

CS2220: Intro to Computational Biology

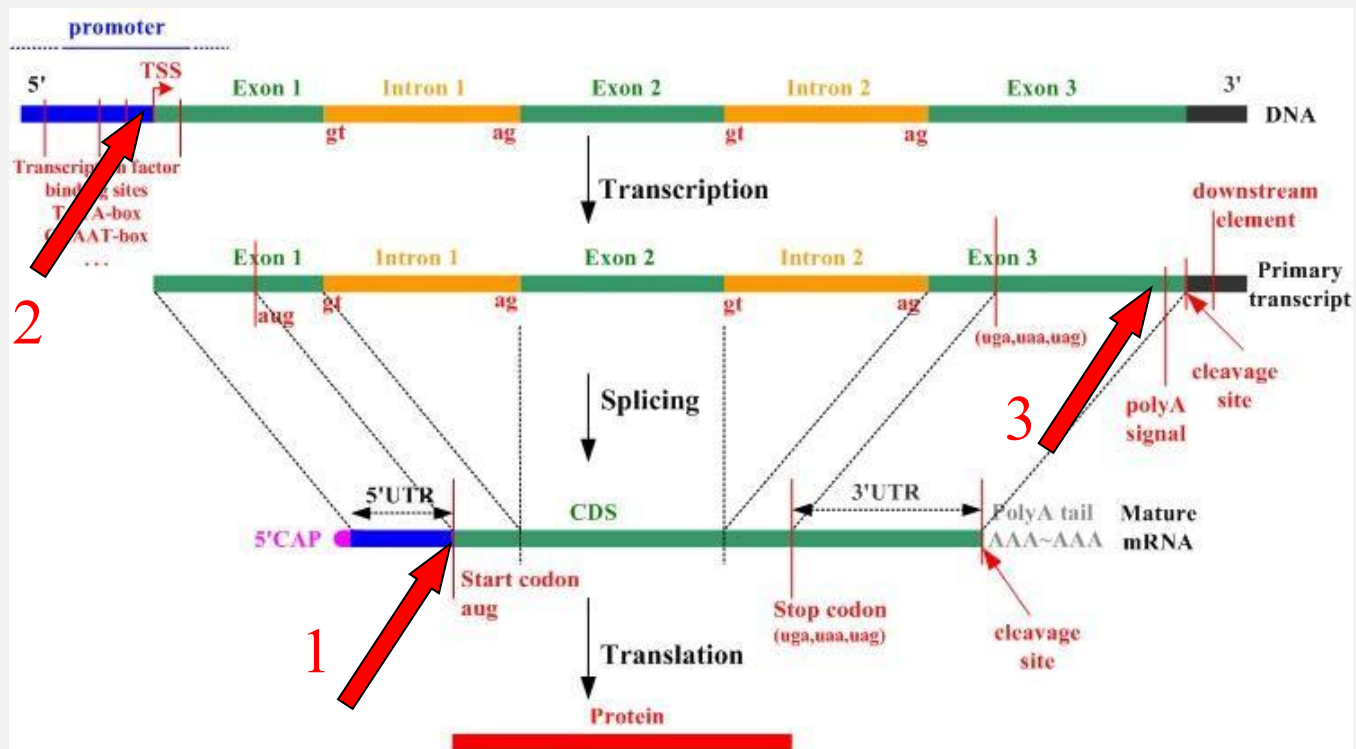
Gene Feature Recognition

Wong Limsoon



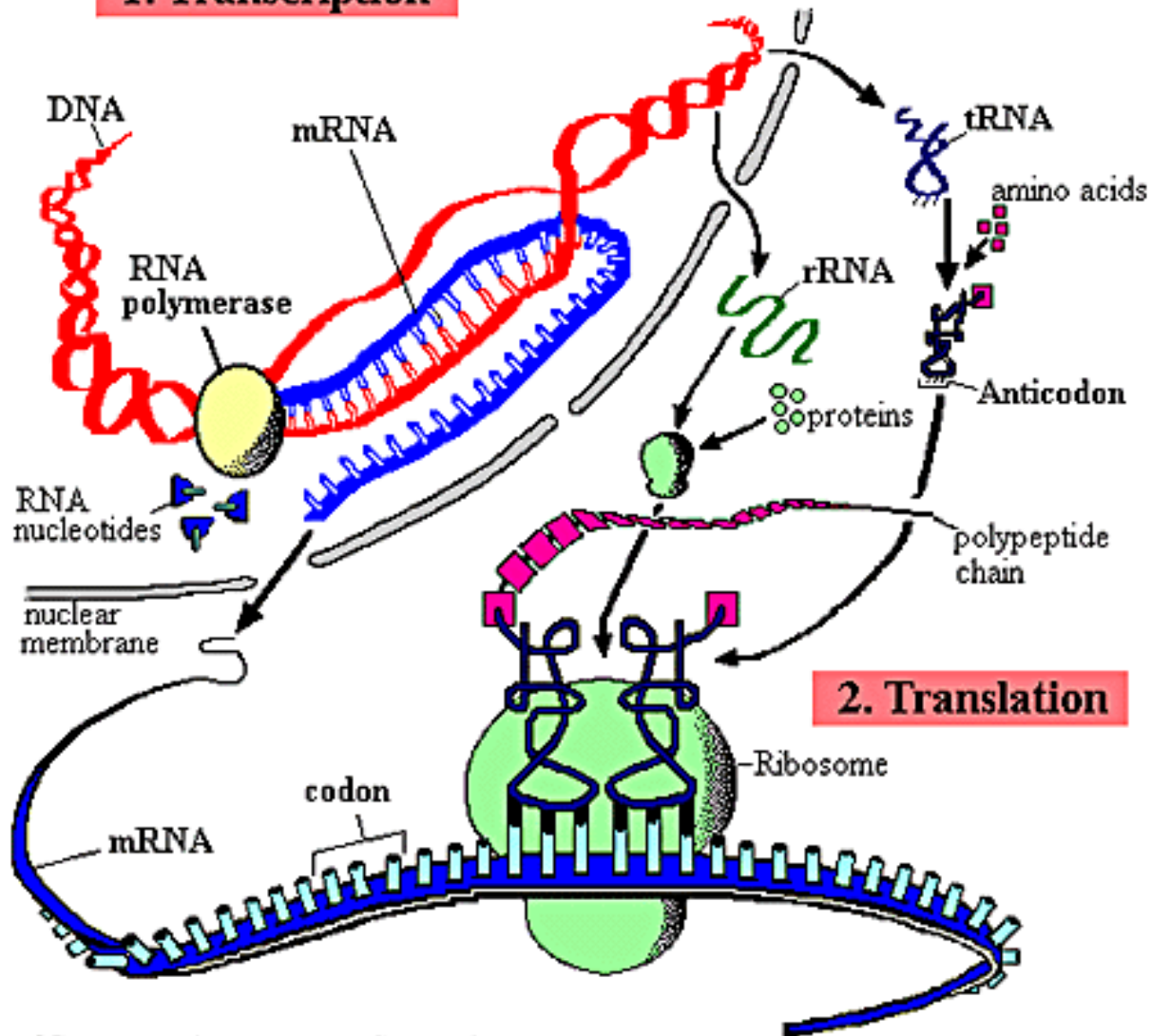
National University of Singapore

Outline



Some relevant biology

1. Transcription



**Central
dogma**

Protein synthesis

Transcription

Synthesize mRNA from one strand of DNA

RNA polymerase temporarily separates double-stranded DNA

It begins transcription at transcription start site

Once RNA polymerase reaches transcription stop site, transcription stops

More “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

$4^3 = 64$ codons

Not 1-to-1 corr to 20 amino acids

Most organisms use the same decoding table

Amino acids can be classified into 4 groups. A single-base change in a codon is usually 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' ATGGTGTCTGTCTGCCCGACAAGGGCAATGTCAAGGCCGCCTGGGGCAAGGTTGGCGGCCACGCTGCAGAGTATGGCGCAGAGGCCCTG 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' GAGAGGATGTTTCCTGAGCTTCCCCACCACCAAGACCTACTTCCCCCACTTCGACCTGAGCCACGGCTCCGCGCAGGTCAAGGGCCACGGC 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' GCGAAGGTGGCCGCCGCGCTGACCAAAGCGGTGGAACACCTGGACGACCTGCCCCGGTGCCCTGTCTGAACTGAGTGACCTGCACGCTCAC 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' AAGCTGCGTGTGGACCCGGTCAACTTCAAGCTTCTGAGCCACTCCCTGCTGGTGACCCTGGCCTCCCACCTCCCCAGTGATTTCACCCCC 360
  ...))).....))).....))).....))).....))).....))).....)))

