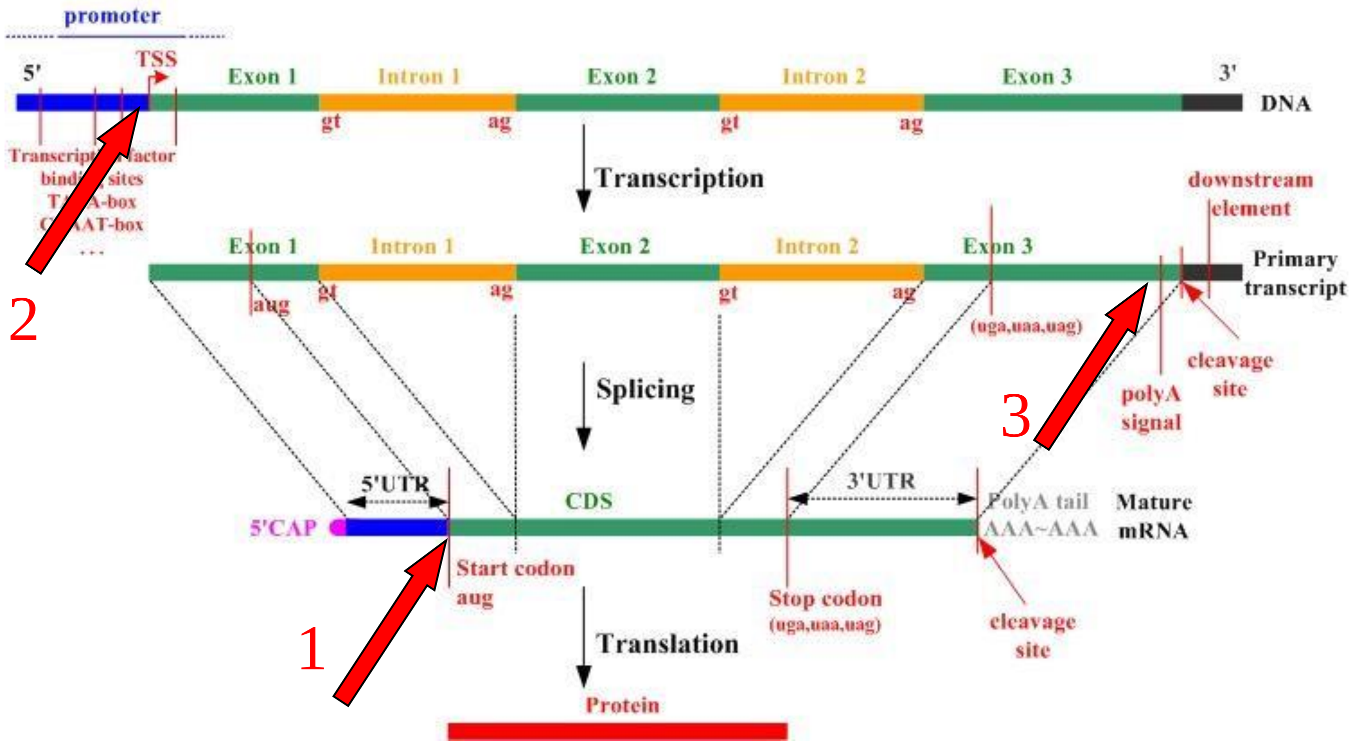
For written notes on this lecture, please read Chapters 4 and 7 of *The Practical Bioinformatician*, and Koh & Wong, “Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences”, *Proc GIW 2007*, pages 73--82

CS2220: Introduction to Computational Biology
Unit 2: Gene Feature Recognition

Wong Limsoon

NUS National University of Singapore

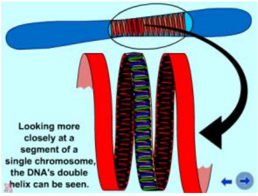
Plan


promoter
5'
TSS
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
3'
DNA
gt
ag
gt
ag
Transcript factor
binding sites
TATA-box
CCAAT-box
...
Transcription
downstream
element
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Primary
transcript
aug
gt
ag
gt
ag
(uga,uaa,uag)
cleavage
site
polyA
signal
Splicing
5'UTR
CDS
3'UTR
PolyA tail
AAA~AAA
Mature
mRNA
5'CAP
Start codon
aug
Stop codon
(uga,uaa,uag)
cleavage
site
Translation
Protein
1
2
3

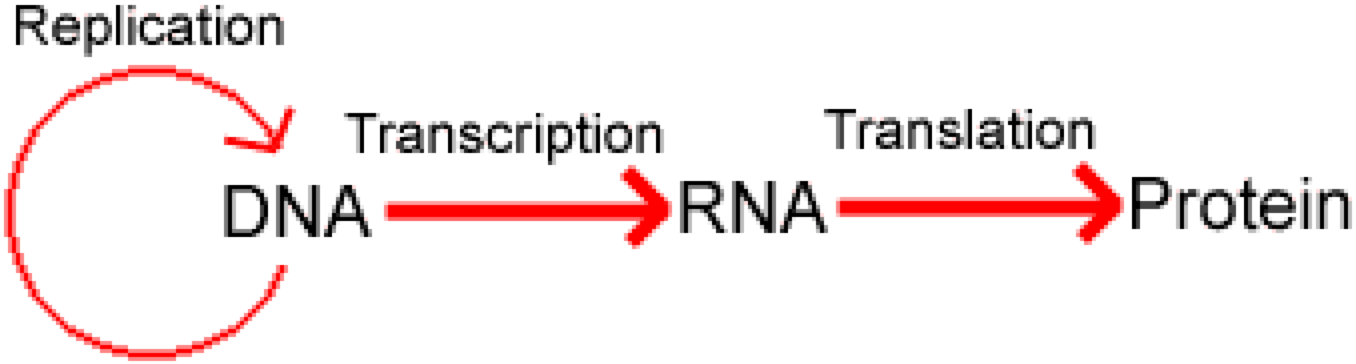
Some relevant biology

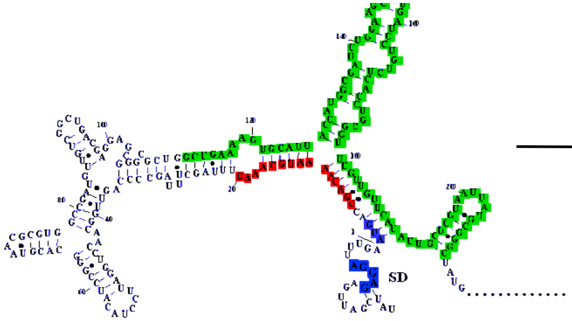
NUS

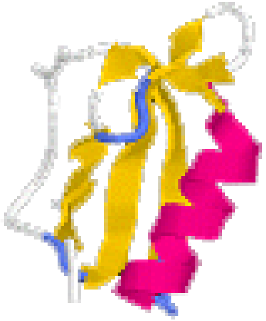
National University
of Singapore


Looking more closely at a segment of a single chromosome, the DNA's double helix can be seen.

Central dogma



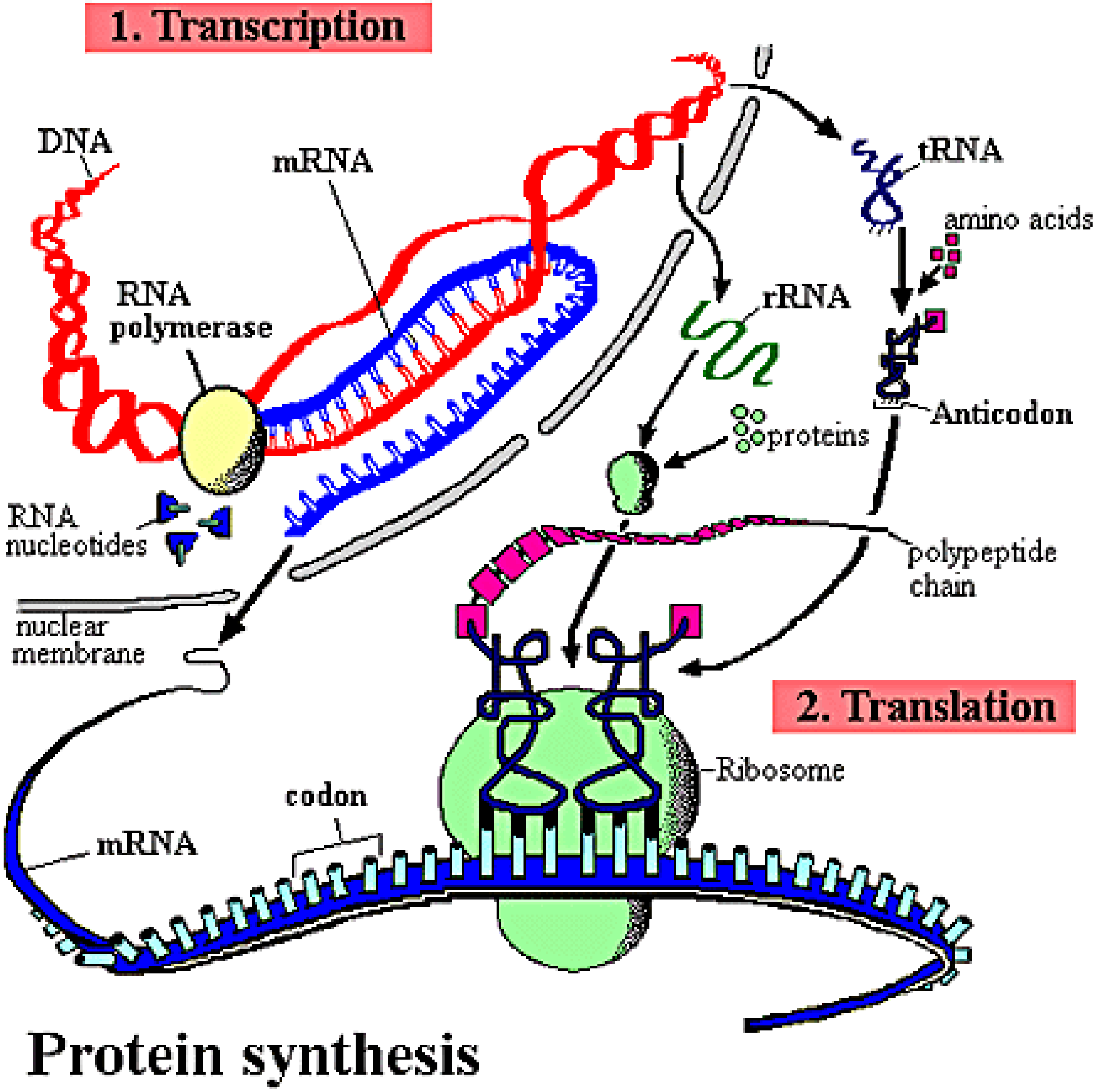




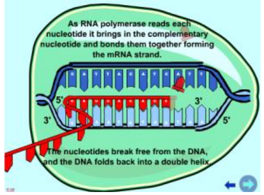
...AATGGTACCGATGACCTG...

...AAUGGUACCGAUGACCUGGAGC...

...TRLRPLLALLALWP...



Players in protein synthesis



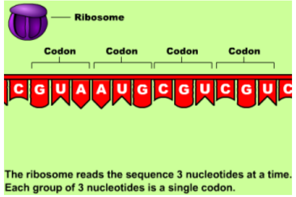
As RNA polymerase reads each nucleotide it brings in the complementary nucleotide and bonds them together forming the mRNA strand.

The nucleotides break free from the DNA, and the DNA folds back into a double helix.

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


Ribosome
Codon
Codon
Codon
Codon
C G U A A U G C G U C G U C
The ribosome reads the sequence 3 nucleotides at a time.
Each group of 3 nucleotides is a single codon.

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

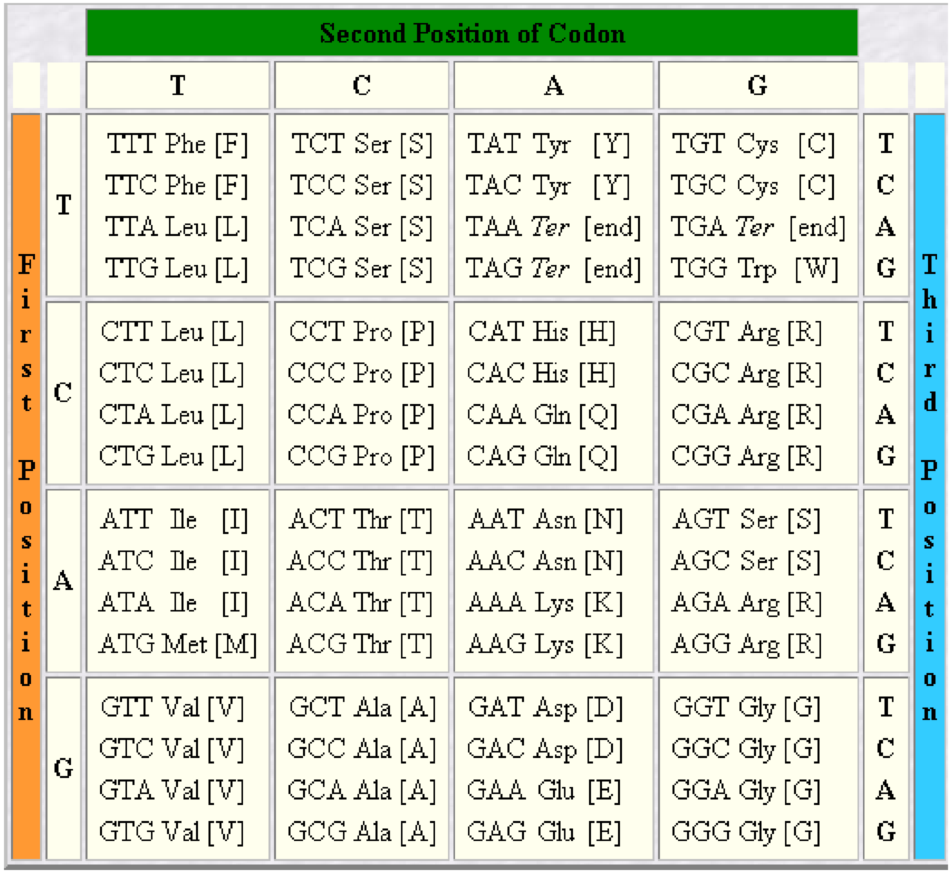
$\Rightarrow$ Codons are not 1-to-1 corr to 20 amino acids

- All organisms use the same decoding table (except some mitochrondrial 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)

