

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

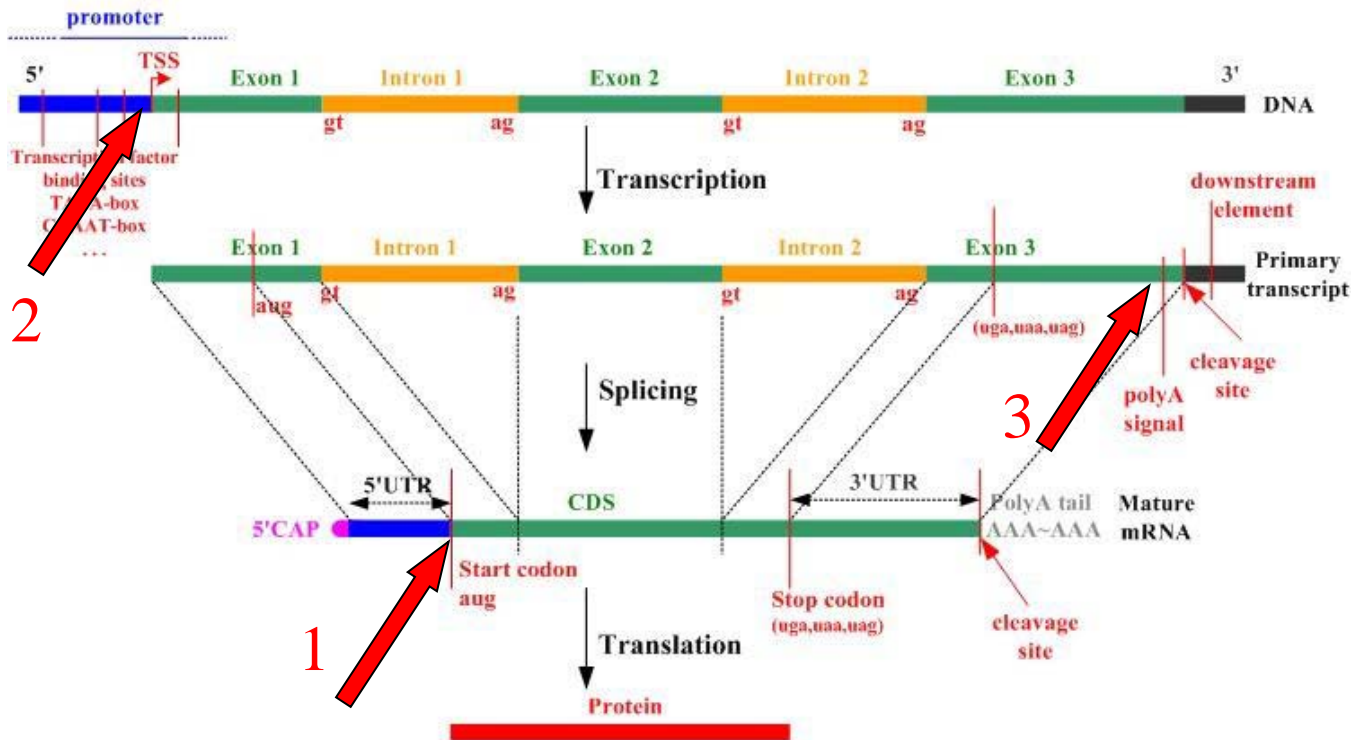
# CS2220: Introduction to Computational Biology

## Lecture 3: Gene Feature Recognition

**Limsoon Wong**



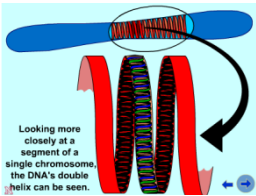
# Plan



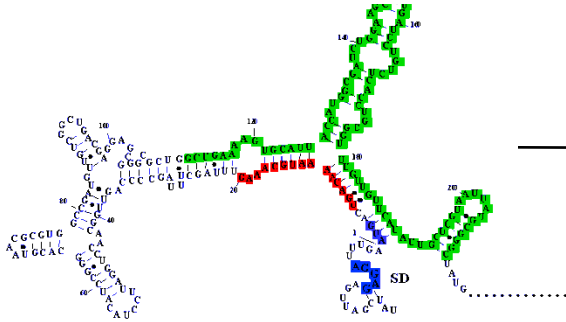
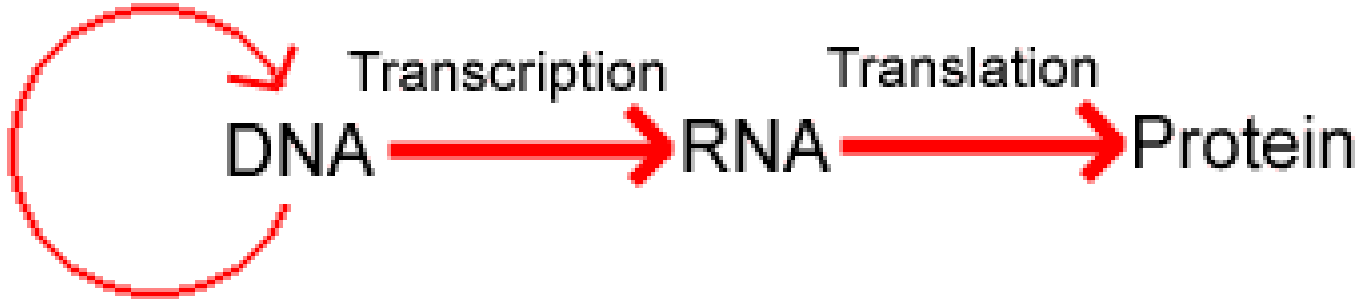
# Some Relevant Biology



# Central Dogma



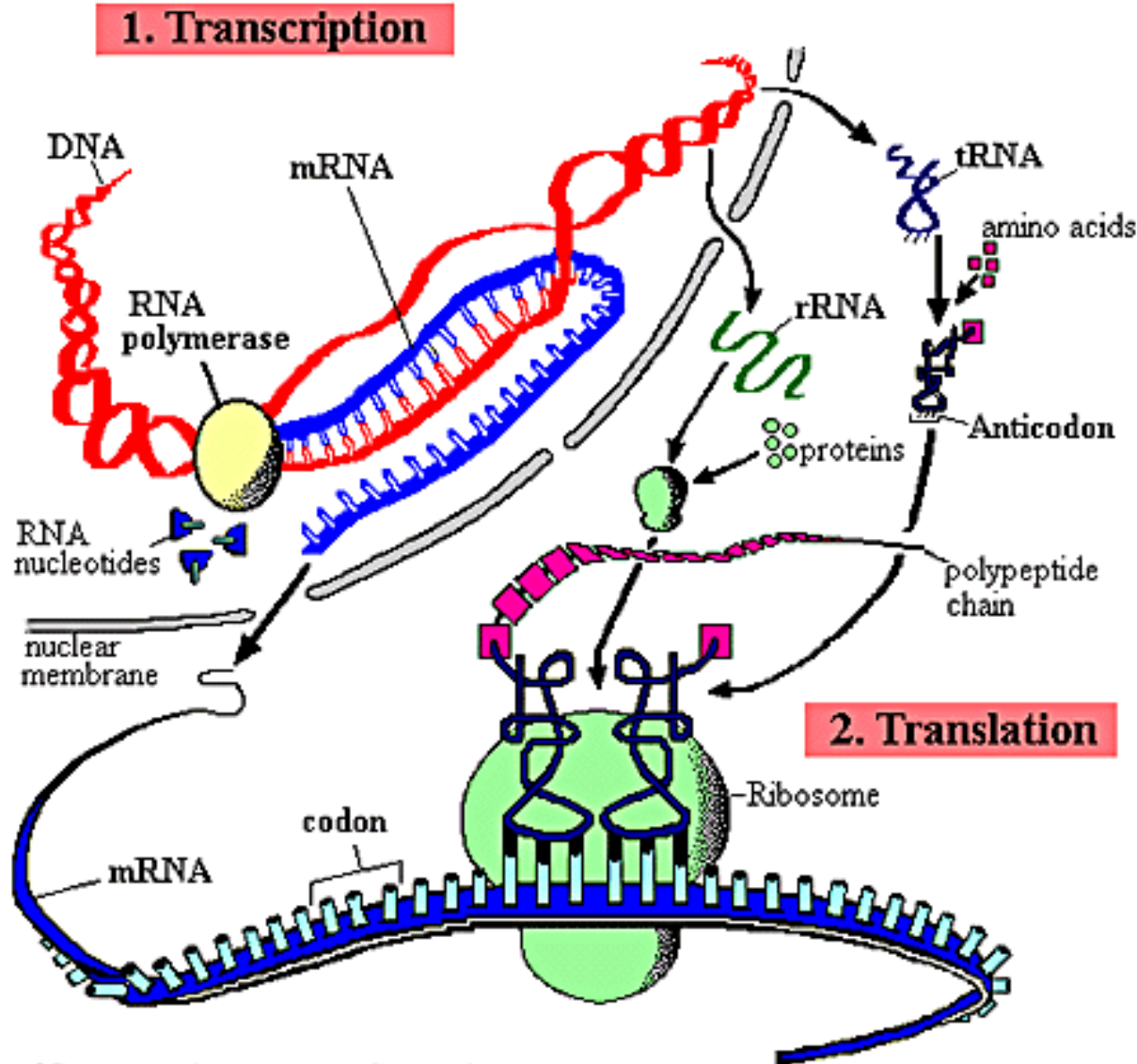
Replication



...AAUGGUACCGAUGACCUGGAGC...

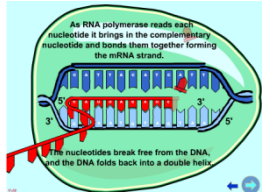
...AATGGTACCGATGACCTG...

...TRLRPLLALLALWP...



