For written notes on this lecture, please read Chapters 4 and 7 of *The Practical Bioinformatician*

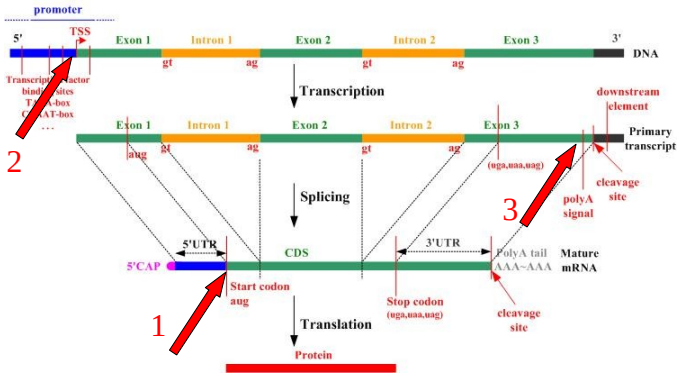
Accurate Recognition of Translation Initiation Sites, Transcription Start Sites, and Polyadenylation Signals in Genomic Sequences

Limsoon Wong

25 May 2007

Based on works of/with Huiqing Liu (TIS), Vlad Bajic (TSS), & Chuan Hock Koh (PAS)

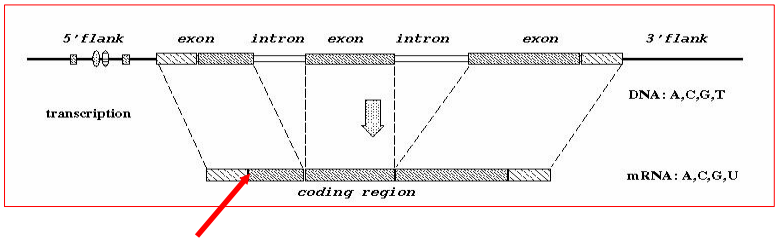
Plan



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
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA     160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA     240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
.............................................................      80
.............................iEEEEEEEEEEEEEEEEEEEEEEEEEE     160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE     240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

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

Approach

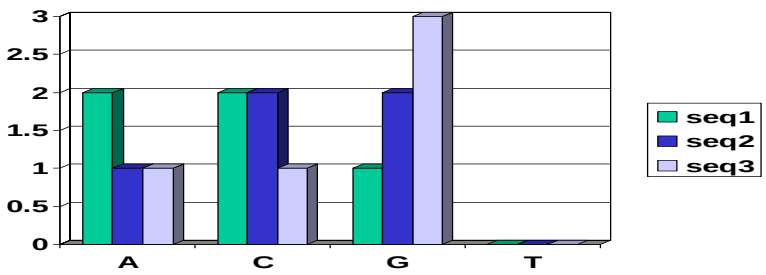
- **Training data gathering**
- **Signal generation**
  - k-grams, distance, domain know-how, ...
- **Signal selection**
  - Entropy, $\chi 2$, CFS, t-test, domain know-how...
- **Signal integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...

