

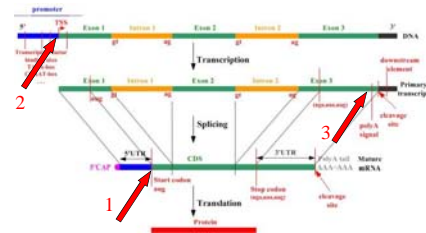
For written notes on this lecture, please read Chapters 4 and 7 of *The Practical Bioinformatician*, and Koh & Wong, "Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences", *Proc GIW 2007*, pages 73–82

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

Limsoon Wong



### Plan

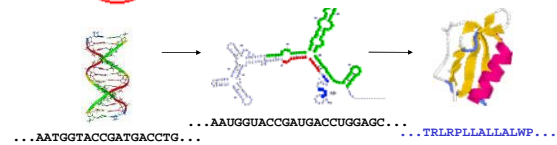
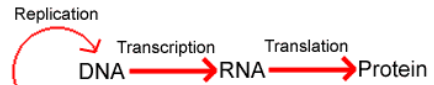


Copyright 2011 © Limsoon Wong

### Some Relevant Biology

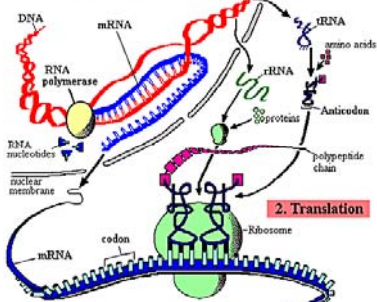


### Central Dogma



Copyright 2011 © Limsoon Wong

#### 1. Transcription



### Players in Protein Synthesis

Protein synthesis

Copyright 2011 © Limsoon Wong



### 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

Copyright 2011 © Limsoon Wong

## 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<sup>3</sup>=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

Copyright 2011 © Limsoon Wong

## Genetic Code

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

		Second Position of Codon				
		I	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T	
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C	
	TTA Leu [L]	TCA Ser [S]	TAA Tyr [amd]	TGA Trp [amd]	A	
	TTG Leu [L]	TGG Ser [S]	TAG Tyr [amd]	TGG Trp [W]	G	
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T	
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CCG Arg [R]	C	
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CCA Arg [R]	A	
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CCG Arg [R]	G	
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T	
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C	
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A	
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G	
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T	
	GTG Val [V]	GCC Ala [A]	GAU Asp [D]	GGC Gly [G]	C	
	GTA Val [V]	GCA Ala [A]	GAA Gln [Q]	GGA Gly [G]	A	
	GTG Val [V]	GCG Ala [A]	GAG Gln [Q]	GGG Gly [G]	G	

Copyright 2011 © Limsoon Wong

## Example

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

```

VIRTUAL RIBOSOME
-----
Translation table: Standard 2000
>Seq:
Reading frame: 1
      N V L S A A A D E G S H V H A A V S E V G S H A A E V S L E A L
1' ATGTTCTCTCTCCGCGCAGAGGCAATGTCAGAGCCGCTGAGAGGCAAGTTGAGGAGGCTGCAAGATATGCGCAGAGGCGCTG 80
   >>>...)))
      E K S F L S F F T T E V F R F P L S H S S A G V S H S S
1' GAGAGAGCTTCTTAACTTCCGACACAGACTACTTCCGCAATTTGACCTGAGCAGAGGCTGCGAGAGGCTGAGAGGCAAGG 180
   >>>...)))
      A E V A A A I T E A V E R L C D D I P G A I S E L S D L R A E
1' GCGAGGTTGCGCCGCTGACAGAGGTTGAGACTTGAAGACTTCCGCTTCCCTTCTTAACTGATGACTTGAAGCTTGA 270
   >>>...)))
      F L S V D F V V F N L S H S L I V T I A S H L P S D F T E
1' AAGTCTGCTTGGAGGCTTAACTTAACTTAACTTAACTTAACTTAACTTAACTTAACTTAACTTAACTTAACTTAACTTAACTTAA 360
   >>>...)))
      A V N A S I D E F L A H V S T V I T S E Y S *
1' GCGTTGCGCCGCTTCCGACAGGCTTCCGACAGGCTTCCGACAGGCTTCCGACAGGCTTCCGACAGGCTTCCGACAGGCTTCC 420
   >>>...)))
Annotation key:
>>> | START CODON (START)
))) | START CODON (ALTERNATIVE)
*** | STOP
  