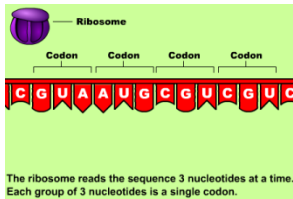
# Players in Protein Synthesis

## 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 → 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



# 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

		Second Position of Codon					
		T	C	A	G		
F i r s t  P o s i t i o n	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T	T h i r d  P o s i t i o n
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C	
		TTA Leu [L]	TCA Ser [S]	TAA <i>Ter</i> [end]	TGA <i>Ter</i> [end]	A	
		TTG Leu [L]	TCG Ser [S]	TAG <i>Ter</i> [end]	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]	CGC Arg [R]	C	
		CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A	
		CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG 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	
		GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C	
		GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A	
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G	



# Example

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

```

VIRTUAL RIBOSOME
----
Translation table: Standard SGC0

>Seq1
Reading frame: 1

  M V L S A A D K G N V K A A W G K V G G H A A E Y G A E A L
5' ATGGTGTCTGTCTGCCCGACAAAGGGCAATGTCAAGGCCGCCTGGGGCAAGGTTGGCGGCCACGCTGCAGAGTATGGCGCAGAGGCCCTG 90
  >>>...))).....

  E R M F L S F P T T K T Y F P H F D L S H G S A Q V K G H G
5' GAGAGGATGTTCCCTGAGCTTCCCCACCACCAAGACCTACTTCCCCCACTTCGACCTGAGCCACGGCTCCGCGCAGGTCAAGGGCCACGGC 180
  .....>>>...))).....

  A K V A A A L T K A V E H L D D L P G A L S E L S D L H A H
5' GCGAAGGTGGCCCGCCGCTGACCAAGCGGTGGAACACCTGGACGACCTGCCCGGTGCCCTGTCTGAACTGAGTGACCTGCACGCTCAC 270
  .....))).....))).....))).....))).....))).....))).....

  K L R V D P V N F K L L S H S L L V T L A S H L P S D F T P
5' AAGCTGCGTGTGGACCCGGTCAACTTCAAGCTTCTGAGCCACTCCCTGCTGGTGACCCCTGGCCCTCCCACCTCCCCAGTGATTTACACCCC 360
  ...))).....))).....))).....))).....))).....))).....

  A V H A S L D K F L A N V S T V L T S K Y R *
5' GCGGTCCACGCCTCCCTGGACAAGTTCTTGGCCAACGTGAGCACCGTGCTGACCTCCAAATACCGTTAA 429
  .....))).....))).....))).....))).....***)

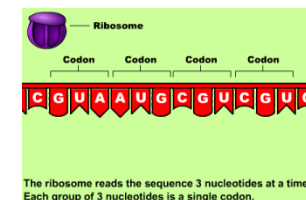
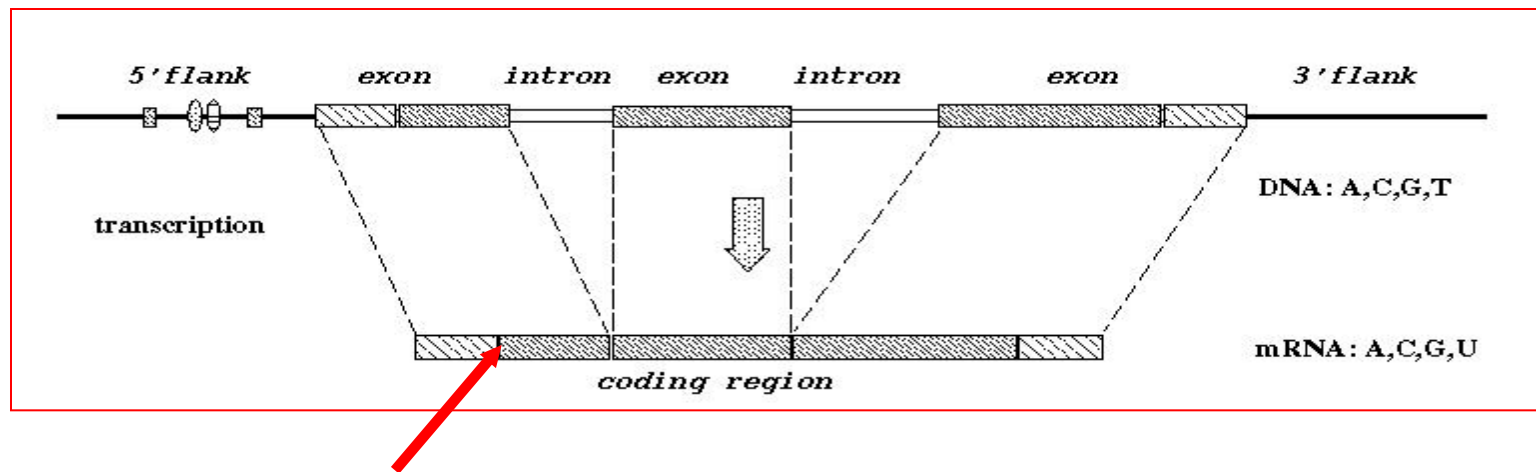
Annotation key:
>>> : START codon (strict)
))) : START codon (alternative)
*** : STOP
  
```

# Recognition of Translation Initiation Sites

**An introduction to the World's simplest TIS  
recognition system**



# Translation Initiation Site



# A Sample cDNA

```

299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTGAGGGCAGCTGTA      160
GGAGGCAGATGGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA      240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
.....                                                                    80
.....iEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE      160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE      240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
  
```

- What makes the second ATG the TIS?

# Approach

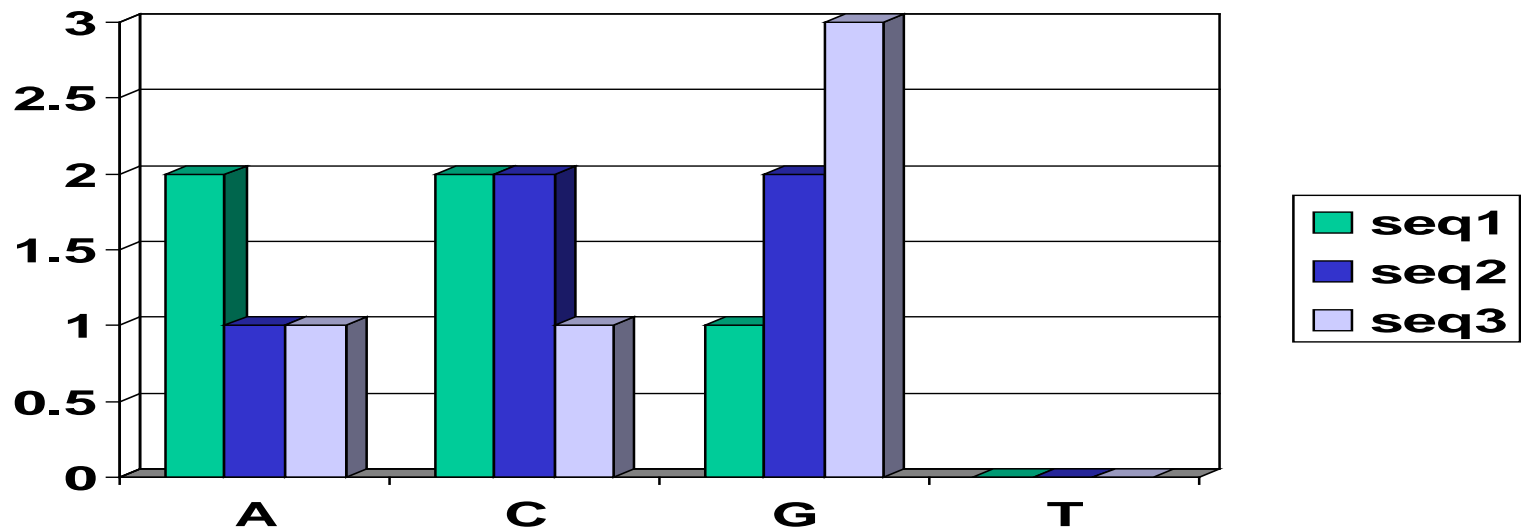
- **Training data gathering**
- **Signal generation**
  - k-grams, distance, domain know-how, ...
- **Signal selection**
  - Entropy,  $\chi^2$ , CFS, t-test, domain know-how...
- **Signal integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...

# 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)**
  - $K = 1, 2, 3, 4, 5, \dots$
  - Window size vs. fixed position
  - Up-stream, downstream vs. any where in window
  - In-frame vs. any frame



# Signal Generation: An Example

```

299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCAAGCCATGGCTGAACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA    160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGGCGCTGGTGCCGAGGA      240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
  
```

- **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?



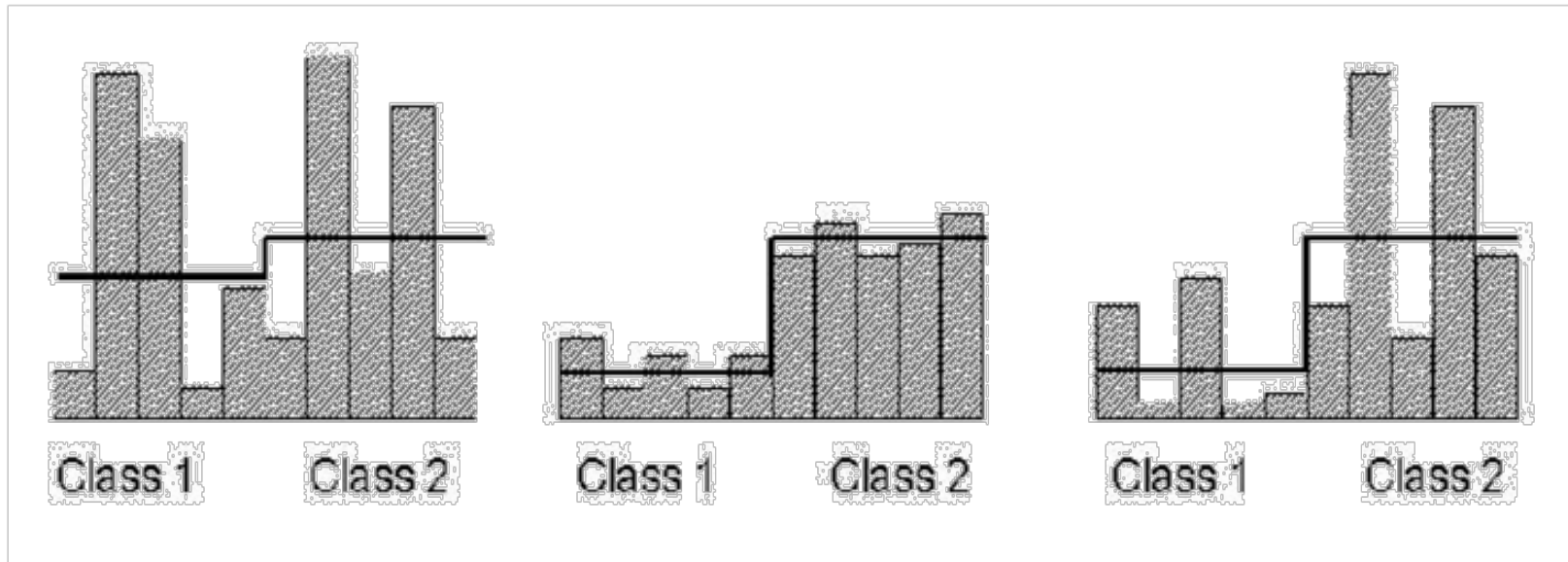


# Too Many Features

- For each value of  $k$ , there are  $4^k * 3 * 2$   $k$ -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 w/ low intra-class distance
- Choose a signal w/ 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  $A_{ij}$  ( $E_{ij} = R_i * C_j / N$ ).