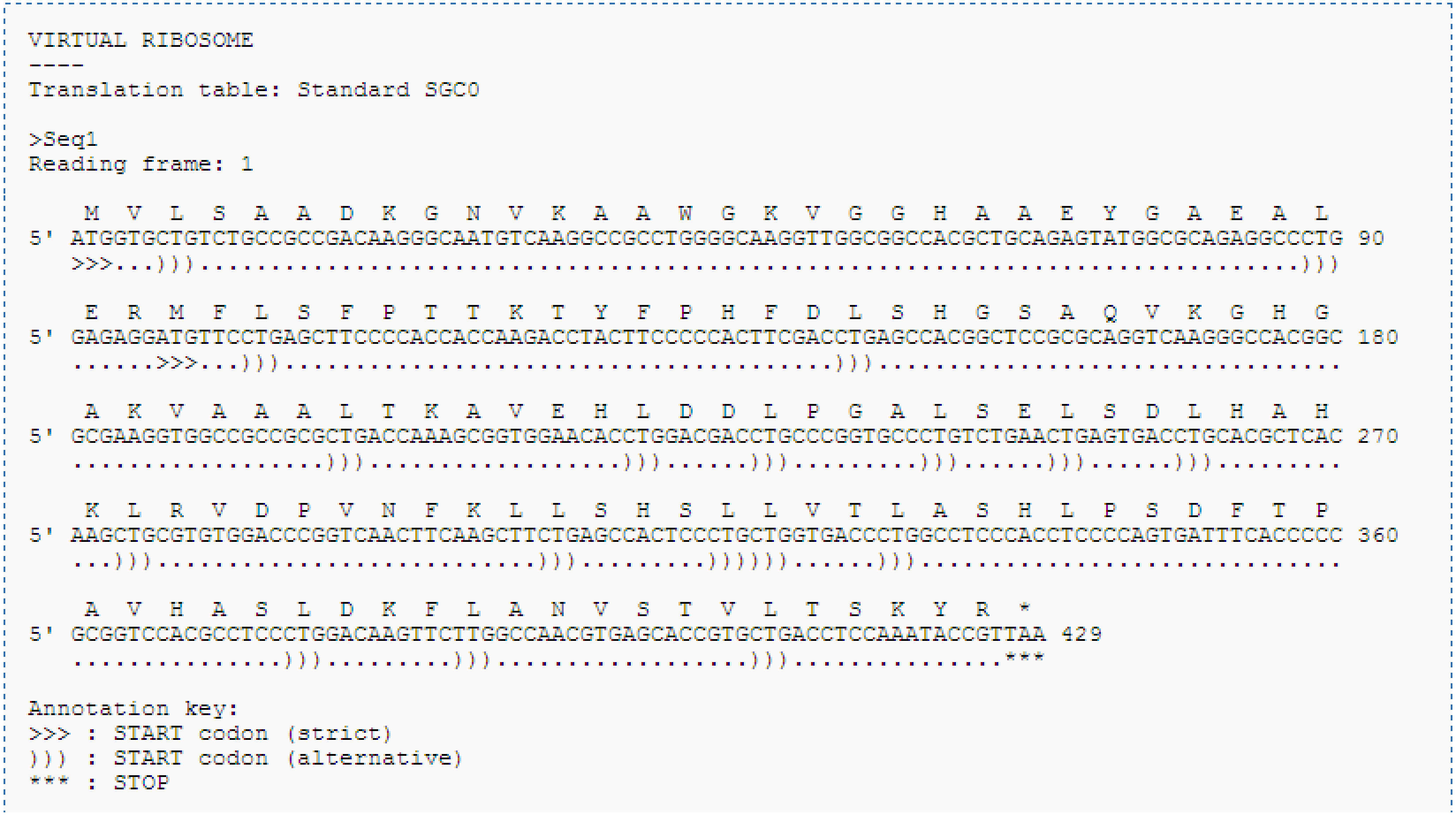
- **Stop codon**
  - TAA
  - TAG
  - TGA



		Second Position of Codon					
		T	C	A	G		
First Position	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T	Third Position
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C	
		TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A	
		TTG Leu [L]	TCG Ser [S]	TAG *Ter* [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:



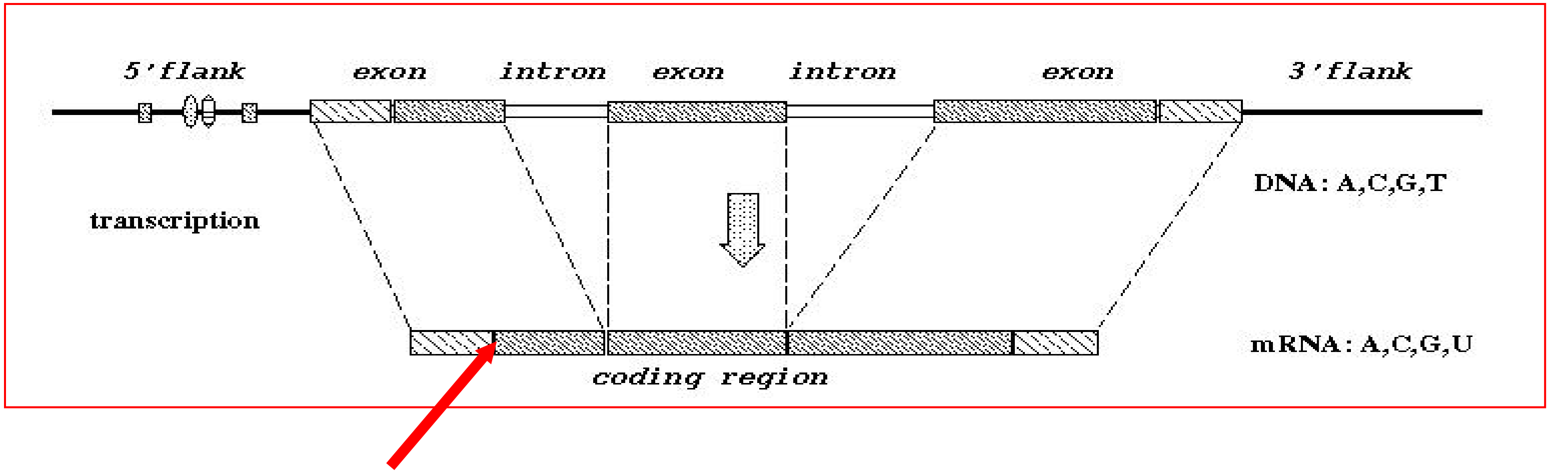
```
VIRTUAL RIBOSOME
----
Translation table: Standard SGC0

>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' ATGGTGCTGTCTGCCGCCGACAAGGGCAATGTCAAGGCCGCCTGGGGCAAGGTTGGCGGCCACGCTGCAGAGTATGGCGCAGAGGCCCTG 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' GAGAGGATGTTCCTGAGCTTCCCCACCACCAAGACCTACTTCCCCCACTTCGACCTGAGCCACGGCTCCGCGCAGGTCAAGGGCCACGGC 180
   ......>>>...))).......................................)))."...............................
```

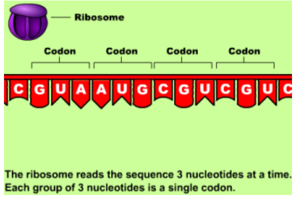
Translation initiation sites

An introduction to the World's simplest TIS recognition system

NUS
National University
of Singapore

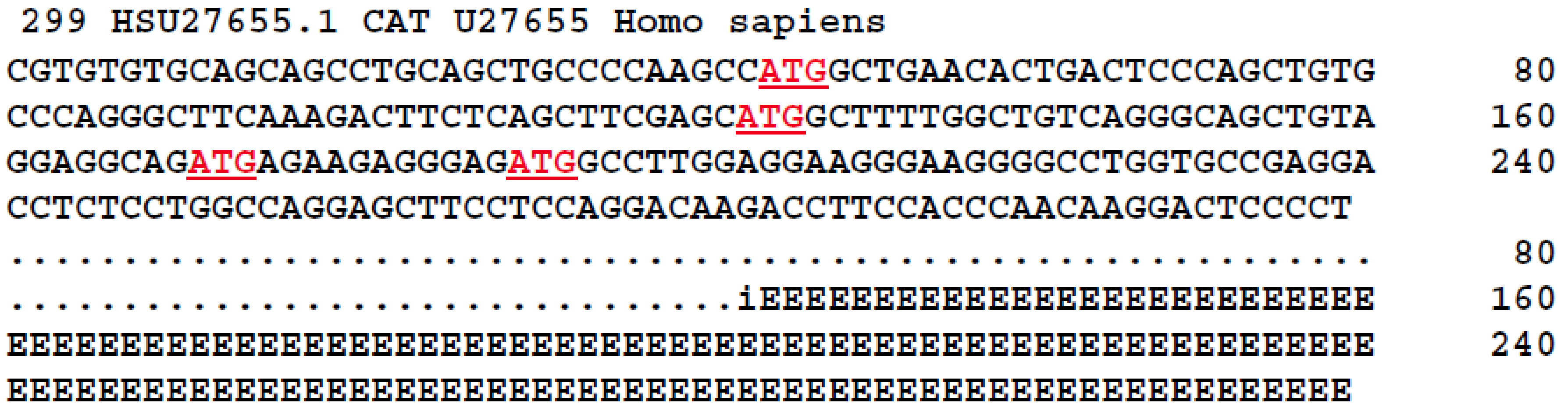
Translation initiation site





The ribosome reads the sequence 3 nucleotides at a time. Each group of 3 nucleotides is a single codon.

A sample cDNA



```
 299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA     160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA     240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
............................................................      80
................................iEEEEEEEEEEEEEEEEEEEEEEEEEEE     160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE     240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

- **What makes the second ATG the TIS?**

Approach

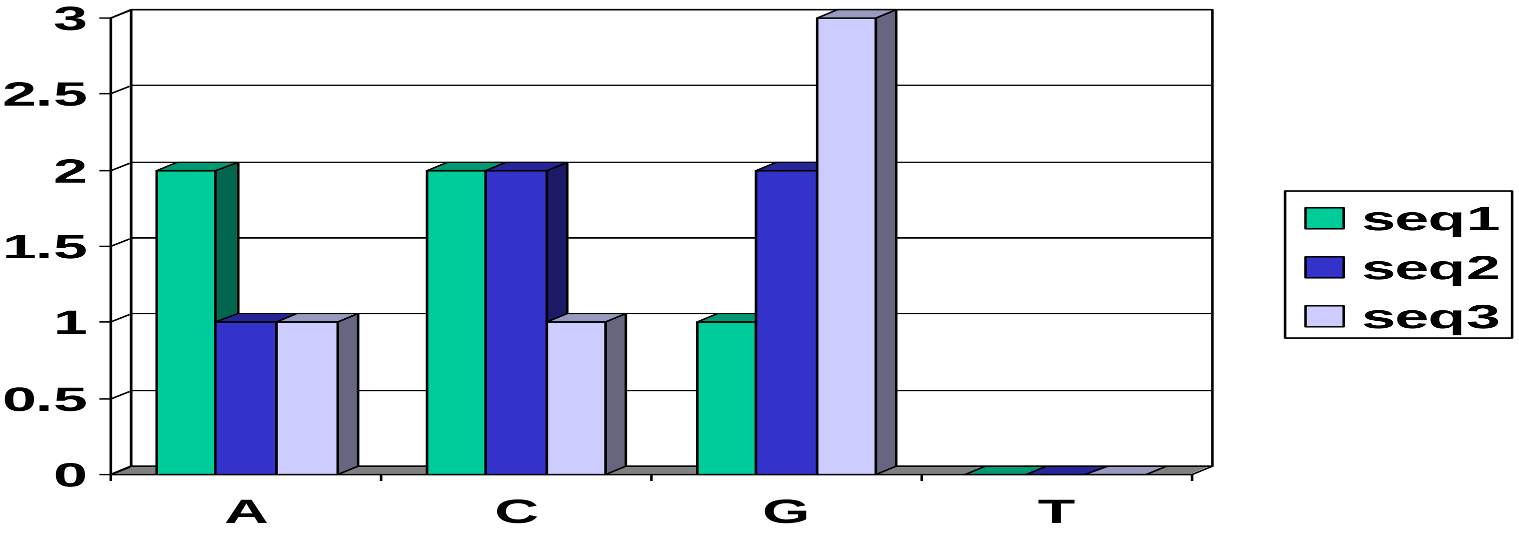
- **Training data gathering**
- **Signal generation**
  - k-grams, distance, domain know-how, ...
- **Signal selection**
  - Entropy, $\chi2$, CFS, t-test, domain know-how...
- **Signal integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...

Training & testing data

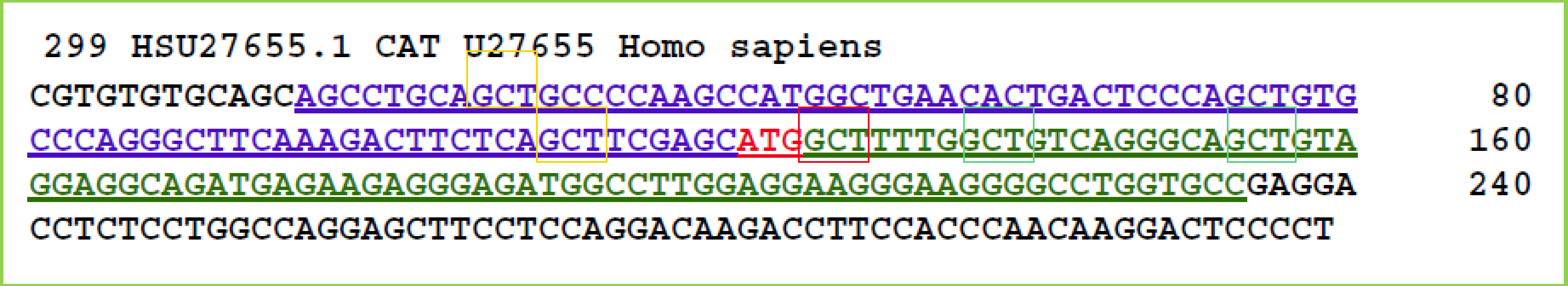
- **Vertebrate dataset of Pedersen & Nielsen [ISMB'97]**
- **3312 sequences**
- **13503 ATG sites**
- **3312 (24.5%) are TIS**
- **10191 (75.5%) are non-TIS**
- **Use for 3-fold x-validation expts**

Signal generation

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



Signal generation: Example



```
 299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG       80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA      160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA      240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
```

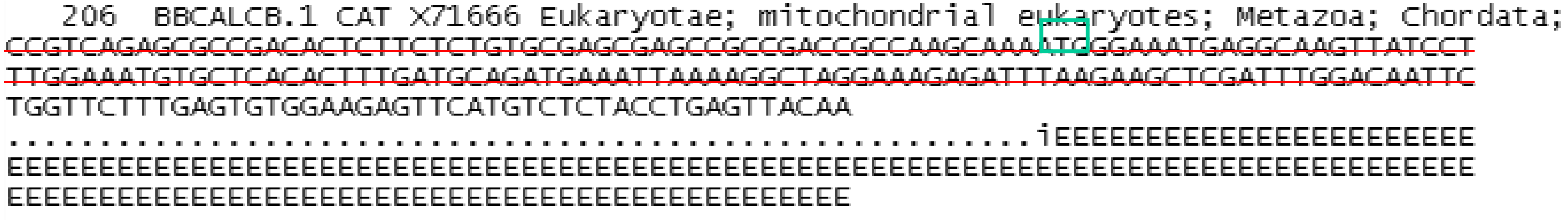
- **Window = ±100 bases**
- **In-frame, downstream**
  - GCT = 1, TTT = 1, ATG = 1…
- **Any-frame, downstream**
  - GCT = 3, TTT = 2, ATG = 2…
- **In-frame, upstream**
  - GCT = 2, TTT = 0, ATG = 0, ...

Exercise: Find the in-frame downstream ATG

Exercise #1

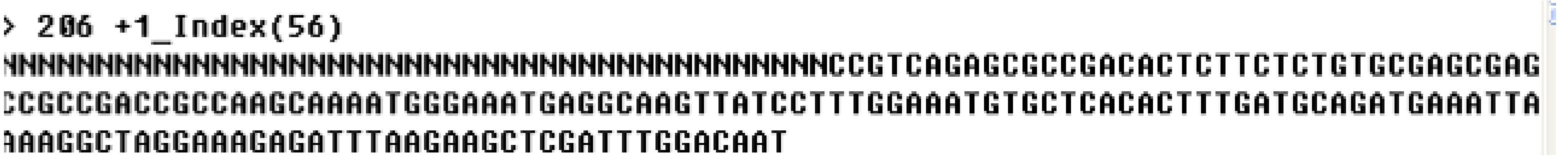
Feature generation - Summary

Raw Data



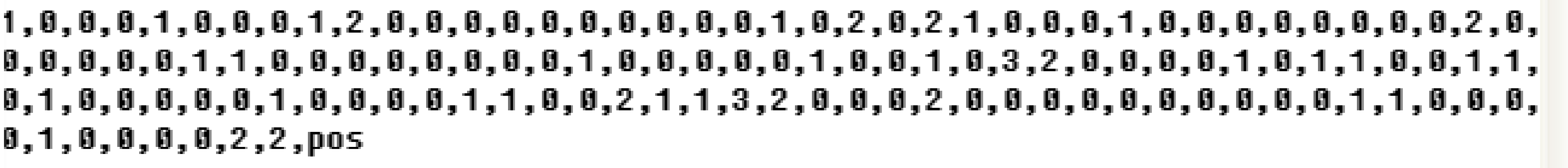
```
   206  BBCALCB.1 CAT X71666 Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
CCGTCAGAGCGCCGACACTCTTCTCTGTGCGAGCGAGCCGCCGACCGCCAAGCAAAATGGGAAATGAGGCAAGTTATCCT
TTGGAAATGTGCTCACACTTTGATGCAGATGAAATTAAAAGGCTAGGAAAGAGATTTAAGAAGCTCGATTTGGACAATTC
TGGTTCTTTGAGTGTGGAAGAGTTCATGTCTCTACCTGAGTTACAA
........................................................iEEEEEEEEEEEEEEEEEEEEEEE
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