```

Copyright 2011 © Limsoon Wong

## Recognition of Translation Initiation Sites

An introduction to the World's simplest TIS recognition system

## Translation Initiation Site

Copyright 2011 © Limsoon Wong

## A Sample cDNA

```

299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCAAGCCATGCTGAACACTGACTCCCAGCTGTG 80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTGGCTGTCCAGGCAGCTGTA 160
GGAGGCAGATGAGAGAGAGGGAGATGGCCTTGAGGAGAGGGAAGGGCCCTGGTCCCGAGGA 240
CCTCTCCTGGCCAGGAGCTTCTCCAGGACAAGACTTCCACCCCAACAGGACTCCCT
.....
.....|EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE 80
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE 160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE 240
  
```

- What makes the second ATG the TIS?

Copyright 2011 © Limsoon Wong

13

### 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, ...

Copyright 2011 © Limsoon Wong

14

### 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...

Copyright 2011 © Limsoon Wong

15

### Signal Generation

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

Copyright 2011 © Limsoon Wong

16

### Signal Generation: An Example

```

299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGGGACTGGCCCAAGCCATGGCTGAACAATGATCTCCCACTGTG 80
CCCAAGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCCAGGGCACTGTG 160
GGAGCCAGATGAAGAGAGGGAGATGGCTTGGAGGAGGGAAAGGGCTGGTCCGAGGA 240
CCTCTCTGGCCAGGAGCTTCTCCAGGACAAAGACTTCCACCCCAACAGGACTCCCT
  
```

- Window =  $\pm 100$  bases
- In-frame, downstream
  - GCT = 1, TTT = 1, ATG = 1... Exercise: Find the in-frame downstream ATG
- Any-frame, downstream
  - GCT = 3, TTT = 2, ATG = 2... Exercise: What are the possible k-grams (k=3) in this sequence?
- In-frame, upstream
  - GCT = 2, TTT = 0, ATG = 0, ...

Copyright 2011 © Limsoon Wong

17

### Feature Generation - Summary

**Raw Data**

```

> 205 |BB|KL1 |L1 |X1 |71655 |H|P|Y|T|A|I |S|P|Y|T|A|I |M|A|T|A|I |L|I|D|A|T|
CGCTCAGAGCCCACTCTCTCTCTGGCCCAAGCCATGGCTGAACAATGATCTCCCACTGTG
TTTGGCTGTCCAGGGCACTGTGGAGCCAGATGAAGAGAGGGAGATGGCTTGGAGGAGGGAAAGGGCTGGTCCGAGGA
TGGTCTTTGAGTGTGGAGTTCATGTCTCTACCTGAGTTACAC.....TTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
  
```

**An ATG segment – positive sample**

```

> 284 |I |I|x(56)
#####CCTCAGACCCCCACACTCTTCTGTCTCCAGCCAG
:CCCCCAGCCCAAAATGGAAATGAGCCACTATCCTTTGGAAATGCTCAGACTTGTATGCAGATGAATTA
:AGGCTAGCAGAGAGATTAGAGAGCTCATTGGACAAAT
  
```

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

```

1,0,0,1,0,0,1,2,0,0,0,0,0,0,0,0,1,0,2,0,2,1,0,0,0,1,0,0,0,0,0,2,0,
4,0,0,0,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,1,1,0,0,1,1,
0,1,0,0,0,0,1,0,0,0,1,1,0,0,2,1,1,0,2,0,0,2,0,0,0,0,0,0,0,1,1,0,0,0,
0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
  
```

Copyright 2011 © Limsoon Wong


18

### Too Many Features

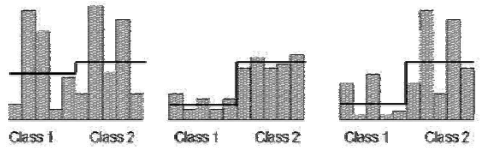
- For each value of k, there are  $4^k \times 3 \times 2$  k-grams
- If we use k = 1, 2, 3, 4, 5, we have  $24 + 96 + 384 + 1536 + 6144 = 8184$  features!
- This is too many for most machine learning algorithms

Copyright 2011 © Limsoon Wong

19

Signal Selection (Basic Idea) 


- Choose a signal w/ low intra-class distance
- Choose a signal w/ high inter-class distance



Class 1    Class 2    Class 1    Class 2    Class 1    Class 2

Copyright 2011 © Limsoon Wong

20

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  $i$ .

