Gene Finding And Gene Structure Prediction
Course 2005

Lorenzo Cerutti
Swiss Institute of Bioinformatics, Lausanne

Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
  - Software
- Homology methods
  - Principle of the method
  - Software
- Performance evaluation
The Central Dogma

Genetic code

<table>
<thead>
<tr>
<th>First letter</th>
<th>Second letter</th>
<th>Third letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td>UU</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td>UU</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>UU</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>UU</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>CU</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>CA</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>CUG</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td>AU</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td>AU</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>AU</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>AU</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>AUG</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td>GU</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>UG</td>
<td>U</td>
<td>UC</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>

### Reading frame

Forward strand:
- frame #1: ATG GTA ACA TGG C..
- frame #2: TGG TAA CAT GGC
- frame #3: GGT AAC ATG GC.

Reverse strand:
- frame #4: GCC ATG TTA CCA T..
- frame #5: CCA TGT TAC CAT
- frame #6: CAT GTT ACC AT.

### Prokaryotic gene structure

- **Simple gene structure**
- **Overlapping genes**
What is gene finding?

- Given a genomic DNA sequence we want to predict regions encoding for a protein: the genes.

- Gene finding consists in:
  - identify the coding potential of a region in a particular frame
  - identify boundaries between coding and non-coding regions
Gene finding: not an easy task!

- DNA sequence signals have a low information content
- DNA signals may vary in different organisms
- difficult to discriminate real signals from noise
- gene structure can be complex (sparse exons, alternative splicing, ...)
- pseudo-genes
- sequence errors (frame shifts, ...)
- Human genome: 3 billion base pairs and 35,000 protein-coding genes

Gene finding strategies (1)

- Ab initio methods:
  - signals: short DNA motifs (promoters, start/stop codons, splice sites, ...)
  - coding statistics: nucleotide compositional bias in coding regions
- Strengths:
  - easy to run and fast execution
  - only require DNA sequence as input
- Weaknesses:
  - prior knowledge is required (training set)
  - high number of mispredicted gene structures
Gene finding strategies (2)

- Homology methods:
  - gene structure is deduced using homologous sequences (ESTs, mRNAs, proteins)
  - accurate results when using close homologous sequences
- Strengths:
  - accurate
- Weaknesses:
  - need of homologous sequences
  - slow execution

Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
  - Software
- Homology methods
  - Principle of the method
  - Software
- Performance evaluation
Overview of ab initio methods

Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
  - Software
- Homology methods
  - Principle of the method
  - Software
- Performance evaluation
Coding statistics

- Inter-genic regions, introns, and exons have different nucleotide contents
- Example: observed stop codons (TAG, TAA, TGA)
  - assuming an uniform random distribution, we expect stop codons every 64/3 codons (∼ 21 codons) in average
  - in coding regions the appearance of stop codons decrease
  - ... but, this measure is sensitive to frame shift errors and can’t detect short coding regions

Coding statistics: dimers frequencies

- Dimer frequencies observed in proteins from Shewanella (avg ∼ 5%):