# Example

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

	obs	exp	$(\text{obs} - \text{exp})^2/\text{exp}$
HM	40	$60 \cdot 50 / 100 = 30$	3.3
HW	20	$60 \cdot 50 / 100 = 30$	3.3
LM	10	$40 \cdot 50 / 100 = 20$	5.0
LW	30	$40 \cdot 50 / 100 = 20$	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?

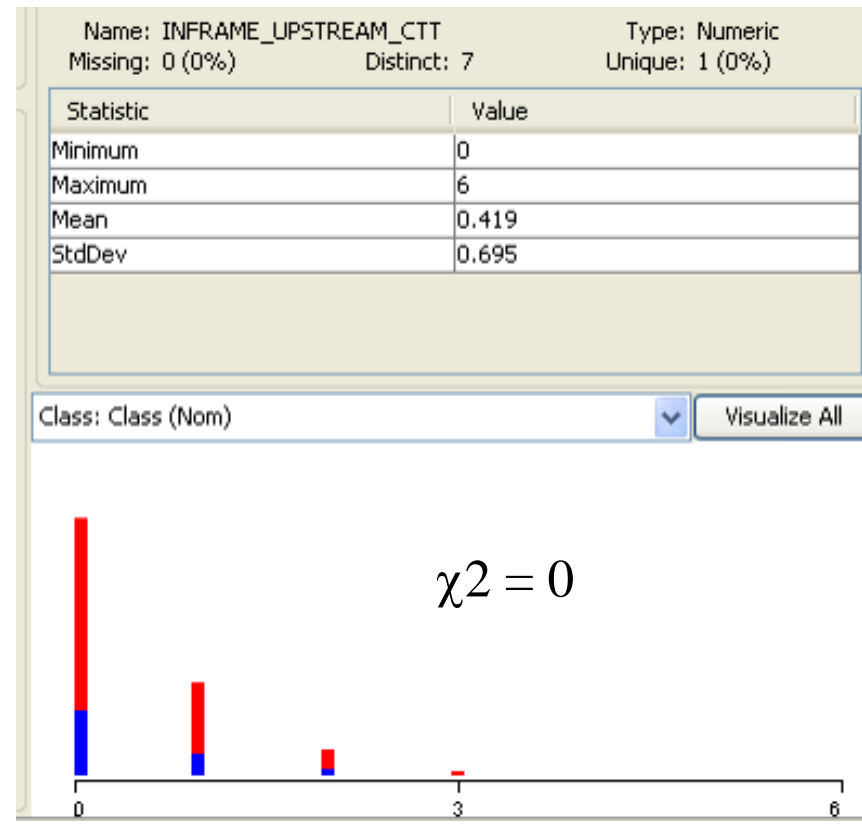
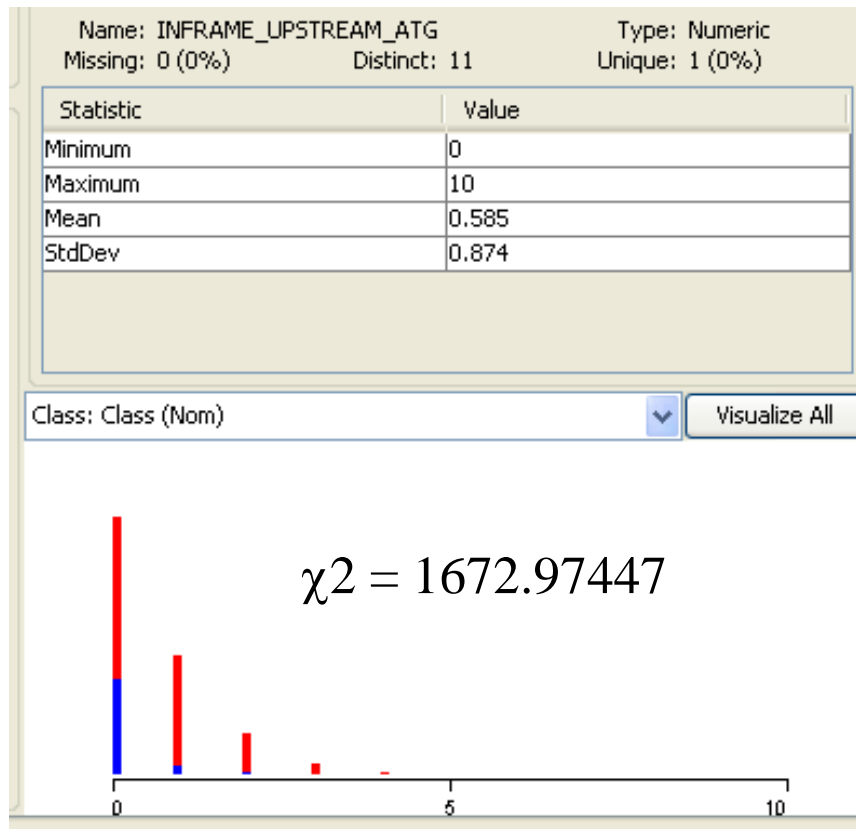
# Signal Selection (e.g., CFS)

- **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?

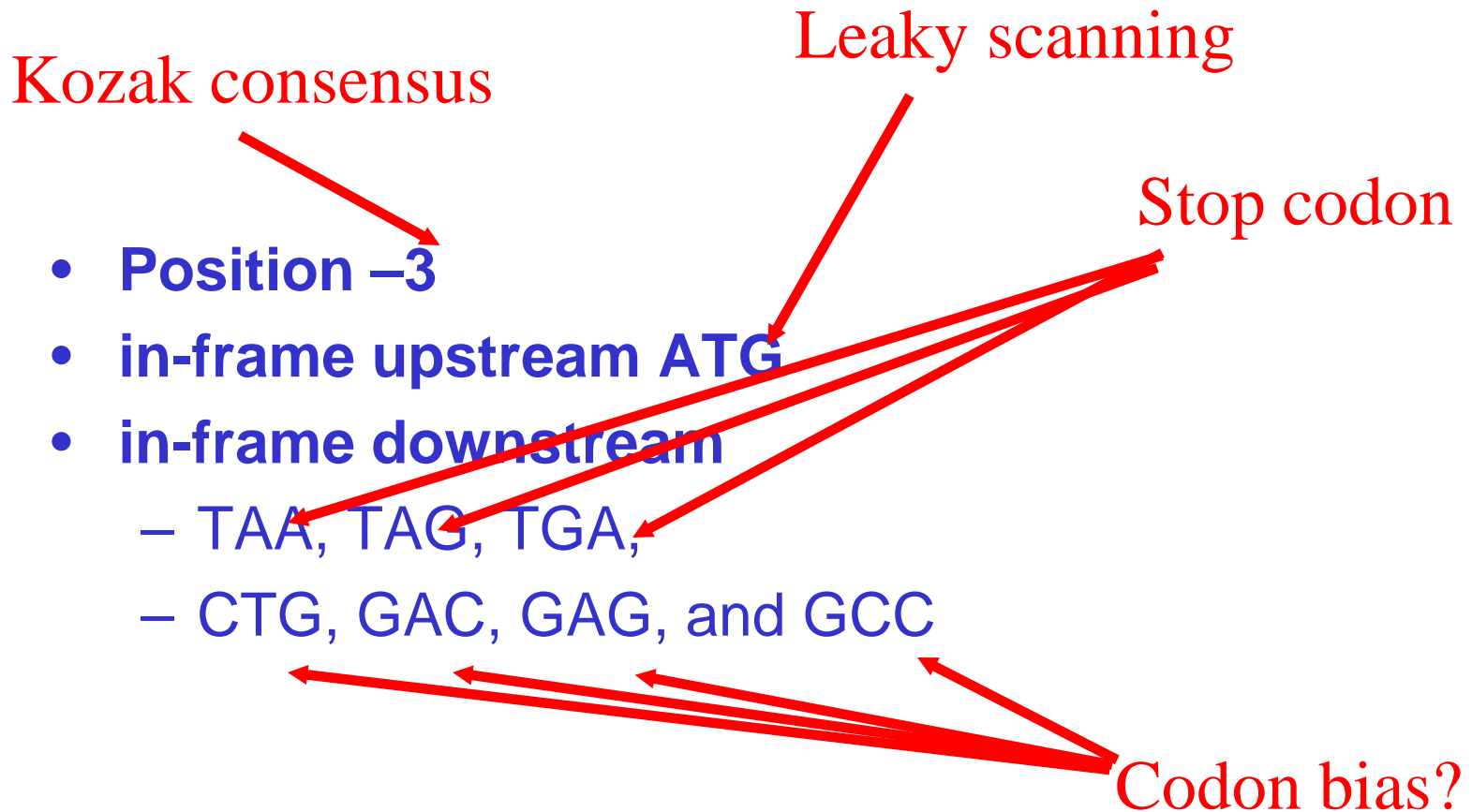


# Distributions of Two Example 3-Grams



- Which is the better one?