Copyright 2011 © Limsoon Wong

21

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  $i$ .

Copyright 2011 © Limsoon Wong

22

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  $A_{ij}$  ( $E_{ij} = R_i * C_j / N$ ).

Copyright 2011 © Limsoon Wong

23

Example 

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


	obs	exp	(obs - exp) <sup>2</sup> /exp
HM	40	60*50/100=30	3.3
HW	20	60*50/100=30	3.3
LM	10	40*50/100=20	5.0
LW	30	40*50/100=20	5.0

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

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

Copyright 2011 © Limsoon Wong

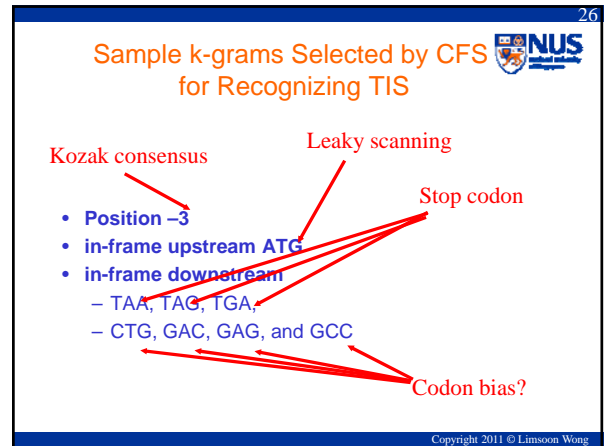
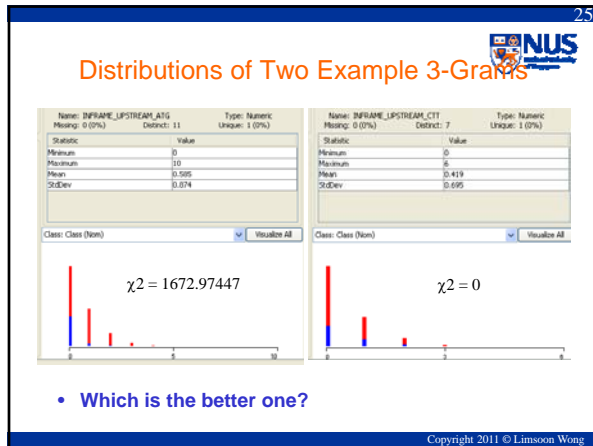
24

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?

Copyright 2011 © Limsoon Wong



- 27
- ### 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, ...**
- Copyright 2011 © Limsoon Wong

28

### Results (3-fold x-validation)

	predicted as positive	predicted as negative
positive	TP	FN
negative	FP	TN

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

	TP/(TP + FN)	TN/(TN + FP)	TP/(TP + FP)	Accuracy
Naïve Bayes	84.3%	86.1%	66.3%	85.7%
SVM	73.9%	93.2%	77.9%	88.5%
Neural Network	77.6%	93.2%	78.8%	89.4%
Decision Tree	74.0%	94.4%	81.1%	89.4%

Copyright 2011 © Limsoon Wong

29

### Improvement by Voting

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

	TP/(TP + FN)	TN/(TN + FP)	TP/(TP + FP)	Accuracy
NB+SVM+NN	79.2%	92.1%	76.5%	88.9%
NB+SVM+Tree	78.8%	92.0%	76.2%	88.8%
NB+NN+Tree	77.6%	94.5%	82.1%	90.4%
SVM+NN+Tree	75.9%	94.3%	81.2%	89.8%
Best of 4	84.3%	94.4%	81.1%	89.4%
Worst of 4	73.9%	86.1%	66.3%	85.7%

Copyright 2011 © Limsoon Wong

30

### 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

	TP/(TP + FN)	TN/(TN + FP)	TP/(TP + FP)	Accuracy
NB	84.3%	86.1%	66.3%	85.7%
SVM	73.9%	93.2%	77.9%	88.5%
NB+Scanning	87.3%	96.1%	87.9%	93.9%
SVM+Scanning	88.5%	96.3%	88.6%	94.4%