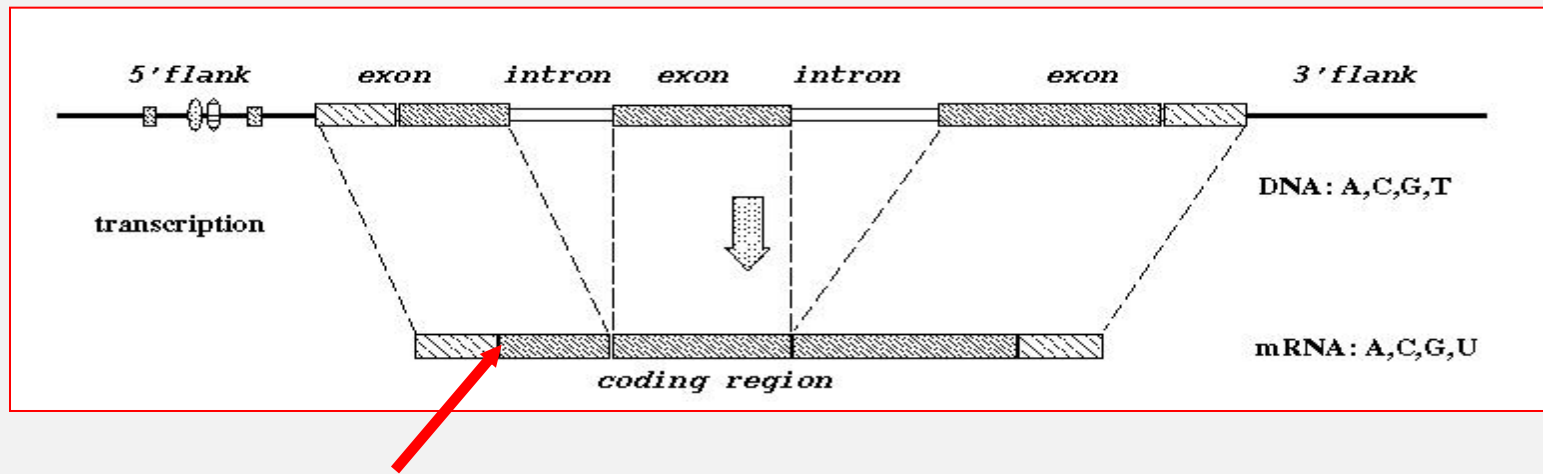
  A V H A S L D K F L A N V S T V L T S K Y R *
5' GCGGTCCACGCTCCCTGGACAAGTTCTTGGCCAACGTGAGCACCGTGCTGACCTCCAAATACCGTTAA 429
  .....))).....))).....))).....))).....***

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


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
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA      160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA      240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
.....iEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE      80
.....iEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE      160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE      240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

What makes the second ATG the TIS?

Recall the knowledge discovery workflow...

Training data gathering

Feature generation

k-grams, distance, domain know-how, ...

Feature selection

Entropy, χ^2 , CFS, t-test, domain know-how...

Feature 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

Feature generation

K-grams (ie., k consecutive letters)

$K = 1, 2, 3, 4, 5, \dots$

Window size vs. fixed position

Up-stream, downstream vs. anywhere in window

In-frame vs. any frame

Exercise

```
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, ...

Find the in-frame downstream ATG

Feature generation - Summary

Raw Data

[illegible]

An ATG segment – positive sample