# Sample k-grams Selected by CFS for Recognizing TIS



# 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)

	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%

---

# Improvement by Voting

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

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

# 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

---

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

---

# 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%*

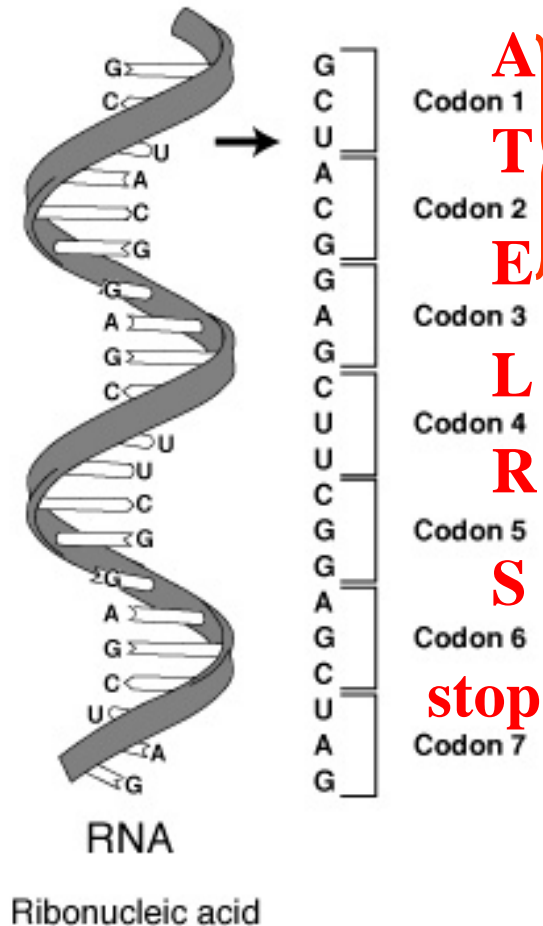
\* result not directly comparable

# 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

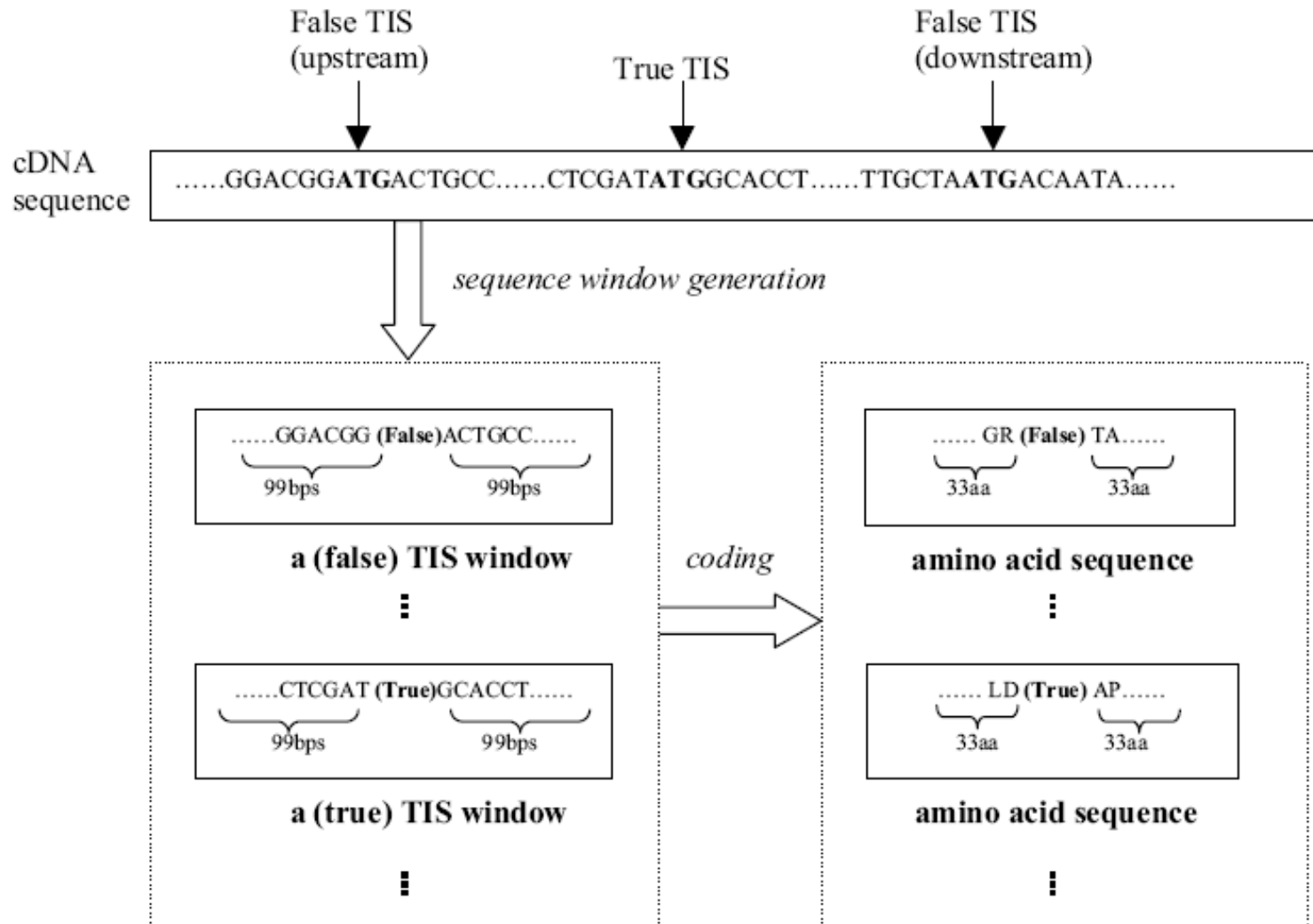


How about using k-grams from the translation?

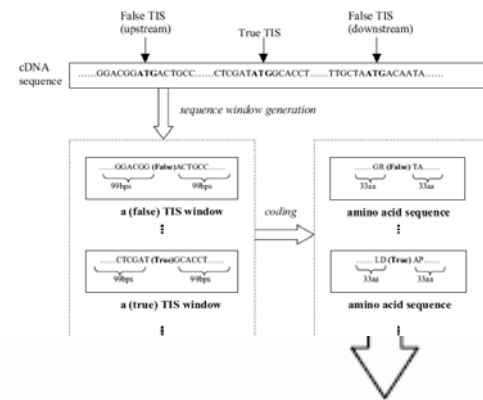
First	U	C	A	G	Last
U	Phe <b>F</b>	Ser <b>S</b>	Tyr <b>Y</b>	Cys <b>C</b>	U
	Phe	Ser	Tyr	Cys	C
	Leu <b>L</b>	Ser	Stop (Ochre)	Stop (Umber)	A
	Leu	Ser	Stop (Amber)	Trp <b>W</b>	G
C	Leu	Pro <b>P</b>	His <b>H</b>	Arg <b>R</b>	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln <b>Q</b>	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile <b>I</b>	Thr <b>T</b>	Asn <b>N</b>	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys <b>K</b>	Arg	A
	Met <b>M</b>	Thr	Lys	Arg	G
G	Val <b>V</b>	Ala <b>A</b>	Asp <b>D</b>	Gly <b>G</b>	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu <b>E</b>	Gly	A
	Val	Ala	Glu	Gly	G

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

# Amino-Acid Features



# 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)	DOWN4-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)

**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**

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

# 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)

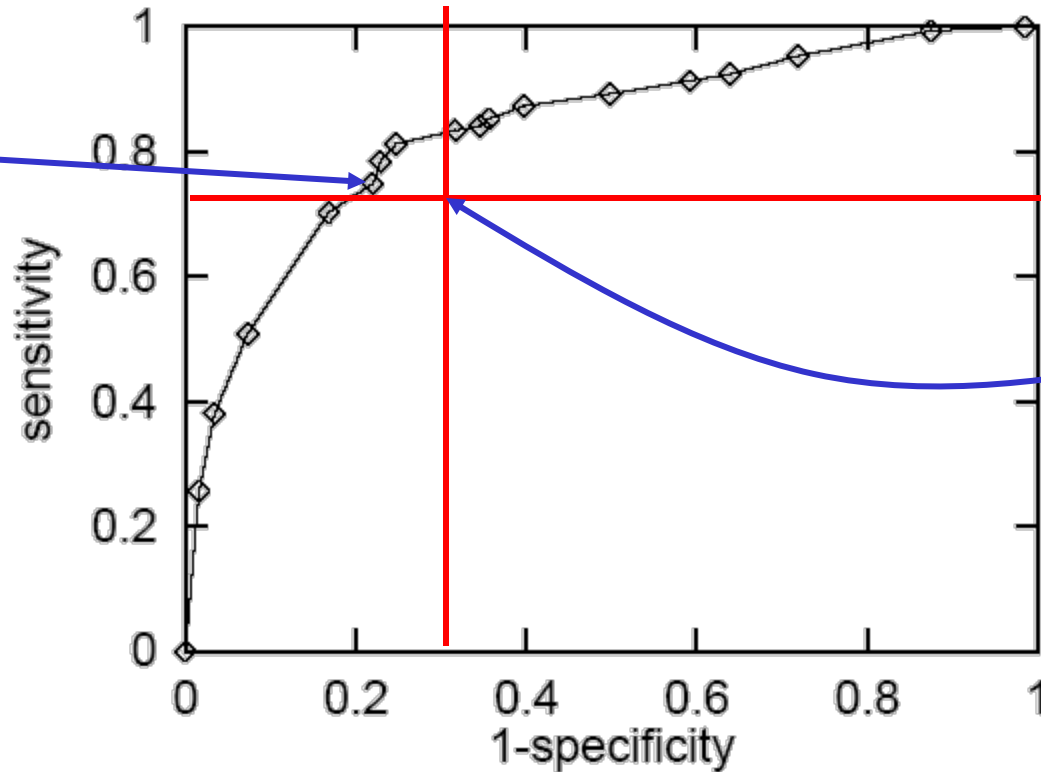
# Validation Results (on Hatzigeorgiou's)