Copyright 2011 © Limsoon Wong

### Performance Comparisons

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

\* result not directly comparable

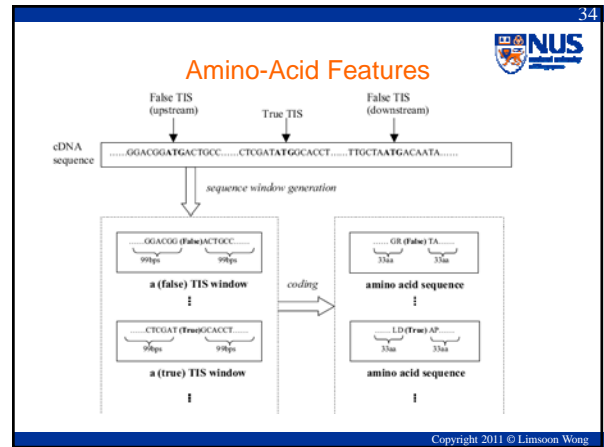
- ### Technique Comparisons
- Pedersen&Nielsen [ISMB'97]**
    - Neural network
    - No explicit features
  - Zien [Bioinformatics'00]**
    - SVM+kernel engineering
    - No explicit features
  - Hatzigeorgiou [Bioinformatics'02]**
    - Multiple neural networks
    - Scanning rule
    - No explicit 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?

First	U	C	A	G	Last
U	Phe	Ser	Thr	Cys	U
Phe	F	S	Y	C	C
Ser	L	Stop (Choker)	Stop (Choker)	A	C
Leu	L	Ser	Stop (Anker)	Trp	W
C	Leu	Pro	Phe	H	R
Leu	Pro	Glu	Arg	A	C
Leu	Pro	Glu	Arg	A	A
A	De	I	Asn	Ser	U
De	Thr	Asn	N	Ser	C
De	Thr	Lys	K	Arg	A
Met	M	Thr	Lys	K	G
G	Val	Ala	Arg	Ob	C
Val	Ala	Ob	E	Ob	A
Val	Ala	Ob	Ob	Ob	G

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



### Amino-Acid Features

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

42 1-gram amino acid patterns	882 2-gram amino acid patterns	3 bio-knowledge patterns	class label
UP-A, UP-R, ..., UP-N, DOWN-A, DOWN-R, ..., DOWN-N (numeric type)	UP-AA, UP-AR, ..., UP-NN, DOWN-AA, DOWN-AR, ..., DOWN-NN (numeric type)	DOWN-G UP3-AorG, UP-ATG (boolean type, Y or N)	True, False

Frequency as values

1, 3, 5, 0, 4, ...	6, 2, 7, 0, 5, ...	N, N, N,	False
↓	↓	↓	↓
6, 5, 7, 9, 0, ...	2, 0, 3, 10, 0, ...	Y, Y, Y,	True
↓	↓	↓	↓

### Amino Acid K-grams Discovered (by entropy)

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

Fold	UP-ATG	DOWN-STOP	UP3-AorG	DOWN-A	DOWN-V	UP-A	DOWN-L	DOWN-D	DOWN-E	UP-G
1	1	2	4	3	6	5	8	9	7	10
2	1	2	3	4	5	6	7	8	9	10
3	1	2	3	4	5	6	8	9	7	10

37

**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)

Copyright 2011 © Limsoon Wong

38

**Validation Results (on Hatzigeorgiou's)**

Algorithm	Sensitivity	Specificity	Precision	Accuracy
SVMs(linear)	96.28%	89.15%	25.31%	89.42%
SVMs(quad)	94.14%	90.13%	26.70%	90.28%
Ensemble Trees	92.02%	92.71%	32.52%	92.68%