```
> 206 +1_Index(56)
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCGTCAGAGCGCCGACACTCTTCTCTGTGCGAGCGAG
CCGCCGACCGCCAAGCAAAATGGGAATGAGGCAAGTTATCCTTTGGAAATGTGCTCACACTTTGATGCAGATGAAATTA
AAAGGCTAGGAAAGAGATTTAAGAAGCTCGATTGGACAAT
```



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


Too many features

K-grams (ie., k consecutive letters)

$K = 1, 2, 3, 4, 5, \dots$

Window size vs. fixed position

Up-stream, downstream vs. anywhere in window

In-frame vs. any frame

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

Why?

If we use $k = 1, 2, 3, 4, 5$, there would be $24 + 96 + 384 + 1536 + 6144 = 8184$ features!

This is too many for most machine learning methods

Most of these 8184 features are irrelevant

They confuses these machine learning methods

Feature selection: Principle

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?

Feature selection: t-statistic

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 .

Feature selection: χ^2

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 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?

Feature selection: CFS

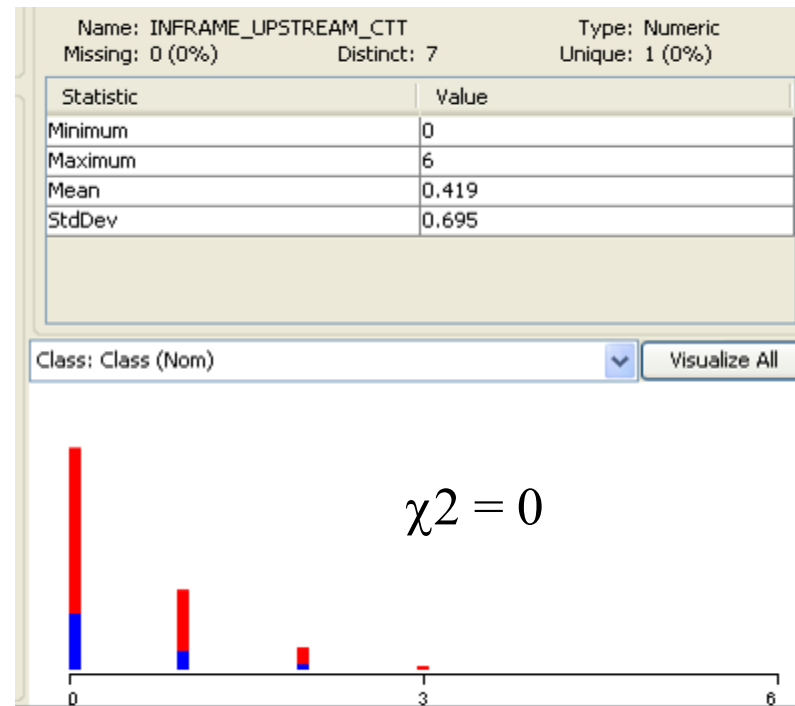
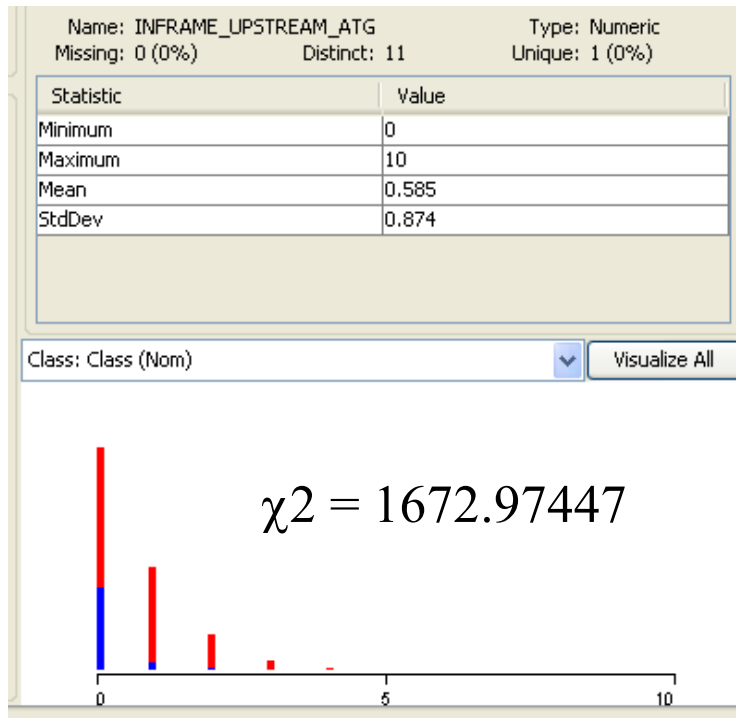
Instead of scoring individual signals, how about scoring a group of signals as a whole?

Correlation-based Feature Selection (CFS)

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?

Exercise: Distributions of two 3-grams



Which is the better one? Why?

Exercise

CFS selected these features for recognizing TIS:

Position -3

in-frame upstream ATG

in-frame downstream

TAA, TAG, TGA,

CTG, GAC, GAG, and GCC

Why would these features be important for recognizing TIS in mRNA?

Feature 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 rule”