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

# Validation Results (on Chr X and Chr 21)

Our  
method



ATGpr

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

# 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



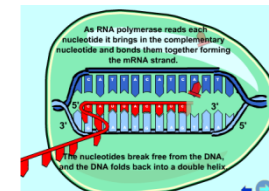
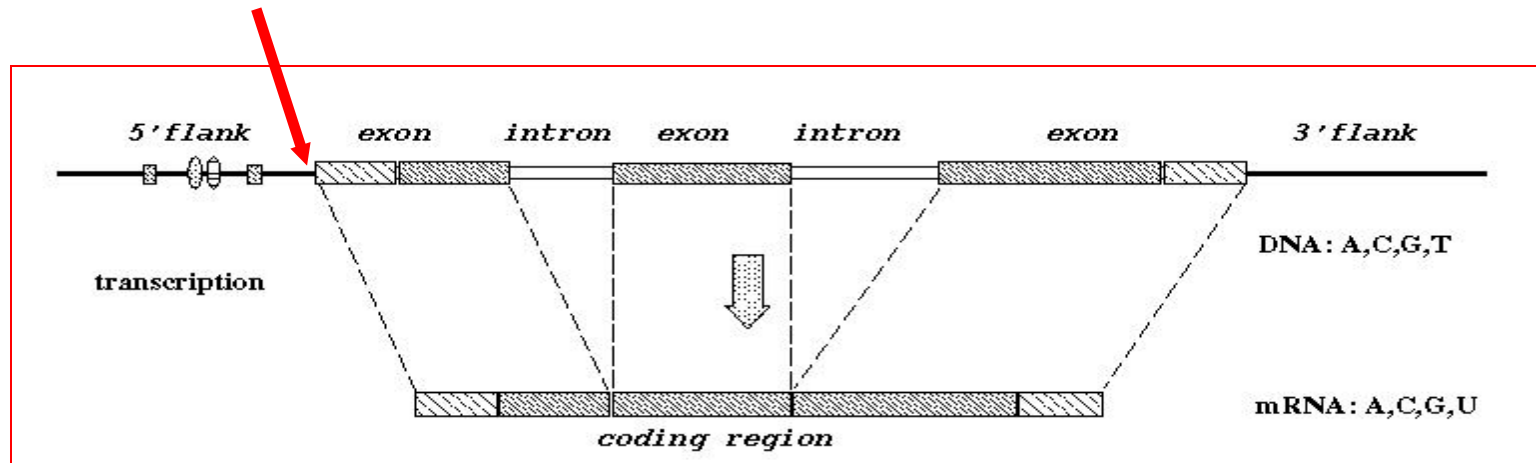


# Recognition of Transcription Start Sites

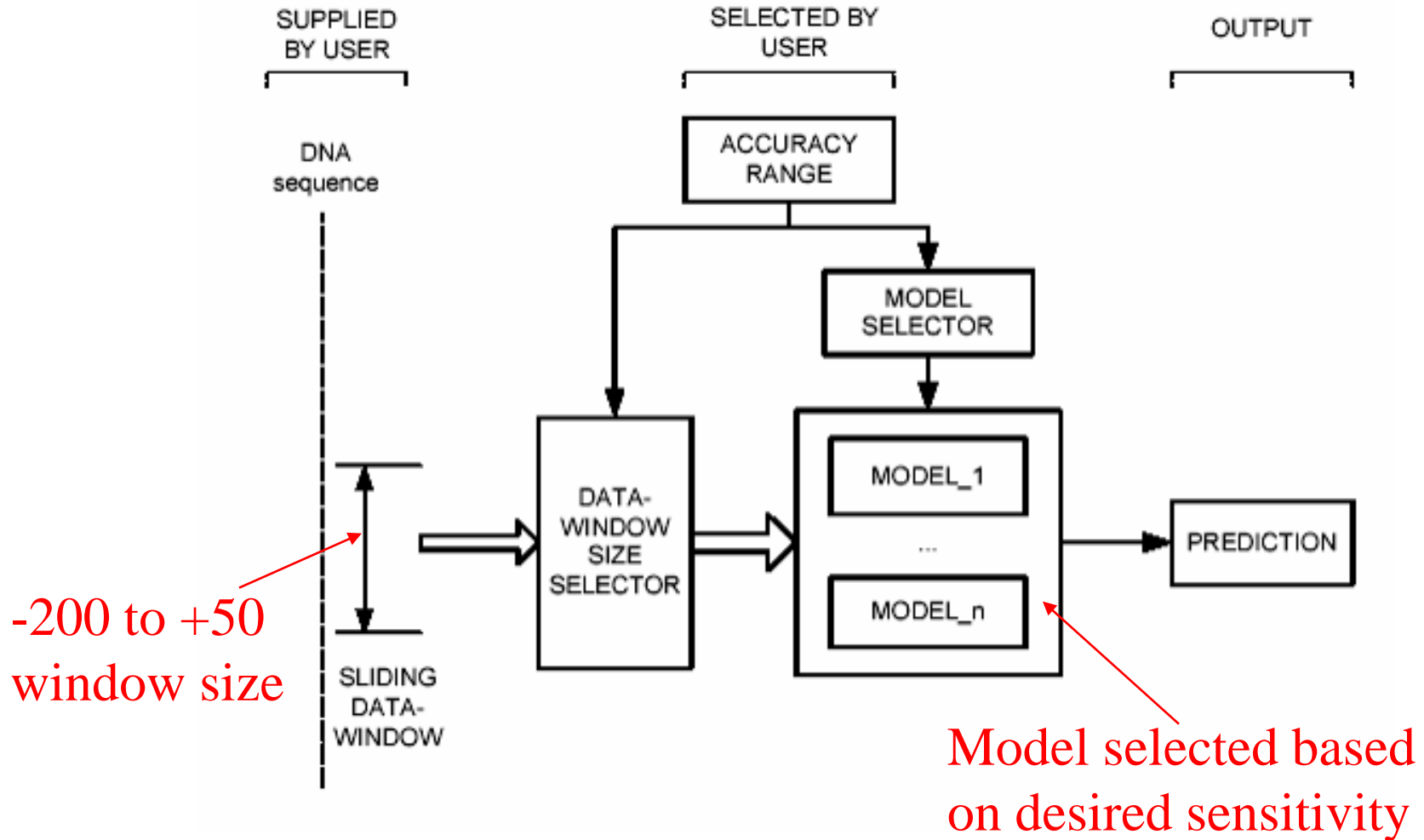
**An introduction to the World's best TSS  
recognition system:  
A heavy tuning approach**