<table>
<thead>
<tr>
<th></th>
<th>ala</th>
<th>asn</th>
<th>arg</th>
<th>asp</th>
<th>cys</th>
<th>glu</th>
<th>gly</th>
<th>glu</th>
<th>his</th>
<th>ile</th>
<th>leu</th>
<th>met</th>
<th>lys</th>
<th>phe</th>
<th>pro</th>
<th>ser</th>
<th>thr</th>
<th>tyr</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>ala</td>
<td>9.5</td>
<td>4.3</td>
<td>4.1</td>
<td>5.3</td>
<td>1.2</td>
<td>6.0</td>
<td>6.5</td>
<td>4.8</td>
<td>2.0</td>
<td>6.5</td>
<td>11.0</td>
<td>2.6</td>
<td>6.0</td>
<td>3.7</td>
<td>3.5</td>
<td>6.2</td>
<td>1.1</td>
<td>5.0</td>
<td>2.7</td>
</tr>
<tr>
<td>arg</td>
<td>7.9</td>
<td>3.9</td>
<td>5.5</td>
<td>5.3</td>
<td>1.1</td>
<td>5.0</td>
<td>5.9</td>
<td>5.5</td>
<td>2.6</td>
<td>6.5</td>
<td>14.4</td>
<td>2.2</td>
<td>5.0</td>
<td>4.7</td>
<td>3.6</td>
<td>5.5</td>
<td>1.4</td>
<td>4.4</td>
<td>4.0</td>
</tr>
<tr>
<td>asn</td>
<td>9.6</td>
<td>4.2</td>
<td>4.9</td>
<td>4.9</td>
<td>1.0</td>
<td>5.3</td>
<td>7.4</td>
<td>5.6</td>
<td>2.3</td>
<td>6.0</td>
<td>10.2</td>
<td>2.0</td>
<td>4.9</td>
<td>3.5</td>
<td>5.1</td>
<td>6.1</td>
<td>1.5</td>
<td>5.8</td>
<td>1.1</td>
</tr>
<tr>
<td>asp</td>
<td>9.3</td>
<td>4.7</td>
<td>6.0</td>
<td>5.1</td>
<td>1.0</td>
<td>6.7</td>
<td>7.0</td>
<td>2.9</td>
<td>1.8</td>
<td>7.1</td>
<td>9.6</td>
<td>2.3</td>
<td>6.3</td>
<td>4.3</td>
<td>3.9</td>
<td>5.9</td>
<td>1.6</td>
<td>5.1</td>
<td>1.6</td>
</tr>
<tr>
<td>cys</td>
<td>8.4</td>
<td>3.1</td>
<td>4.8</td>
<td>5.4</td>
<td>1.7</td>
<td>5.6</td>
<td>8.1</td>
<td>5.2</td>
<td>4.2</td>
<td>5.4</td>
<td>10.2</td>
<td>1.8</td>
<td>3.8</td>
<td>4.2</td>
<td>4.6</td>
<td>5.3</td>
<td>1.6</td>
<td>4.4</td>
<td>2.4</td>
</tr>
<tr>
<td>glu</td>
<td>9.4</td>
<td>3.6</td>
<td>5.8</td>
<td>4.5</td>
<td>0.8</td>
<td>4.9</td>
<td>5.8</td>
<td>7.0</td>
<td>2.6</td>
<td>5.9</td>
<td>12.7</td>
<td>2.4</td>
<td>5.0</td>
<td>4.0</td>
<td>3.5</td>
<td>5.4</td>
<td>1.1</td>
<td>5.0</td>
<td>2.8</td>
</tr>
<tr>
<td>gly</td>
<td>10.3</td>
<td>3.9</td>
<td>4.9</td>
<td>4.4</td>
<td>0.9</td>
<td>4.5</td>
<td>7.0</td>
<td>6.8</td>
<td>2.7</td>
<td>5.5</td>
<td>12.4</td>
<td>2.0</td>
<td>4.1</td>
<td>3.9</td>
<td>3.8</td>
<td>5.8</td>
<td>1.4</td>
<td>5.3</td>
<td>1.0</td>
</tr>
<tr>
<td>his</td>
<td>8.1</td>
<td>3.9</td>
<td>4.8</td>
<td>5.1</td>
<td>1.2</td>
<td>6.0</td>
<td>6.4</td>
<td>4.6</td>
<td>2.4</td>
<td>4.8</td>
<td>10.1</td>
<td>2.7</td>
<td>5.8</td>
<td>4.8</td>
<td>4.8</td>
<td>5.8</td>
<td>1.4</td>
<td>5.1</td>
<td>1.7</td>
</tr>
<tr>
<td>ile</td>
<td>7.3</td>
<td>3.9</td>
<td>4.7</td>
<td>5.4</td>
<td>1.5</td>
<td>6.9</td>
<td>6.9</td>
<td>5.6</td>
<td>3.0</td>
<td>6.2</td>
<td>10.8</td>
<td>1.6</td>
<td>4.8</td>
<td>5.0</td>
<td>5.2</td>
<td>6.8</td>
<td>1.7</td>
<td>4.9</td>
<td>2.2</td>
</tr>
<tr>
<td>leu</td>
<td>11.5</td>
<td>4.9</td>
<td>7.8</td>
<td>6.5</td>
<td>2.1</td>
<td>6.9</td>
<td>7.2</td>
<td>3.6</td>
<td>2.1</td>
<td>5.3</td>
<td>8.6</td>
<td>1.8</td>
<td>5.3</td>
<td>3.2</td>
<td>4.2</td>
<td>7.0</td>
<td>0.9</td>
<td>5.8</td>
<td>2.9</td>
</tr>
<tr>
<td>lys</td>
<td>10.4</td>
<td>4.3</td>
<td>6.2</td>
<td>5.2</td>
<td>1.1</td>
<td>5.2</td>
<td>6.8</td>
<td>3.7</td>
<td>2.0</td>
<td>5.6</td>
<td>10.4</td>
<td>2.3</td>
<td>5.3</td>
<td>3.8</td>
<td>4.5</td>
<td>7.4</td>
<td>1.0</td>
<td>6.2</td>
<td>2.6</td>
</tr>
<tr>
<td>met</td>
<td>10.4</td>
<td>3.8</td>
<td>6.2</td>
<td>5.2</td>
<td>1.0</td>
<td>5.3</td>
<td>6.6</td>
<td>5.3</td>
<td>2.6</td>
<td>5.2</td>
<td>11.1</td>
<td>1.9</td>
<td>4.7</td>
<td>2.8</td>
<td>4.6</td>
<td>6.0</td>
<td>1.2</td>
<td>5.8</td>
<td>2.6</td>
</tr>
<tr>
<td>phe</td>
<td>10.8</td>
<td>3.8</td>
<td>4.8</td>
<td>4.6</td>
<td>0.7</td>
<td>4.6</td>
<td>7.0</td>
<td>4.9</td>
<td>1.7</td>
<td>4.7</td>
<td>11.4</td>
<td>2.8</td>
<td>5.2</td>
<td>3.3</td>
<td>5.1</td>
<td>7.4</td>
<td>0.9</td>
<td>6.3</td>
<td>2.0</td>
</tr>
<tr>
<td>pro</td>
<td>9.6</td>
<td>5.2</td>
<td>3.7</td>
<td>6.5</td>
<td>1.2</td>
<td>6.4</td>
<td>7.9</td>
<td>2.7</td>
<td>1.9</td>
<td>6.7</td>
<td>7.4</td>
<td>2.5</td>
<td>5.0</td>
<td>3.9</td>
<td>3.6</td>
<td>8.0</td>
<td>1.3</td>
<td>5.8</td>
<td>3.3</td>
</tr>
<tr>
<td>ser</td>
<td>8.4</td>
<td>4.6</td>
<td>3.6</td>
<td>5.4</td>
<td>0.7</td>
<td>7.6</td>
<td>5.4</td>
<td>5.2</td>
<td>2.3</td>
<td>6.1</td>
<td>11.2</td>
<td>2.4</td>
<td>5.5</td>
<td>4.2</td>
<td>2.8</td>
<td>6.5</td>
<td>1.4</td>
<td>5.4</td>
<td>2.9</td>
</tr>
<tr>
<td>thr</td>
<td>9.1</td>
<td>3.7</td>
<td>6.5</td>
<td>5.0</td>
<td>1.0</td>
<td>5.4</td>
<td>7.2</td>
<td>5.2</td>
<td>2.6</td>
<td>6.0</td>
<td>11.4</td>
<td>2.2</td>
<td>4.5</td>
<td>4.1</td>
<td>4.1</td>
<td>4.6</td>
<td>1.5</td>
<td>5.0</td>
<td>1.2</td>
</tr>
<tr>
<td>trp</td>
<td>9.1</td>
<td>3.7</td>
<td>4.2</td>
<td>5.6</td>
<td>0.9</td>
<td>5.7</td>
<td>7.5</td>
<td>5.7</td>
<td>2.2</td>
<td>5.5</td>
<td>12.2</td>
<td>2.0</td>
<td>4.2</td>
<td>3.5</td>
<td>5.9</td>
<td>6.2</td>
<td>1.1</td>
<td>5.3</td>
<td>2.6</td>
</tr>
<tr>
<td>tyr</td>
<td>7.9</td>
<td>3.6</td>
<td>6.5</td>
<td>4.9</td>
<td>1.2</td>
<td>4.5</td>
<td>7.1</td>
<td>7.0</td>
<td>2.6</td>
<td>3.0</td>
<td>11.7</td>
<td>1.6</td>
<td>4.5</td>
<td>4.7</td>
<td>4.9</td>
<td>6.4</td>
<td>1.5</td>
<td>4.6</td>
<td>1.4</td>
</tr>
<tr>
<td>val</td>
<td>9.6</td>
<td>4.4</td>
<td>4.1</td>
<td>5.9</td>
<td>1.0</td>
<td>6.2</td>
<td>6.4</td>
<td>3.4</td>
<td>1.8</td>
<td>6.5</td>
<td>10.2</td>
<td>2.5</td>
<td>5.2</td>
<td>3.7</td>
<td>3.8</td>
<td>7.2</td>
<td>1.1</td>
<td>6.1</td>
<td>2.7</td>
</tr>
</tbody>
</table>
Coding statistics: dimers to dicodons