An ATG segment – positive sample



```
> 206 +1_Index(56)
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCGTCAGAGCGCCGACACTCTTCTCTGTGCGAGCGAG
CCGCCGACCGCCAAGCAAAATGGGAAATGAGGCAAGTTATCCTTTGGAAATGTGCTCACACTTTGATGCAGATGAAATTA
AAAGGCTAGGAAAGAGATTTAAGAAGCTCGATTTGGACAAT
```

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



```
1,0,0,0,1,0,0,0,1,2,0,0,0,0,0,0,0,0,0,0,1,0,2,0,2,1,0,0,0,1,0,0,0,0,0,0,0,0,2,0,
0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,1,0,3,2,0,0,0,0,1,0,1,1,0,0,1,1,
0,1,0,0,0,0,0,1,0,0,0,0,1,1,0,0,2,1,1,3,2,0,0,0,2,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,
0,1,0,0,0,0,2,2,pos
```

Too many features

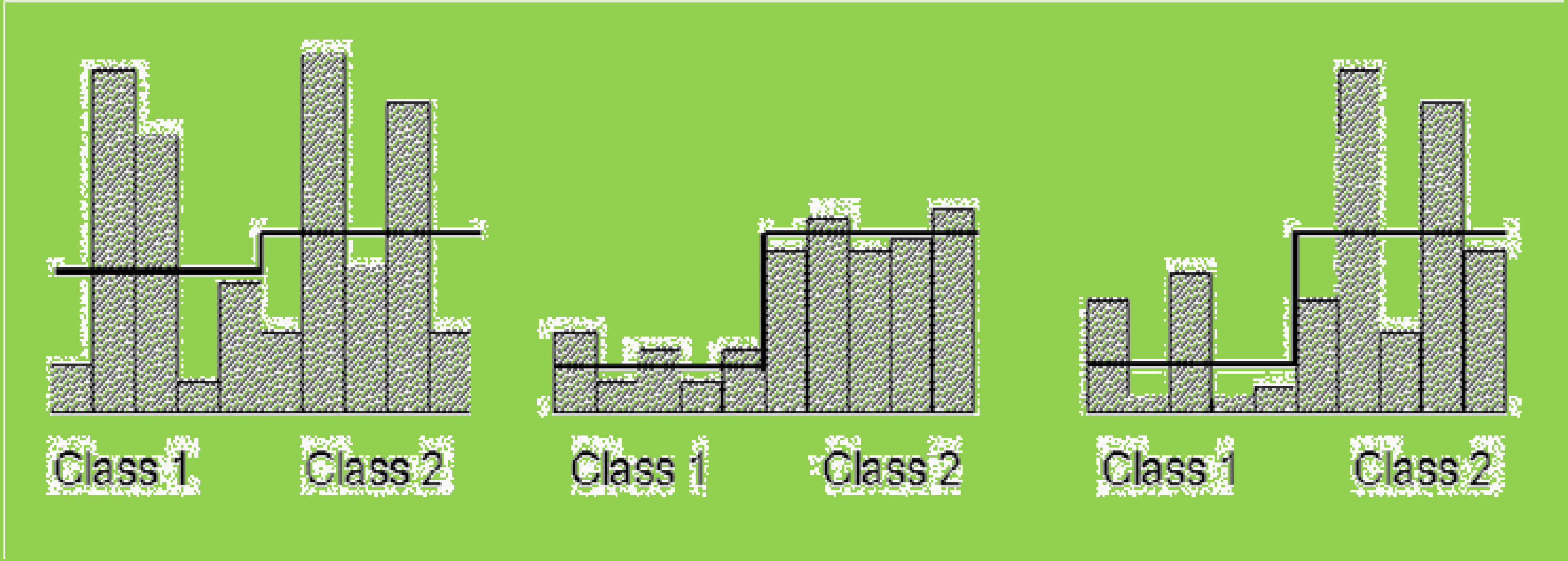
- For each value of k, there are 4^k * 3 * 2 k-grams

- 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

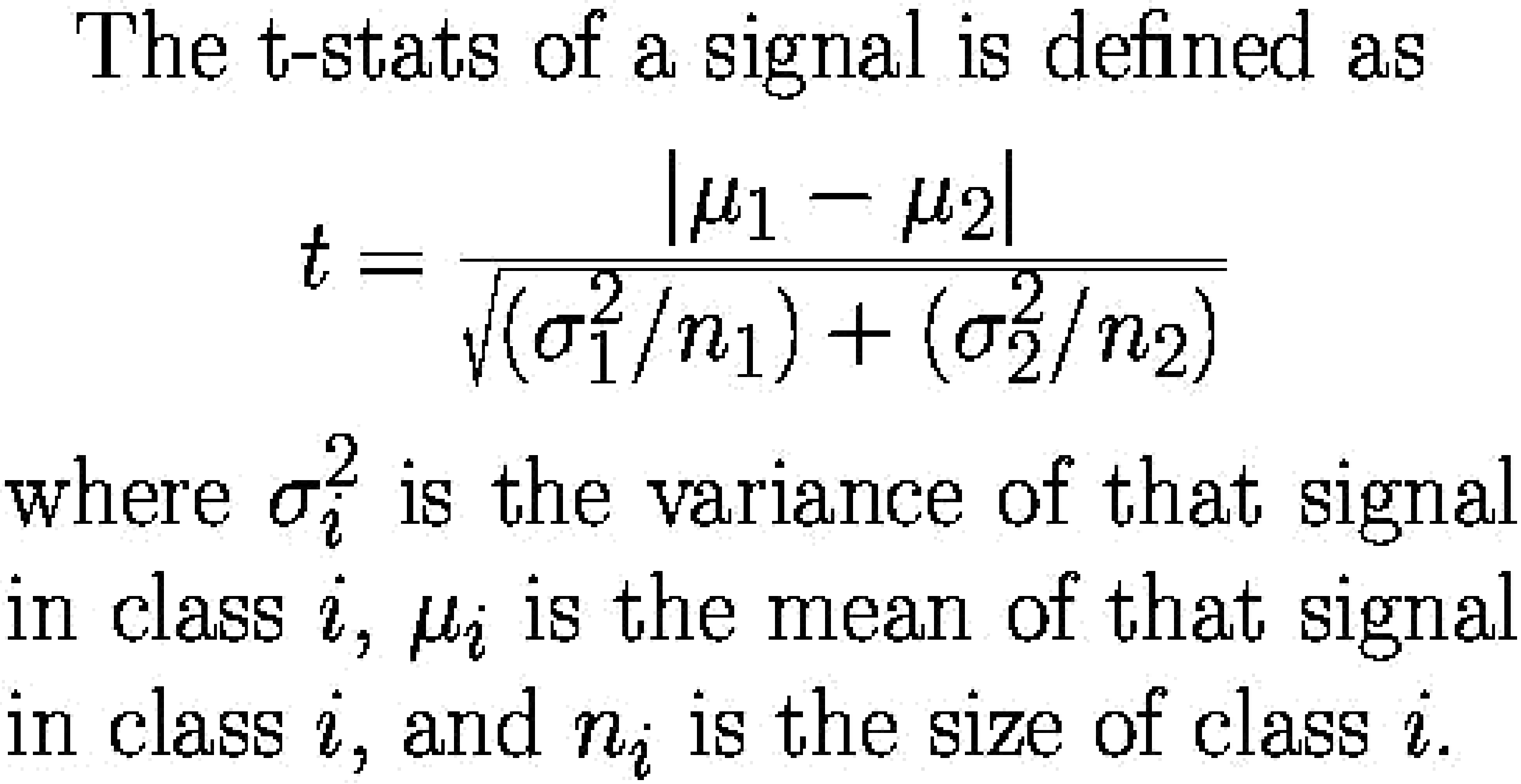
- **Choose a signal w/ low intra-class distance**
- **Choose a signal w/ high inter-class distance**



- **Which of these three features are best for distinguishing Class 1 from Class 2? Why?**

Exercise #2

Signal selection: t-statistics

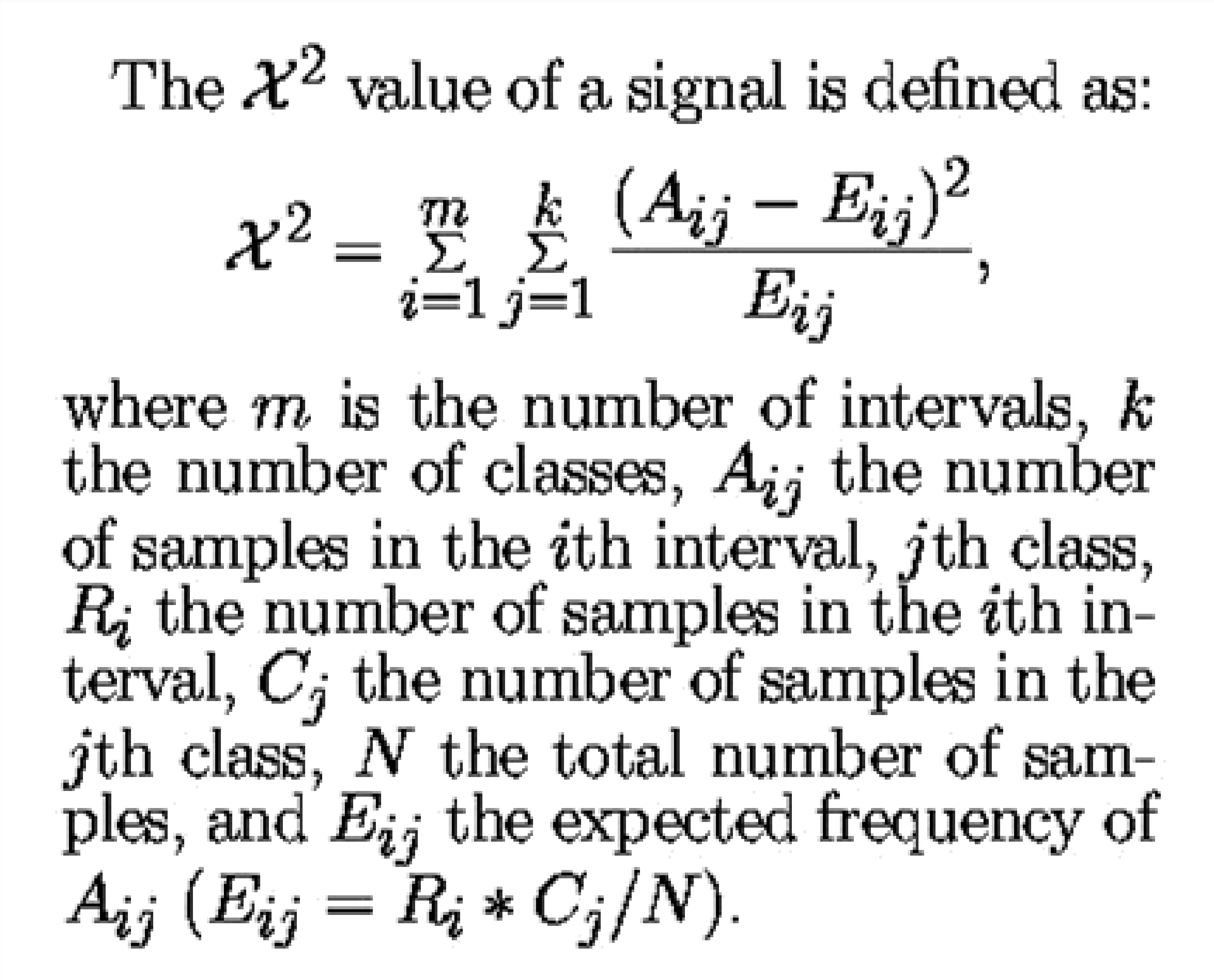


The t-stats of a signal is defined as

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

where σ_i^2 is the variance of that signal in class i, μ_i is the mean of that signal in class i, and n_i is the size of class i.

Signal selection: χ2



The χ^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 ith interval, jth class, R_i the number of samples in the ith interval, C_j the number of samples in the jth 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	(obs – exp)^2/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

χ2=16.6
P = 0.00004,
df = 1
So weight and
sex are not indep

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

Exercise #3

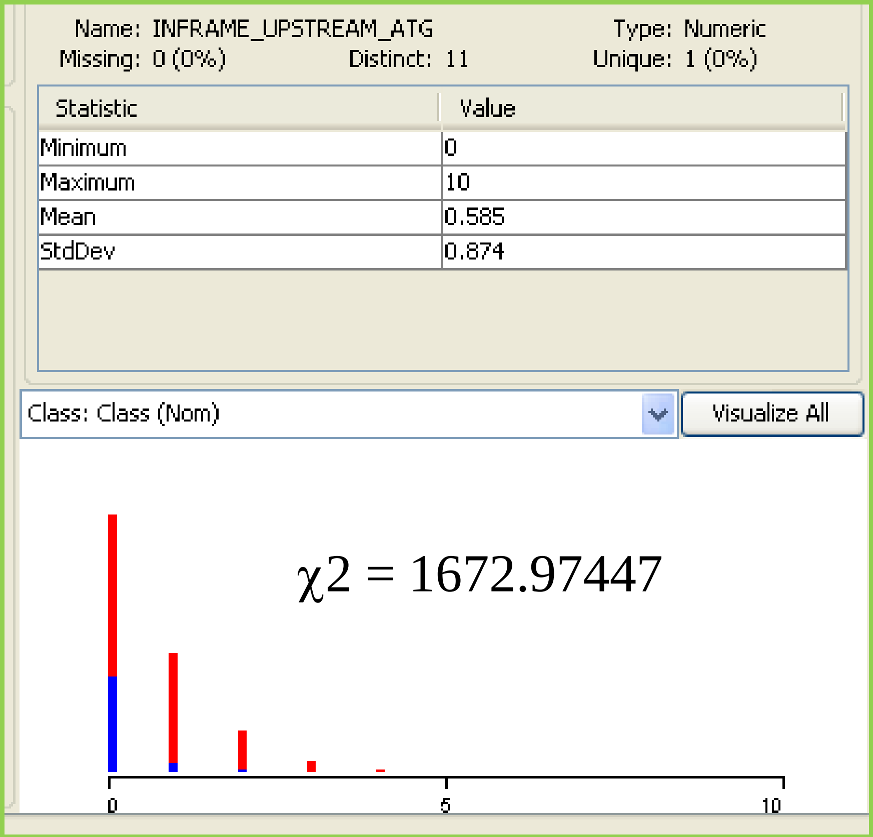
Signal selection: CFS

- **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

- **What is the main challenge in implementing CFS?**

Distributions of two 3-grams

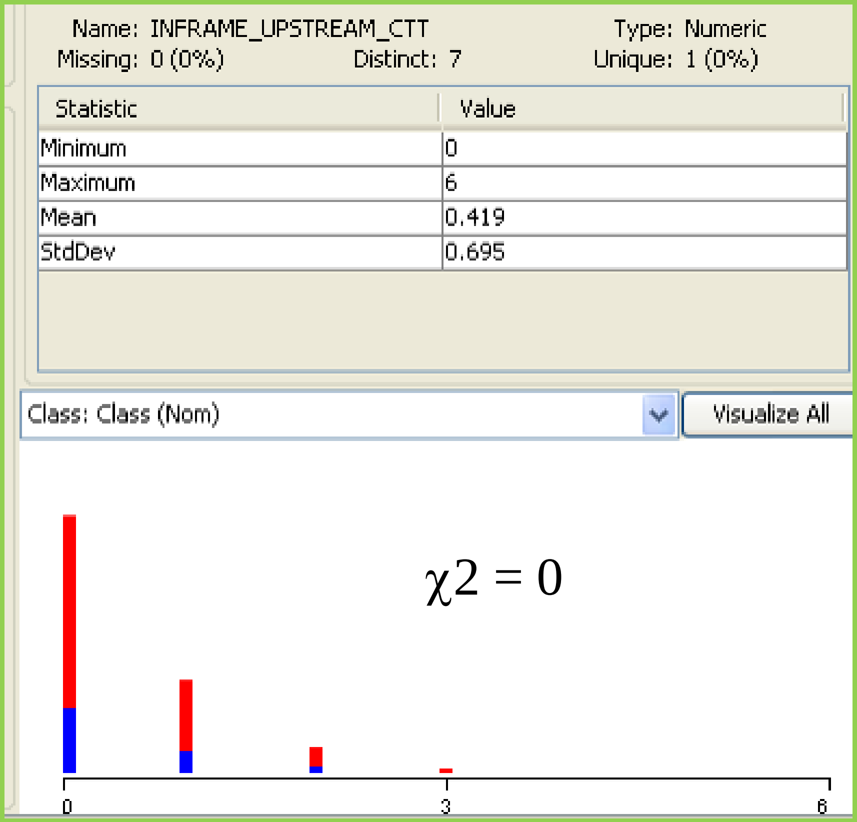


Name: INFRAME_UPSTREAM_ATG Type: Numeric
Missing: 0 (0%) Distinct: 11 Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	10
Mean	0.585
StdDev	0.874