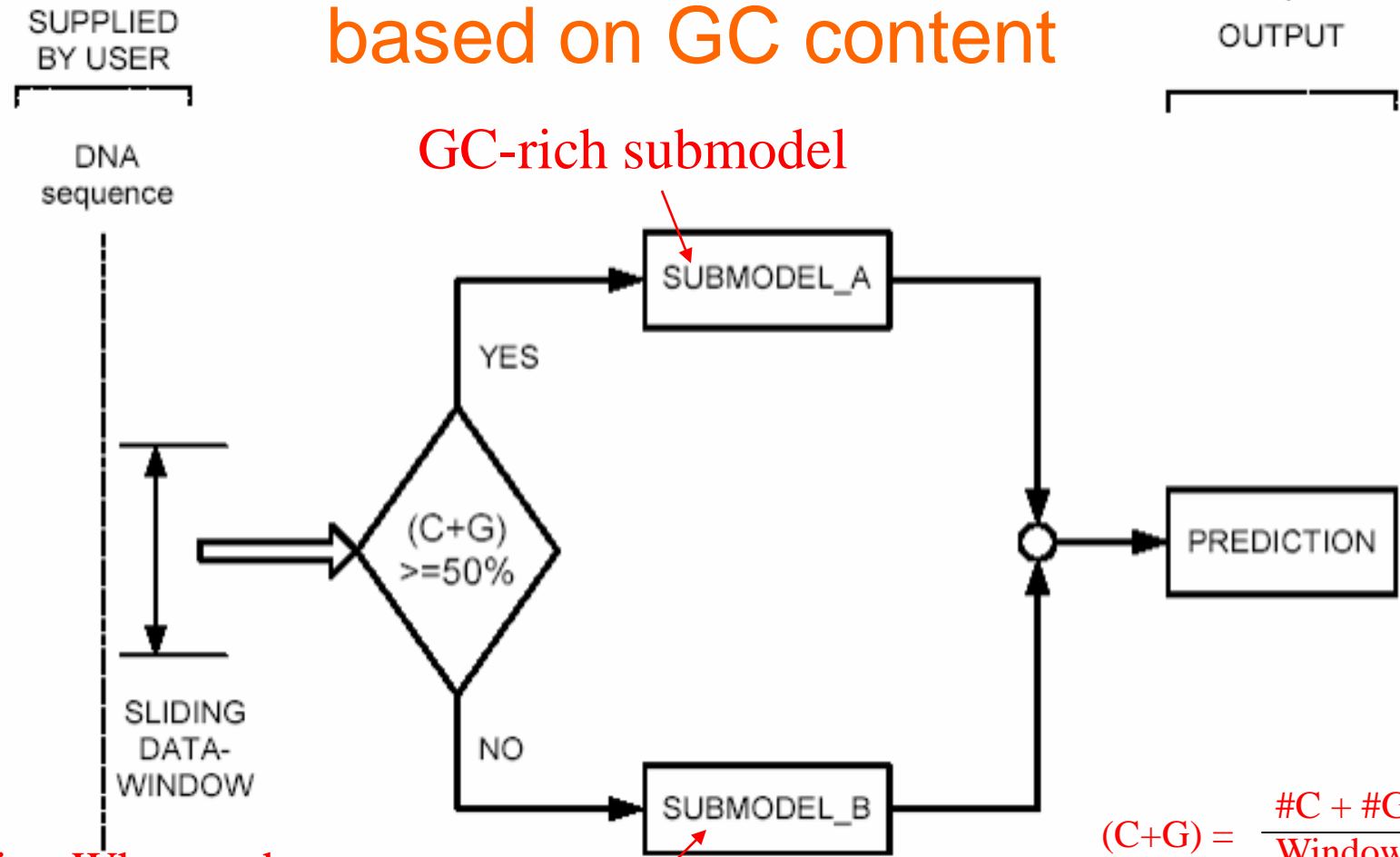
# Transcription Start Site



# Structure of Dragon Promoter Finder

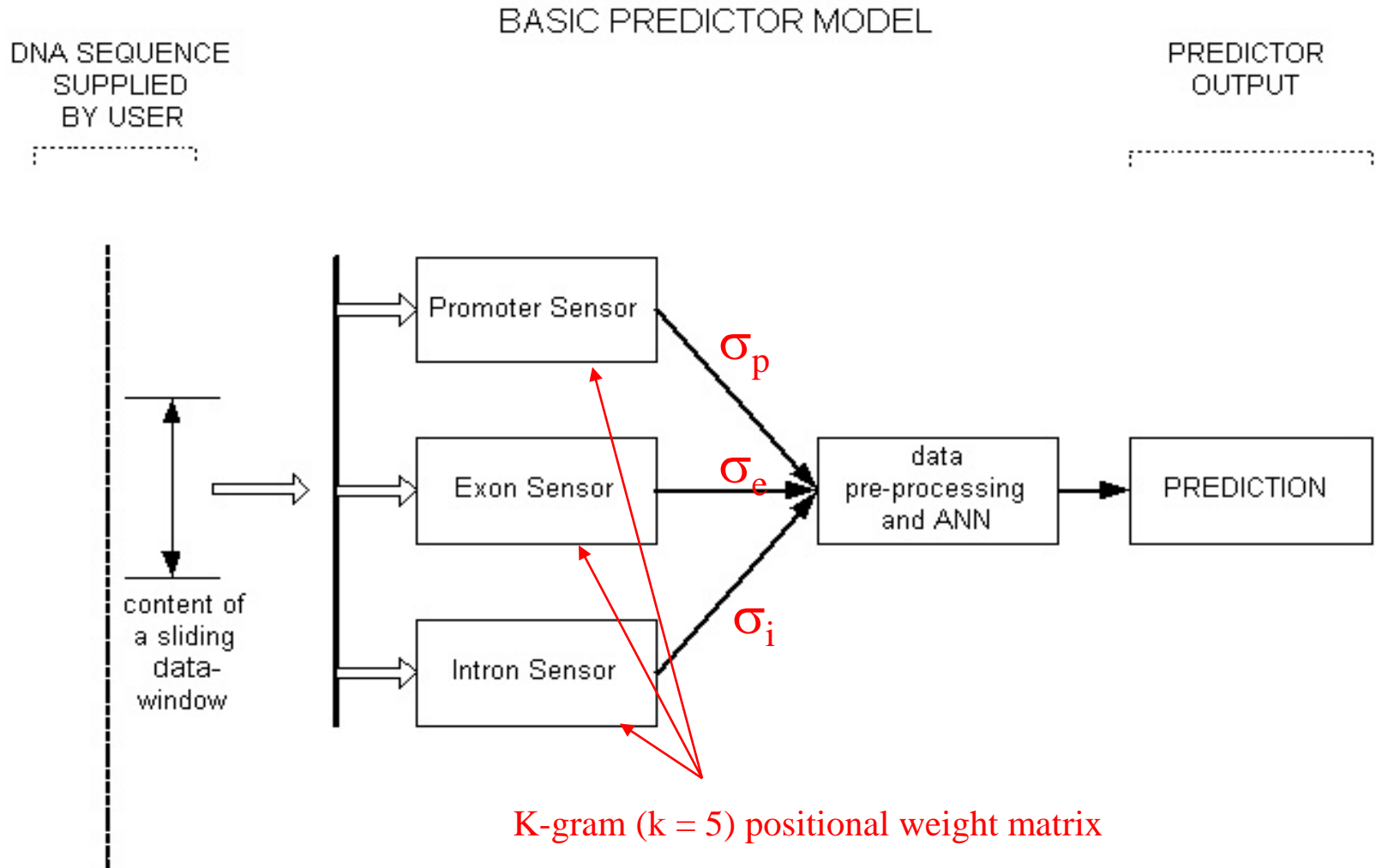


# Each model has two submodels based on GC content



Exercise: Why are the submodels based on GC content?

# Data Analysis Within Submodel



# 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

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

Window size

Pentamer at  $i^{\text{th}}$  position in input

Frequency of  $j^{\text{th}}$  pentamer at  $i^{\text{th}}$  position in training window

$j^{\text{th}}$  pentamer at  $i^{\text{th}}$  position in training window

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> = AGTGTACCTG
  - Seq<sub>3</sub> = AGTTCGTATG
- **Then**

1-mer	pos1	pos2	pos3	pos4	pos5	pos6	pos7	pos8	pos9	pos10
<b>A</b>	<b>3/3</b>	<b>0/3</b>	<b>0/3</b>							
<b>C</b>	<b>0/3</b>	<b>1/3</b>	<b>1/3</b>		<b>Exercise: Fill in the rest of the table</b>					
<b>G</b>	<b>0/3</b>	<b>2/3</b>	<b>0/3</b>							
<b>T</b>	<b>0/3</b>	<b>0/3</b>	<b>2/3</b>							

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> = AGTGTACCTG
  - 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		Exercise: Fill in the rest of the table				
...	...	...	...						
TT	0/3	0/3	1/3				1/3		



# Data Preprocessing & ANN

Tuning parameters

$$s_E = \text{sat}(\sigma_p - \sigma_e, a_e, b_e)$$

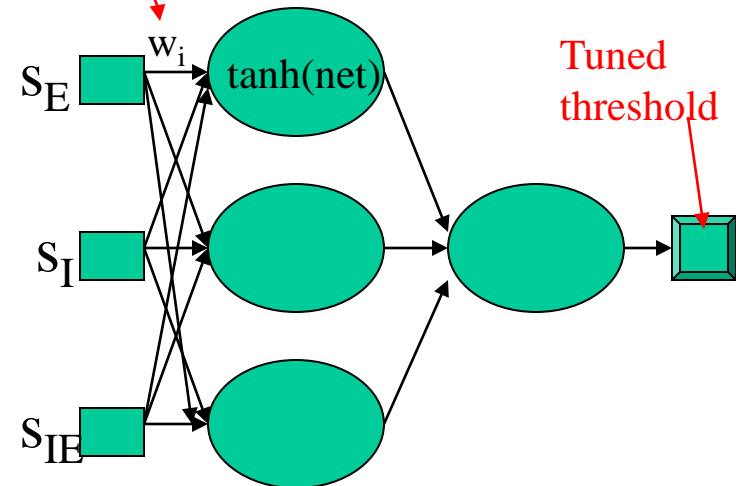
$$s_I = \text{sat}(\sigma_p - \sigma_i, a_i, b_i)$$

$$s_{EI} = \text{sat}(\sigma_e - \sigma_i, a_{ei}, b_{ei}),$$

where the function *sat* is defined by

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

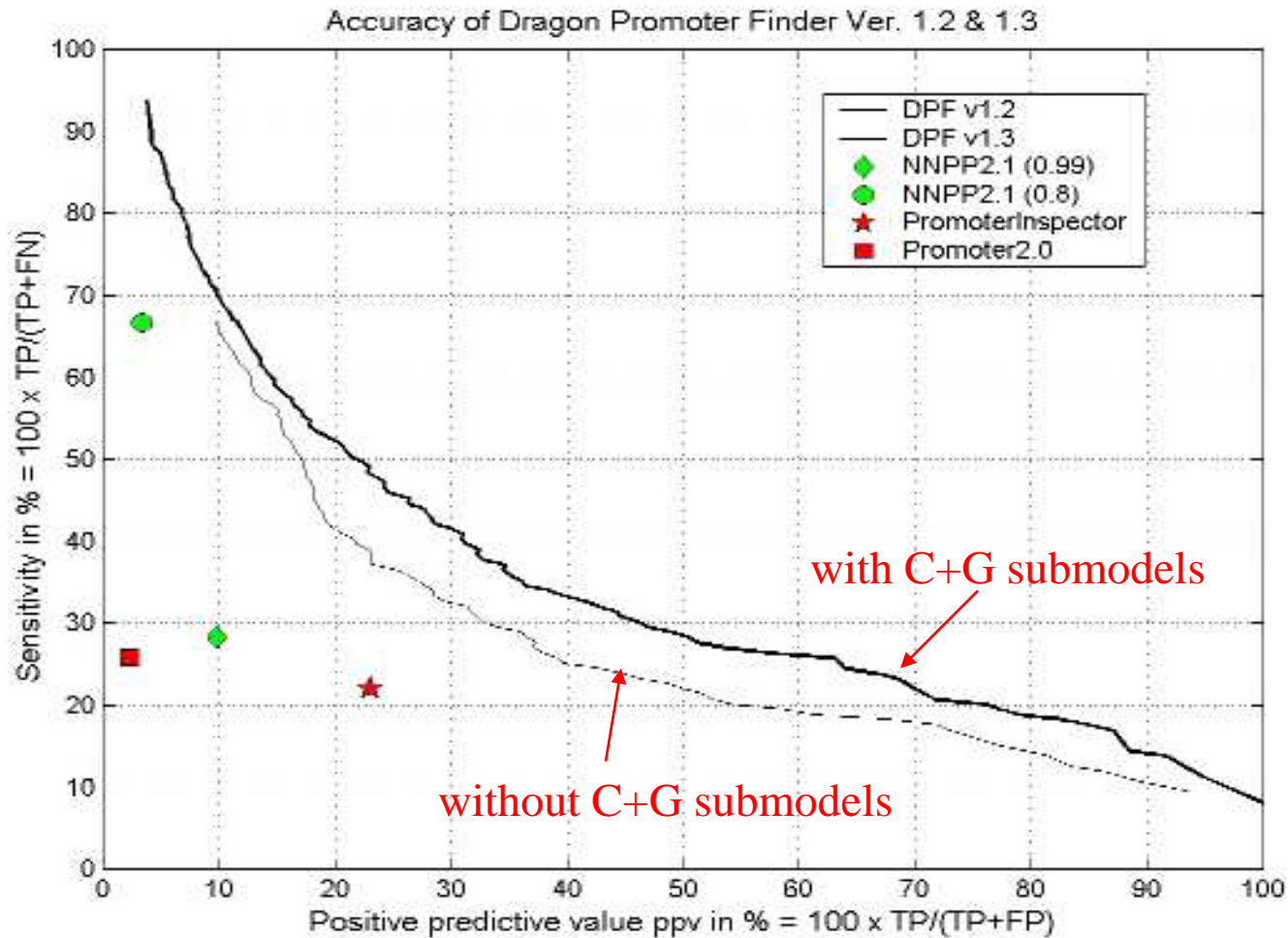
Simple feedforward ANN  
 trained by the Bayesian  
 regularisation method