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

Copyright 2011 © Limsoon Wong

39

**Validation Results (on Chr X and Chr 21)**

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

Copyright 2011 © Limsoon Wong

40

**About the Inventor: Huiqing Liu**

- **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

Copyright 2011 © Limsoon Wong

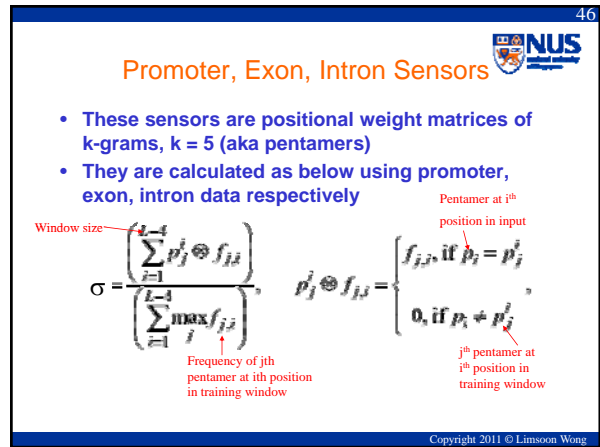
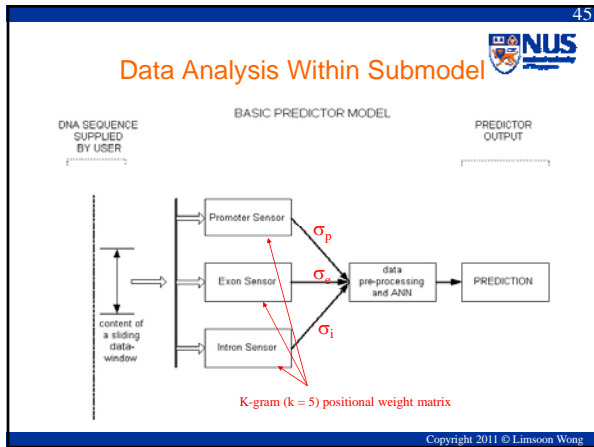
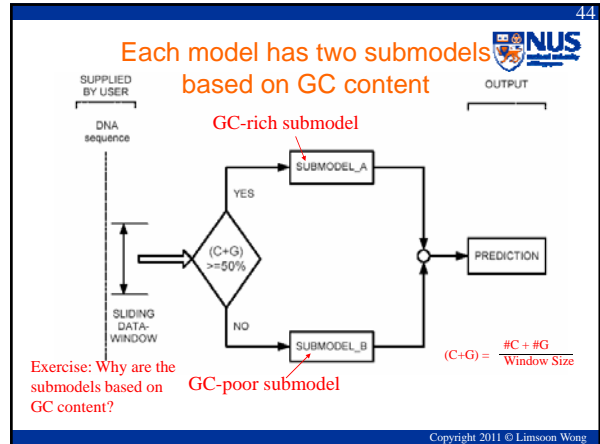
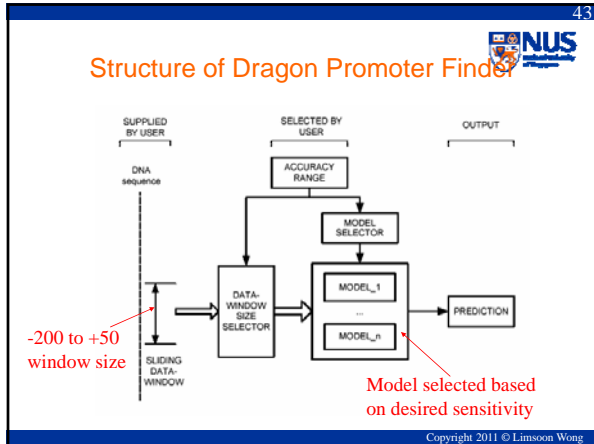
**Recognition of Transcription Start Sites**

An introduction to the World's best TSS recognition system:  
A heavy tuning approach

42

**Transcription Start Site**

Copyright 2011 © Limsoon Wong



47

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

- Give me 3 DNA seq of length 10:
  - Seq<sub>1</sub> = ACCGAGTTCT
  - Seq<sub>2</sub> = AGTGACCTG
  - Seq<sub>3</sub> = AGTTCGTATG
- Then

1-mer	pos1	pos2	pos3	pos4	pos5	pos6	pos7	pos8	pos9	pos10
A	3/3	0/3	0/3							
C	0/3	1/3	1/3							
G	0/3	2/3	0/3							
T	0/3	0/3	2/3							

Exercise: Fill in the rest of the table

Copyright 2011 © Limsoon Wong

48

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

- Give me 3 DNA seq of length 10:
  - Seq<sub>1</sub> = ACCGAGTTCT
  - Seq<sub>2</sub> = AGTGACCTG
  - Seq<sub>3</sub> = AGTTCGTATG
- Then

