For written notes on this lecture, please read Chapters 4 and 7 of *The Practical Bioinformatician*

CS2220: Introduction to Computational Biology
Lecture 4: Gene Feature Recognition

Limsoon Wong
3 February 2006
Central Dogma of Molecular Biology
What is a gene?
Central Dogma

Replication → Transcription → Translation

DNA → RNA → Protein

...AAUGGUACCAGUGCCUGGAGC...
...AATGGTACCGATGACCTG...
...TRLRPLLALLALWP...
Transcription: DNA → nRNA
Splicing: nRNA → mRNA

RNA synthesis and processing
**Translation: mRNA → protein**

<table>
<thead>
<tr>
<th>Codon 1</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>F</td>
<td>S</td>
<td>Tyr</td>
<td>Y</td>
<td>Cys</td>
</tr>
<tr>
<td>Ser</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Leu</td>
<td>L</td>
<td>Ser</td>
<td>Stop (Ochre)</td>
<td>Stop (Amber)</td>
<td>A</td>
</tr>
<tr>
<td>Ser</td>
<td>Stop (Amber)</td>
<td>Trp</td>
<td>W</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codon 2</th>
<th>C</th>
<th>Leu</th>
<th>Pro</th>
<th>His</th>
<th>Arg</th>
<th>R</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>Pro</td>
<td>Gln</td>
<td>Arg</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codon 3</th>
<th>A</th>
<th>Ile</th>
<th>Thr</th>
<th>Asn</th>
<th>N</th>
<th>Ser</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>Thr</td>
<td>Lys</td>
<td>Arg</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codon 4</th>
<th>M</th>
<th>Met</th>
<th>Thr</th>
<th>Lys</th>
<th>Arg</th>
<th>G</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Codon 5</th>
<th>G</th>
<th>Val</th>
<th>Ala</th>
<th>Asp</th>
<th>D</th>
<th>Gly</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Codon 6</th>
<th>G</th>
<th>Val</th>
<th>Ala</th>
<th>Glu</th>
<th>E</th>
<th>Gly</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
<td>Gly</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What does DNA data look like?

• A sample GenBank record from NCBI
What does protein data look like?

- A sample GenPept record from NCBI
Recognition of Translation Initiation Sites

An introduction to the World's simplest TIS recognition system
Translation Initiation Site

DNA: A, C, G, T
mRNA: A, C, G, U

5' flank   exon   intron   exon   intron   exon   3' flank

transcription

coding region
A Sample cDNA

299 HSU27655.1 CAT U27655 Homo sapiens
CGTGCTGCAGCAGCCTGAGCTGCCCAAGCCCCTCTGAGCTGACCTGACTCCCAGCTGTG 80
CCCAGGGCTTCAAAGACTTCTCAGCTCTCGAGCAATGTCTTTTGGCTGTCAGGCGAGCTGTA 160
GGAGGCAGATGAGAAGAGGGAGATGGAAGGAGGGGACCTGGTGCCCGAGGA 240
CCTCTCTGGCCAGGAGCTTCTCCAGGACAAGACTTCCACCACCAACGAAGACTCTCCCT 80
...............................................................iEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE 160
EEEEE,iEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE 240
EEEEE,iEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE

• What makes the second ATG the TIS?
Approach

• Training data gathering
• Signal generation
  – k-grams, distance, domain know-how, ...
• Signal selection
  – Entropy, $\chi^2$, CFS, t-test, domain know-how...
• Signal integration
  – SVM, ANN, PCL, CART, C4.5, kNN, ...
Training & Testing Data

- Vertebrate dataset of Pedersen & Nielsen [ISMB’97]
- 3312 sequences
- 13503 ATG sites
- 3312 (24.5%) are TIS
- 10191 (75.5%) are non-TIS
- Use for 3-fold x-validation expts
Signal Generation

• **K-grams (i.e., k consecutive letters)**
  - $K = 1, 2, 3, 4, 5, \ldots$
  - Window size vs. fixed position
  - Up-stream, down-stream vs. any where in window
  - In-frame vs. any frame
Signal Generation: An Example

Window = ±100 bases
• In-frame, downstream
  – GCT = 1, TTT = 1, ATG = 1…
• Any-frame, downstream
  – GCT = 3, TTT = 2, ATG = 2…
• In-frame, upstream
  – GCT = 2, TTT = 0, ATG = 0, ...