Class: Class (Nom)

$\chi 2 = 1672.97447$



Name: INFRAME_UPSTREAM_CTT Type: Numeric
Missing: 0 (0%) Distinct: 7 Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	6
Mean	0.419
StdDev	0.695

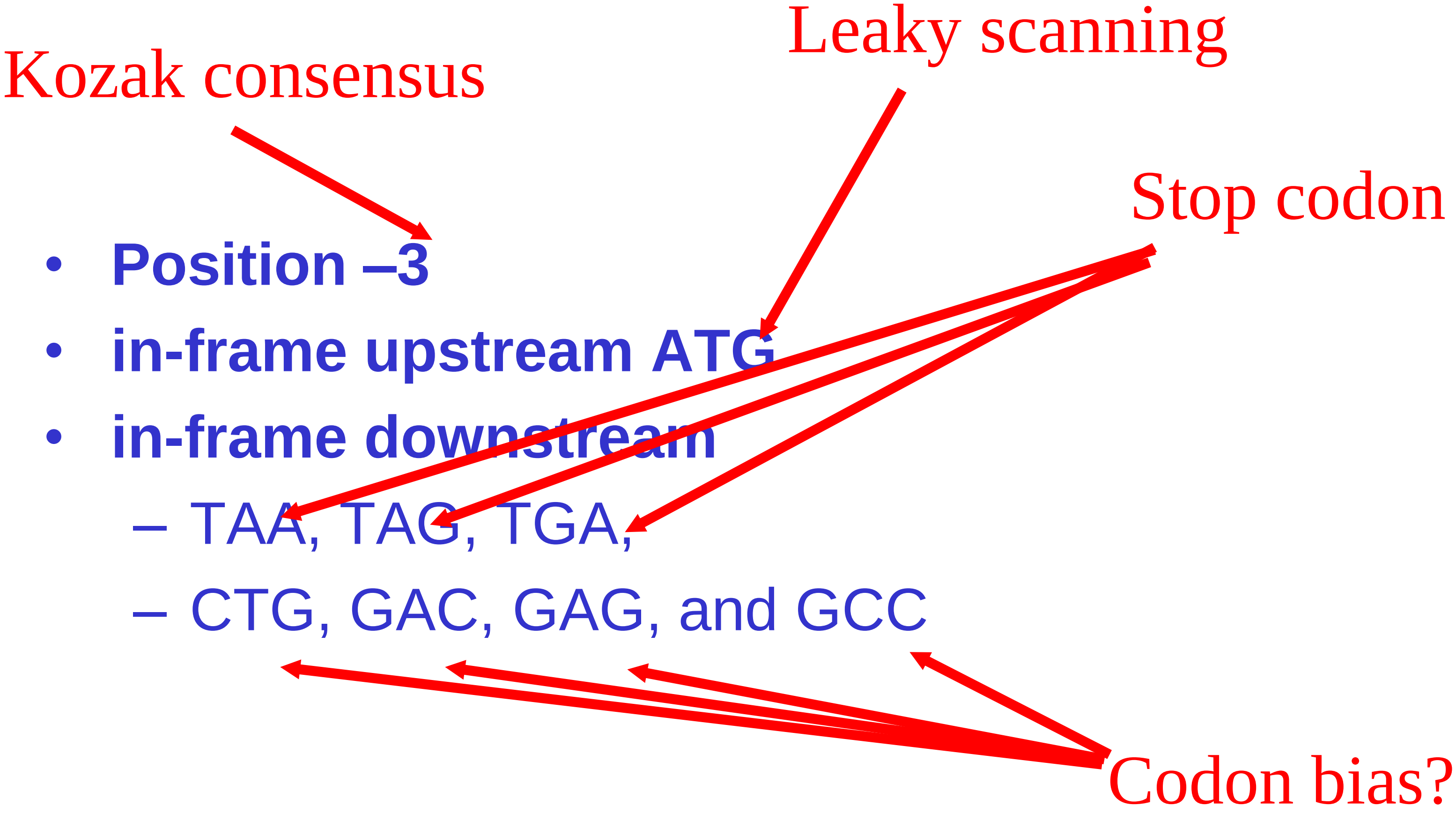
Class: Class (Nom)

$\chi 2 = 0$

- **Which is the better one? Why?**

Exercise #4

Sample k-grams selected by CFS for recognizing TIS



Kozak consensus

Leaky scanning

Stop codon

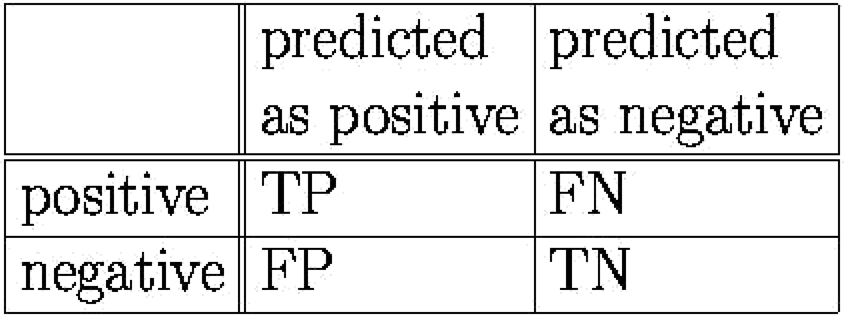
- **Position –3**
- **in-frame upstream ATG**
- **in-frame downstream**
  - TAA, TAG, TGA,
  - CTG, GAC, GAG, and GCC

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



	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

	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%

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%

Performance comparison

	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%*

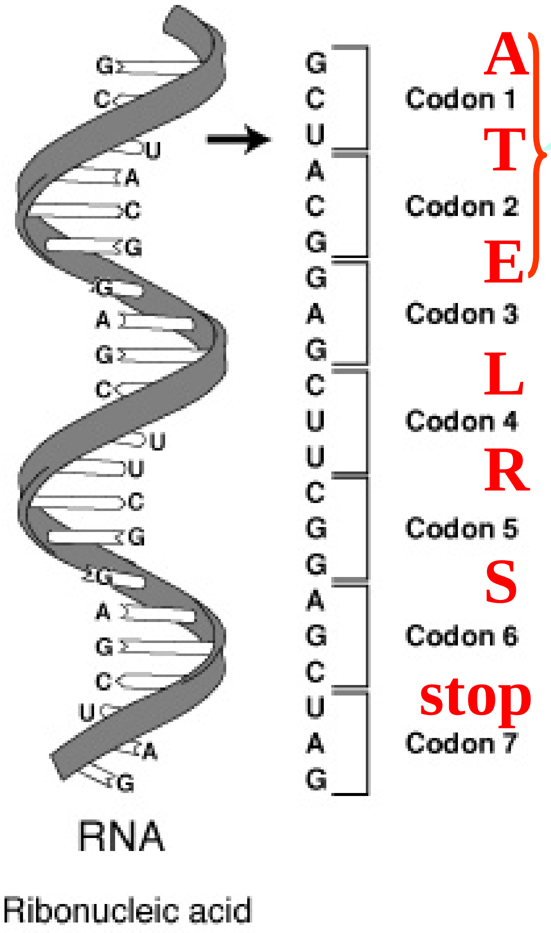
* result not directly comparable

Technique comparison

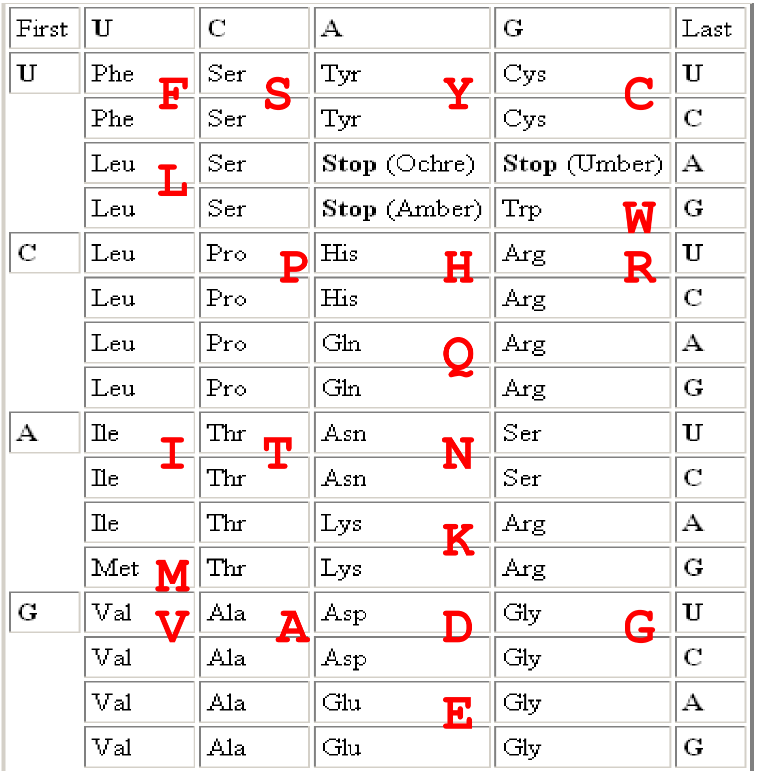
- **Pedersen&Nielsen [ISMB'97]**
  - Neural network
  - No explicit features
- **Zien [Bioinformatics'00]**
  - SVM+kernel engineering
  - No explicit features
- **Hatzigeorgiou [Bioinformatics'02]**
  - Multiple neural networks
  - Scanning rule
  - No explicit features

- **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



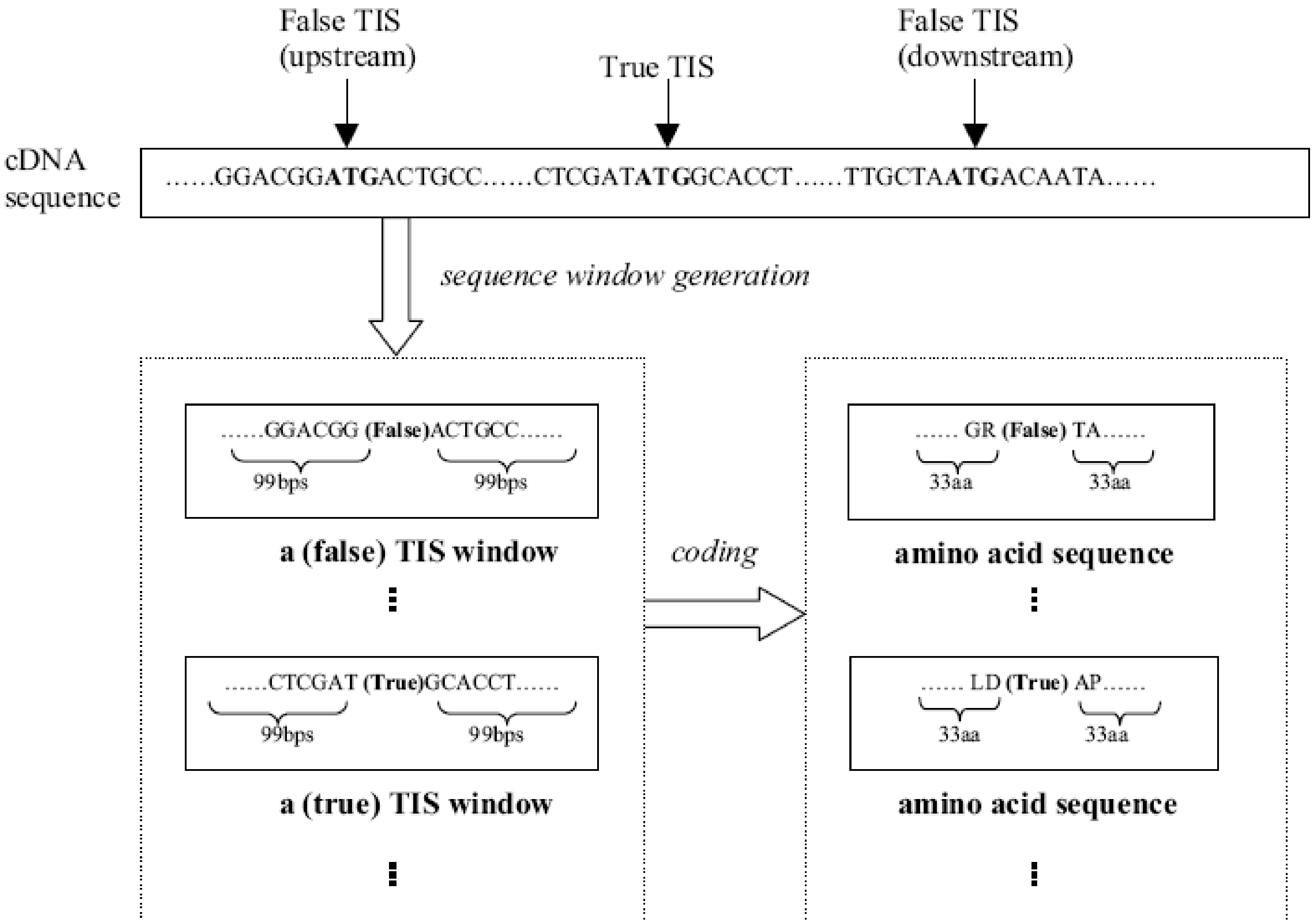
How about using k-grams from the translation?