Apply Naïve Bayes or SVM left-to-right until first ATG predicted as positive. That's the TIS; skip the rest

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 [Bioinform'00]

SVM + kernel engineering

No explicit features

Hatzigeorgiou [Bioinform'02]

Multiple neural networks

Scanning rule

No explicit features

Our approach

Explicit feature generation

Explicit feature selection

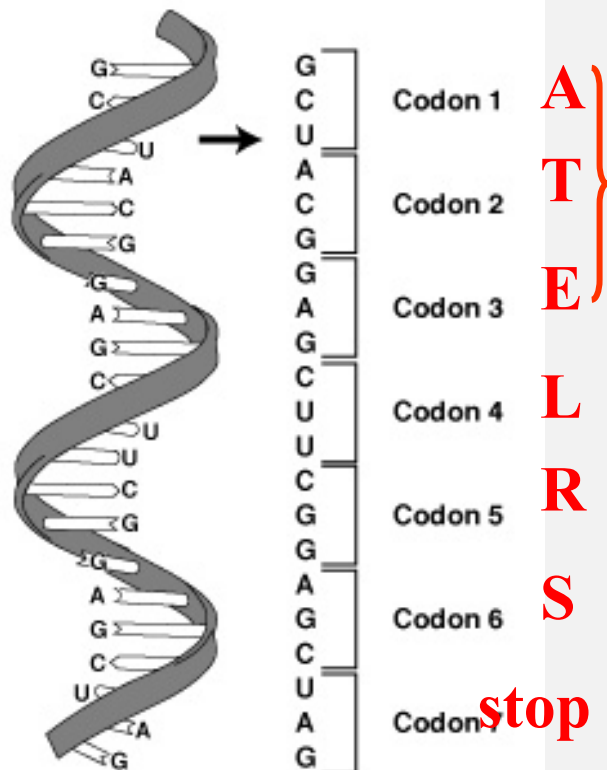
Use any machine learning method w/o any form of complicated tuning

Scanning rule is useful when predicting TIS for mRNA

Exercise

Should the scanning rule be used when predicting TIS on whole chromosome?

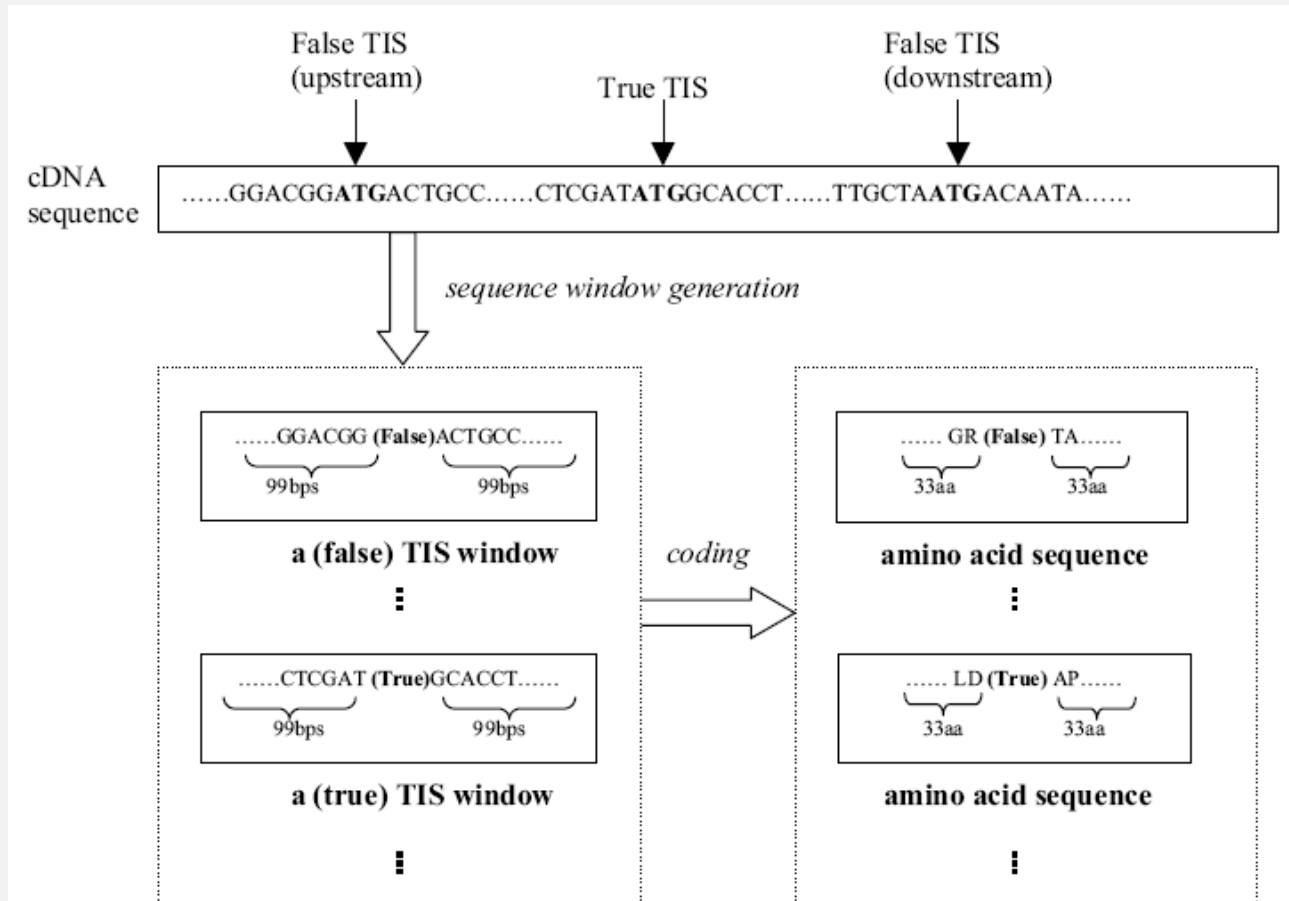
mRNA → protein



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

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

Sample k-grams selected by CFS for recognizing TIS:

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

From 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:

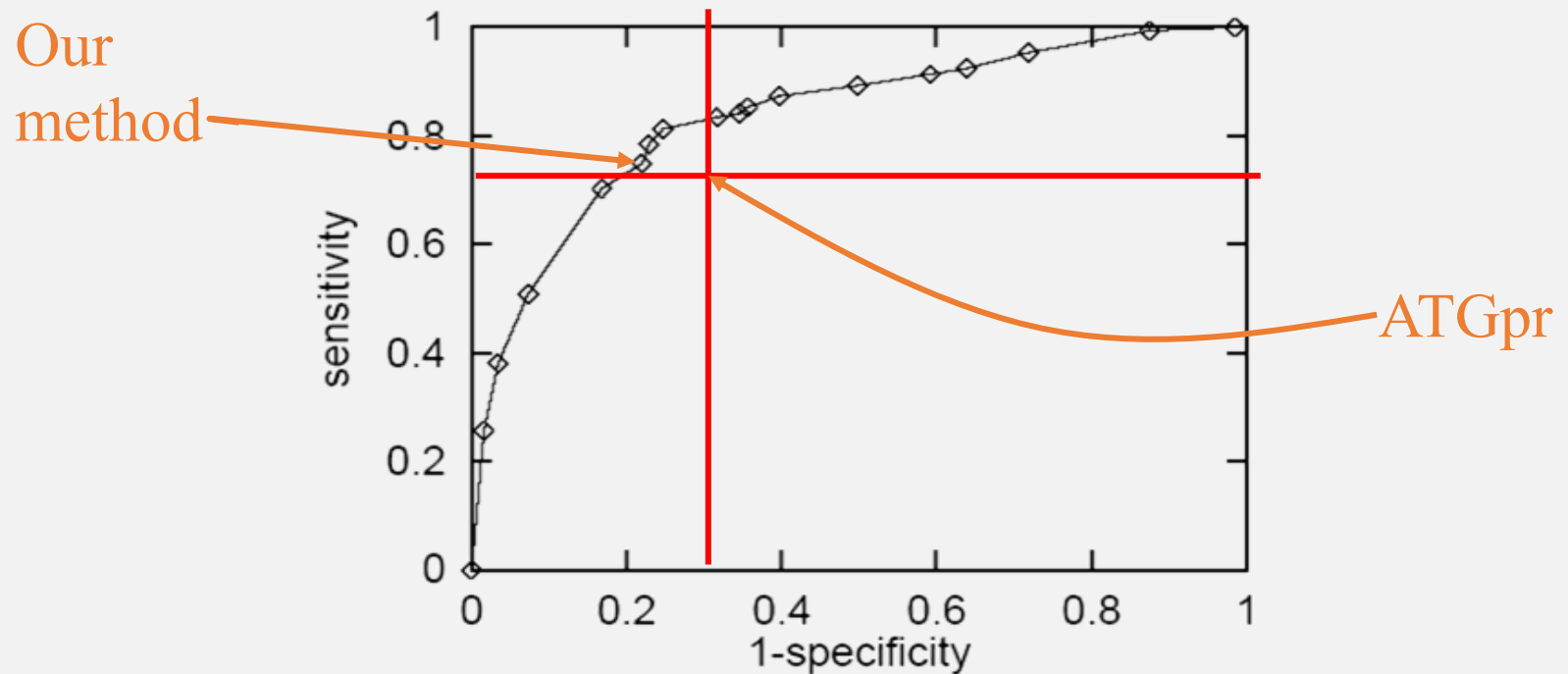
Well-characterized human gene sequences from chromosome X (565 TIS) and chromosome 21 (180 TIS)

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

Liu Huiqing

PhD, NUS, 2004

*Director of Translational Bioinformatics
at Daiichi Sankyo*

Asian Innovation Gold Award 2003

*New Jersey Cancer Research Award
for Scientific Excellence 2008*

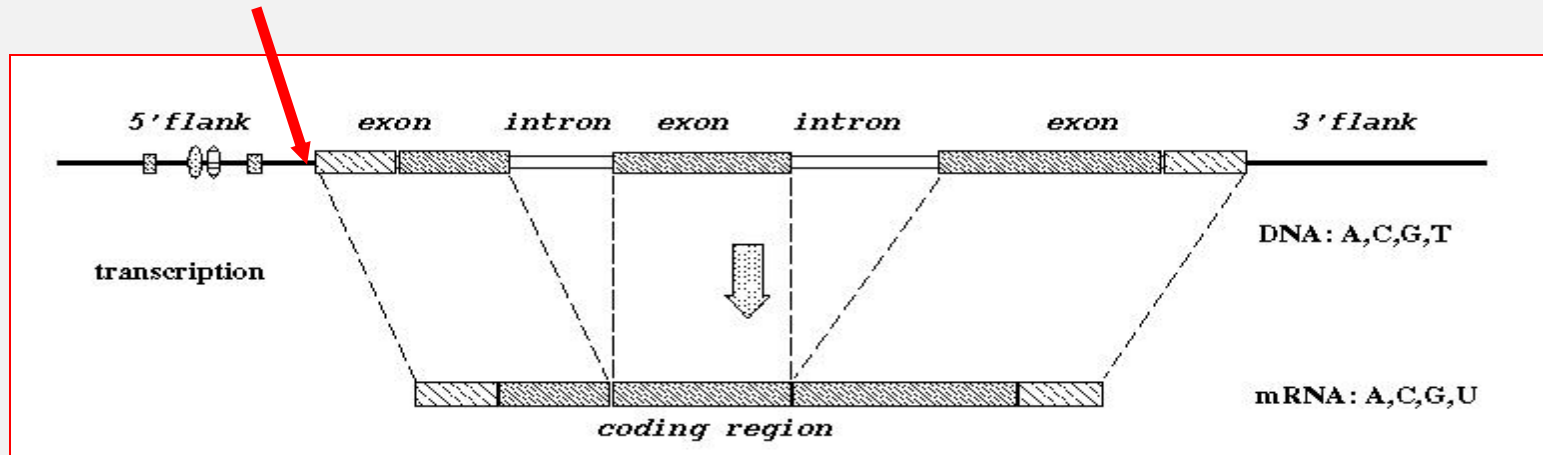
Gallo Prize 2008



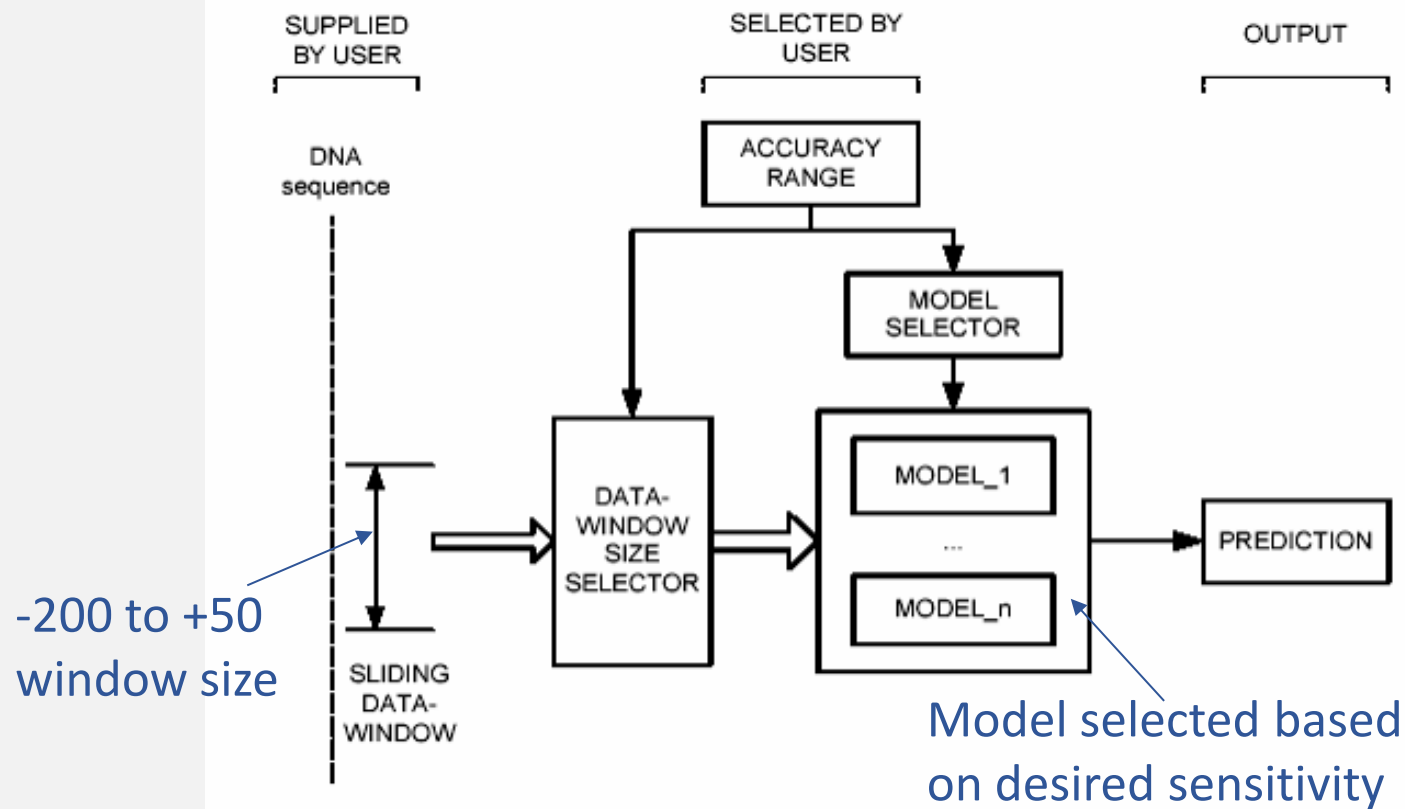
Recognition of Transcription Start Sites

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

Transcription start site



Structure of Dragon Promoter Finder

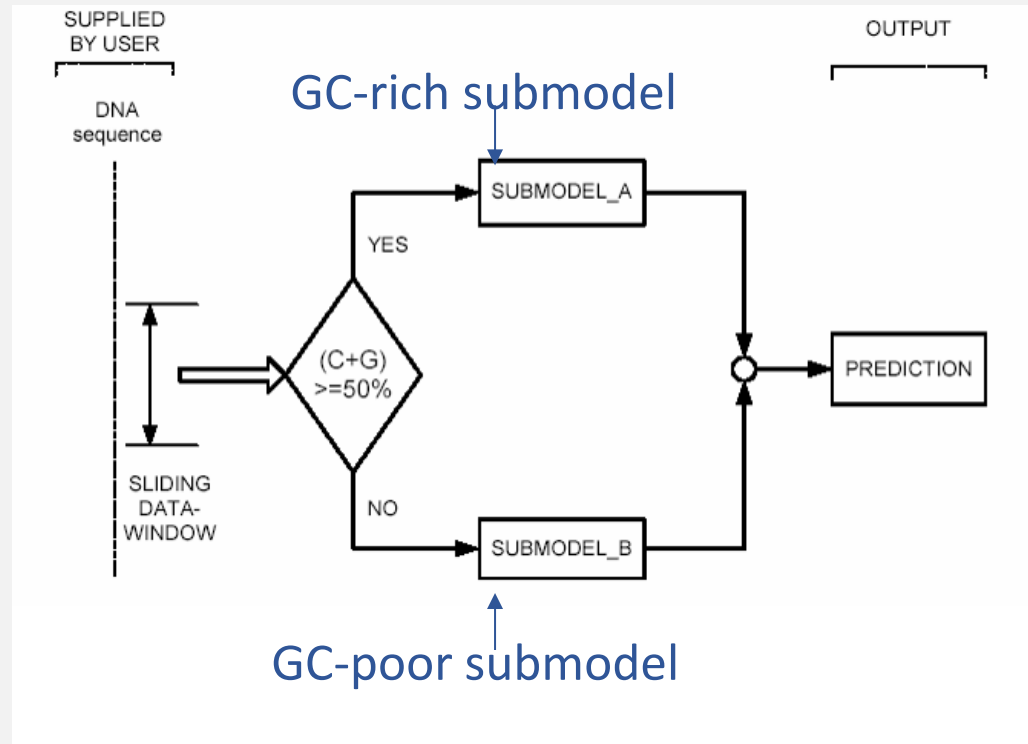


Each model has two submodels based on GC content

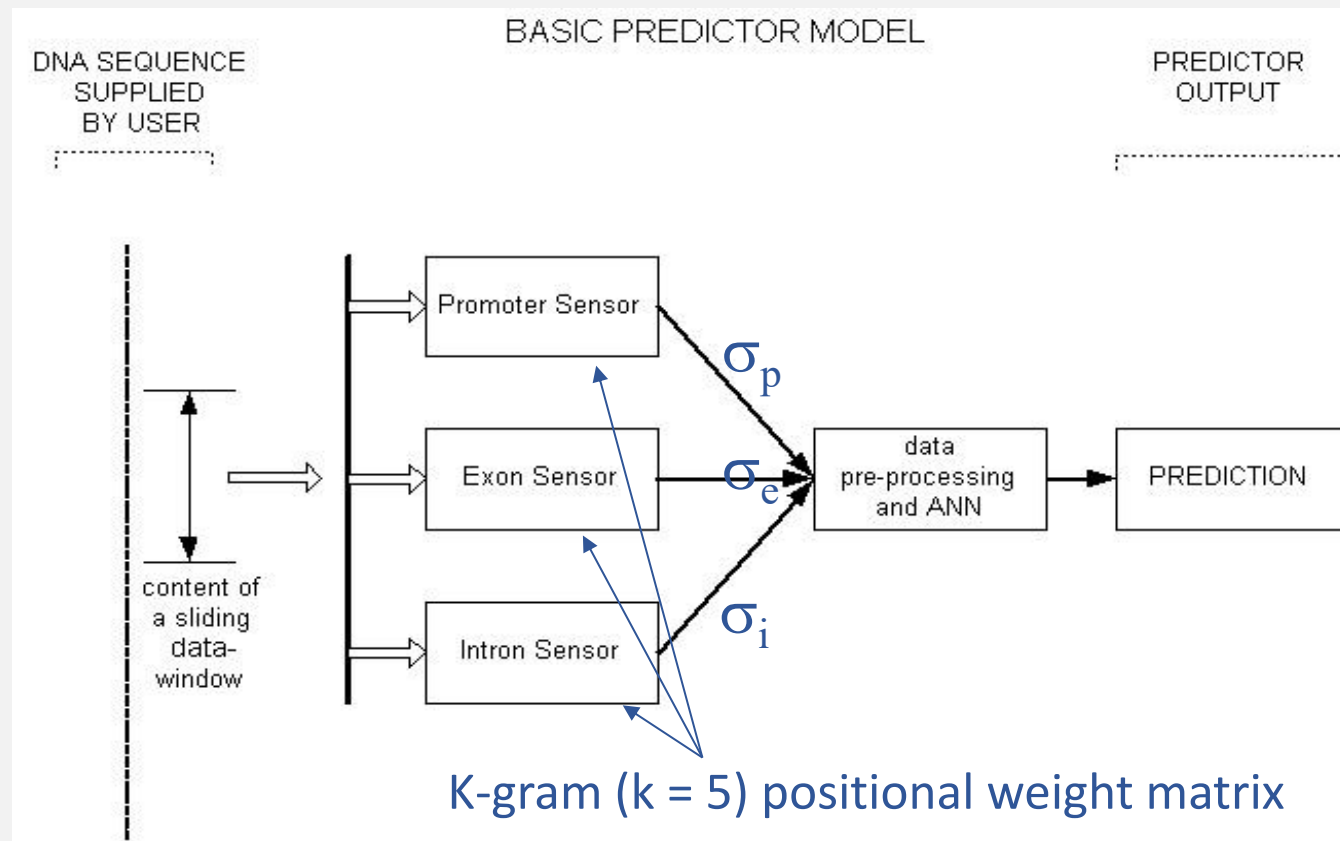
GC content