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

2-mer	pos1	pos2	pos3	pos4	pos5	pos6	pos7	pos8	pos9
AA	0/3	0/3	0/3						
AC	1/3	0/3	0/3						
...	...	...	...						
TT	0/3	0/3	1/3					1/3	

Exercise: Fill in the rest of the table

Copyright 2011 © Limsoon Wong



49

## Data Preprocessing & ANN

Tuning parameters

$$s_E = \text{sat}(\sigma_E - \sigma_{E_c}, a_E, b_E)$$

$$s_I = \text{sat}(\sigma_I - \sigma_{I_c}, a_I, b_I)$$

$$s_{EI} = \text{sat}(\sigma_{EI} - \sigma_{EI_c}, a_{EI}, b_{EI})$$

where the function  $\text{sat}$  is defined by

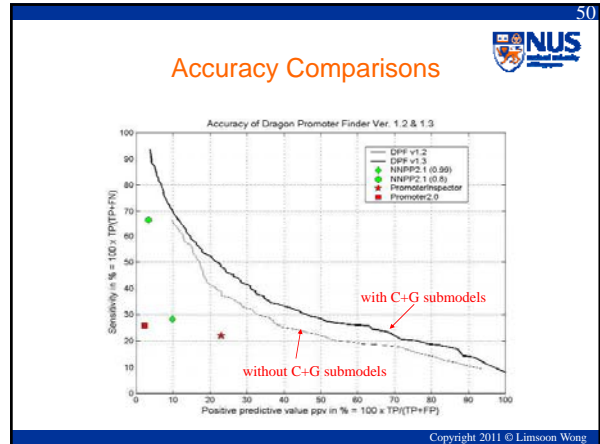
$$\text{sat}(x, a, b) = \begin{cases} a, & \text{if } x > a \\ x, & \text{if } b \leq x \leq a \\ b, & \text{if } b > x \end{cases}$$

Simple feedforward ANN trained by the Bayesian regularisation method

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

$\text{net} = \sum s_i * w_i$

Copyright 2011 © Limsoon Wong



51

## 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

Copyright 2011 © Limsoon Wong

52

## 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

Copyright 2011 © Limsoon Wong

53

## 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.

Copyright 2011 © Limsoon Wong

54


## About the Inventor: Vlad Bajic

- **Vladimir B. Bajic**
  - Principal Scientist, I<sup>2</sup>R, 2001-2006
  - Currently Director & Professor, Computational Bioscience Research Center, KAUST

Copyright 2011 © Limsoon Wong

## Recognition of Poly-A Signal Sites

A twist to the "feature generation, feature selection, feature integration" approach



## Eukaryotic Pre-mRNA Processing

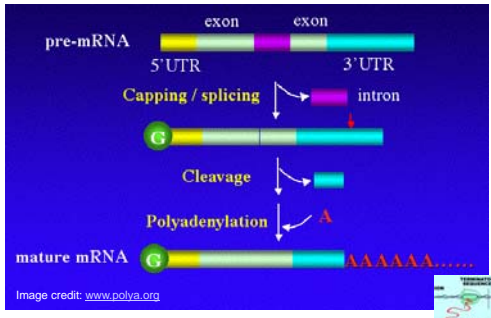



Image credit: [www.polya.org](http://www.polya.org)  
Copyright 2011 © Limsoon Wong