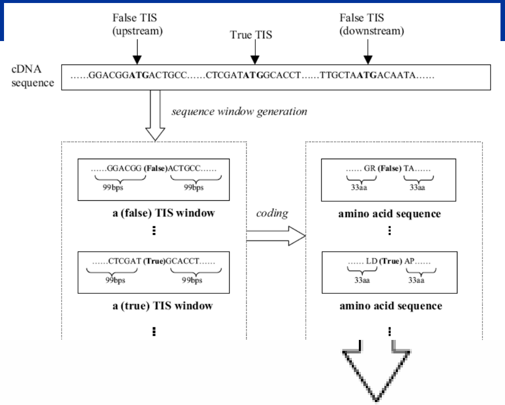
First	U	C	A	G	Last
U	Phe F	Ser S	Tyr Y	Cys C	U
	Phe	Ser	Tyr	Cys	C
	Leu L	Ser	Stop (Ochre)	Stop (Umber)	A
	Leu	Ser	Stop (Amber)	Trp W	G
C	Leu	Pro P	His H	Arg R	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln Q	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile I	Thr T	Asn N	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys K	Arg	A
	Met M	Thr	Lys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

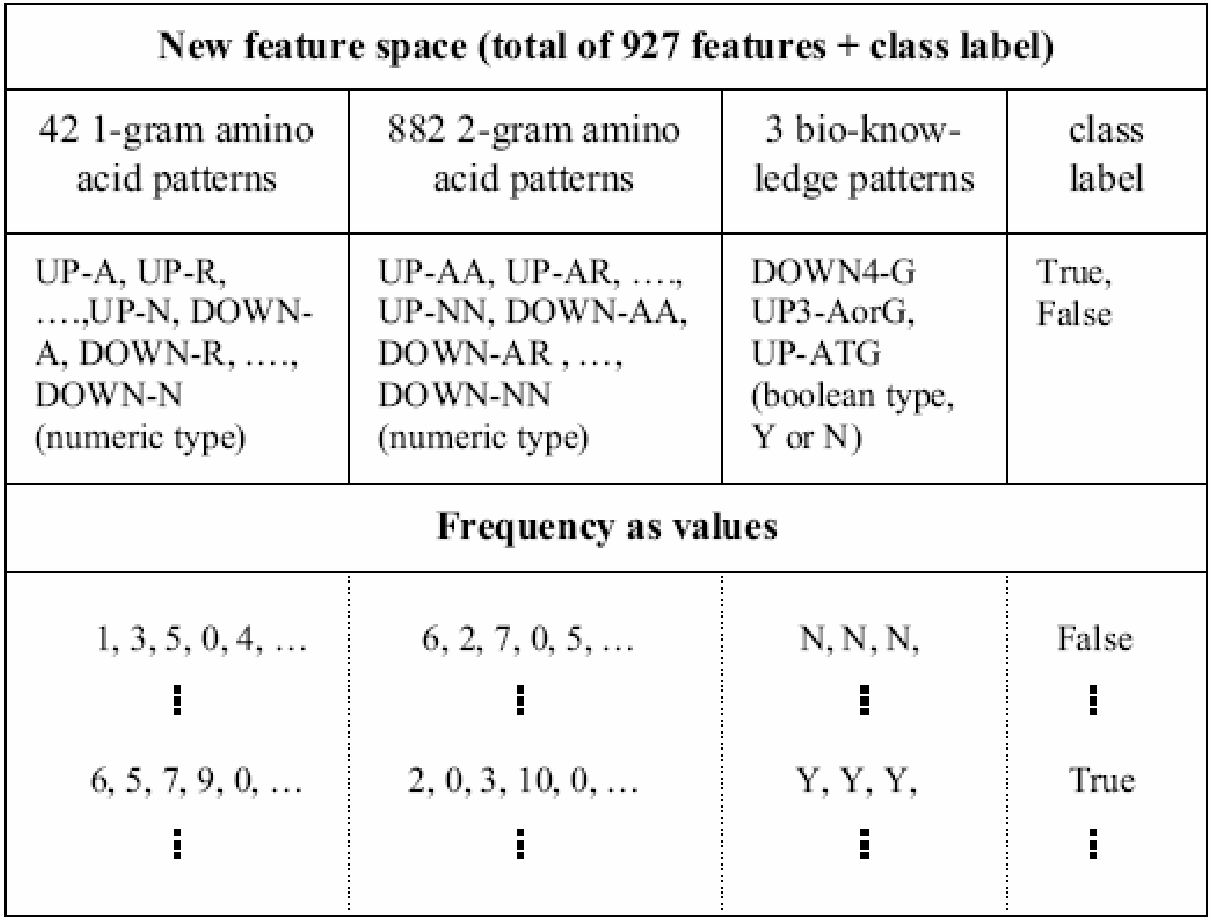
Exercise: List the first 10 amino acid in our example sequence

Amino-acid features


False TIS
(upstream)
True TIS
False TIS
(downstream)
cDNA
sequence
......GGACGGATGACTGCC......CTCGATATGGCACCT......TTGCTAATGACAATA......
sequence window generation
......GGACGG (False)ACTGCC......
99bps
99bps
a (false) TIS window
......CTCGAT (True)GCACCT......
99bps
99bps
a (true) TIS window
coding
...... GR (False) TA......
33aa
33aa
amino acid sequence
...... LD (True) AP......
33aa
33aa
amino acid sequence

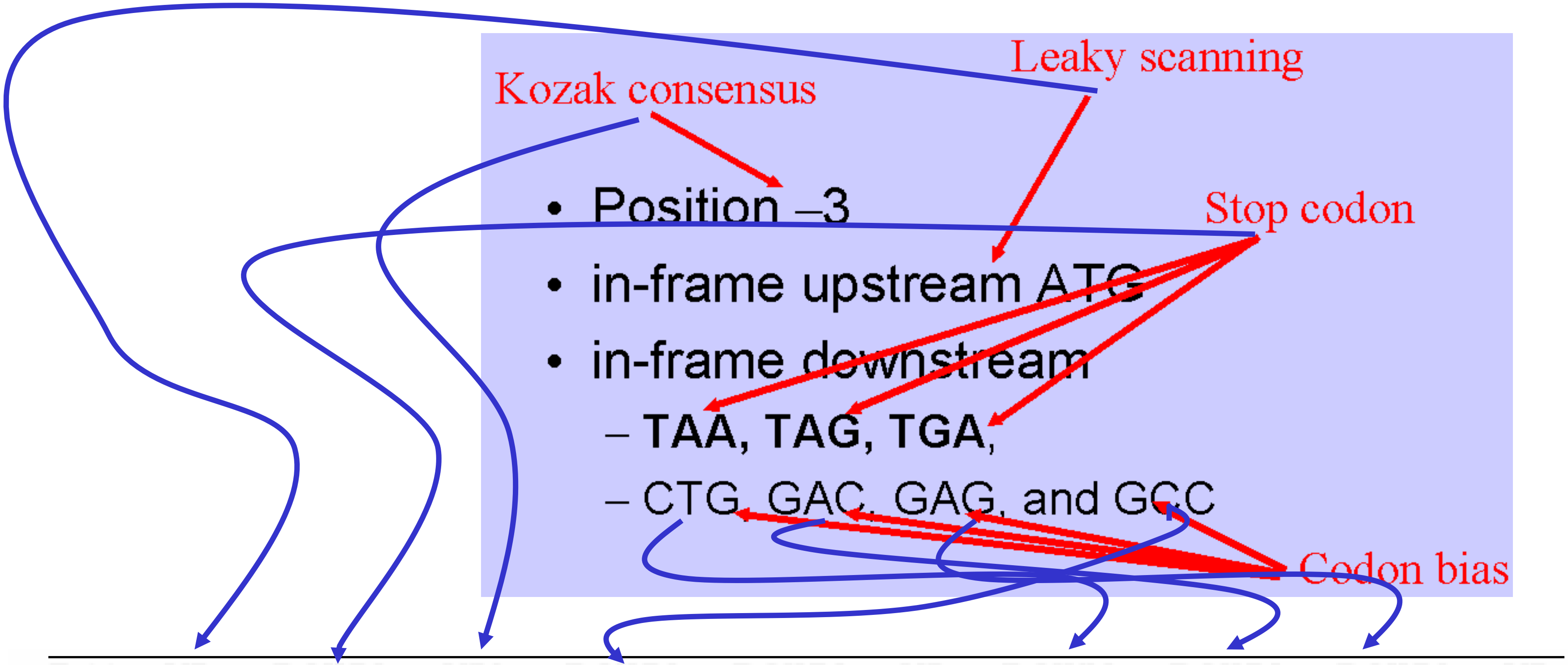
Amino-acid features





New feature space (total of 927 features + class label)			
42 1-gram amino acid patterns	882 2-gram amino acid patterns	3 bio-know-ledge 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

Leaky scanning

Stop codon

Codon bias

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

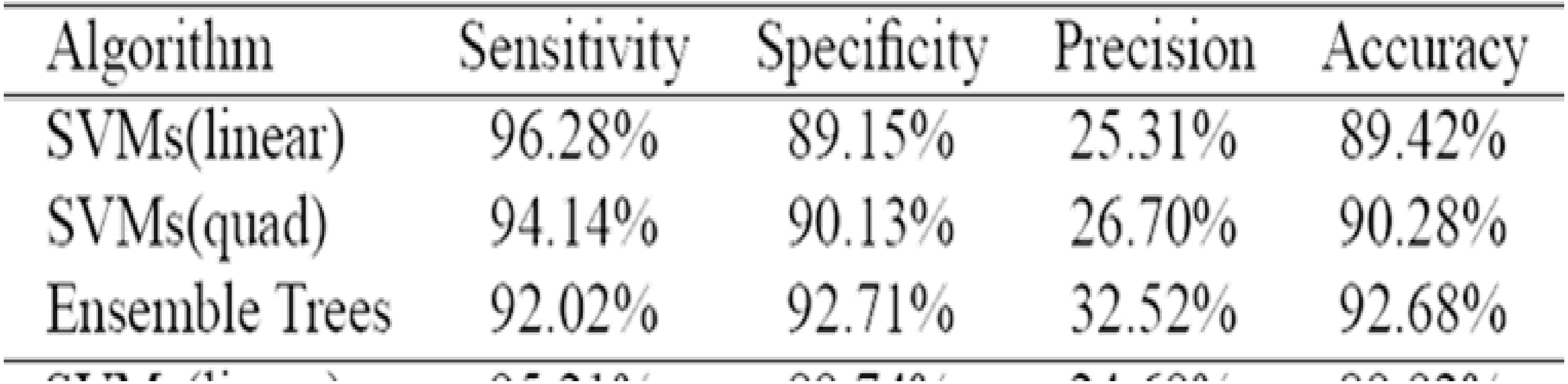
Fold	UP- ATG	DOWN- STOP	UP3- AorG	DOWN- A	DOWN- V	UP- A	DOWN- L	DOWN- D	DOWN- E	UP- G
1	1	2	4	3	6	5	8	9	7	10
2	1	2	3	4	5	6	7	8	9	10
3	1	2	3	4	5	6	8	9	7	10

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)

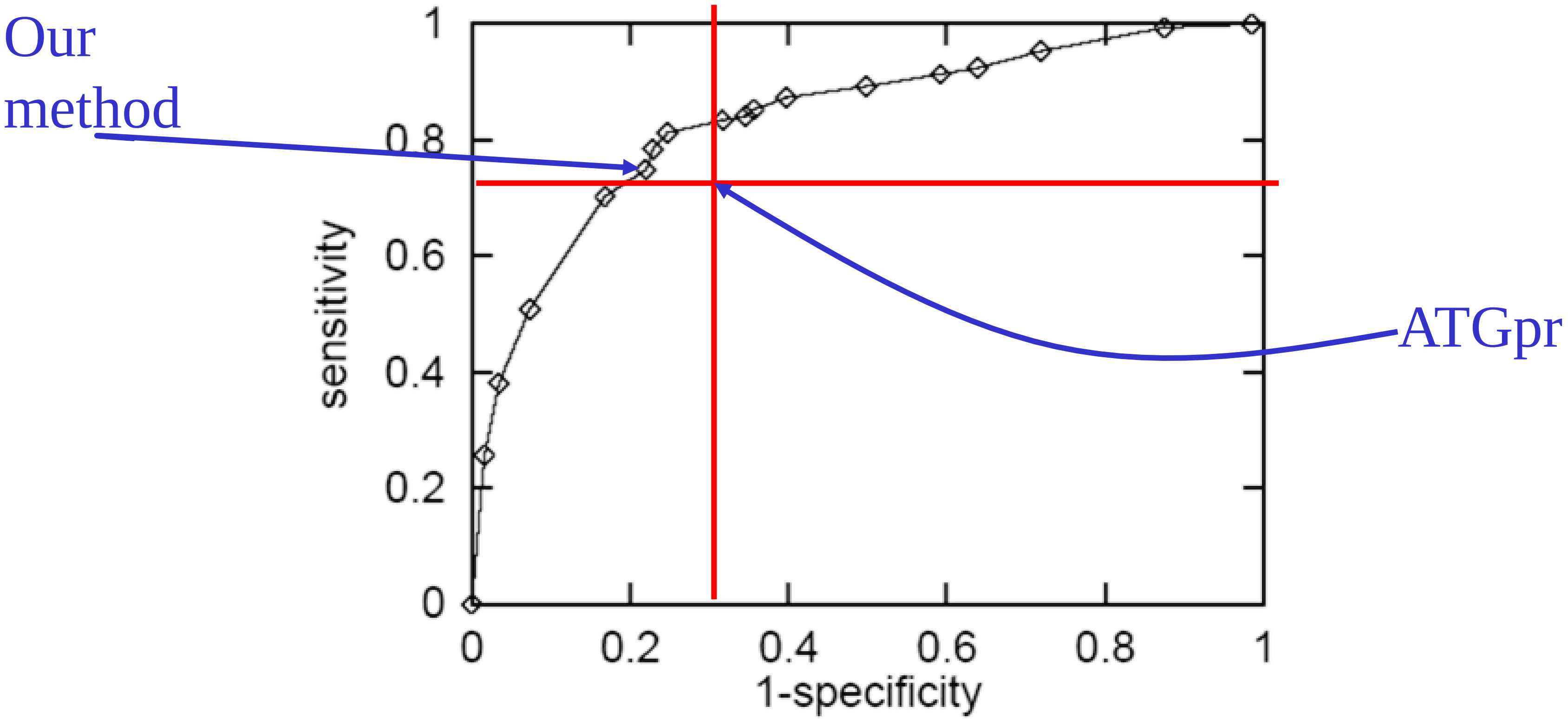
Validation results, on Hatzigeorgiou's



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 & 21



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

About the inventor: Huiqing Liu

- **Huiqing Liu**
  - PhD, NUS, 2004
  - Currently PI at Incyte
  - Asian Innovation Gold Award 2003
  - New Jersey Cancer Research Award for Scientific Excellence 2008
  - Gallo Prize 2008

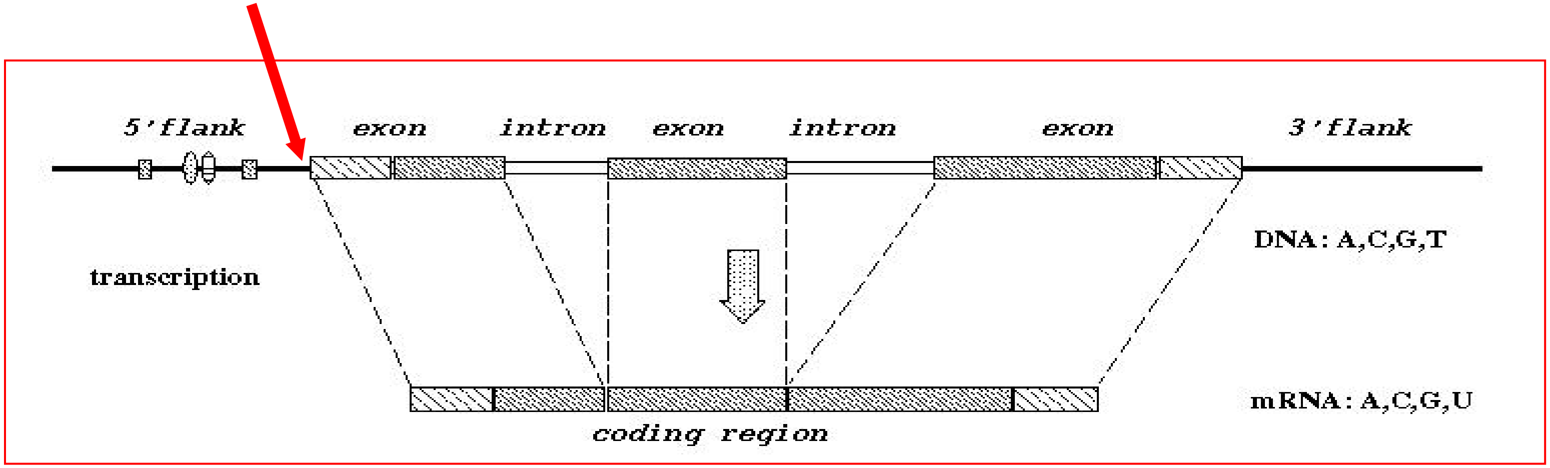
Recognition of Transcription Start Sites

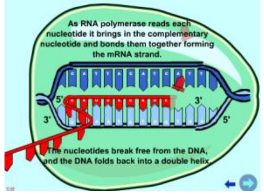
An introduction to the World’s best TSS recognition system of its time:

A heavy tuning approach

NUS
National University of Singapore

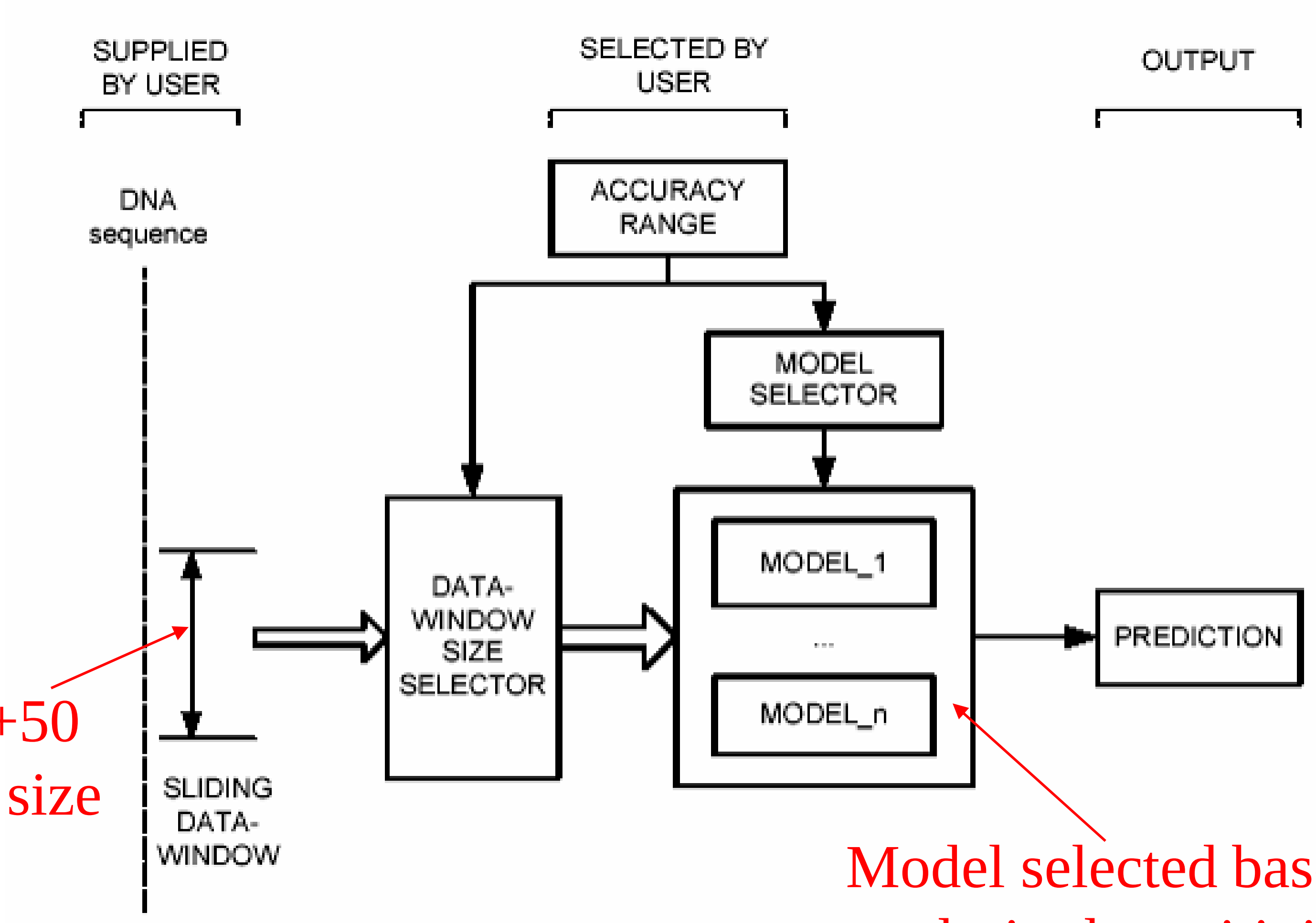
Transcription start site




As RNA polymerase reads each nucleotide it brings in the complementary nucleotide and bonds them together forming the mRNA strand.

The nucleotides break free from the DNA, and the DNA folds back into a double helix.

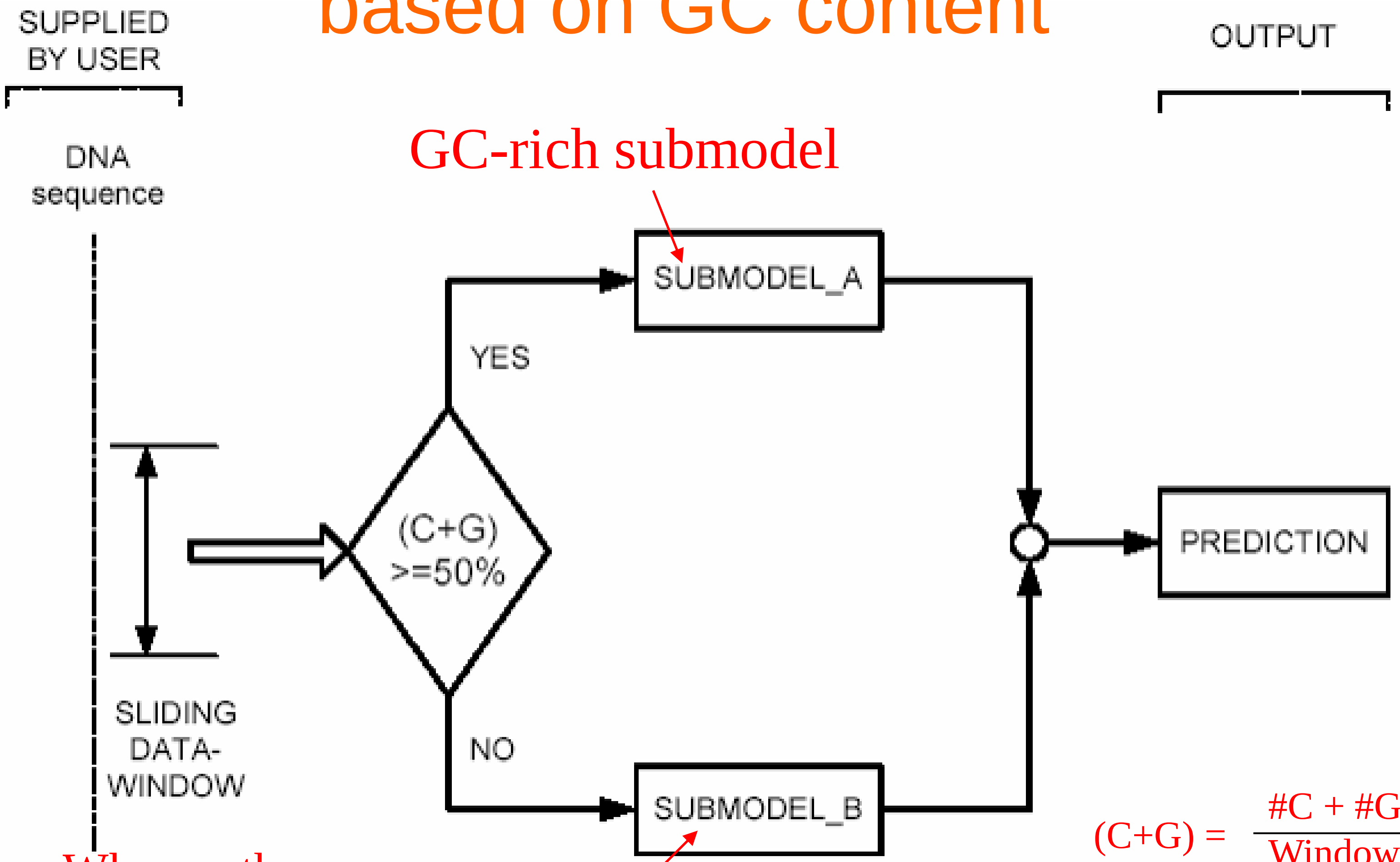
Structure of Dragon Promoter Finder



-200 to +50 window size

Model selected based on desired sensitivity

Each model has two submodels based on GC content



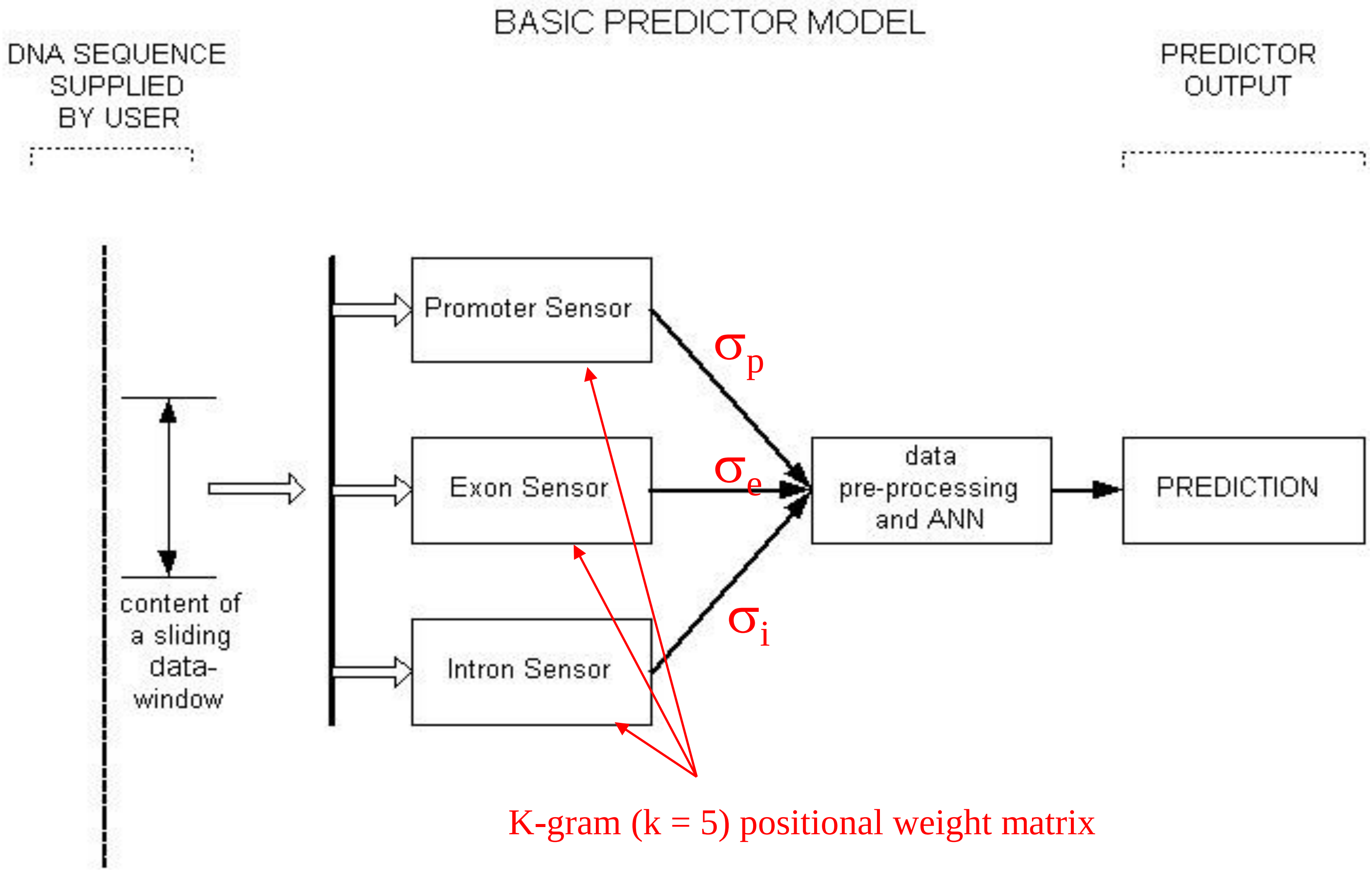
GC-rich submodel

GC-poor submodel

$$(C+G) = \frac{\#C + \#G}{\text{Window Size}}$$

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

Window size

Pentamer at ith position in input

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

Frequency of jth pentamer at ith position in training window

jth pentamer at ith position in training window

Just to make sure you know what I mean….

- **Given 3 DNA seq of length 10:**
  - Seq_1 = ACCGAGTTCT
  - Seq_2 = AGTGTACCTG
  - Seq_3 = 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

Exercise #5

Just to make sure you know what I mean…

- **Given 3 DNA seq of length 10:**
  - Seq_1 = ACCGAGTTCT
  - Seq_2 = AGTGTACCTG
  - Seq_3 = 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

Exercise #6

Data preprocessing & ANN

Tuning parameters

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

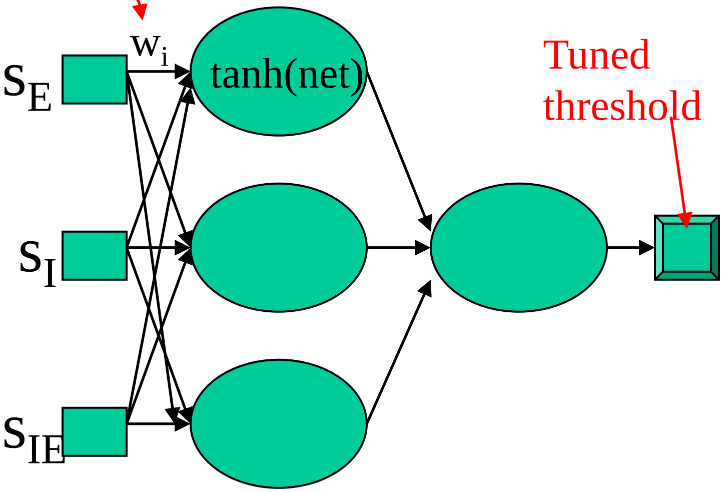
$$s_I = sat(\sigma_p - \sigma_i, a_i, b_i),$$

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

where the function *sat* is defined by

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

Simple feedforward ANN trained by the Bayesian regularisation method

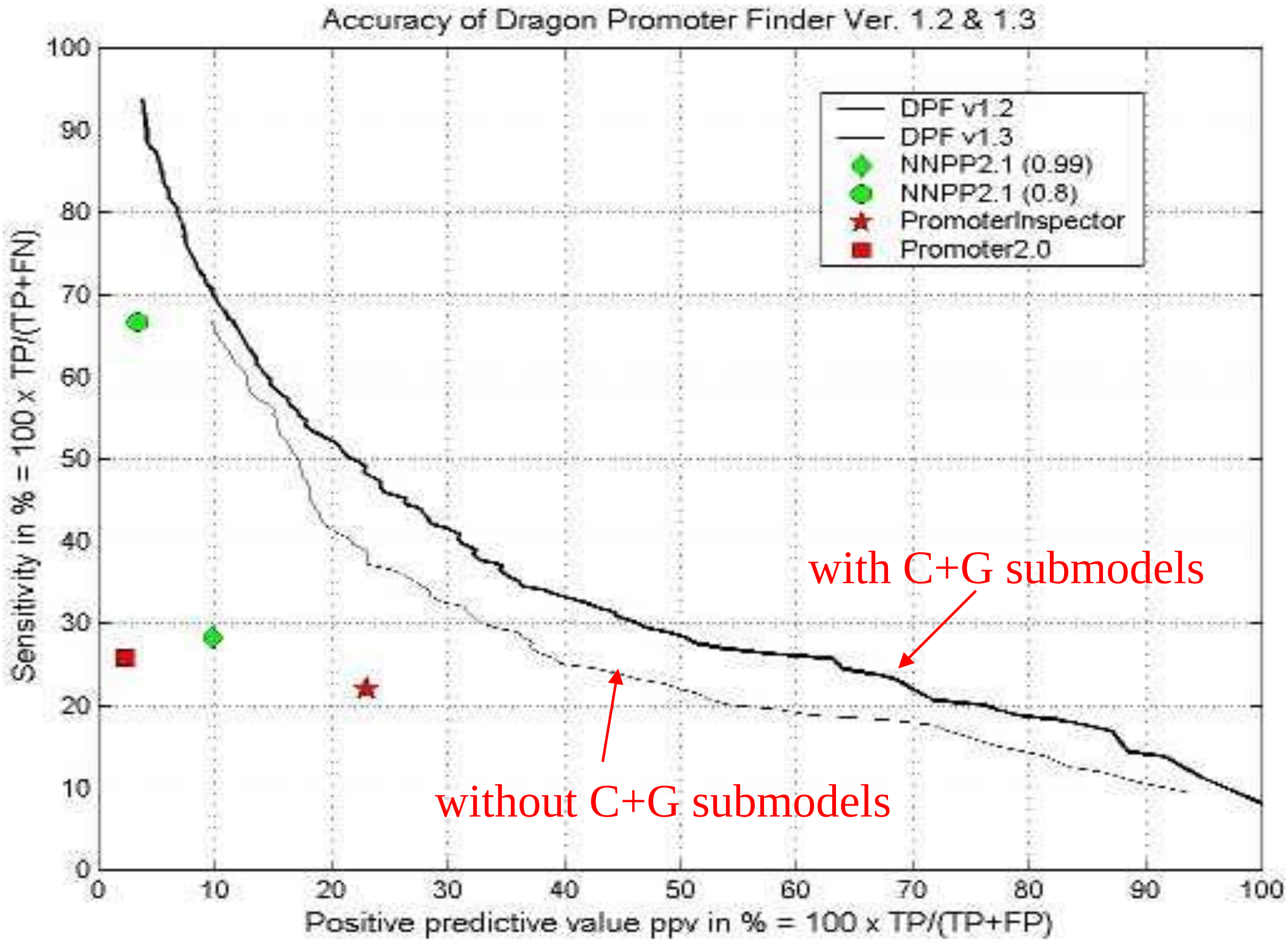


Tuned threshold

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

$$net = \Sigma\, s_i * w_i$$

Accuracy comparison



with C+G submodels

without C+G submodels

Training data criteria & preparation

- **Contain both positive and negative sequences**
- **Sufficient diversity, resembling different transcription start mechanisms**
- **Sufficient diversity, resembling different non-promoters**
- **Sanitized as much as possible**

- **TSS taken from**
  - 793 vertebrate promoters from EPD
  - -200 to +50 bp of TSS
- **non-TSS taken from**
  - GenBank,
  - 800 exons
  - 4000 introns,
  - 250 bp,
  - non-overlapping,
  - <50% identities

Tuning data preparation

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

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

- **Non-TSS taken from**
  - 1600 human 3’UTR seqs
  - 500 human exons
  - 500 human introns
  - 250 bp
  - no overlap

Testing data criteria & preparation

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

- 159 TSS from 147 human and human virus seqs
- cummulative length of more than 1.15Mbp
- Taken from GENESCAN, Geneld, Genie, etc.