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, …
  - 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
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA     160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA     240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
```

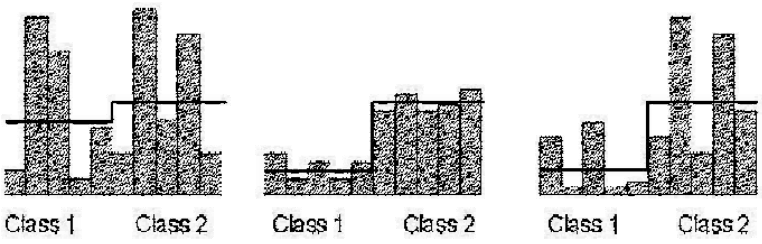
- **Window = ±100 bases**
- **In-frame, downstream**
  - GCT = 1, TTT = 1, ATG = 1…
- **Any-frame, downstream**
  - GCT = 3, TTT = 2, ATG = 2…
- **In-frame, upstream**
  - GCT = 2, TTT = 0, ATG = 0, ...

Too Many Signals

- **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!**
- **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 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 (e.g., $\chi 2$)

The χ^2 value of a signal is defined as:

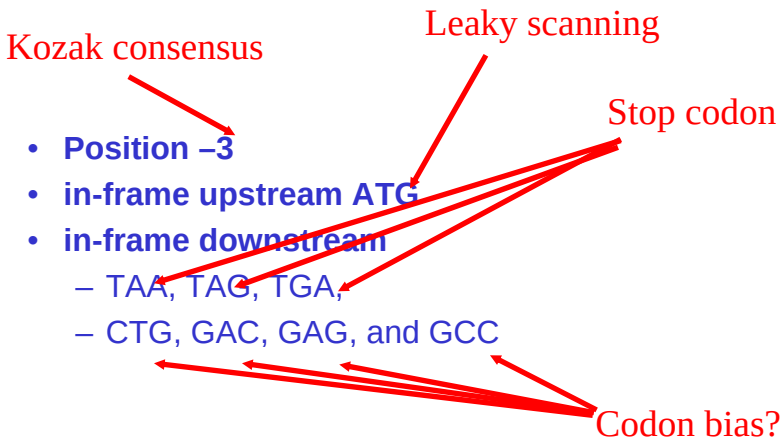
$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

where m is the number of intervals, k the number of classes, A_{ij} the number of samples in the 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} ($E_{ij} = R_i * C_j / N$).

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

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

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

	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 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 Comparisons

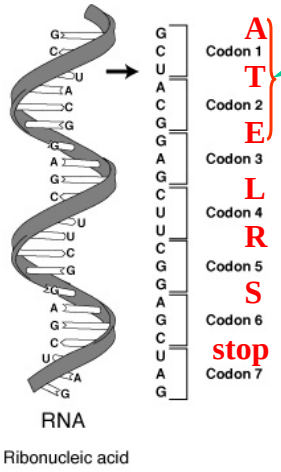
	TP/(TP + FN)	TN/(TN + FP)	TP/(TP + FP)	Accuracy
NB	84.3%	86.1%	66.3%	85.7%
Decision Tree	74.0%	94.4%	81.1%	89.4%
NB+NN+Tree	77.6%	94.5%	82.1%	90.4%
SVM+Scanning	88.5%	96.3%	88.6%	94.4%*
Pedersen&Nielsen	78%	87%	-	85%
Zien	69.9%	94.1%	-	88.1%
Hatzigeorgiou	-	-	-	94%*

* 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