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

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

Window size $\rightarrow$

$$\sigma = \frac{\left(\sum_{i=1}^{L-4} p_j^i \otimes f_{j,i} \right)}{\left(\sum_{i=1}^{L-4} \max_j f_{j,i} \right)},$$

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

Pentamer at i^{th} position in input

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

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

Just making sure you know what I mean

3 DNA seq of length 10:

Seq₁ = ACCGAGTTCT

Seq₂ = AGTGTACCTG

Seq₃ = AGTTCGTATG

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							

Just making sure you know what I mean

3 DNA seq of length 10:

Seq₁ = ACCGAGTTCT

Seq₂ = AGTGTACCTG

Seq₃ = AGTTCGTATG

Exercise:

How many rows should this 2-mer 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			1/3			
...	...	...	...						
TT	0/3	0/3	1/3				1/3		

Feature generation & integration by ANN

Tuning parameters

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

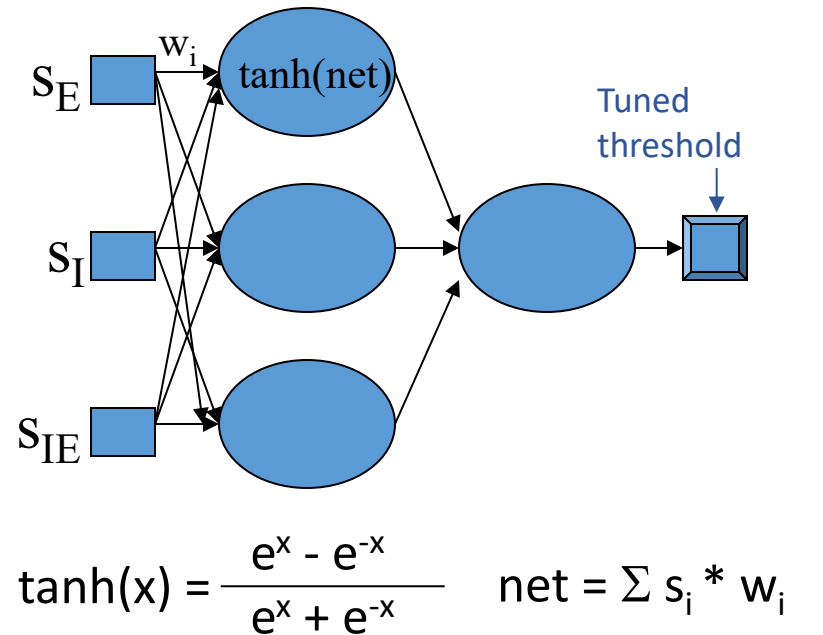
$$s_I = \text{sat}(\sigma_p - \sigma_i, a_i, b_i),$$

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

where the function sat is defined by

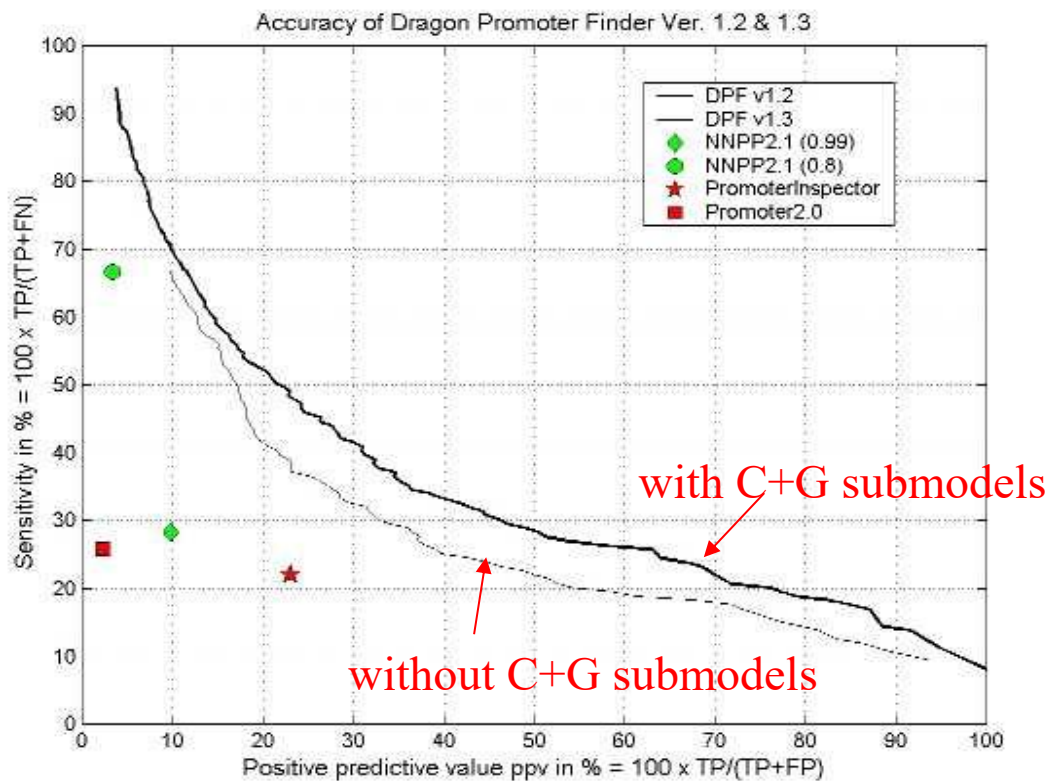
$$\text{sat}(x, a, b) = \begin{cases} a, & \text{if } x > a \\ x, & \text{if } b \leq x \leq a \\ b, & \text{if } b > x \end{cases}$$

Feature generation



Feature integration by ANN

Accuracy comparison



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 from EPD

793 vertebrate promoters

200 to +50 bp of TSS

non-TSS from GenBank

800 exons

4000 introns,

250 bp,

non-overlapping,

<50% identities

Tuning data preparation

To tune adjustable system parameters in Dragon, a separate tuning data set was needed

TSS from

