Some Relevant Biology

Central Dogma

Players in Protein Synthesis

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 → G, G → C, & 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
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 ⇒ 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

Recognition of Translation Initiation Sites

An introduction to the World’s simplest TIS recognition system

Translation Initiation Site

A Sample cDNA

- 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: 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, ...

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

Too Many Features

- For each value of k, there are \( 4^k \cdot 3 \cdot 2 \) 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-stat of a signal is defined as

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

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 \(i\).

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)/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>

\[ \chi^2 = 16.6 \]

\[ P = 0.00004, \text{ df } = 1 \]

So weight and sex are not independent

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

Signal Selection (e.g., \(\chi^2\))

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

\[ \chi^2 = \sum \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \]

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, \(R_j\) the number of samples in the \(j\)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 \(A_{ij}\) (\(E_{ij} = R_j * C_j / N\)).

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?
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
  - TAG, TAC, 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)

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

<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>

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>Rest 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>86.5%</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>Technique</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>94.5%*</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
- Our approach
  - Explicit feature generation
  - Explicit feature selection
  - Use any machine learning method w/o any form of complicated tuning
  - Scanning rule is optional

Amino-Acid Features

Exercise: List the first 10 amino acids in our example sequence

Amino Acid K-grams

Discovered (by entropy)
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)

- Using top 100 features selected by entropy and trained on Pedersen & Nielsen’s dataset

Validation Results (on Chr X and Chr 21)

- Using top 100 features selected by entropy and trained on Pedersen & Nielsen’s

About the Inventor: Huiqing Liu

- PhD, NUS, 2004
- Currently Senior Scientist at Centocor
- Asian Innovation Gold Award 2003
- New Jersey Cancer Research Award for Scientific Excellence 2008
- Gallo Prize 2008

Recognition of Transcription Start Sites

An introduction to the World’s best TSS recognition system:
A heavy tuning approach
Structure of Dragon Promoter Finder

Data Analysis Within Submodel

Promoter, Exon, Intron Sensors

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

Each model has two submodels based on GC content

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

- Model selected based on desired sensitivity
- Window size: -200 to +50

- Exercise: Why are the submodels based on GC content?

- Pentamer at ith position in put Window size
- Frequency of jth pentamer at ith position in training window

- K-gram (k = 5) positional weight matrix

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

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

- Exercise: Fill in the rest of the table

<table>
<thead>
<tr>
<th>A</th>
<th>0/3</th>
<th>0/3</th>
<th>0/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>G</td>
<td>0/3</td>
<td>2/3</td>
<td>0/3</td>
</tr>
<tr>
<td>T</td>
<td>0/3</td>
<td>0/3</td>
<td>2/3</td>
</tr>
</tbody>
</table>
Data Preprocessing & ANN

Tuning parameters

Simple feedforward ANN trained by the Bayesian regularization method

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

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

Accuracy Comparisons

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

- 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.

About the Inventor: Vlad Bajic

- Vladimir B. Bajic
  - Principal Scientist, PR, 2001-2006
  - Currently Director & Professor, Computational Bioscience Research Center, KAUST
Recognition of Poly-A Signal Sites

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

Eukaryotic Pre-mRNA Processing

- Pre-mRNA
- 5'UTR
- 3'UTR
- Capping
- Splicing
- Intron
- Cleavage
- Polyadenylation
- Mature mRNA

Polyadenylation in Eukaryotes

- **Addition of poly(A) tail to RNA**
  - Begins as transcription finishes
  - 3'-most segment of newly-made RNA is cleaved off
  - Poly(A) tail is then synthesized at 3' end
- **Poly(A) tail is imp't for nuclear export, translation & stability of mRNA**
- Tail is shortened over time. When short enough, the mRNA is degraded

Poly-A Signals in Human (Gautheret et al., 2000)

<table>
<thead>
<tr>
<th>Site</th>
<th>Frequency</th>
<th>p-value</th>
<th>Localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAUAAA</td>
<td>59 (37)</td>
<td>0.002</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AUUAAA</td>
<td>89 (24)</td>
<td>0.001</td>
<td>S 0.17 ± 0.2</td>
</tr>
<tr>
<td>UAUAAA</td>
<td>145 (16)</td>
<td>0.002</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>UAACAA</td>
<td>25 (10)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>CAUAAA</td>
<td>36 (10)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AUAUA</td>
<td>52 (9)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAAGA</td>
<td>10 (5)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAGCA</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAGCG</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAUGA</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAUUC</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAUCG</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAACG</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAAGG</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAUUG</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAAGT</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAUUU</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAUAU</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAAGU</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAACU</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAUCG</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAACU</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
<tr>
<td>AAUAG</td>
<td>5 (3)</td>
<td>0.001</td>
<td>S 0.16 ± 0.7</td>
</tr>
</tbody>
</table>

Poly-A Signals in Arabidopsis

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)
Approach on Arab PAS Sites (II)

- 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*, 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</th>
<th>UME 1</th>
<th>UME 2</th>
<th>PASA LI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>CDS</td>
<td>65%</td>
<td>0.34</td>
<td>68%</td>
</tr>
<tr>
<td>TSS</td>
<td>58%</td>
<td>0.47</td>
<td>53%</td>
</tr>
<tr>
<td>Introns</td>
<td>64%</td>
<td>0.39</td>
<td>71%</td>
</tr>
</tbody>
</table>

About the Inventor: Koh Chuan Hock

- Koh Chuan Hock
  - BComp (CB), NUS, 2008
  - Currently PhD candidate at SOC

Concluding Remarks...

- 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
Acknowledgements

- The slides for PAS site prediction are adapted from slides given to me by Koh Chuan Hock

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)


References (PAS Recognition)


References (Feature Selection)