- A bias is observed for dimers in proteins
- The dimer bias is reflected in dicodons: coding and non-coding regions have different dicodons bias.
- The bias in the observed dicodon (hexamer) frequencies can be used to predict coding regions in genomic sequences.
- All gene finding programs use this information!

Coding statistics: scoring (1)

Let $f_{abc,a'b'c'}^c$ denote the observed frequency for dicodon $abc, a'b'c'$ in a set of known coding regions, and let $f_{abc,a'b'c'}^n$ denote the observed frequency for the same dicodon in non-coding regions.

The score of dicodon $abc, a'b'c'$ in being coding is defined as:

$$P(abc, a'b'c') = \log\left(\frac{f_{abc,a'b'c'}^c}{f_{abc,a'b'c'}^n}\right)$$


**Coding statistics: scoring (2)**

- Properties of $P(abc, a' b' c')$:
  - if $P(abc, a' b' c') = 0$: dicodon $abc, a' b' c'$ has the same frequencies in coding and non-coding regions
  - if $P(abc, a' b' c') > 0$: dicodon $abc, a' b' c'$ is observed more frequently in coding regions
  - if $P(abc, a' b' c') < 0$: dicodon $abc, a' b' c'$ is observed more frequently in non-coding regions

**Coding statistics: scoring (3)**

- Assume $S = a_1 b_1 c_1, a_2 b_2 c_2, ..., a_{n+1} b_{n+1} c_{n+1}$ is a coding region with unknown reading frame.

- We can calculate the score of each frame of being coding:

  \[ P_1 = P(a_1 b_1, c_1 a_2 b_2 c_2) + P(a_3 b_3 c_3, a_4 b_4 c_4) + ... + P(a_{n-1} b_{n-1} c_{n-1}, a_n b_n c_n) \]

  \[ P_2 = P(b_1 c_1 a_2, b_2 c_2 a_3) + P(b_3 c_3 a_4, b_4 c_4 a_5) + ... + P(b_{n-1} c_{n-1} a_n, b_n c_n a_{n+1}) \]

  \[ P_3 = P(c_1 a_2 b_2, c_2 a_3 b_3) + P(c_3 a_4 b_4, c_4 a_5 b_5) + ... + P(c_{n-1} a_n b_n, c_n a_{n+1}, b_{n+1}) \]
Coding statistics: example

- \( f^c(ACG, TAG) = 0.000, \ f^b(ACG, TAG) = 0.062 \)
- \( f^c(CGT, AGC) = 0.068, \ f^b(CGT, AGC) = 0.019 \)
- \( f^c(GTA, GCT) = 0.021, \ f^b(GTA, GCT) = 0.026 \)
- \( P(ACG, TAG) = -\infty \) (special case STOP codon)
- \( P(CGT, AGC) = \log(0.068 / 0.019) = 1.3 \)
- \( P(GTA, GCT) = \log(0.021 / 0.026) = -0.2 \)

Coding statistics: scoring (4)

- Procedure for predicting coding regions using coding statistics:
  - find all ORFs of the sequence (start/stop regions)
  - slide through the ORFs with a window of 60bp and find good scoring regions
Coding statistics: limitations

- Which solution is the best?
- Where are the coding region boundaries?
- Where we put the score cutoff to consider a region as coding?
  - low cutoff ⇒ high number of false positives
  - high cutoff ⇒ high number of false negatives

Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
  - Software
- Homology methods
  - Principle of the method
  - Software
- Performance evaluation
Signals

- Detection of signals in DNA sequences helps in detecting the correct coding regions
- A number of signals can be used:
  - promoter regions
  - acceptor/donor sites for splicing
  - intron branching points
  - poly-adenilation
  - ...

