For written notes on this lecture, please read Chapters 4 and 7 of The Practical Bioinformatician, and Koh & Wong, “Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences”, Proc GIW 2007, pages 73--82

CS2220: Introduction to Computational Biology
Lecture 3: Gene Feature Recognition

Limsoon Wong
30 January 2009

Plan
Some Relevant Biology

Central Dogma

Replication ➔ Transcription ➔ Translation

DNA ➔ RNA ➔ Protein

...AATGGTACCGATGACCTG...

...TRLRPLLALLALWP......AAUGGUACCGAUGACCUGGAGC...

...AAUGGUACCGAUGACCUGGAGC...

...TRLRPLLALLALWP...

Copyright 2009 © Limsoon Wong
Players in Protein Synthesis

1. Transcription

- Synthesize mRNA from one strand of DNA
  - An enzyme RNA polymerase temporarily separates double-stranded DNA
  - It begins transcription at transcription start site
  - A → A, C→C, G→G, & T→U
  - Once RNA polymerase reaches transcription stop site, transcription stops

- Additional “steps” for Eukaryotes
  - Transcription produces pre-mRNA that contains both introns & exons
  - 5’ cap & poly-A tail are added to pre-mRNA
  - RNA splicing removes introns & mRNA is made
  - mRNA are transported out of nucleus

2. Translation

Protein synthesis
Translation

- Synthesize protein from mRNA
- Each amino acid is encoded by consecutive seq of 3 nucleotides, called a codon
- The decoding table from codon to amino acid is called genetic code
- $4^3=64$ diff codons
  \[ \Rightarrow \text{Codons are not 1-to-1 corr to 20 amino acids} \]
- All organisms use the same decoding table (except some mitochondrial genes)
- Amino acids can be classified into 4 groups. A single-base change in a codon is usu insufficient to cause a codon to code for an amino acid in diff group

Genetic Code

- **Start codon**
  - ATG (code for M)
- **Stop codon**
  - TAA
  - TAG
  - TGA
Example

Example of computational translation - notice the indication of (alternative) start-codons:

VIRTUAL RIBOSOME

---

Translation table: Standard 8060

Segment:

```
+-------------------+
| AUG   UUU   UGA   |
| (start)           |
+-------------------+
```

Reading frame: 1

```
NM LSA A B GON UVH E A M W G G H A E Y S A E L
9* ATG GGU CUC UUC GAA AGG CUG CAA AAG GUG CGU AGU AGG CUG CGU GAG CUG
999...|||-----------------------------|

ER H F L S P T T T T Y F H P D L M H O S A G V W O H S 9*
9* GBA GAG UG T CUC UUC GAA AGG CUG CAA AAG GUG CGU GAG CUG
999...|||-----------------------------|

A Y V A A A L T K A V E N L D D L P G A L S E L S D L H A S
9* CUA CUG UGG GUG CGU AGU CUU AAG GUG CGU CGU AGU AGG CUG
999...|||-----------------------------|

L R V D P V H F K L S H N S L L V T L A S H L E S D F T P
9* AAG CUG UGG GUG CGU AGU CUU AAG GUG CGU CGU AGU AGG CUG
999...|||-----------------------------|

A Y K A S L D P F L A S V S Y T U L T I S P Y R *
9* CGG UCG CAC UUC GAA GGG CUC AAG GUG CUG CAA AAG GUG CGU 429
999...|||-----------------------------|

Annotation key:

>>> | START CDS (90%) |
| ( ) | START codon (alternative) |
| *** | STOP |
```

Recognition of Translation Initiation Sites

An introduction to the World's simplest TIS recognition system
A Sample cDNA

299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAAGCTGCTGCCCAAGCCCATGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCAGCATG
GCTTTTGGCTGTCAGGGCAGCTGTA     160
GGAGGCAGATG
AGAAGAGGGAGATG
GCCTTGGAGGAAGGGAAGGGGCCTGGTGCCAGGA     240
CCTCTCTGGCCCCAGGATTTCCCTCCAGGACAGCTTCTCCACCACAAAAGGACCTCCCT
............................................................    80
............................................................    160
............................................................    240

• 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 (ie., k consecutive letters)**
  – \( K = 1, 2, 3, 4, 5, \ldots \)
  – Window size vs. fixed position
  – Up-stream, downstream vs. any where in window
  – In-frame vs. any frame

![Signal Generation Diagram]

Signal Generation: An Example

• **Window = \( \pm 100 \) 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, ...