*20 full-length gene seqs
with known TSS*

-200 to +50 bp of TSS

no overlap with EPD

Non-TSS from

1600 human 3'UTR seqs

500 human exons

500 human introns

250 bp

no overlap

Testing data criteria & preparation

Seqs should be from the training or evaluation of other systems (no bias!)

Seqs should be disjoint from training and tuning data sets

Seqs should have TSS

Seqs should be cleaned to remove redundancy, <50% identities

159 TSS from 147 human and human virus seqs

Cumulative length of more than 1.15Mbp

Taken from GENESCAN, Geneld, Genie, etc.

About the inventor: Vlad Bajic

Vladimir B. Bajic

Principal Scientist, I²R, 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

Eukaryotic pre-mRNA processing

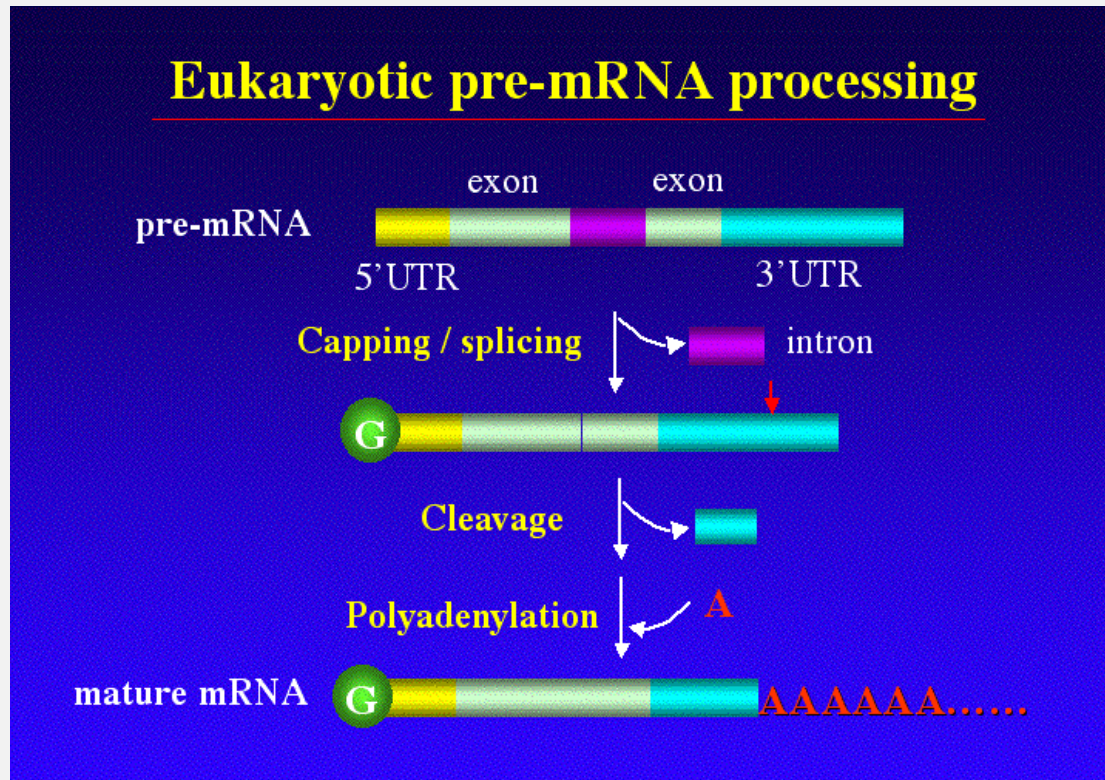


Image credit: www.polya.org

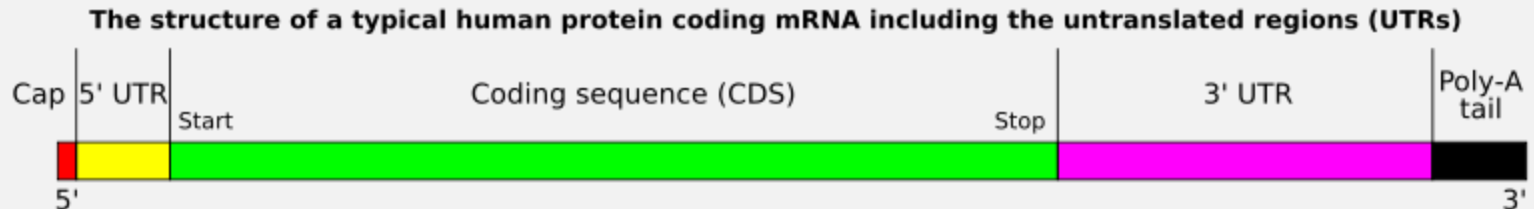
Polyadenylation in eukaryotes

Add 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 tail is short enough,
the mRNA is degraded*

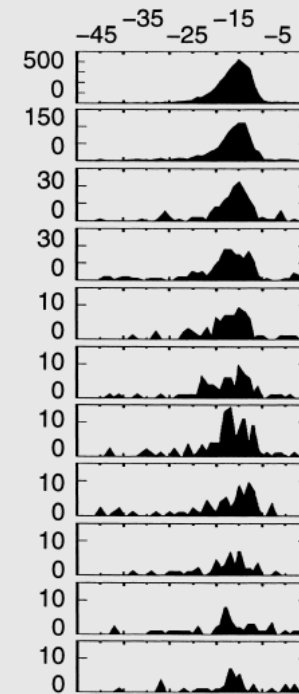


Source: Wikipedia

Poly-A signals in human

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

Hexamer	Observed (expected) ^a	% sites	p^b	Position average $\pm$ SD	Location ^c
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^{-57}	-16 ± 5.9	30
UAUAAA	180 (53)	3.2	4×10^{-45}	-18 ± 7.8	30
CAUAAA	76 (23)	1.3	1×10^{-18}	-17 ± 5.9	10
GAUAAA	72 (21)	1.3	2×10^{-18}	-18 ± 6.9	10
AAUAUA	96 (33)	1.7	2×10^{-19}	-18 ± 6.9	10
AAUACA	70 (16)	1.2	5×10^{-23}	-18 ± 8.7	10
AAUAGA	43 (14)	0.7	1×10^{-9}	-18 ± 6.3	10
AAAAAG	49 (11)	0.8	5×10^{-17}	-18 ± 8.9	10
ACUAAA	36 (11)	0.6	1×10^{-08}	-17 ± 8.1	10



Beaudoing et al., *Genome Research*, 10:1001-1010, 2000

Poly-A signals in Arabidopsis

In human, 58.2% of PAS is AAUAAA

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