Signals: limitations

- DNA sequence signals have a low information content and can be degenerated
- ... to use together with coding statistics
Methods for signal detection

- **Pattern**: flexible consensus string
- **Weight matrix**: position specific scoring matrix
- **HMMs**: Hidden Markov Models
- **NN**: Neural networks (trained with TP/TN!)
  - The perceptron for acceptor site (Horton and Kanehisa, 1992):
    \[
    \begin{array}{cccccccc}
    T & A & C & G & C & G & T & A \\
    0100 & 0001 & 0001 & 0010 & 0010 & 1000 & 0100 & 0010 \\
    \end{array}
    \]
    \[
    \begin{array}{cccccccc}
    w1 & w2 & w3 & w4 & w5 & w6 & w7 & w8 \\
    1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    \end{array}
    \]
    
    \[
    \begin{array}{cccccccc}
    1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    \end{array}
    \]
    ~0 => false
    ~1 => true

Acceptor/Donor signals (1)

- Distribution observed for donor sites in human:
  - Table

<table>
<thead>
<tr>
<th></th>
<th>-14</th>
<th>-13</th>
<th>-12</th>
<th>-11</th>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>11.1</td>
<td>12.7</td>
<td>3.2</td>
<td>4.8</td>
<td>12.7</td>
<td>8.7</td>
<td>16.7</td>
<td>16.7</td>
<td>12.7</td>
<td>9.5</td>
<td>26.2</td>
<td>6.3</td>
<td>100</td>
<td>0.0</td>
<td>21.4</td>
</tr>
<tr>
<td>c</td>
<td>36.5</td>
<td>30.9</td>
<td>19.1</td>
<td>23.0</td>
<td>34.9</td>
<td>39.7</td>
<td>40.5</td>
<td>40.5</td>
<td>36.5</td>
<td>33.3</td>
<td>68.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
<tr>
<td>g</td>
<td>9.5</td>
<td>10.3</td>
<td>15.1</td>
<td>12.7</td>
<td>8.7</td>
<td>9.5</td>
<td>16.7</td>
<td>4.8</td>
<td>2.4</td>
<td>6.3</td>
<td>13.5</td>
<td>0.0</td>
<td>0.0</td>
<td>100</td>
<td>62.7</td>
</tr>
<tr>
<td>t</td>
<td>38.9</td>
<td>41.3</td>
<td>58.7</td>
<td>55.6</td>
<td>42.1</td>
<td>40.5</td>
<td>30.9</td>
<td>37.3</td>
<td>44.4</td>
<td>47.6</td>
<td>27.0</td>
<td>25.4</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
</tbody>
</table>

- Distribution observed for acceptor sites in human:
  - Table
Acceptor/Donor signals (2)

- Information content:

\[ I_j = \left| \sum_i -f(i,j) \ast \log(f(i,j)/q(i)) \right| \]

where \( i = \{a, c, g, t\} \), \( j \) is the position (column), \( f(i,j) \) is the observed frequency for symbol \( i \) at position \( j \), and \( q(i) \) is the distribution of symbol \( i \) (in our case \( q(i) = 0.25 \)).

- A column with evenly distributed nucleotides has a low information content
- A column with unevenly distributes nucleotides has a higher information content

Acceptor/Donor signals (3)

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>34.0</td>
<td>60.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>52.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>c</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.5</td>
</tr>
<tr>
<td>g</td>
<td>18.3</td>
<td>12.5</td>
<td>80.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>t</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.3</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

\[ I_{-3} = \left| \sum_i -f(i,-3) \ast \log_2(f(i,-3)/0.25) \right| \]

\[ = \left| -0.34 \ast \log_2(0.34/0.25) \right| \\
-0.363 \ast \log_2(0.363/0.25) \\
-0.183 \ast \log_2(0.183/0.25) \\
-0.114 \ast \log_2(0.114/0.25) \right| \]

\[ = 0.13 \]

\[ I_{+1} = \left| \sum_i -f(i,+1) \ast \log_2(f(i,+1)/0.25) \right| = 2 \]
### Acceptor/Donor signals (4)

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>34.0</td>
<td>60.4</td>
<td>9.2</td>
<td>0.0</td>
<td>0.0</td>
<td>52.6</td>
<td>71.3</td>
<td>7.1</td>
<td>16.0</td>
</tr>
<tr>
<td>c</td>
<td>36.3</td>
<td>12.9</td>
<td>3.3</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
<td>7.6</td>
<td>5.5</td>
<td>16.5</td>
</tr>
<tr>
<td>g</td>
<td>18.3</td>
<td>12.5</td>
<td>80.3</td>
<td>100</td>
<td>0.0</td>
<td>41.9</td>
<td>11.8</td>
<td>81.4</td>
<td>20.9</td>
</tr>
<tr>
<td>t</td>
<td>11.4</td>
<td>14.2</td>
<td>7.3</td>
<td>0.0</td>
<td>100</td>
<td>2.5</td>
<td>9.3</td>
<td>5.9</td>
<td>46.2</td>
</tr>
</tbody>
</table>