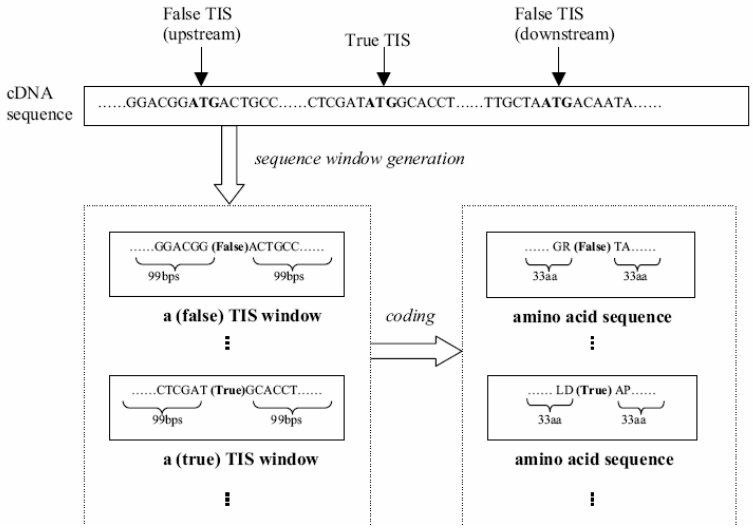
G C U A C G G A G C U U C G G A G C U A G

Codon 1 **A**
Codon 2 **T**
Codon 3 **E**
Codon 4 **L**
Codon 5 **R**
Codon 6 **S**
Codon 7 **stop**

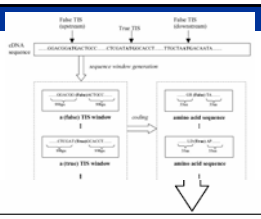
How about using k-grams from the translation?

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

Amino-Acid Features

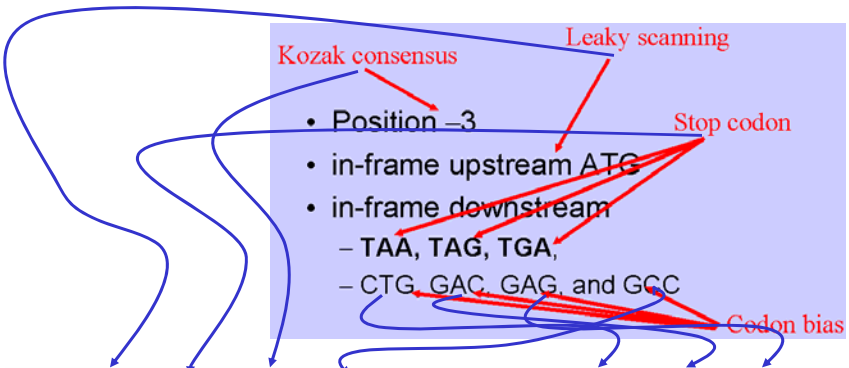


Amino-Acid Features



New feature space (total of 927 features + class label)			
42 1-gram amino acid patterns	882 2-gram amino acid patterns	3 bio-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, 5, 7, 9, 0, … ⋮	6, 2, 7, 0, 5, … ⋮ 2, 0, 3, 10, 0, … ⋮	N, N, N, ⋮ Y, Y, Y, ⋮	False ⋮ True ⋮

Amino Acid K-grams Discovered (by entropy)



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

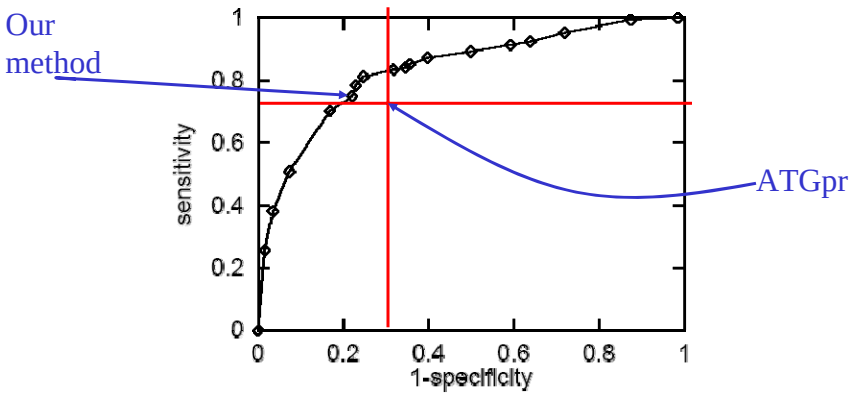
- **A. Hatzigeorgiou:**
  - 480 fully sequenced human cDNAs
  - 188 left after eliminating sequences similar to training set (Pedersen & Nielsen's)
  - 3.42% of ATGs are TIS
- **Our own:**
  - well characterized human gene sequences from chromosome X (565 TIS) and chromosome 21 (180 TIS)

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)



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

References (TIS Recognition)

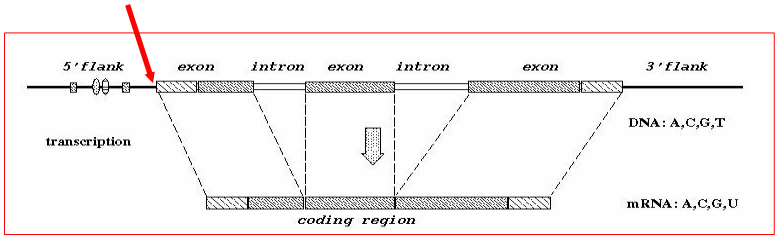
- A. G. Pedersen, H. Nielsen, "Neural network prediction of translation initiation sites in eukaryotes", *ISMB* 5:226--233, 1997
- A. Zien et al., "Engineering support vector machine kernels that recognize translation initiation sites", *Bioinformatics* 16:799--807, 2000
- A. G. Hatzigeorgiou, "Translation initiation start prediction in human cDNAs with high accuracy", *Bioinformatics* 18:343--350, 2002