Exercise: Find the in-frame downstream ATG
Exercise: What are the possible k-grams (k=3) in this sequence?
Too Many Signals

• For each value of $k$, there are $4^k \times 3 \times 2$ $k$-grams

• If we use $k = 1, 2, 3, 4, 5$, we have $24 + 96 + 384 + 1536 + 6144 = 8184$ features!

• This is too many for most machine learning algorithms
Signal Selection (Basic Idea)

- Choose a signal with low intra-class distance
- Choose a signal with high inter-class distance
Signal Selection (e.g., t-statistics)

The t-stats of a signal is defined as

\[ t = \frac{|\mu_1 - \mu_2|}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \]

where \( \sigma_i^2 \) is the variance of that signal in class \( i \), \( \mu_i \) is the mean of that signal in class \( i \), and \( n_i \) is the size of class \( i \).
Signal Selection (e.g., MIT-correlation)

The MIT-correlation value of a signal is defined as

\[ MIT = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2} \]

where \( \sigma_i \) is the standard deviation of that signal in class \( i \) and \( \mu_i \) is the mean of that signal in class \( i \).
Signal Selection (e.g., $\chi^2$)

The $\chi^2$ value of a signal is defined as:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

where $m$ is the number of intervals, $k$ the number of classes, $A_{ij}$ the number of samples in the $i$th interval, $j$th class, $R_i$ the number of samples in the $i$th interval, $C_j$ the number of samples in the $j$th class, $N$ the total number of samples, and $E_{ij}$ the expected frequency of $A_{ij}$ ($E_{ij} = R_i \times C_j / N$).
Signal Selection (e.g., CFS)

• Instead of scoring individual signals, how about scoring a group of signals as a whole?

• CFS
  – Correlation-based Feature Selection
  – A good group contains signals that are highly correlated with the class, and yet uncorrelated with each other

Exercise: What is the main challenge in implementing CFS?
Sample k-grams Selected by CFS for Recognizing TIS

- Position –3
- in-frame upstream ATG
- in-frame downstream
  - TAA, TAG, TGA,
  - CTG, GAC, GAG, and GCC

Kozak consensus

Leaky scanning

Stop codon

Codon bias?
Signal Integration

• **kNN**
  – Given a test sample, find the k training samples that are most similar to it. Let the majority class win

• **SVM**
  – Given a group of training samples from two classes, determine a separating plane that maximises the margin of error

• **Naïve Bayes, ANN, C4.5, ...**
### Results (3-fold x-validation)

<table>
<thead>
<tr>
<th></th>
<th>predicted as positive</th>
<th>predicted as negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

**Exercise:**
What is TP/(TP+FP)?

<table>
<thead>
<tr>
<th>Method</th>
<th>TP/(TP + FN)</th>
<th>TN/(TN + FP)</th>
<th>TP/(TP + FP)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>84.3%</td>
<td>86.1%</td>
<td>66.3%</td>
<td>85.7%</td>
</tr>
<tr>
<td>SVM</td>
<td>73.9%</td>
<td>93.2%</td>
<td>77.9%</td>
<td>88.5%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>77.6%</td>
<td>93.2%</td>
<td>78.8%</td>
<td>89.4%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>74.0%</td>
<td>94.4%</td>
<td>81.1%</td>
<td>89.4%</td>
</tr>
</tbody>
</table>
## Improvement by Voting

- Apply any 3 of Naïve Bayes, SVM, Neural Network, & Decision Tree. Decide by majority

<table>
<thead>
<tr>
<th></th>
<th>TP/(TP + FN)</th>
<th>TN/(TN + FP)</th>
<th>TP/(TP + FP)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB+SVM+NN</td>
<td>79.2%</td>
<td>92.1%</td>
<td>76.5%</td>
<td>88.9%</td>
</tr>
<tr>
<td>NB+SVM+Tree</td>
<td>78.8%</td>
<td>92.0%</td>
<td>76.2%</td>
<td>88.8%</td>
</tr>
<tr>
<td>NB+NN+Tree</td>
<td>77.6%</td>
<td>94.5%</td>
<td>82.1%</td>
<td>90.4%</td>
</tr>
<tr>
<td>SVM+NN+Tree</td>
<td>75.9%</td>
<td>94.3%</td>
<td>81.2%</td>
<td>89.8%</td>
</tr>
<tr>
<td>Best of 4</td>
<td>84.3%</td>
<td>94.4%</td>
<td>81.1%</td>
<td>89.4%</td>
</tr>
<tr>
<td>Worst of 4</td>
<td>73.9%</td>
<td>86.1%</td>
<td>66.3%</td>
<td>85.7%</td>
</tr>
</tbody>
</table>
Improvement by Scanning