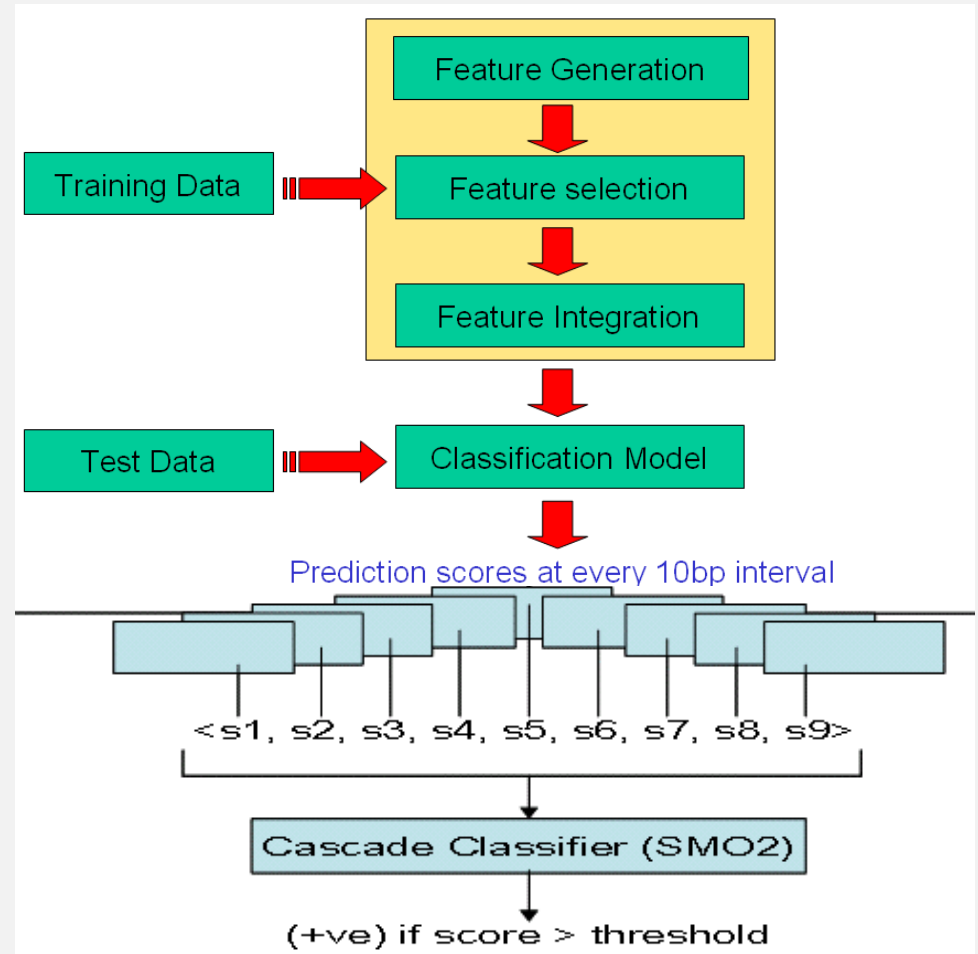
Hexamer	Observed (expected) ^a	% sites	<i>p</i> ^b	Position average $\pm$ SD	Location ^c
AAUAAA	3286 (317)	58.2	0	-16 ± 4.7	
AUUAAA	843 (112)	14.9	0	-17 ± 5.3	
AGUAAA	156 (32)	2.7	6×10^{-57}	-16 ± 5.9	
UAUAAA	180 (53)	3.2	4×10^{-45}	-18 ± 7.8	
CAUAAA	76 (23)	1.3	1×10^{-18}	-17 ± 5.9	
GAUAAA	72 (21)	1.3	2×10^{-18}	-18 ± 6.9	
AAUAUA	96 (33)	1.7	2×10^{-19}	-18 ± 6.9	
AAUACA	70 (16)	1.2	5×10^{-23}	-18 ± 8.7	
AAUAGA	43 (14)	0.7	1×10^{-9}	-18 ± 6.3	
AAAAAG	49 (11)	0.8	5×10^{-17}	-18 ± 8.9	
ACUAAA	36 (11)	0.6	1×10^{-08}	-17 ± 8.1	

Beaudoing et al.,
Genome Research,
10:1001-1010, 2000

In contrast PAS in Arabidopsis is highly degenerate

E.g., only 10% of Arabidopsis PAS is AAUAAA!

Cascade classifier approach on Arab PAS sites

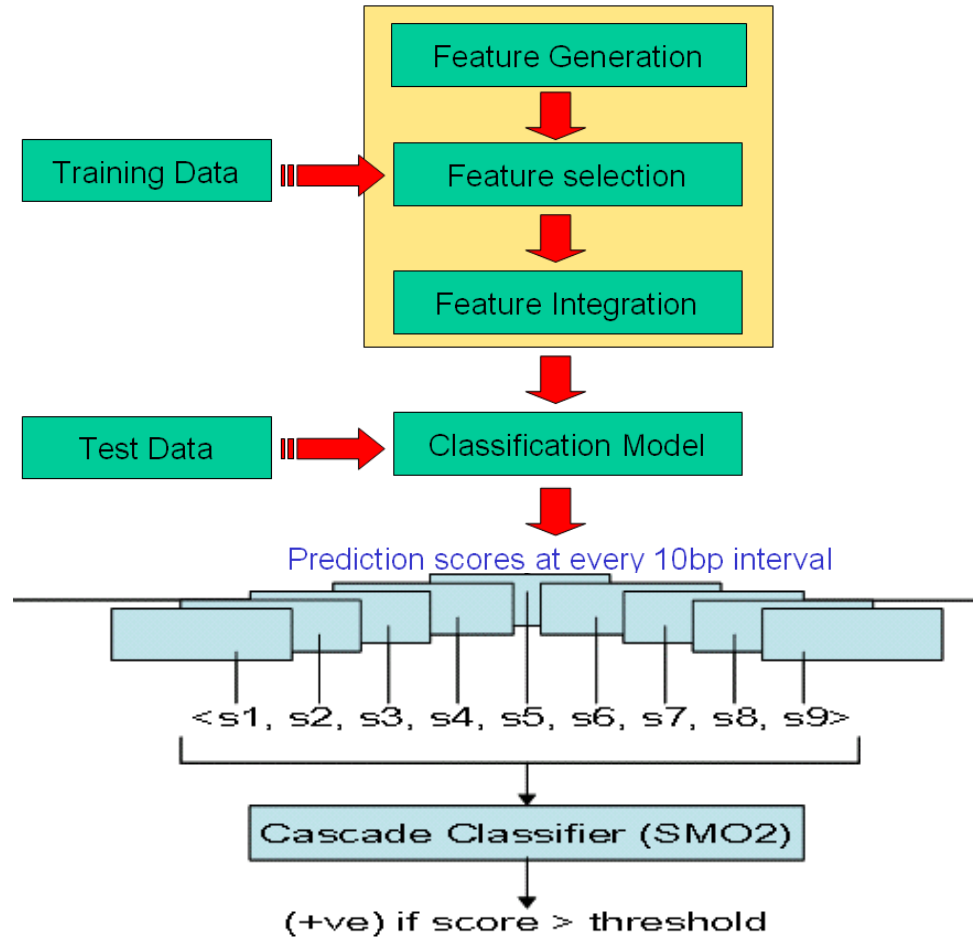


Exercise

Why is the cascade classifier used?

What do you think the prediction scores $\langle s_1, \dots, s_9 \rangle$ look like for PAS sites?

What do you think these scores look like for non-PAS sites?



Data collection

Dataset #1 from Hao Han, 811 +ve seq (-200/+200)

Dataset #2 from Hao Han, 9742 -ve seq (-200/+200)

Dataset #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, selection, & integration

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: χ^2

Feature integration & cascade: SVM

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

*Data Science Mgr at
Indeed Inc, Japan*

Retired in 2023 to relax!



Concluding remarks...

What we have learned

Gene feature recognition applications: TIS, TSS, PAS

General methodology: “Feature generation, feature selection, feature integration”

Important tactics

Multiple models to optimize overall performance

Feature transformation (DNA $\rightarrow$ amino acid)