## 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**



Source: Wikipedia  
Copyright 2011 © Limsoon Wong

## Poly-A Signals in Human (Gautheret et al., 2006)

Hexamer	Observed (expected)*	% sites	p <sup>h</sup>	Position average ± SD	Location†
AAUAAA	3286 (317)	58.2	0	-16 ± 4.7	500
AUUAAA	843 (112)	14.9	0	-17 ± 5.3	150
AGUAAA	156 (32)	2.7	6 × 10 <sup>-27</sup>	-16 ± 5.9	90
UAUAAA	180 (53)	3.2	4 × 10 <sup>-22</sup>	-18 ± 7.8	10
CAUAAA	76 (23)	1.3	1 × 10 <sup>-16</sup>	-17 ± 5.9	10
GAUAAA	72 (21)	1.3	2 × 10 <sup>-16</sup>	-18 ± 6.9	10
AAUAAU	96 (33)	1.7	2 × 10 <sup>-17</sup>	-18 ± 6.9	10
AAUACA	70 (16)	1.2	5 × 10 <sup>-23</sup>	-18 ± 8.7	10
AAUAGA	43 (14)	0.7	1 × 10 <sup>-24</sup>	-18 ± 6.3	10
AAAAAG	49 (11)	0.8	5 × 10 <sup>-17</sup>	-18 ± 8.9	10
ACUAAA	36 (11)	0.6	1 × 10 <sup>-26</sup>	-17 ± 6.1	10
AAGAAA	62 (10)	1.1	9 × 10 <sup>-28</sup>	-19 ± 11	10
AAUGAA	49 (10)	0.8	4 × 10 <sup>-18</sup>	-20 ± 10	10
UUUAAA	69 (20)	1.2	3 × 10 <sup>-18</sup>	-17 ± 12	10
AAAAACA	29 (5)	0.5	8 × 10 <sup>-13</sup>	-20 ± 10	10
GGGGCU	22 (3)	0.3	9 × 10 <sup>-12</sup>	-24 ± 13	0

Copyright 2011 © Limsoon Wong

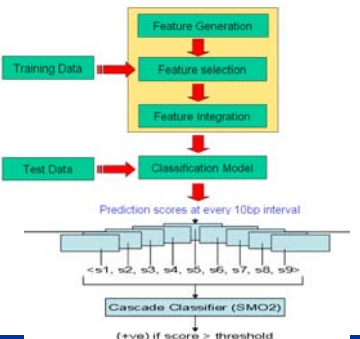
## Poly-A Signals in Arabidopsis

Hexamer	Observed (expected)*	% sites	p <sup>h</sup>	Position average ± SD	Location†
AAUAAA	3286 (317)	58.2	0	-16 ± 4.7	500
AUUAAA	843 (112)	14.9	0	-17 ± 5.3	150
AGUAAA	156 (32)	2.7	6 × 10 <sup>-27</sup>	-16 ± 5.9	90
UAUAAA	180 (53)	3.2	4 × 10 <sup>-22</sup>	-18 ± 7.8	10
CAUAAA	76 (23)	1.3	1 × 10 <sup>-16</sup>	-17 ± 5.9	10
GAUAAA	72 (21)	1.3	2 × 10 <sup>-16</sup>	-18 ± 6.9	10
AAUAAU	96 (33)	1.7	2 × 10 <sup>-17</sup>	-18 ± 6.9	10
AAUACA	70 (16)	1.2	5 × 10 <sup>-23</sup>	-18 ± 8.7	10
AAUAGA	43 (14)	0.7	1 × 10 <sup>-24</sup>	-18 ± 6.3	10
AAAAAG	49 (11)	0.8	5 × 10 <sup>-17</sup>	-18 ± 8.9	10
ACUAAA	36 (11)	0.6	1 × 10 <sup>-26</sup>	-17 ± 6.1	10
AAGAAA	62 (10)	1.1	9 × 10 <sup>-28</sup>	-19 ± 11	10
AAUGAA	49 (10)	0.8	4 × 10 <sup>-18</sup>	-20 ± 10	10
UUUAAA	69 (20)	1.2	3 × 10 <sup>-18</sup>	-17 ± 12	10
AAAAACA	29 (5)	0.5	8 × 10 <sup>-13</sup>	-20 ± 10	10
GGGGCU	22 (3)	0.3	9 × 10 <sup>-12</sup>	-24 ± 13	0

**In contrast to human, PAS in Arab is highly degenerate. E.g., only 10% of Arab PAS is AAUAAA!**

Copyright 2011 © Limsoon Wong

## Approach on Arab PAS Sites (I)



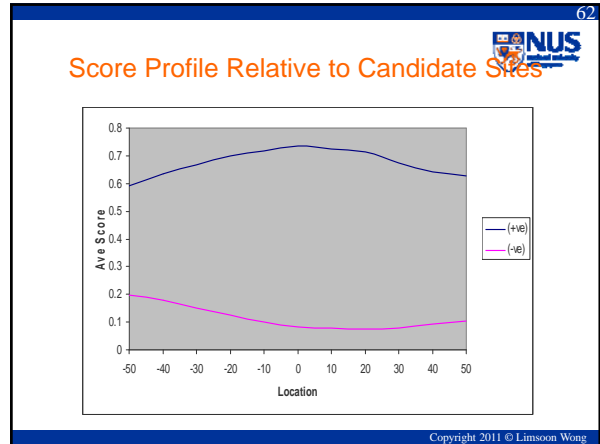
Copyright 2011 © Limsoon Wong

61

**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\*7, etc)
  - Freq of features above in 3 diff windows: (-110/+5), (-35/+15), (-50/+30)
- **Feature selection**
  - $\chi^2$
- **Feature integration & Cascade**
  - SVM

