - There is a distinct nucleotide usage pattern in the exons and intron/non-coding regions.
    - This is a result of the organisms codon use bias.
  - CpG islands and core promoter elements upstream of the genes
  - Poly-A tail at 3’ UTR.
  - There are weak sequence consensuses at the intron-exon & exon-intron boundaries.

![Diagram showing donor and acceptor sites](image-url)
Donor-Acceptor Sites

(A) exon

(B) bits

(C) bits

HUMAN

ARABIDOPSIS
Distribution of introns, exons
Approach

- Effective gene prediction algorithms
  - capture the various known signals using different “sensors”.
  - combine these signals based on the overall gene structure.
Approach

- These sensors are implemented using different models.
  - Exon: interpolated/inhomogeneous Markov chain
  - Acceptor/Donor sites: Profile HMMs or PSSMs
  - Intron: random model
- The sensors are combined together using different approaches
  - Dynamic programming
  - Hidden Markov model
  - Neural networks
- There is a notion of “parsing” the DNA sequence to identify its various components
  - annotation
GENSCAN

- It couples the various sensors using an HMM.
  - Individual sensors are implemented using probabilistic models
    - profile HMMs
    - IIMM
    - background probabilities
- Main elements
  - $E_{\text{init}}$, $E_{\text{term}}$ models for initial and final exons
  - $I_0$, $I_1$, $I_2$ models for introns
    - subscripts use to model where in the last codon of the previous exon the intro starts
      - codon phase
      - $I_1$ means that the intron starts after the first base of the last codon
  - $E_0$, $E_1$, $E_2$ models for internal exons
    - subscripts capture the same codon phase and this is why we can only transition from $I_i$ to $E_i$.
- It has been shown to be very effective.
A somewhat more detailed view of the HMM model
Combining the signals using Dynamic Programming

- Dynamic programming is a popular way of combining the signals of the various sensors
  - GeneFinder, GeneParser, Gaze, etc.

- Goal of the dynamic programming approach is to identify a “parse” of a DNA string into its various components (introns, exons, acceptor/donor sites, etc) such that it maximizes the score of the overall “parse”.

Example of Dynamic Programming-Based Gene Prediction

- We will focus on the following model of gene structure:

  - Consider a DNA sequence $X$ of length $L$ that we need to “parse”.

```
  ATG → Initial Exon → Internal Exon → Final Exon → STOP Codon
            |             |               |
            V             |               V
                           |               V
                          |               Introns
                           |               V
                          V               |               V
      Single Gene Exon

```
## DP-Based Gene Prediction

### Definitions:

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>internal_exon(i, j)</code></td>
<td>The score of $X_{i...j}$ representing an internal exon.</td>
</tr>
<tr>
<td><code>initial_exon(i, j)</code></td>
<td>The score of $X_{i...j}$ representing an initial exon.</td>
</tr>
<tr>
<td><code>terminal_exon(i, j)</code></td>
<td>The score of $X_{i...j}$ representing a terminal exon.</td>
</tr>
<tr>
<td><code>single_exon(i, j)</code></td>
<td>The score of $X_{i...j}$ representing a single-exon gene.</td>
</tr>
<tr>
<td><code>intron(i, j)</code></td>
<td>The score of $X_{i...j}$ representing an intron.</td>
</tr>
<tr>
<td><code>intergenic(i, j)</code></td>
<td>The score of $X_{i...j}$ representing an intergenic region.</td>
</tr>
</tbody>
</table>

- $S(j)$: Translation starts at $j$.
- $D(j)$: Donor at position $j$ ($j$ first intron base).
- $A(j)$: Acceptor at position $j$ ($j$ first exon base).
- $T(j)$: Stop codon at position $j$.

These represent the value of the best “parse” satisfying the specified condition.

So, what we are after:

$$\max_{i<L} \{ T(i) + \text{intergenic}(i, L) \}$$
DP-Based Gene Prediction

\[
\max_{i < L} \{ T(i) + \text{intergenic}(i, L) \}
\]

\[
T(j) = \max_{i < j} \left\{ \begin{array}{l}
A(i) + \text{terminal_exon}(i, j - 1) \\
S(i) + \text{single_exon}(i, j - 1)
\end{array} \right. \\
S(j) = \text{intergenic}(1, j - 1)
\]

\[
A(j) = \max_{i < j} \{ D(i) + \text{intron}(i, j - 1) \}
\]

\[
D(j) = \max_{i < j} \left\{ \begin{array}{l}
S(i) + \text{initial_exon}(i, j - 1) \\
A(i) + \text{internal_exon}(i, j - 1)
\end{array} \right.
\]

- **S(j)**: Translation starts at \( j \).
- **D(j)**: Donor at position \( j \) (\( j \) first intron base).
- **A(j)**: Acceptor at position \( j \) (\( j \) first exon base).
- **T(j)**: Stop codon at position \( j \).
Evaluation of Accuracy

**Sensitivity (SN)**  
Fraction of exons (coding nucleotides) whose boundaries are predicted exactly (that are predicted as coding)

**Specificity (Sp)**  
Fraction of the predicted exons (coding nucleotides) that are exactly correct (that are coding)

**Correlation Coefficient (CC)**  
Combined measure of Sensitivity & Specificity  
Range: -1 (always wrong) → +1 (always right)

\[
Sn = \frac{TP}{TP + FN} \\
Sp = \frac{TP}{TP + FP} \\
CC = \frac{(TP \times TN) - (FN \times FP)}{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}^{\frac{1}{2}}
\]
Evaluate

- Exons/introns
  - Boundaries (much harder)
- Promoters
- Entire gene structures
- Point is - different granularities
Personalized Medicine?
Readings

* Handout given during the lecture. Gene Prediction by Mount. Has good biological perspective and material for prokaryotic prediction.

* L. Stein, Genome Annotation: From Sequence to Biology, Nature Reviews Genetics, 2001. Excellent review. Defines the various annotation problems very clearly as well as the various solutions.

* A. Krogh, Gene Finding: Putting the parts together. Explains the dynamic programming method used widely in GeneParser and GeneFinding algorithm


* C. Burge and S. Karlin, Prediction of Complete Gene Structures in Human Genomic DNA (GENSCAN), JMB, 1997. Paper on GENSCAN. Be sure to read the METHODS section in this paper. Explains how HMMs are used to annotate and model entire gene structures.

Check the website for further optional readings!