### Acceptor/Donor signals (5)

<table>
<thead>
<tr>
<th></th>
<th>-14</th>
<th>-13</th>
<th>-12</th>
<th>-11</th>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>11.1</td>
<td>12.7</td>
<td>3.2</td>
<td>4.8</td>
<td>12.7</td>
<td>8.7</td>
<td>16.7</td>
<td>16.7</td>
<td>12.7</td>
<td>9.5</td>
<td>26.2</td>
<td>6.3</td>
<td>100</td>
<td>0.0</td>
<td>21.4</td>
</tr>
<tr>
<td>c</td>
<td>36.5</td>
<td>30.9</td>
<td>19.1</td>
<td>23.0</td>
<td>34.9</td>
<td>39.7</td>
<td>34.9</td>
<td>40.5</td>
<td>40.5</td>
<td>36.5</td>
<td>33.3</td>
<td>68.2</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
<tr>
<td>g</td>
<td>9.5</td>
<td>10.3</td>
<td>15.1</td>
<td>12.7</td>
<td>8.7</td>
<td>9.5</td>
<td>16.7</td>
<td>4.8</td>
<td>2.4</td>
<td>6.3</td>
<td>13.5</td>
<td>0.0</td>
<td>0.0</td>
<td>100</td>
<td>62.7</td>
</tr>
<tr>
<td>t</td>
<td>38.9</td>
<td>41.3</td>
<td>58.7</td>
<td>55.6</td>
<td>42.1</td>
<td>40.5</td>
<td>30.9</td>
<td>37.3</td>
<td>44.4</td>
<td>47.6</td>
<td>27.0</td>
<td>25.4</td>
<td>0.0</td>
<td>0.0</td>
<td>7.9</td>
</tr>
</tbody>
</table>
Detection of signals

- Build model (i.e. weigh matrix) for splicing sites, poly-A sites, ...
- Take in consideration positions with high information content
- Transform frequencies to scores
- Scan the sequence with the weight matrix

<table>
<thead>
<tr>
<th></th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.4</td>
<td>1.3</td>
<td>-1.4</td>
<td>inf</td>
<td>-inf</td>
<td>1.1</td>
<td>1.5</td>
<td>-1.8</td>
<td>-0.6</td>
</tr>
<tr>
<td>c</td>
<td>0.5</td>
<td>-1.0</td>
<td>-2.9</td>
<td>inf</td>
<td>inf</td>
<td>-3.2</td>
<td>-1.7</td>
<td>-2.2</td>
<td>-0.6</td>
</tr>
<tr>
<td>g</td>
<td>-0.5</td>
<td>-1.0</td>
<td>1.7</td>
<td>2</td>
<td>inf</td>
<td>0.7</td>
<td>-1.3</td>
<td>1.7</td>
<td>-0.3</td>
</tr>
<tr>
<td>t</td>
<td>-1.1</td>
<td>-0.8</td>
<td>-1.8</td>
<td>inf</td>
<td>2</td>
<td>-3.3</td>
<td>-1.4</td>
<td>-2.1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

CAGGTAAAGT: $0.5 + 1.3 + 1.7 + 2 + 2 + 1.1 + 1.5 + 1.7 + 0.9 = 12.7$

TCCGTTCGA: $-1.1 - 1 - 2.9 + 2 + 2 - 3.3 - 1.7 - 2.2 - 0.6 = -8.8$

EPD: Eukaryotic promoter database

- Annotated non-redundant collection of eukaryotic POL II promoters
- Contains promoters for which the transcription site has been determined experimentally
- WEB access: www.epd.isb-sib.ch
EPD entry

General information about the entry
Entry name: EPD_NC2
Entry type: standard
Promoter type: multiple
Accession number: EPD_NC2
Description of the gene: Nuclei.
Creation date: 19-SEP-1993 (Rel. 56)
Last revision: 13-SEP-2004 (Rel. 80)
Taxonomic division: VRT
Organism: Homo sapiens (human)
Keywords: Nuclei genes, Phosphorylation, Methylation, DNA-binding.

Reference with other entries
Biological group: Biological group 212: Mammalian nuclei.
Alternative promoter: none.
Neighbouring gene(s): none.

ORF strategy
GENOME
NT_00463.13
NT_004639
[8538464, 9974646]

EPD
Ns NC2

DNA References
AC017104 [EMBL GeSeq_DB] AG017104 [EMBL GeSeq_DB]
M60501 [EMBL GenBank_DB] [EMBL GenBank_DB]
A339561 [EMBL GeSeq_DB] A339560 [EMBL GeSeq_DB]
A339560 [EMBL GeSeq_DB] A339560 [EMBL GeSeq_DB]
A339560 [EMBL GeSeq_DB] A339560 [EMBL GeSeq_DB]

SWISSPROT
PL1555 [PEGMS\_HUMAN]

References
MEDLINE:2522
1. Siddhara S., Siddhara O.W., Vining F., Pollet H.B., ferres A.L.
   "Genomic organization of the human nuclear protein gene"
MEDLINE:1178528
2. Suzuki T., Yamauchi R., Nakano K., Sugano S.
   "MDF: database of mouse transcriptional start sites and full-length cDNAs".
   Nucleic Acids Res. 30:328-331 (2002).
MEDLINE:942157
3. Beane S., Beane S., Beane S., Beane S.
   "The mammalian genome collection"

Sequence
Nucleotide sequence:
[Insert nucleotide sequence information]

Method(s):
Nucleic sequence (1)
Primer extension with homologous sequence (1)
Northern blot (1)
Southern blot (1)
Bacterial expression (1)

Taxonomy
6.1. Chromosomal genes
6.1.2. Structural proteins
6.1.2.3. RNA-binding proteins
6.1.2.3.1. Nucleic RNA-binding proteins

Supplementary information
[Insert supplementary information]

Localization
[Insert localization information]
Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
  - Software
- Homology methods
  - Principle of the method
  - Software
- Performance evaluation

Coding statistics and signals
Coding statistics and signals

- A number of methods exist to integrate predicted signals and coding signals.
- All these methods are classifiers based on machine learning theory.
- Training sets are required to extract coding statistic and signal information.