- Apply Naïve Bayes or SVM left-to-right until first ATG predicted as positive. That’s the TIS
- Naïve Bayes & SVM models were trained using TIS vs. Up-stream ATG

<table>
<thead>
<tr>
<th></th>
<th>TP/(TP + FN)</th>
<th>TN/(TN + FP)</th>
<th>TP/(TP + FP)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>84.3%</td>
<td>86.1%</td>
<td>66.3%</td>
<td>85.7%</td>
</tr>
<tr>
<td>SVM</td>
<td>73.9%</td>
<td>93.2%</td>
<td>77.9%</td>
<td>88.5%</td>
</tr>
<tr>
<td>NB+Scanning</td>
<td>87.3%</td>
<td>96.1%</td>
<td>87.9%</td>
<td>93.9%</td>
</tr>
<tr>
<td>SVM+Scanning</td>
<td>88.5%</td>
<td>96.3%</td>
<td>88.6%</td>
<td>94.4%</td>
</tr>
</tbody>
</table>
### Performance Comparisons

<table>
<thead>
<tr>
<th>Method</th>
<th>TP/(TP + FN)</th>
<th>TN/(TN + FP)</th>
<th>TP/(TP + FP)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>84.3%</td>
<td>86.1%</td>
<td>66.3%</td>
<td>85.7%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>74.0%</td>
<td>94.4%</td>
<td>81.1%</td>
<td>89.4%</td>
</tr>
<tr>
<td>NB+NN+Tree</td>
<td>77.6%</td>
<td>94.5%</td>
<td>82.1%</td>
<td>90.4%</td>
</tr>
<tr>
<td>SVM+Scanning</td>
<td>88.5%</td>
<td>96.3%</td>
<td>88.6%</td>
<td>94.4%*</td>
</tr>
<tr>
<td>Pedersen&amp;Nielsen</td>
<td>78%</td>
<td>87%</td>
<td>-</td>
<td>85%</td>
</tr>
<tr>
<td>Zien</td>
<td>69.9%</td>
<td>94.1%</td>
<td>-</td>
<td>88.1%</td>
</tr>
<tr>
<td>Hatzigeorgiou</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>94%*</td>
</tr>
</tbody>
</table>

* result not directly comparable
Technique Comparisons

• Pedersen & Nielsen [ISMB’97]
  – Neural network
  – No explicit features

• Zien [Bioinformatics’00]
  – SVM+kernel engineering
  – No explicit features

• Hatzigeorgiou
  [Bioinformatics’02]
  – Multiple neural networks
  – Scanning rule
  – No explicit features

• Our approach
  – Explicit feature generation
  – Explicit feature selection
  – Use any machine learning method w/o any form of complicated tuning
  – Scanning rule is optional
mRNA→protein

How about using k-grams from the translation?

<table>
<thead>
<tr>
<th>First</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>F</td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>S</td>
<td>Leu</td>
<td>Ser</td>
<td>Stop (Ochre)</td>
<td>Stop (Amber)</td>
<td>A</td>
</tr>
<tr>
<td>Y</td>
<td>Leu</td>
<td>Ser</td>
<td>Stop (Amber)</td>
<td>Trp</td>
<td>G</td>
</tr>
<tr>
<td>C</td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
<td>U</td>
</tr>
<tr>
<td>P</td>
<td>Leu</td>
<td>Pro</td>
<td>His</td>
<td>Arg</td>
<td>C</td>
</tr>
<tr>
<td>H</td>
<td>Leu</td>
<td>Pro</td>
<td>Gin</td>
<td>Arg</td>
<td>A</td>
</tr>
<tr>
<td>Q</td>
<td>Leu</td>
<td>Pro</td>
<td>Gin</td>
<td>Arg</td>
<td>G</td>
</tr>
<tr>
<td>O</td>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
<td>U</td>
</tr>
<tr>
<td>N</td>
<td>Ile</td>
<td>Thr</td>
<td>Asn</td>
<td>Ser</td>
<td>C</td>
</tr>
<tr>
<td>K</td>
<td>Ile</td>
<td>Thr</td>
<td>Lys</td>
<td>Arg</td>
<td>A</td>
</tr>
<tr>
<td>M</td>
<td>Met</td>
<td>Thr</td>
<td>Lys</td>
<td>Arg</td>
<td>G</td>
</tr>
<tr>
<td>D</td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
<td>U</td>
</tr>
<tr>
<td>A</td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
<td>Gly</td>
<td>C</td>
</tr>
<tr>
<td>V</td>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
<td>Gly</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>Val</td>
<td>Ala</td>
<td>Glu</td>
<td>Gly</td>
<td>G</td>
</tr>
</tbody>
</table>