Copyright 2011 © Limsoon Wong



63

**Validation Results**

SN 0	SMO 1		SMO 2		PASS 1.0	
	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
Control Sequences						
CDS	90%	0.36	94%	0.34	95%	3.7
5'UTR	79%	0.42	83%	0.49	78%	5.5
Intron	64%	0.59	71%	0.67	63%	6.3

Table 2. Equal error-rate points of SMO1, SMO2, and PASS 1.0 for SN\_10.

SN 10	SMO 1		SMO 2		PASS 1.0	
	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
Control Sequences						
CDS	94%	0.36	96%	0.31	96%	4
5'UTR	86%	0.53	89%	0.6	81%	5.7
Intron	73%	0.63	77%	0.77	67%	6.6

Table 3. Equal error-rate points of SMO1, SMO2, and PASS 1.0 for SN\_30.


SN 30	SMO 1		SMO 2		PASS 1.0	
	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
Control Sequences						
CDS	97%	0.44	97%	0.37	97%	4.3
5'UTR	90%	0.42	92%	0.67	84%	6.2
Intron	79%	0.75	83%	0.81	72%	6.8

Copyright 2011 © Limsoon Wong

64

**About the Inventor: Koh Chuan Hock**

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



Copyright 2011 © Limsoon Wong

**Concluding Remarks...**



66

**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 → amino acid)
  - Classifier cascades


Copyright 2011 © Limsoon Wong

Any Question?



68

## Acknowledgements




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

Copyright 2011 © Limsoon Wong

69

## References (TIS Recognition)




- A. G. Pedersen, H. Nielsen, "Neural network prediction of translation initiation sites in eukaryotes", *ISMB* 5:226--233, 1997
- 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

Copyright 2011 © Limsoon Wong

70

## References (TSS Recognition)




- V.B.Bajic et al., "Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates", *J. Mol. Graph. & Mod.* 21:323--332, 2003
- J.W.Fickett, A.G.Hatzigeorgiou, "Eukaryotic promoter recognition", *Gen. Res.* 7:861--878, 1997
- M.Scherf et al., "Highly specific localisation of promoter regions in large genome sequences by PromoterInspector", *JMB* 297:599--606, 2000
- V. B. Bajic and A. Chong. "Tuning the Dragon Promoter Finder System for Human Promoter Recognition", *The Practical Bioinformatician*, Chapter 7, pages 157--165, 2004

Copyright 2011 © Limsoon Wong

71

## References (PAS Recognition)




- Q. Li et al., "Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures". *Plant Physiology*, 138:1457-1468, 2005
- J. E. Tabaska, M. Q. Zhang, "Detection of polyadenylation signals in human DNA sequences". *Gene*, 231:77-86, 1999
- M. Legendre, D. Gautheret, "Sequence determinants in human polyadenylation site selection". *BMC Genomics*, 4:7, 2003
- B. Tian et al., "Prediction of mRNA polyadenylation sites by support vector machine". *Bioinformatics*, 22:2320-2325, 2006
- C. H. Koh, L. Wong. "Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences". *Proc. GIW 2007*, pages 73--82

Copyright 2011 © Limsoon Wong

72

## References (Feature Selection)



- M. A. Hall, "Correlation-based feature selection machine learning", PhD thesis, Dept of Comp. Sci., Univ. of Waikato, New Zealand, 1998
- U. M. Fayyad, K. B. Irani, "Multi-interval discretization of continuous-valued attributes", *IJCAI* 13:1022-1027, 1993
- H. Liu, R. Sentiono, "Chi2: Feature selection and discretization of numeric attributes", *IEEE Intl. Conf. Tools with Artificial Intelligence* 7:338--391, 1995

Copyright 2011 © Limsoon Wong