299 HSU27655.1 CAT U27655 Homo sapiens
CGGTGTCACGACCGCTGAGCTCCCCCCAAGCCTGACCTGACTCCCCAATGCT
CCCAGGCTTCAGACTCAGGCACTGTTA 80
GGAGGCAAGAGAGAGAGATGGCTTGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA...
Feature Generation - Summary

Raw Data

206 88CALCB.1 CAT X72666 Eukaryotes; mitochondrial eukaryotes; Metazoa; Chordata; chordata | 206 88CALCB.1 CAT X72666 Eukaryotes; mitochondrial eukaryotes; Metazoa; Chordata; chordata | 206 88CALCB.1 CAT X72666 Eukaryotes; mitochondrial eukaryotes; Metazoa; Chordata; chordata | 206 88CALCB.1 CAT X72666 Eukaryotes; mitochondrial eukaryotes; Metazoa; Chordata; chordata | 206 88CALCB.1 CAT X72666 Eukaryotes; mitochondrial eukaryotes; Metazoa; Chordata; chordata

An ATG segment – positive sample

A feature vector --- upstream/downstream inframe 3 grams

Too Many Signals

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

• If we use $k = 1, 2, 3, 4, 5$, we have $2^4 + 96 + 384 + 1536 + 6144 = 8184$ 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-stat 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 \( A_{ij} \) (\( E_{ij} = R_i \cdot C_j / N \)).
Example

- Suppose you have a sample of 50 men and 50 women and the following weight distribution is observed:

<table>
<thead>
<tr>
<th></th>
<th>obs</th>
<th>exp</th>
<th>(obs – exp)^2/exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>HM</td>
<td>40</td>
<td>60*50/100 = 30</td>
<td>3.3</td>
</tr>
<tr>
<td>HW</td>
<td>20</td>
<td>60*50/100 = 30</td>
<td>3.3</td>
</tr>
<tr>
<td>LM</td>
<td>10</td>
<td>40*50/100 = 20</td>
<td>5.0</td>
</tr>
<tr>
<td>LW</td>
<td>30</td>
<td>40*50/100 = 20</td>
<td>5.0</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 16.6 \]
\[ P = 0.00004, \]
\[ df = 1 \]
So weight and sex are not indep

- Is weight a good attribute for distinguishing men from women?

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?
Distributions of Two Example 3-Grams

\[ \chi^2 = 1672.97447 \]

\[ \chi^2 = 0 \]

- Which is the better one?

Sample k-grams Selected by CFS for Recognizing TIS

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

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>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>

Exercise: What is TP/(TP+FP)?
### 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></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
Exercise: List the first 10 amino acid in our example sequence
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>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>i</td>
<td>False</td>
</tr>
</tbody>
</table>

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

Amino Acid K-grams Discovered (by entropy)

- Kozak consensus
- Leaky scanning
- 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-AarG</th>
<th>DOWN-A</th>
<th>DOWN-AR</th>
<th>UP-A</th>
<th>DOWN-AL</th>
<th>DOWN-D</th>
<th>DOWN-DE</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>4</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>9</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)

Validation Results (on Hatzigeorgiou’s)

<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
Our method

Validation Results (on Chr X and Chr 21)

- 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
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

- **GC-rich submodel**
- **GC-poor submodel**

(C+G) = \#C + \#G

Window Size

Exercise: Why are the submodels based on GC content?

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

\[ \sigma = \frac{\sum_{j=1}^{L-4} p_j \otimes f_{j,i}}{\sum_{j=1}^{L-4} \max_j f_{j,i}} \]

\[ p_j^i \otimes f_{j,i} = \begin{cases} f_{j,i}, & \text{if } p_j^i = p_j^i \\ 0, & \text{if } p_j^i \neq p_j^i \end{cases} \]

**Window size**

Just to make sure you know what I mean ...

- Give me 3 DNA seq of length 10:
  - Seq₁ = ACCGAGTTCT
  - Seq₂ = AGTGTACCTG
  - Seq₃ = AGTTCGTATG

- Then

<table>
<thead>
<tr>
<th>1-mer</th>
<th>pos1</th>
<th>pos2</th>
<th>pos3</th>
<th>pos4</th>
<th>pos5</th>
<th>pos6</th>
<th>pos7</th>
<th>pos8</th>
<th>pos9</th>
<th>pos10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3/3</td>
<td>0/3</td>
<td>0/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0/3</td>
<td>2/3</td>
<td>0/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0/3</td>
<td>0/3</td>
<td>2/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Exercise: Fill in the rest of the table.
Just to make sure you know what I mean …