Exercise: List the first 10 amino acid in our example sequence
Amino-Acid Features

General Diagram:
- **cDNA sequence**: ......GGACGGATGACTGCC......CTCGATATGGCACCT......TTGCTAATGGACAAATA......
- **Sequence window generation**:
  - **False TIS (upstream)**
  - **True TIS**
  - **False TIS (downstream)**

**Coding**:
- **a (false) TIS window**
  - 99bps
  - 99bps
- **a (true) TIS window**
  - 99bps
  - 99bps

**Amino acid sequence**:
- **GR (False) TA**
  - 33aa
  - 33aa
- **LD (True) AP**
  - 33aa
  - 33aa

Copyright 2006 © Limsoon Wong
## Amino-Acid Features

### New feature space (total of 927 features + class label)

<table>
<thead>
<tr>
<th>42 1-gram amino acid patterns</th>
<th>882 2-gram amino acid patterns</th>
<th>3 bio-knowledge patterns</th>
<th>class label</th>
</tr>
</thead>
</table>

#### Frequency as values

<table>
<thead>
<tr>
<th>1, 3, 5, 0, 4, ...</th>
<th>6, 2, 7, 0, 5, ...</th>
<th>N, N, N, ...</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>6, 5, 7, 9, 0, ...</td>
<td>2, 0, 3, 10, 0, ...</td>
<td>Y, Y, Y, ...</td>
<td>True</td>
</tr>
</tbody>
</table>

---

Copyright 2006 © Limsoon Wong

---

[Diagram of DNA sequence and amino acid features]
Amino Acid K-grams
Discovered (by entropy)

- Position -3
- in-frame upstream ATG
- in-frame downstream
  - TAA, TAG, TGA,
  - CTG, GAC, GAG, and GCC

<table>
<thead>
<tr>
<th>Fold</th>
<th>UP-ATG</th>
<th>DOWN-STOP</th>
<th>UP3-AorG</th>
<th>DOWN-A</th>
<th>DOWN-V</th>
<th>UP-A</th>
<th>DOWN-L</th>
<th>DOWN-D</th>
<th>DOWN-E</th>
<th>UP-G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>
Independent Validation Sets

• **A. Hatzigeorgiou:**
  – 480 fully sequenced human cDNAs
  – 188 left after eliminating sequences similar to training set (Pedersen & Nielsen’s)
  – 3.42% of ATGs are TIS

• **Our own:**
  – well characterized human gene sequences from chromosome X (565 TIS) and chromosome 21 (180 TIS)
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVMs(linear)</td>
<td>96.28%</td>
<td>89.15%</td>
<td>25.31%</td>
<td>89.42%</td>
</tr>
<tr>
<td>SVMs(quad)</td>
<td>94.14%</td>
<td>90.13%</td>
<td>26.70%</td>
<td>90.28%</td>
</tr>
<tr>
<td>Ensemble Trees</td>
<td>92.02%</td>
<td>92.71%</td>
<td>32.52%</td>
<td>92.68%</td>
</tr>
</tbody>
</table>

- Using top 100 features selected by entropy and trained on Pedersen & Nielsen’s dataset
Validation Results (on Chr X and Chr 21)

Our method

• Using top 100 features selected by entropy and trained on Pedersen & Nielsen’s
Recognition of
Transcription Start Sites

An introduction to the World’s best TSS recognition system:
A heavy tuning approach
Transcription Start Site

5'flank  exon  intron  exon  intron  exon  3'flank

transcription
coding region

DNA: A, C, G, T
mRNA: A, C, G, U

Copyright 2006 © Limsoon Wong
Structure of Dragon Promoter Finder

-200 to +50 window size

Model selected based on desired sensitivity
Each model has two submodels based on GC content

Exercise: Why are the submodels based on GC content?