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

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

# 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**
- **cummulative length of more than 1.15Mbp**
- **Taken from GENESCAN, Geneld, Genie, etc.**

# 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



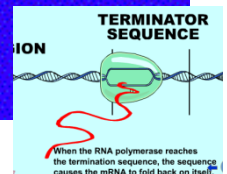
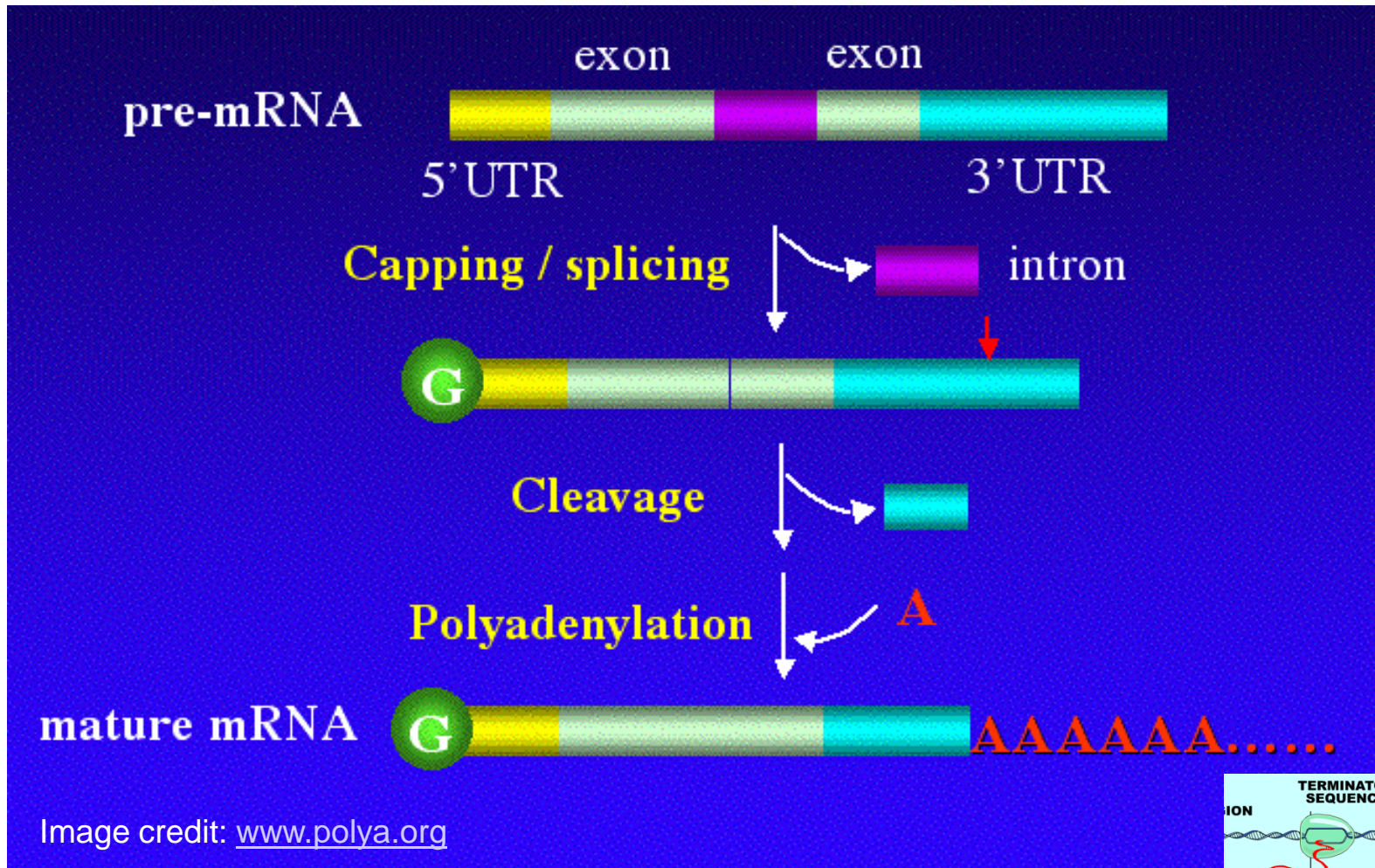
# Recognition of Poly-A Signal Sites

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





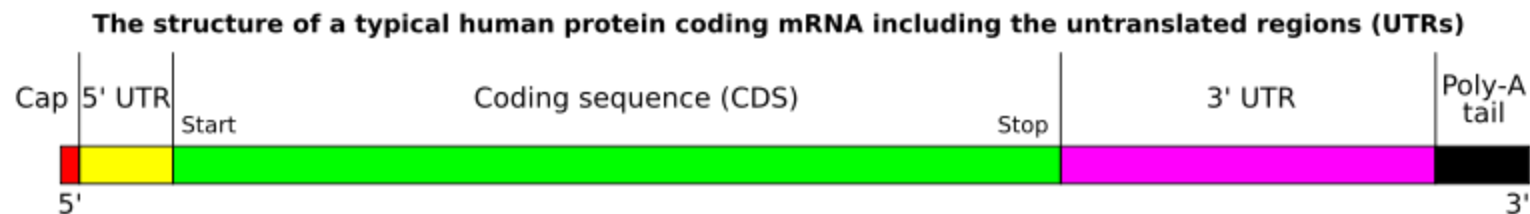
# Eukaryotic Pre-mRNA Processing





# Polyadenylation in Eukaryotes

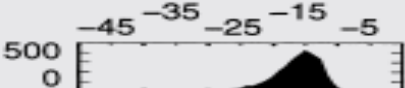
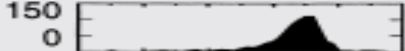



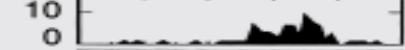

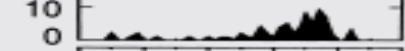




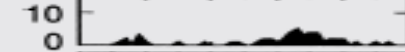



- **Addition of poly(A) tail to RNA**
  - Begins as transcription finishes
  - 3'-most segment of newly-made RNA is cleaved off
  - Poly(A) tail is then synthesized at 3' end
- **Poly(A) tail is impt for nuclear export, translation & stability of mRNA**
- **Tail is shortened over time. When short enough, the mRNA is degraded**



Source: Wikipedia

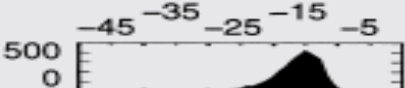
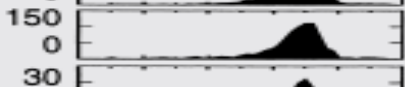

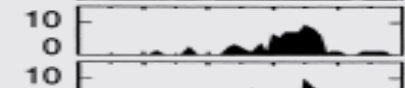


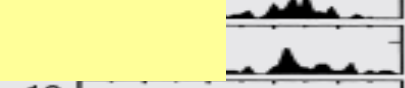

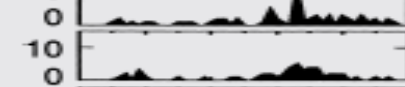

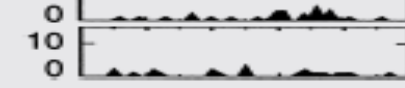


# Poly-A Signals in Human (Gautheret et al., 2000)

**Table 2. Most Significant Hexamers in 3' Fragments: Clustered Hexamers**

Hexamer	Observed (expected) <sup>a</sup>	% sites	$p^b$	Position average $\pm$ SD	Location <sup>c</sup>
<b>AAUAAA</b>	3286 (317)	58.2	0	$-16 \pm 4.7$	
<b>AUUAAA</b>	843 (112)	14.9	0	$-17 \pm 5.3$	
<b>AGUAAA</b>	156 (32)	2.7	$6 \times 10^{-57}$	$-16 \pm 5.9$	
<b>UAUAAA</b>	180 (53)	3.2	$4 \times 10^{-45}$	$-18 \pm 7.8$	
<b>CAUAAA</b>	76 (23)	1.3	$1 \times 10^{-16}$	$-17 \pm 5.9$	
<b>GAUAAA</b>	72 (21)	1.3	$2 \times 10^{-16}$	$-18 \pm 6.9$	
<b>AAUAUA</b>	96 (33)	1.7	$2 \times 10^{-19}$	$-18 \pm 6.9$	
<b>AAUACA</b>	70 (16)	1.2	$5 \times 10^{-23}$	$-18 \pm 8.7$	
<b>AAUAGA</b>	43 (14)	0.7	$1 \times 10^{-9}$	$-18 \pm 6.3$	
<b>AAAAAG</b>	49 (11)	0.8	$5 \times 10^{-17}$	$-18 \pm 8.9$	
<b>ACUAAA</b>	36 (11)	0.6	$1 \times 10^{-08}$	$-17 \pm 8.1$	
<b>AAGAAA</b>	62 (10)	1.1	$9 \times 10^{-26}$	$-19 \pm 11$	
<b>AAUGAA</b>	49 (10)	0.8	$4 \times 10^{-16}$	$-20 \pm 10$	
<b>UUUAAA</b>	69 (20)	1.2	$3 \times 10^{-16}$	$-17 \pm 12$	
<b>AAAACA</b>	29 (5)	0.5	$8 \times 10^{-12}$	$-20 \pm 10$	
<b>GGGGCU</b>	22 (3)	0.3	$9 \times 10^{-12}$	$-24 \pm 13$	