• Give me 3 DNA seq of length 10:
  – Seq₁ = ACCGAGTTCT
  – Seq₂ = AGTGTACCTG
  – Seq₃ = AGTTCGTATG

• Then

<table>
<thead>
<tr>
<th>2-mer</th>
<th>pos1</th>
<th>pos2</th>
<th>pos3</th>
<th>pos4</th>
<th>pos5</th>
<th>pos6</th>
<th>pos7</th>
<th>pos8</th>
<th>pos9</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>0/3</td>
<td>0/3</td>
<td>0/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>1/3</td>
<td>0/3</td>
<td>0/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td>……</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TT</td>
<td>0/3</td>
<td>0/3</td>
<td>1/3</td>
<td>1/3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Exercise: How many rows should this 2-mer table have? How many rows should the pentamer table have?

Data Preprocessing & ANN

Tuning parameters

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

where the function \text{sat} is defined by

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

Simple feedforward ANN trained by the Bayesian regularisation method

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

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

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.
Recognition of Poly-A Signal Sites

A twist to the “feature generation, feature selection, feature integration” approach

Eukaryotic Pre-mRNA Processing

54
In contrast to human, PAS in Arab is highly degenerate. E.g., only 10% of Arab PAS is AAUAAA!
Approach on Arab PAS Sites (I)

- **Data collection**
  - #1 from Hao Han, 811
    - +ve seq (-200/+200)
  - #2 from Hao Han, 9742
    - -ve seq (-200/+200)
  - #3 from Qingshun Li,
    - 6209 (+ve) seq (-300/+100)
    - 1581 (-ve) intron (-300/+100)
    - 1501 (-ve) coding (-300/+100)
    - 864 (-ve) 5’utr (-300/+100)

- **Feature generation**
  - 3-grams, compositional features (4U/1N, G/U*7, etc)
  - Freq of features above in 3 diff windows: (-110/+5), (-35/+15), (-50/+30)

- **Feature selection**
  - $\chi^2$

- **Feature integration & Cascade**
  - SVM
Score Profile Relative to Candidate Sites

Validation Results

<table>
<thead>
<tr>
<th>SN_10</th>
<th>SMO 1</th>
<th>SMO 2</th>
<th>PASS 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SN &amp; SF</td>
<td>Threshold</td>
<td>SN &amp; SF</td>
</tr>
<tr>
<td>CDS</td>
<td>94%</td>
<td>0.36</td>
<td>90%</td>
</tr>
<tr>
<td>5'UTR</td>
<td>96%</td>
<td>0.51</td>
<td>95%</td>
</tr>
<tr>
<td>Intex</td>
<td>73%</td>
<td>0.65</td>
<td>77%</td>
</tr>
</tbody>
</table>

Table 2. Equal error rate points of SMO1, SMO2, and PASS 1.0 for SN_10.

<table>
<thead>
<tr>
<th>SN_30</th>
<th>SMO 1</th>
<th>SMO 2</th>
<th>PASS 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SN &amp; SF</td>
<td>Threshold</td>
<td>SN &amp; SF</td>
</tr>
<tr>
<td>CDS</td>
<td>97%</td>
<td>0.44</td>
<td>97%</td>
</tr>
<tr>
<td>5'UTR</td>
<td>96%</td>
<td>0.62</td>
<td>92%</td>
</tr>
<tr>
<td>Intex</td>
<td>79%</td>
<td>0.75</td>
<td>53%</td>
</tr>
</tbody>
</table>

Table 3. Equal error rate points of SMO1, SMO2, and PASS 1.0 for SN_30.
Concluding Remarks…

What have we learned?

- **Gene feature recognition applications**
  - TIS, TSS, PAS

- **General methodology**
  - “Feature generation, feature selection, feature integration”

- **Important tactics**
  - Multiple models to optimize overall performance
  - Feature transformation (DNA \(\rightarrow\) amino acid)
  - Classifier cascades
Any Question?

Acknowledgements

• The slides for PAS site prediction are adapted from slides given to me by Koh Chuan Hock
References (TIS Recognition)


• A. Zien et al., "Engineering support vector machine kernels that recognize translation initiation sites", *Bioinformatics* 16:799--807, 2000

• A. G. Hatzigeorgiou, "Translation initiation start prediction in human cDNAs with high accuracy", *Bioinformatics* 18:343--350, 2002

• 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 (PAS Recognition)


References (Feature Selection)