About the inventor: Vlad Bajic

- **Vladimir B. Bajic**
  - – Principal Scientist, I^{2}R, 2001-2006
  - – Director & Professor, Computational Bioscience Research Center, KAUST
  - – Passed away in 2019

Recognition of Poly-A signal sites

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

NUS
National University of Singapore

Eukaryotic pre-mRNA processing

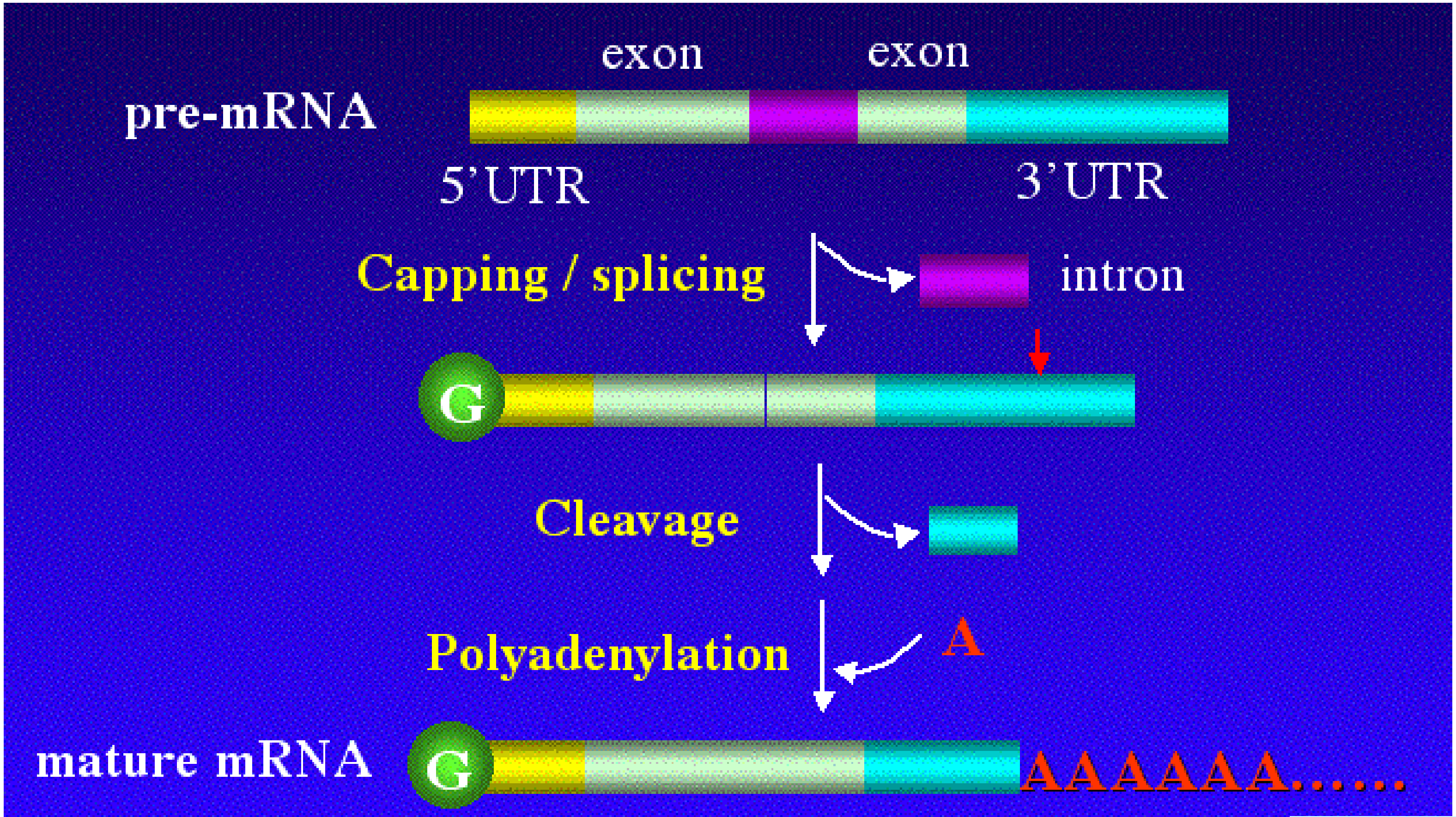
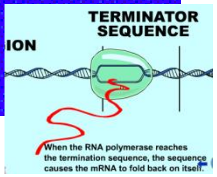
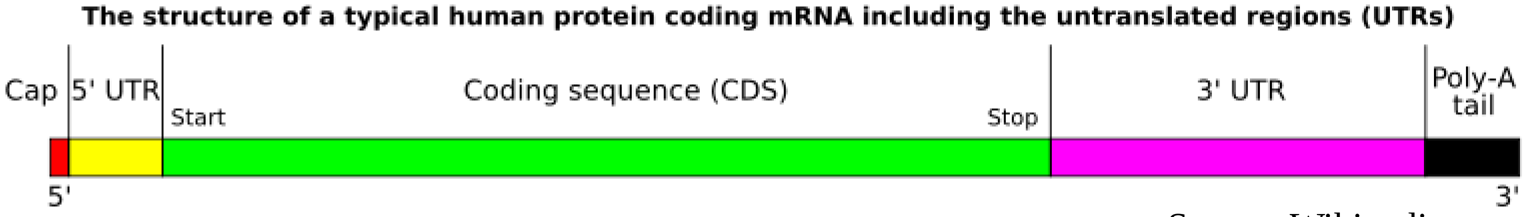


Image credit: www.polya.org


TERMINATOR SEQUENCE
ION
When the RNA polymerase reaches the termination sequence, the sequence causes the mRNA to fold back on itself.

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**


The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5' UTR
Start
Coding sequence (CDS)
Stop
3' UTR
Poly-A tail
5'
3'

Source: Wikipedia

Poly-A signals in human (Gautheret et al., 2000)

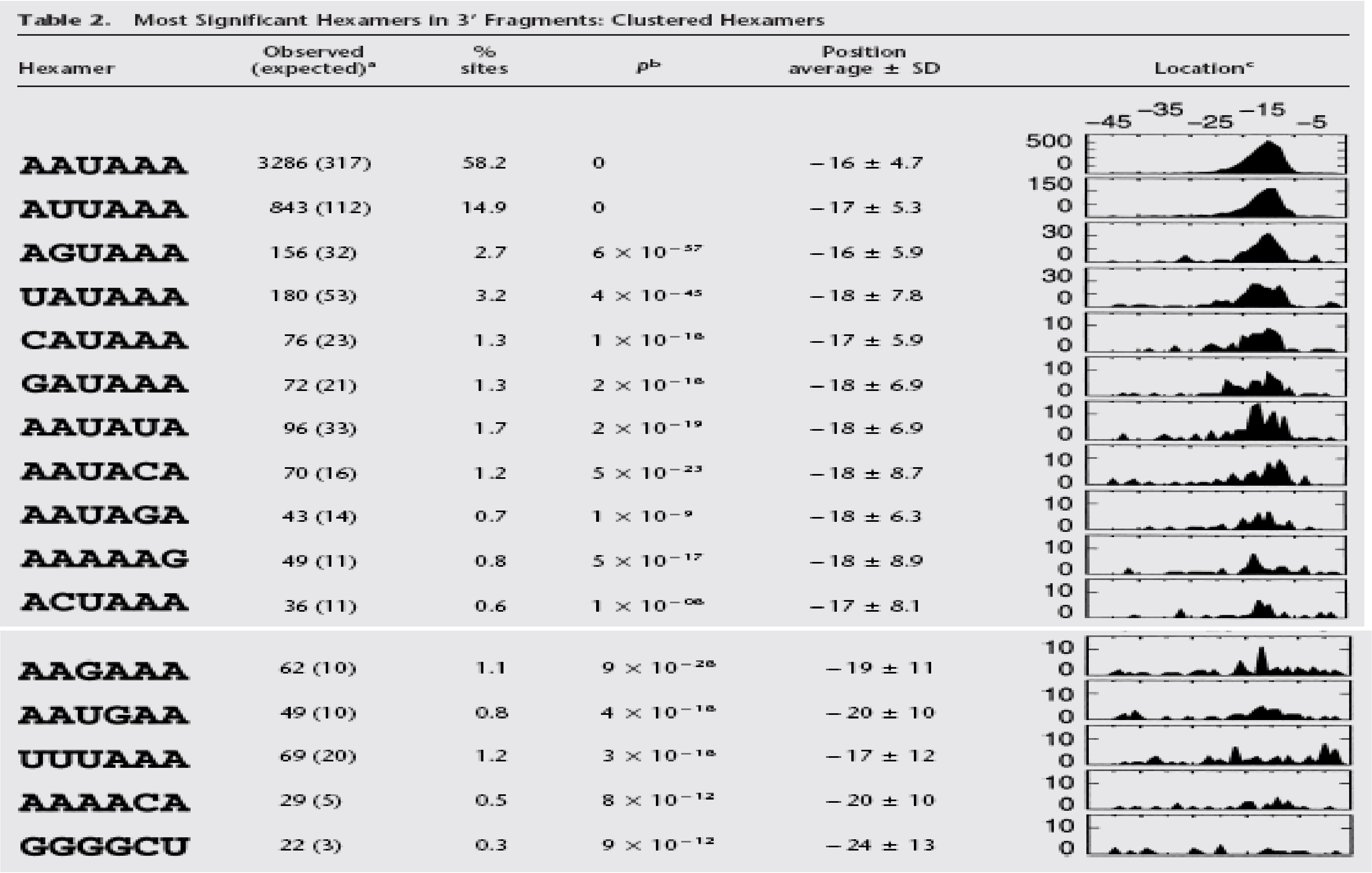


Table 2. Most Significant Hexamers in 3′ Fragments: Clustered Hexamers

Hexamer	Observed (expected)[a]	% sites	P^b	Position average ± SD	Location[c]
					−45 −35 −25 −15 −5
AAUAAA	3286 (317)	58.2	0	−16 ± 4.7	500 0
AUUAAA	843 (112)	14.9	0	−17 ± 5.3	150 0
AGUAAA	156 (32)	2.7	6×10^{-57}	−16 ± 5.9	30 0
UAUAAA	180 (53)	3.2	4×10^{-45}	−18 ± 7.8	30 0
CAUAAA	76 (23)	1.3	1×10^{-18}	−17 ± 5.9	10 0
GAUAAA	72 (21)	1.3	2×10^{-18}	−18 ± 6.9	10 0
AAUAUA	96 (33)	1.7	2×10^{-19}	−18 ± 6.9	10 0
AAUACA	70 (16)	1.2	5×10^{-23}	−18 ± 8.7	10 0
AAUAGA	43 (14)	0.7	1×10^{-9}	−18 ± 6.3	10 0
AAAAAG	49 (11)	0.8	5×10^{-17}	−18 ± 8.9	10 0
ACUAAA	36 (11)	0.6	1×10^{-08}	−17 ± 8.1	10 0
AAGAAA	62 (10)	1.1	9×10^{-28}	−19 ± 11	10 0
AAUGAA	49 (10)	0.8	4×10^{-18}	−20 ± 10	10 0
UUUAAA	69 (20)	1.2	3×10^{-18}	−17 ± 12	10 0
AAAACA	29 (5)	0.5	8×10^{-12}	−20 ± 10	10 0
GGGGCU	22 (3)	0.3	9×10^{-12}	−24 ± 13	10 0