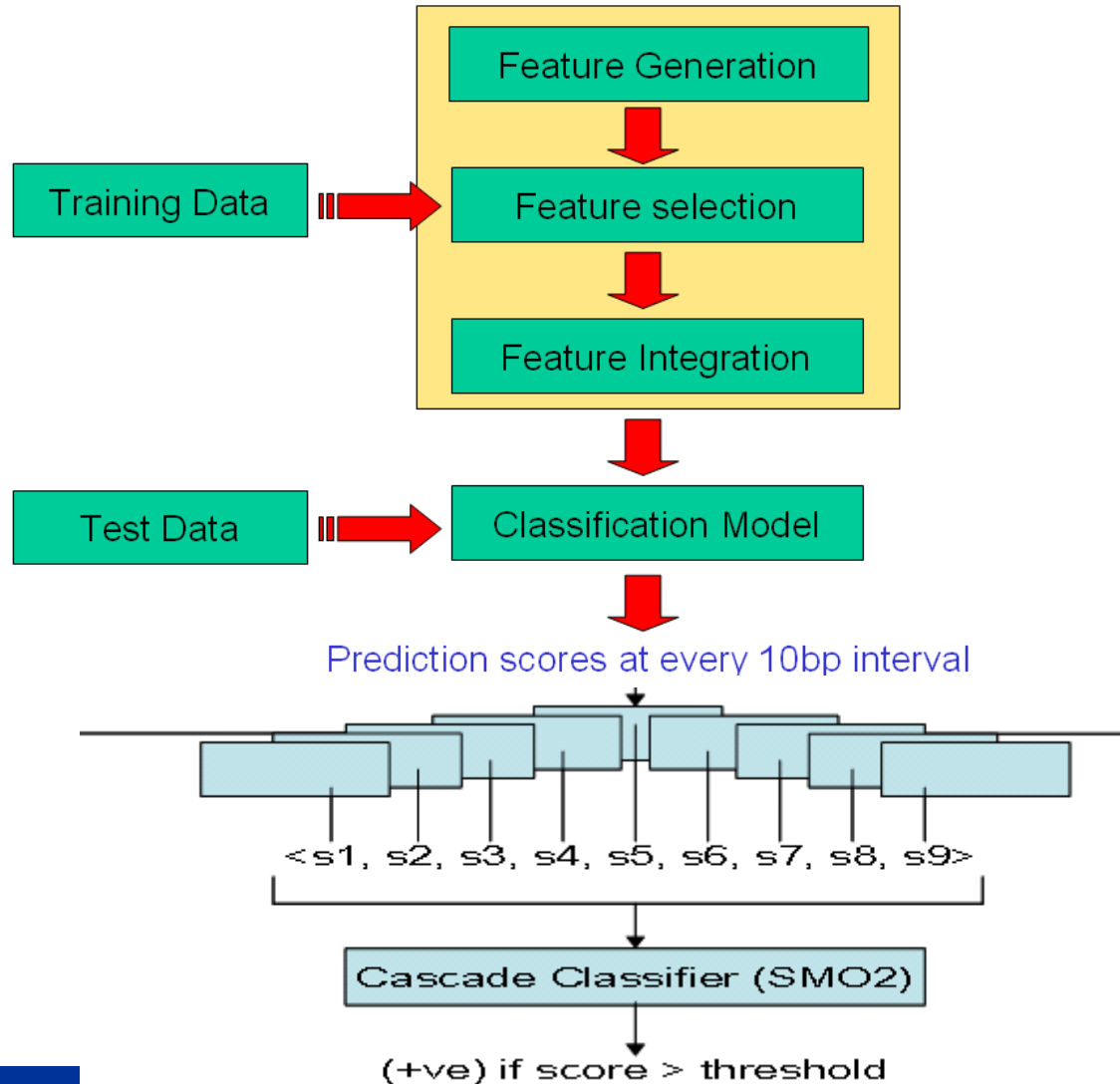
# Poly-A Signals in Arabidopsis

Table 2. Most Significant Hexamers in 3' Fragments: Clustered Hexamers

Hexamer	Observed (expected) <sup>a</sup>	% sites	$p^b$	Position average $\pm$ SD	Location <sup>c</sup>
<b>AAUAAA</b>	3286 (317)	58.2	0	$-16 \pm 4.7$	
<b>AUUAAA</b>	843 (112)	14.9	0	$-17 \pm 5.3$	
<b>AGUAAA</b>	156 (32)	2.7	$6 \times 10^{-57}$	$-16 \pm 5.9$	
<b>UAUAAA</b>	180 (53)	3.2	$4 \times 10^{-45}$	$-18 \pm 7.8$	
<b>CAUAAA</b>	76 (23)	1.3	$1 \times 10^{-16}$	$-17 \pm 5.9$	
<b>GAUAAA</b>	72				
<b>AAUAUA</b>	96				
<b>AAUACA</b>	70				
<b>AAUAGA</b>	43				
<b>AAAAAG</b>	49				
<b>ACUAAA</b>	36 (11)	0.6	$1 \times 10^{-06}$	$-17 \pm 8.1$	
<b>AAGAAA</b>	62 (10)	1.1	$9 \times 10^{-26}$	$-19 \pm 11$	
<b>AAUGAA</b>	49 (10)	0.8	$4 \times 10^{-16}$	$-20 \pm 10$	
<b>UUUAAA</b>	69 (20)	1.2	$3 \times 10^{-16}$	$-17 \pm 12$	
<b>AAAACA</b>	29 (5)	0.5	$8 \times 10^{-12}$	$-20 \pm 10$	
<b>GGGGCU</b>	22 (3)	0.3	$9 \times 10^{-12}$	$-24 \pm 13$	

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

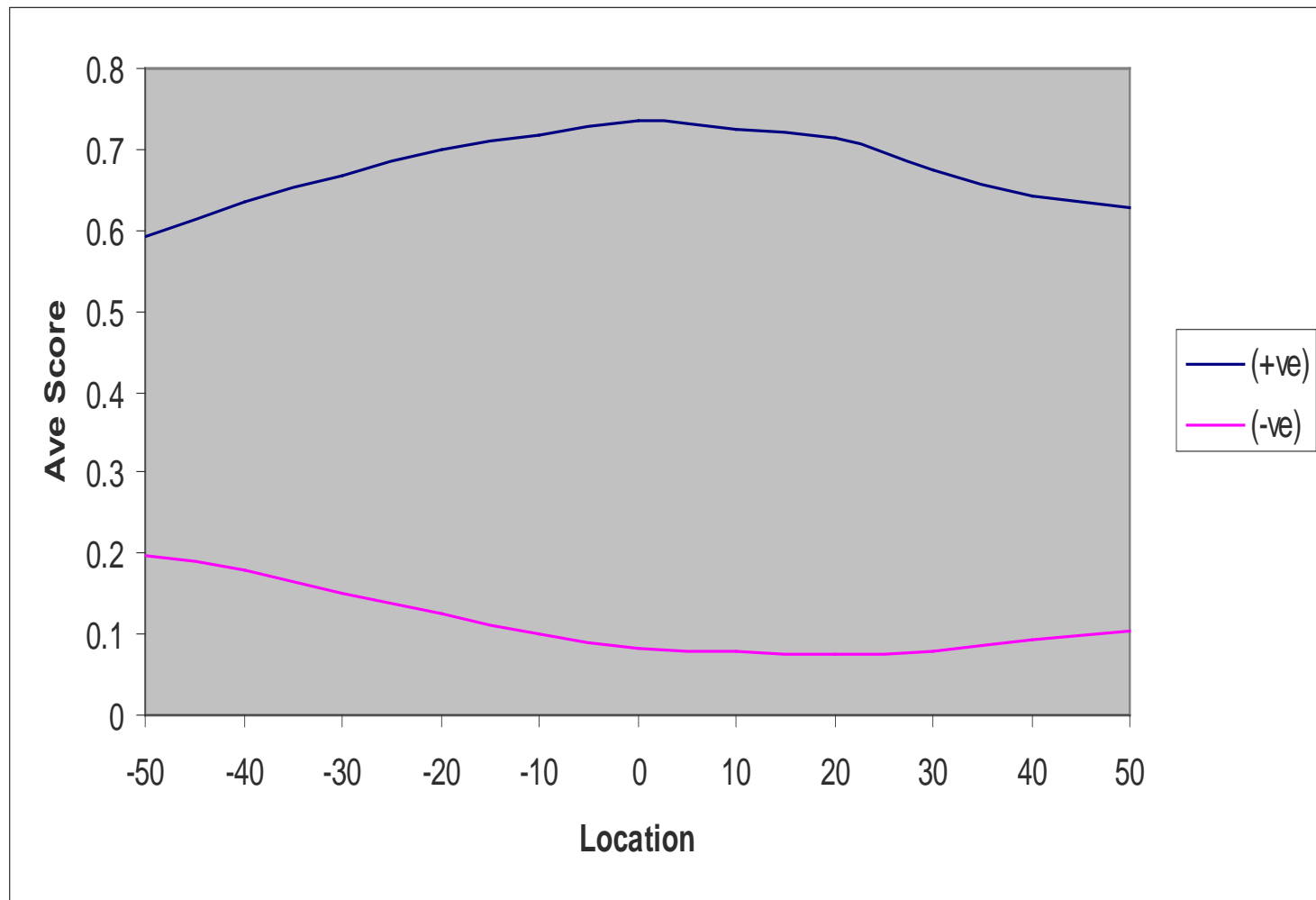
# Approach on Arab PAS Sites (I)



# 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\*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

SN_0	SMO 1		SMO 2		PASS 1.0	
	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
Control Sequences						
CDS	90%	0.26	94%	0.24	95%	3.7
5'UTR	79%	0.42	85%	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<sub>10</sub>.

SN <sub>10</sub>	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%	0.68	77%	0.77	67%	6.6

Table 3. Equal-error-rate points of SMO1, SMO2, and PASS 1.0 for SN<sub>30</sub>.

SN <sub>30</sub>	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%	0.62	92%	0.67	84%	6.2
Intron	79%	0.75	83%	0.81	72%	6.8

# About the Inventor: Koh Chuan Hock

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





# 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 → 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. 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

## 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

## 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

# 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