\[(C+G) = \frac{\#C + \#G}{\text{Window Size}}\]
Data Analysis Within Submodel

K-gram (k = 5) positional weight matrix
Promoter, Exon, Intron Sensors

- These sensors are positional weight matrices of k-grams, $k = 5$ (aka pentamers)
- They are calculated as below using promoter, exon, intron data respectively

\[
\sigma = \left( \frac{\sum_{i=1}^{L-4} p_j^i \otimes f_{j,i}}{\sum_{i=1}^{L-4} \max_j f_{j,i}} \right), \quad p_j^i \otimes f_{j,i} = \begin{cases} f_{j,i}, & \text{if } p_i = p_j^i \\ 0, & \text{if } p_i \neq p_j^i \end{cases}
\]
Data Preprocessing & ANN

Tuning parameters

\[ s_E = sat(\sigma_p - \sigma_e, a_e, b_e), \]
\[ s_I = sat(\sigma_p - \sigma_i, a_i, b_i), \]
\[ s_{EI} = sat(\sigma_e - \sigma_i, a_{ei}, b_{ei}), \]

where the function \( sat \) is defined by

\[ sat(x, a, b) = \begin{cases} a, & \text{if } x > a \\ x, & \text{if } b \leq x \leq a \\ b, & \text{if } b > x \end{cases} \]

Simple feedforward ANN trained by the Bayesian regularisation method

\[ \text{tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \]

\[ \text{net} = \sum s_i \cdot w_i \]
Accuracy Comparisons

Accuracy of Dragon Promoter Finder Ver. 1.2 & 1.3

Sensitivity in % = $100 \times \frac{TP}{TP+FN}$

Positive predictive value ppv in % = $100 \times \frac{TP}{TP+FP}$

- DPF v1.2
- DPF v1.3
- NNPP2.1 (0.99)
- NNPP2.1 (0.8)
- PromoterInspector
- Promoter2.0

- with C+G submodels
- without C+G submodels
Training Data Criteria & Preparation

- Contain both positive and negative sequences
- Sufficient diversity, resembling different transcription start mechanisms
- Sufficient diversity, resembling different non-promoters
- Sanitized as much as possible
- **TSS taken from**
  - 793 vertebrate promoters from EPD
  - -200 to +50 bp of TSS
- **non-TSS taken from**
  - GenBank,
  - 800 exons
  - 4000 introns,
  - 250 bp,
  - non-overlapping,
  - <50% identities
Tuning Data Preparation

• To tune adjustable system parameters in Dragon, we need a separate tuning data set

• **TSS taken from**
  – 20 full-length gene seqs with known TSS
  – -200 to +50 bp of TSS
  – no overlap with EPD

• **Non-TSS taken from**
  – 1600 human 3’UTR seqs
  – 500 human exons
  – 500 human introns
  – 250 bp
  – no overlap
Testing Data Criteria & Preparation

- Seqs should be from the training or evaluation of other systems (no bias!)
- Seqs should be disjoint from training and tuning data sets
- Seqs should have TSS
- Seqs should be cleaned to remove redundancy, <50% identities

- 159 TSS from 147 human and human virus seqs
- Cumulative length of more than 1.15Mbp
- Taken from GENESCAN, Geneld, Genie, etc.
Accuracy on Human Chromosome 22

<table>
<thead>
<tr>
<th>Human chromosome 22 (known genes)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Se</td>
<td>Ppv</td>
</tr>
<tr>
<td>49%</td>
<td>48%</td>
<td></td>
</tr>
<tr>
<td>58%</td>
<td>42%</td>
<td></td>
</tr>
<tr>
<td>64%</td>
<td>33%</td>
<td></td>
</tr>
<tr>
<td>74%</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>80%</td>
<td>23%</td>
<td></td>
</tr>
</tbody>
</table>
Other Gene Features
Other Gene Features
Any Question?
References (TIS Recognition)

- J. Li et al., “Techniques for Recognition of Translation Initiation Sites”, *The Practical Bioinformatician*, Chapter 4, pages 71—90, 2004
References (TSS Recognition)

• M. Scherf et al., “Highly specific localisation of promoter regions in large genome sequences by PromoterInspector”, JMB 297:599--606, 2000
References (Feature Selection)


References (Misc.)


• D. J. States, W. Gish, “Combined use of sequence similarity and codon bias for coding region identification”, *JCB* 1:39--50, 1994

• G. D. Stormo et al., “Use of Perceptron algorithm to distinguish translational initiation sites in E. coli”, *NAR* 10:2997--3011, 1982