Classifier cascades

Acknowledgements

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

Good to read for TIS recognition

Pedersen & Nielsen, “Neural network prediction of translation initiation sites in eukaryotes”, *ISMB* 5:226-233, 1997

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

Hatzigeorgiou, “Translation initiation start prediction in human cDNAs with high accuracy”, *Bioinformatics* 18:343-350, 2002

Li et al., “Techniques for Recognition of Translation Initiation Sites”, *The Practical Bioinformatician*, Chapter 4, pages 71-90, 2004

<https://www.comp.nus.edu.sg/~wongls/psZ/practical-bioinformatician/ch4-wlstis/ch4-wlstis.pdf>

Good to read for TSS recognition

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

Fickett & Hatzigeorgiou, “Eukaryotic promoter recognition”, *Gen. Res.* 7:861-878, 1997

Scherf et al., “Highly specific localisation of promoter regions in large genome sequences by PromoterInspector”, *JMB* 297:599-606, 2000

Bajic & Chong, “Tuning the Dragon Promoter Finder System for Human Promoter Recognition”, *The Practical Bioinformatician*, Chapter 7, pages 157-165, 2004 <https://www.comp.nus.edu.sg/~wongls/psZ/practical-bioinformatician/ch7-bajicdragon/ch7-bajicdragon.pdf>

Good to read for PAS recognition

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

Tabaska & Zhang, “Detection of polyadenylation signals in human DNA sequences”. *Gene*, 231:77-86, 1999

Legendre & Gautheret, “Sequence determinants in human polyadenylation site selection”. *BMC Genomics*, 4:7, 2003

Tian et al., “Prediction of mRNA polyadenylation sites by support vector machine”. *Bioinformatics*, 22:2320-2325, 2006

Koh & Wong. “Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences”. *Proc. GIW 2007*, pages 73-82