Generalized HMMs

![Diagram of Genomic DNA, Exon, Intron, Begin, and End states with Predicted gene structure]
Generalized HMMs (2)

Example: GENSCAN model
Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
  - Software
- Homology methods
  - Principle of the method
  - Software
- Performance evaluation

**GENSCAN: form**

Organism: [Vertebrate] | Suboptimal exon cutoff (optional): 1.00 |
---
Sequence name (optional): 
Print options: Predicted peptides only |
Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):  
Browse... 
Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored): 

To have the results mailed to you, enter your email address here (optional): 

Run GENSCAN  Clear Input...
**GENSCAN: output (1)**

- **WEB server:** [http://genes.mit.edu/GENSCAN.html](http://genes.mit.edu/GENSCAN.html)
- **Models for vertebrates, Arabidopsis, Maize**

<table>
<thead>
<tr>
<th>Gn.Ex</th>
<th>Type</th>
<th>S</th>
<th>Begin</th>
<th>End</th>
<th>Len</th>
<th>Fr</th>
<th>Ph</th>
<th>I/Ac</th>
<th>Do/T</th>
<th>CodRg</th>
<th>P</th>
<th>Tscr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Prom</td>
<td>+</td>
<td>1653</td>
<td>1692</td>
<td>40</td>
<td>-</td>
<td></td>
<td>1</td>
<td>75</td>
<td>151</td>
<td>0.925</td>
<td>12.64</td>
</tr>
<tr>
<td>1.01</td>
<td>Init</td>
<td>+</td>
<td>5215</td>
<td>5266</td>
<td>52</td>
<td>0</td>
<td>1</td>
<td>83</td>
<td>75</td>
<td>163</td>
<td>0.895</td>
<td>15.02</td>
</tr>
<tr>
<td>1.02</td>
<td>Intr</td>
<td>+</td>
<td>5395</td>
<td>5562</td>
<td>168</td>
<td>2</td>
<td>0</td>
<td>89</td>
<td>75</td>
<td>163</td>
<td>0.895</td>
<td>15.02</td>
</tr>
<tr>
<td>1.03</td>
<td>Intr</td>
<td>+</td>
<td>11738</td>
<td>11899</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>74</td>
<td>113</td>
<td>101</td>
<td>0.990</td>
<td>11.15</td>
</tr>
<tr>
<td>1.04</td>
<td>Intr</td>
<td>+</td>
<td>12188</td>
<td>12424</td>
<td>237</td>
<td>0</td>
<td>0</td>
<td>71</td>
<td>86</td>
<td>197</td>
<td>0.662</td>
<td>15.39</td>
</tr>
<tr>
<td>1.05</td>
<td>Intr</td>
<td>+</td>
<td>14288</td>
<td>14623</td>
<td>336</td>
<td>0</td>
<td>0</td>
<td>82</td>
<td>98</td>
<td>263</td>
<td>0.986</td>
<td>22.19</td>
</tr>
<tr>
<td>1.06</td>
<td>Intr</td>
<td>+</td>
<td>17003</td>
<td>17203</td>
<td>201</td>
<td>0</td>
<td>0</td>
<td>116</td>
<td>86</td>
<td>102</td>
<td>0.976</td>
<td>12.06</td>
</tr>
<tr>
<td>1.07</td>
<td>Intr</td>
<td>+</td>
<td>17741</td>
<td>17859</td>
<td>119</td>
<td>0</td>
<td>2</td>
<td>78</td>
<td>109</td>
<td>51</td>
<td>0.984</td>
<td>6.38</td>
</tr>
<tr>
<td>1.08</td>
<td>Intr</td>
<td>+</td>
<td>18197</td>
<td>18264</td>
<td>68</td>
<td>1</td>
<td>2</td>
<td>103</td>
<td>72</td>
<td>81</td>
<td>0.541</td>
<td>5.70</td>
</tr>
</tbody>
</table>

>02:36:44|GENSCAN_predicted_peptide_1|448_aa
MCKAIILRLILLIIHQQVQTLQGKIVLGLGSVKEAEFPCTESKQKTVFPTWKFSDFQA
KILQSQGKVLIRGGSQPSQDFPRFSSKMGAVGSFPLIINLKMHDGQTYICEILRNKCEE
...

**Gn.Ex:** gene number, exon number (for reference)

**Type:**
- Init = Initial exon (ATG to 5' splice site)
- Intr = Internal exon (3' splice site to 5' splice site)
- Term = Terminal exon (3' splice site to stop codon)
- Sngl = Single-exon gene (ATG to stop)
- Prom = Promoter (TATA box / initiation site)
- PlyA = poly-A signal (consensus: AATAAA)

**S:** DNA strand (+ = input strand; - = opposite strand)

**Begin:** beginning of exon or signal (numbered on input strand)

**End:** end point of exon or signal (numbered on input strand)

**Len:** length of exon or signal (bp)

**Fr:** reading frame (a forward strand codon ending at x has frame x mod 3)

**Ph:** net phase of exon (exon length modulo 3)

**I/Ac:** initiation signal or 3' splice site score (tenth bit units)

**Do/T:** 5' splice site or termination signal score (tenth bit units)

**CodRg:** coding region score (tenth bit units)

**P:** probability of exon (sum over all parses containing exon)

**Tscr:** exon score (depends on length, I/Ac, Do/T and CodRg scores)
HMMgene

- WEB server: http://www.cbs.dtu.dk/services/HMMgene/
- Can return sub-optimal predictions to help identifying alternative splicing
- Accept annotation from user
- Human and worm models

HMMgene: form

Submission of a local file (HTML 8.0 or higher)
Organism:
- Human (and other vertebrates)
- C. elegans
File in FASTA format
Options:
- Predict signals
- Best prediction
- File with annotation (optional)