Poly-A signals in Arabidopsis

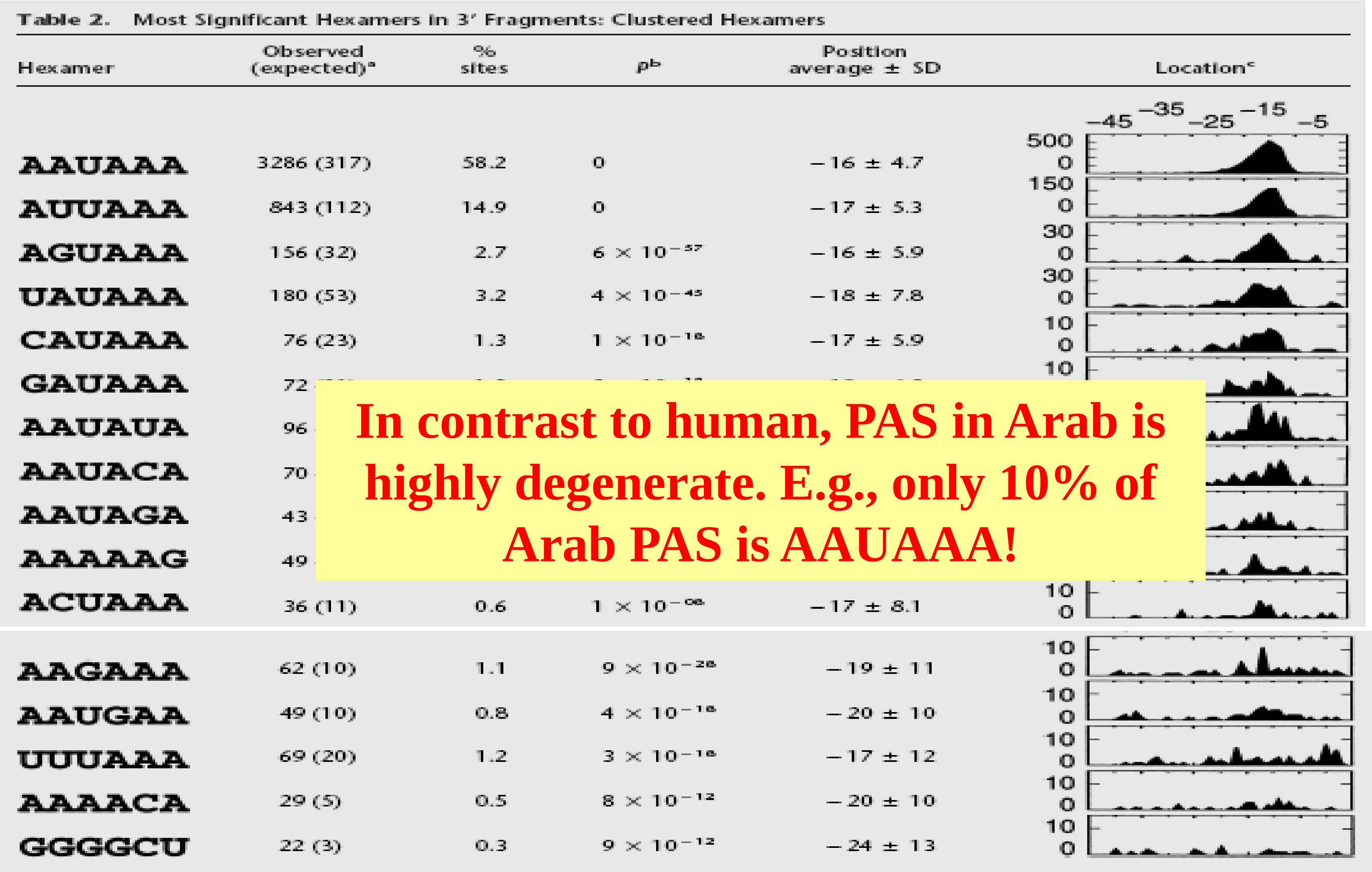
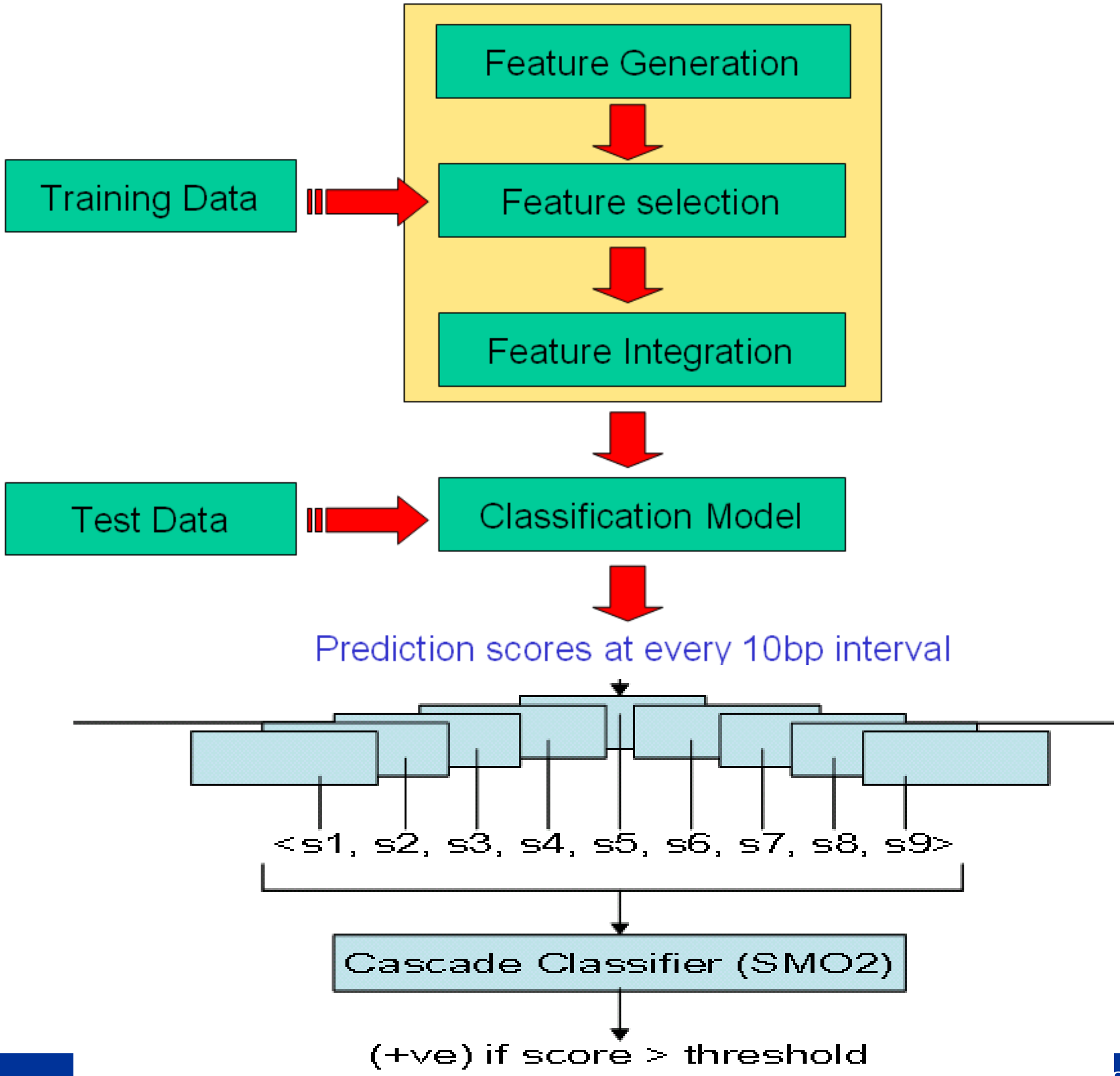


Table 2. Most Significant Hexamers in 3′ Fragments: Clustered Hexamers

Hexamer	Observed (expected)[a]	% sites	P^b	Position average ± SD	Location[c]
					−45 −35 −25 −15 −5
AAUAAA	3286 (317)	58.2	0	−16 ± 4.7	500 0
AUUAAA	843 (112)	14.9	0	−17 ± 5.3	150 0
AGUAAA	156 (32)	2.7	6×10^{-57}	−16 ± 5.9	30 0
UAUAAA	180 (53)	3.2	4×10^{-45}	−18 ± 7.8	30 0
CAUAAA	76 (23)	1.3	1×10^{-16}	−17 ± 5.9	10 0
GAUAAA	72				10
AAUAUA	96				
AAUACA	70				
AAUAGA	43				
AAAAAG	49				
ACUAAA	36 (11)	0.6	1×10^{-08}	−17 ± 8.1	10 0
AAGAAA	62 (10)	1.1	9×10^{-28}	−19 ± 11	10 0
AAUGAA	49 (10)	0.8	4×10^{-16}	−20 ± 10	10 0
UUUAAA	69 (20)	1.2	3×10^{-16}	−17 ± 12	10 0
AAAACA	29 (5)	0.5	8×10^{-12}	−20 ± 10	10 0
GGGGCU	22 (3)	0.3	9×10^{-12}	−24 ± 13	10 0

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)


Feature Generation
Training Data
Feature selection
Feature Integration
Test Data
Classification Model
Prediction scores at every 10bp interval
<s1, s2, s3, s4, s5, s6, s7, s8, s9>
Cascade Classifier (SMO2)
(+ve) if score > threshold

Approach on Arab PAS sites (II)

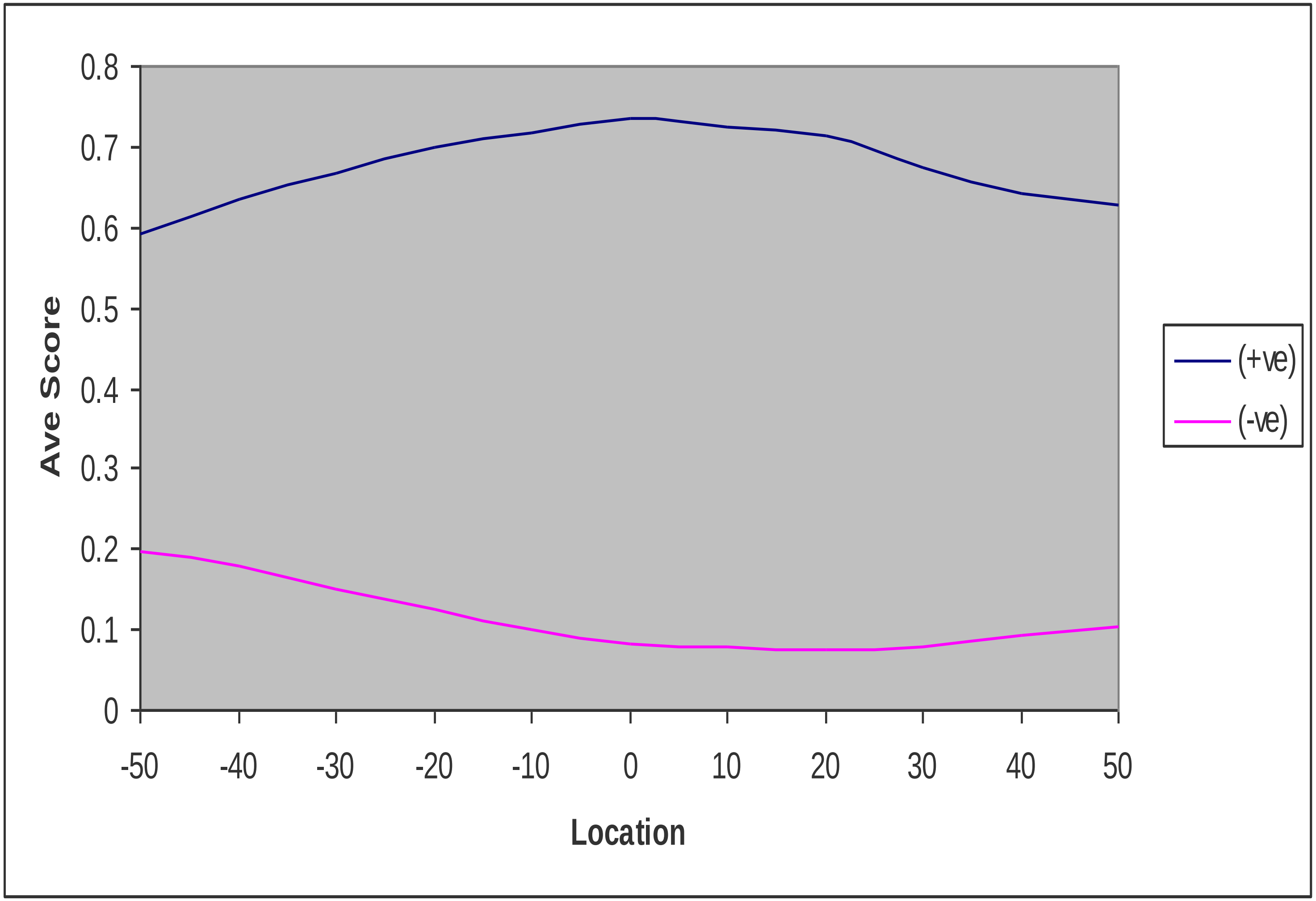
- **Data collection**
  - #1 from Hao Han, 811 +ve seq (-200/+200)
  - #2 from Hao Han, 9742 –ve seq (-200/+200)
  - #3 from Qingshun Li,
    - **6209 (+ve) seq (-300/+100)**
    - **1581 (-ve) intron (-300/+100)**
    - **1501 (-ve) coding (-300/+100)**
    - **864 (-ve) 5'utr (-300/+100)**

- **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)

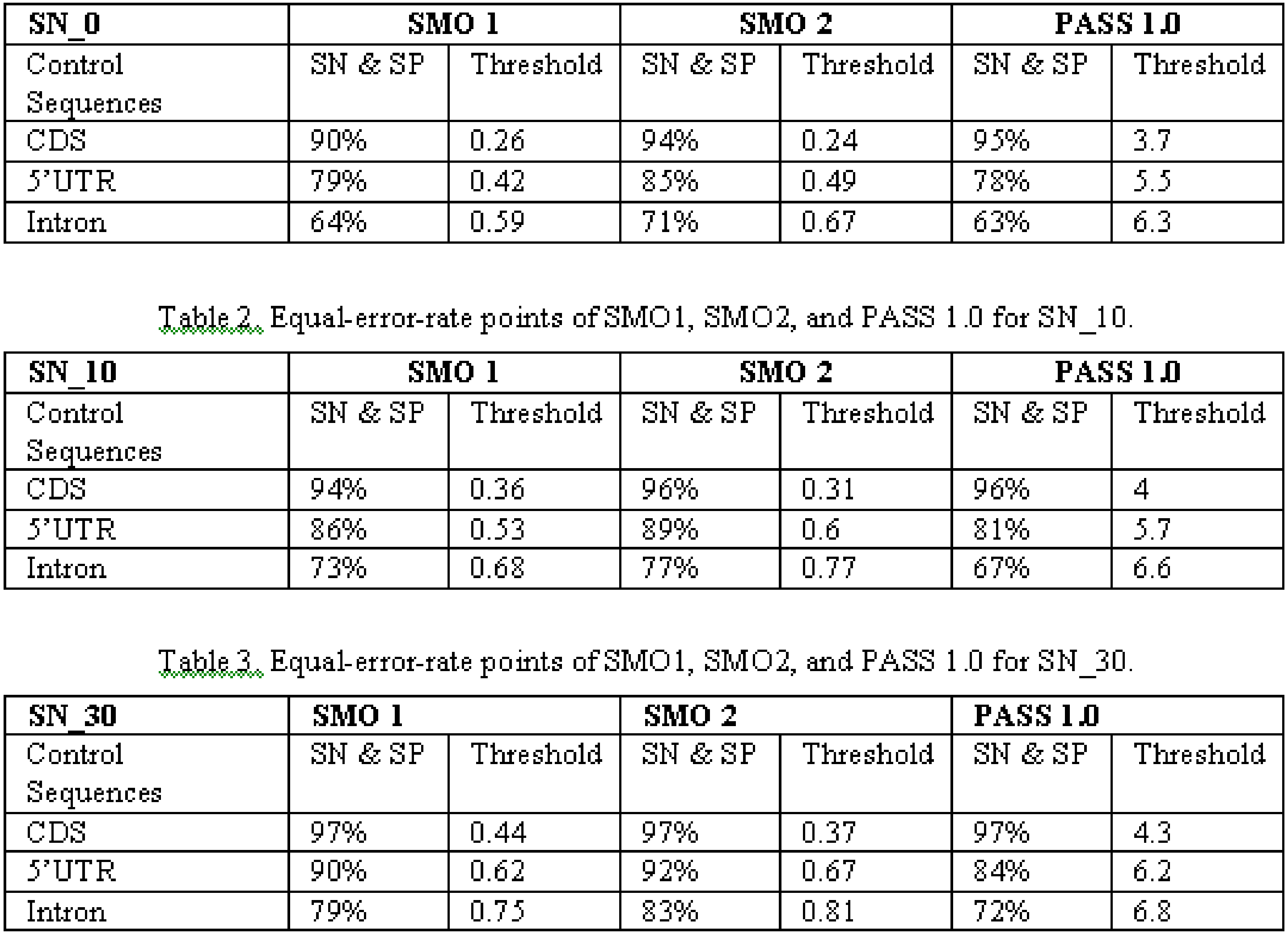
- **Feature selection**
  - $\chi 2$

- **Feature integration & Cascade**
  - SVM

Score profile relative to candidate sites



Validation results



SN_0	SMO 1		SMO 2		PASS 1.0	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
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_10.

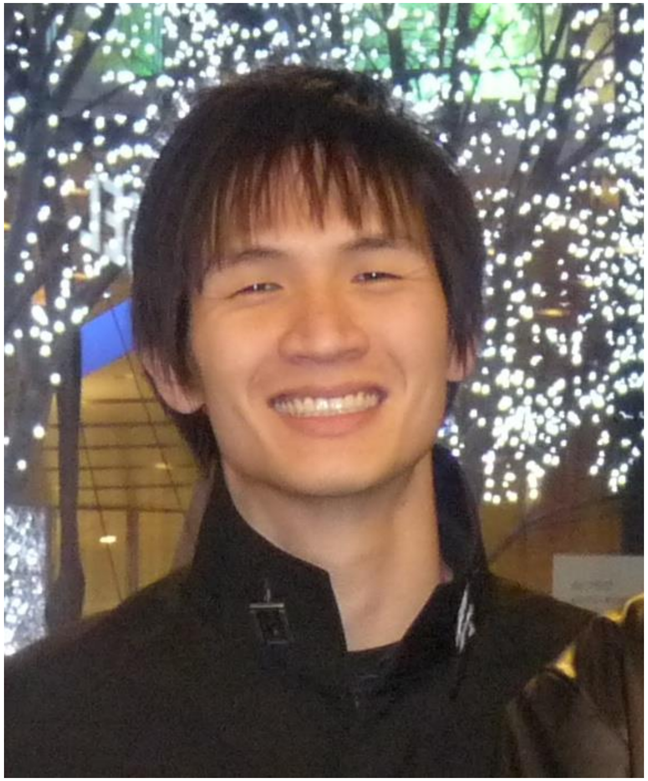
SN_10	SMO 1		SMO 2		PASS 1.0	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
CDS	94%	0.36	96%	0.31	96%	4
5'UTR	86%	0.53	89%	0.6	81%	5.7
Intron	73%	0.68	77%	0.77	67%	6.6

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

SN_30	SMO 1		SMO 2		PASS 1.0	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
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
  - – PhD, NUS, 2012
  - – Currently Data Science Mgr at Indeed Inc, Japan



Concluding remarks…

NUS
National University
of Singapore

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?

NUS
National University of Singapore

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