Submission by pasting sequences:
Organism:
- Human (and other vertebrates)
- C. elegans
Sequence(s) in FASTA format
Options:
- Predict signals
- Best prediction

Annotation (optional)
HMMgene: output

# SEQ  Sequence 00000 (-) A:5406 C:4748 G:4754 T:5092
Sequence HMMgene1.1a firstex 17618 17828 0.578 - 1 bestparse:cds_1
Sequence HMMgene1.1a exon_1 17049 17101 0.560 - 0 bestparse:cds_1
Sequence HMMgene1.1a exon_2 14517 14607 0.659 - 1 bestparse:cds_1
Sequence HMMgene1.1a exon_3 13918 13973 0.718 - 0 bestparse:cds_1
Sequence HMMgene1.1a exon_4 12441 12508 0.751 - 2 bestparse:cds_1
Sequence HMMgene1.1a lastex 7045 7222 0.893 - 0 bestparse:cds_1
Sequence HMMgene1.1a CDS 7045 17828 0.180 - . bestparse:cds_1
Sequence HMMgene1.1a DON 19837 19838 0.001 - 1
Sequence HMMgene1.1a START 19732 19734 0.024 - .
Sequence HMMgene1.1a ACC 19712 19713 0.001 - 0
Sequence HMMgene1.1a DON 19688 19689 0.006 - 1
Sequence HMMgene1.1a DON 19686 19687 0.004 - 0
...

------------ ----- ----
position prob strand and frame

Symbols: firstex = first exon; exon$\_$n$ = internal exon; lastex = last exon;
singleex = single exon gene; CDS = coding region

GRAILexp

- WEB server: http://compbio.ornl.gov/grailexp/
- Based on a neural network that incorporates signals and coding statistics
- Can use homology information!
- Human, Mouse, Drosophila, and Arabidopsis models
- Bacterial models

![Diagram of neural network structure]
**GRAILexp: form**

Select organism: Mouse [Mus musculus]
Select output type: Human-Readable Text

- **Perceval Exon Candidates**
  (Locate Grail.tsv using an improved version of the Grail.3 search tool)
- **Oligo EST/mRNA/cDNA Alignments**
  (Search from the selected EST/mRNA databases and build exons based on similarities with the sequences in these databases)

Select database(s) to search:

- **Gene Models**
  (Assemble complete gene structures from the above selected options, i.e. Perceval exons candidates and/or Oligo EST/mRNA (alignments))
- **Cpg Islands**
  (Find Cpg Islands using Grail.3)
- **Repetitive Elements**
  (Locate repetitive elements using a BLAST-based method against the Repbase database)

**DNA Sequence** (Raw or FASTA format, paste in box or upload file):

---

**GRAILexp: output**

Gene 1, Variant 1  
Strand: +  
Bounds: 1814-6614  
Exons: 5

Top-Scoring Reference: AP038421 (2560 bp) (99% id, 2833-6614)  
>human|AP038421|baylor_ht|AP038421|AP038421 Homo sapiens GPI-linked anchoor protein (GFRA1) mRNA, complete cds

Reference Path: CA487395.1 (881 bp) (97%, 1814-5295)  
AP038421 (2560 bp) (99%, 2833-6614)

---Index---- --------Exons-------- ---------CDS--------- -Ph- -Fr- -Len- -Scr-  
1.1.1 1814 1892 ... ... 0 1 79 100  
1.1.2 2833 2954 ... ... 1 2 122 100  
1.1.3 3842 4127 4088 4127 0 1 286 99  
1.1.4 5002 5295 5002 5295 1 2 286 99  
1.1.5 6531 6614 6531 6614 1 1 84 100  
...

>GrailEXP Gene 1, Var 1 mRNA|Similar to AP038421  
atgaacttggacatcagcaaagatcccgagcactgccggctggctcctagaccggtctcccgacccagtg  
...

>GrailEXP Gene 1, Var 1 protein|Derived from similarity to AP038421  
MFLATLYFALPLLDLLLSAEVSGGDRLDCVKASDQCLKEQSCSTKYRTLRQCVAGKETNFSLASGLEAKDECRSAMEALQKSLVNCRCXEMKZXEMCTYKQSGDRLRESSIESIIKIVPSYEPUNSRSLSDIFRUVFPF1EX  
...
Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
  - Software
- Homology methods
  - Genewise
  - Sim4 and BLAST
- Performance evaluation

Homology methods: principle

Genomic DNA sequence

Detect signals

Alignment with homologous sequence (mRNA, EST, protein)

Genomic DNA sequence

[Diagram showing genomic DNA sequence with detected signals and alignment with homologous sequence.]

LC-SIB-2005 – p.52/7
Genewise uses HMMs to align DNA sequences to protein sequences.

**Principle:**
- combine two HMMs:
  1. HMM to translate DNA sequence to aa
  2. HMM to align translated sequence to homologous protein
- add transitions to deal with frame-shifts
- add intron model

Good performances, but requires good homologous sequences (>70%) and a lot of CPU

**WEB server:** http://www.ebi.ac.uk/Wise2/
Genewise: simplified model

Genewise: form

<table>
<thead>
<tr>
<th>Your Email</th>
<th>Results</th>
<th>Output for alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>interactive</td>
<td>Parameters</td>
</tr>
</tbody>
</table>

Output for gene predictions

- Gene structure
- Translation
- cDNA
- EMRBL feature table format

Sequence 1: paste a Protein Sequence in fasta format OR upload a file:

Seq. 1 Upload file: [Browse...]
Genewise: perfect match

seq1
| 249 TDRRIGCLLS | GLDSSLVAATLLX |
| TDRRIGCLLS | GLDSSLVAATLLX |
| TDRRIGCLLS | G:G[ggg] |
| GLDSSLVAATLLX |

seq1
| 12930 agaaagtcttGGTGAAGT Intron 4 | TAGGGgtgtatgggacta |
| caggtggttc | <1-----[12961:13408]-1> |
| gtacgttcccta |
| acagtcctaa |
| cgcccgttctggg |

Gene 2979 19554
Exon 2979 3227 phase 0
Exon 7315 7552 phase 0
Exon 12416 12601 phase 1
Exon 12859 12960 phase 1
Exon 13409 13536 phase 1
Exon 14999 15125 phase 0
Exon 16356 16462 phase 1
Exon 18601 18756 phase 0
Exon 19348 19554 phase 0

Genewise: frame shift

seq1
| 249 TDRRIGCLLS | GLDSSLVAATLLX |
| TDRRIGCLLS | S |
| TDRRIGCLLS | G:G[ggg] |
| GLDSSLVAATLLX |

seq1
| 12930 agaaagtcttGGTGAAGT Intron 4 | TAGGGgtgtatgggacta |
| caggtggttc | <1-----[12961:13408]-1> |
| gtacgttcccta |
| acagtcctaa |
| cgcccgttctggg |

Gene 1
Gene 2979 12953
Exon 2979 3227 phase 0
Exon 7315 7552 phase 0
Exon 12416 12601 phase 1
Exon 12859 12953 phase 1
Genewise: mismatch

... seq1 249 TDRR--CLLS GLDSSLVAATLLK
   TDRR CLLS GLDSSLVAATLLK
   TDRRIGCLLS G:G[ggg] GLDSSLVAATLLK
... seq1 12910 agaaagtcttGGTGAAGT Intron 4 TAGGGgtgtatgggacta
caggtgggct <l-----[12961:13408]-l> gtacgttccctta
acctgccctaa cgcccggttctggg
...

Gene 1
Gene 2979 19554
   Exon 2979 3227 phase 0
   Exon 7315 7552 phase 0
   Exon 12416 12601 phase 1
   Exon 12859 12960 phase 1
   Exon 13409 13556 phase 0
   Exon 14999 15125 phase 0
   Exon 16356 16462 phase 1
   Exon 18601 18756 phase 0
   Exon 19348 19554 phase 0

Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
- Software
- Homology methods
  - Genewise
  - Sim4 and BLAST
- Performance evaluation
sim4

- sim4 aligns cDNA to genomic sequences
- sim4 performs standard dynamic programming, but:
  - models splice sites
  - introns are treated as a special kind of gaps with low penalties
- sim4 performs very well, but needs strong similarity between the sequences

sim4 output

...
**BLAST**

- Can be used to find genomic regions similar to ESTs, cDNA, proteins
- A hit doesn’t mean necessarily an exon. Need of post-processing
- Indicates the rough position of exons

  ![Diagram showing AG and GT, ideal BLAST, and real BLAST]

- ...but BLAST is fast! can reduce the search space for other programs

---

**Trimming with BLAST**

1. Protein sequence
2. cDNA sequence
3. BLAST vs genomic
4. Get best BLAST HSPs (trimming)
5. GeneWise
6. sim4
Outline

- Introduction
- Ab initio methods
  - Coding statistics
  - Signal detection
  - Integration of signal detection and coding statistics
  - Software
- Homology methods
  - Genewise
  - Sim4 and BLAST
- Performance evaluation

Evaluation of performances (1)

- **Sensitivity**: $S_n$ is the proportion of coding nucleotides correctly predicted as coding:

  $$S_n = \frac{TP}{TP + FN}$$

- **Specificity**: $S_p$ is the proportion on nucleotides predicted as coding that are actually coding:

  $$S_p = \frac{TP}{TP + FP}$$
Evaluation of performances (2)

- **Correlation coefficient**: CC is a single measure that captures both specificity and sensitivity:

\[
CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}
\]

- **Approximate correlation**: AC is similar to CC, but defined under any circumstance:

\[
AC = (ACP - 0.5) \times 2
\]

where

\[
ACP = \frac{1}{4} \left( \frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN} \right)
\]

Benchmark

- **Evaluation of some programs** (Rogic et al., 2001)

<table>
<thead>
<tr>
<th>Program</th>
<th>No. of sequences</th>
<th>(S_n)</th>
<th>(S_p)</th>
<th>AC</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGENES</td>
<td>195</td>
<td>0.86</td>
<td>0.88</td>
<td>0.84±0.19</td>
<td>0.83</td>
</tr>
<tr>
<td>GeneMark.hmm</td>
<td>195</td>
<td>0.87</td>
<td>0.89</td>
<td>0.84±0.18</td>
<td>0.83</td>
</tr>
<tr>
<td>MZEF</td>
<td>119</td>
<td>0.70</td>
<td>0.73</td>
<td>0.68±0.21</td>
<td>0.66</td>
</tr>
<tr>
<td>GENSCAN</td>
<td>195</td>
<td>0.95</td>
<td>0.90</td>
<td>0.91±0.12</td>
<td>0.91</td>
</tr>
<tr>
<td>HMMgene</td>
<td>195</td>
<td>0.93</td>
<td>0.93</td>
<td>0.91±0.13</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Comments

- HMMgene and GENSCAN perform the best
- Some program’s accuracy depend on G+C content, except for HMMgene and GENSCAN
- Accuracy decrease for short (<70) and long (>200) exons
- Internal exons are more likely to be correctly predicted
- Initial and terminal exons are the most likely to be missed

Limits

- Existing predictors are for protein coding regions
- Predictions work fine for "typical" genes:
  - partial gene are often missed
  - training sets may be biased
  - atypical genes use others grammars
... coffee!