- **Huiqing Liu, Limsoon Wong.** Data Mining Tools for Biological Sequences**. *Journal of Bioinformatics & Computational Biology*, 1(1):139--168, April 2003**

Recognition of Transcription Start Sites

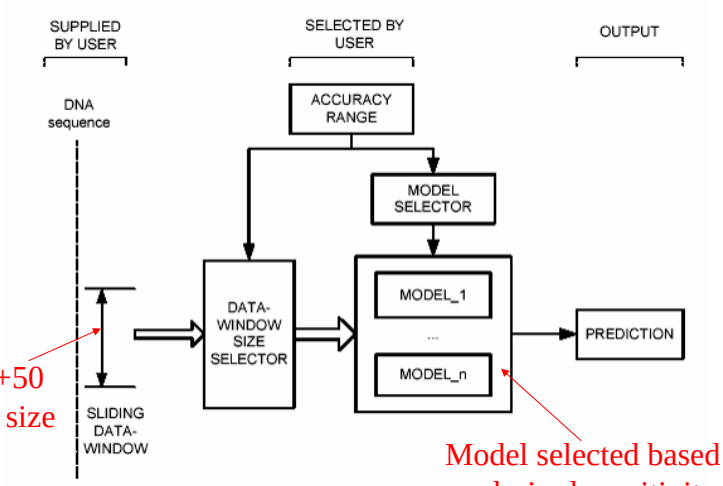
An introduction to the World's best TSS recognition system:

A heavy tuning approach

Transcription Start Site



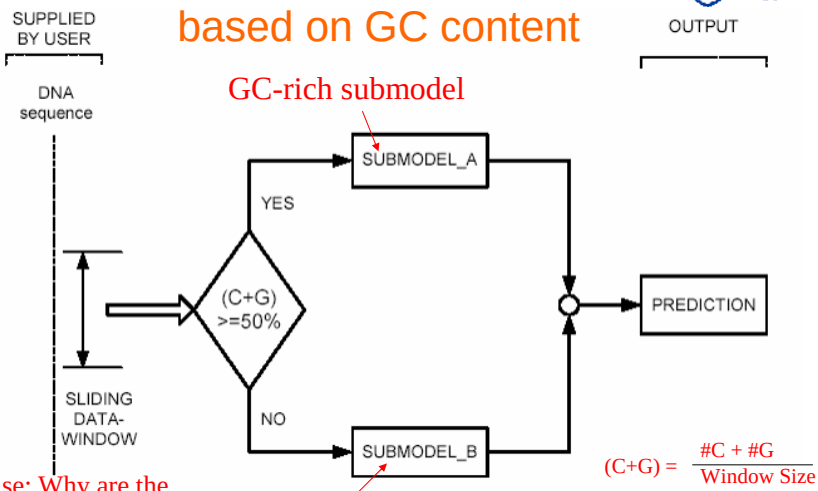
Structure of Dragon Promoter Finder



-200 to +50 window size

Model selected based on desired sensitivity

Each model has two submodels based on GC content

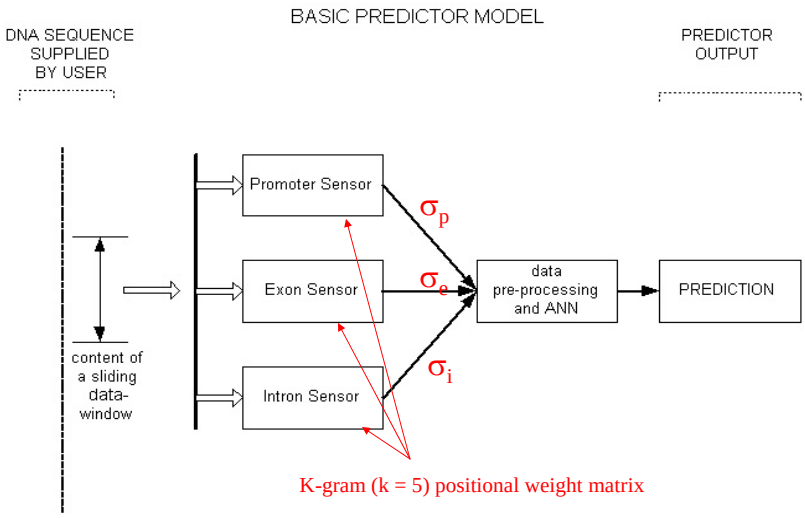


$(C+G) = \frac{\#C + \#G}{\text{Window Size}}$

Exercise: Why are the submodels based on GC content?

GC-poor submodel

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

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

Pentamer at i^{th} position in input

Frequency of jth pentamer at ith position in training window

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

Just to make sure you know what I mean …

- **Give me 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		Exercise: Fill in the rest of the table					
G	0/3	2/3	0/3							
T	0/3	0/3	2/3							

Just to make sure you know what I mean …

- **Give me 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		Exercise: Fill in the rest of the table				
…	…	…	…						
TT	0/3	0/3	1/3				1/3		

Data Preprocessing & ANN

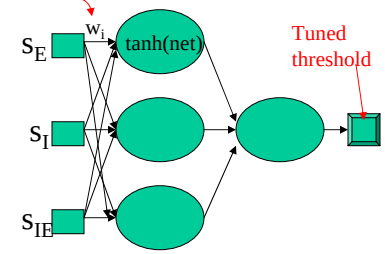
Tuning parameters

$$s_E = sat(\sigma_p - \sigma_e, a_e, b_e),$$
$$s_I = sat(\sigma_p - \sigma_i, a_i, b_i),$$
$$s_{EI} = sat(\sigma_e - \sigma_i, a_{ei}, b_{ei}),$$

where the function sat is defined by

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

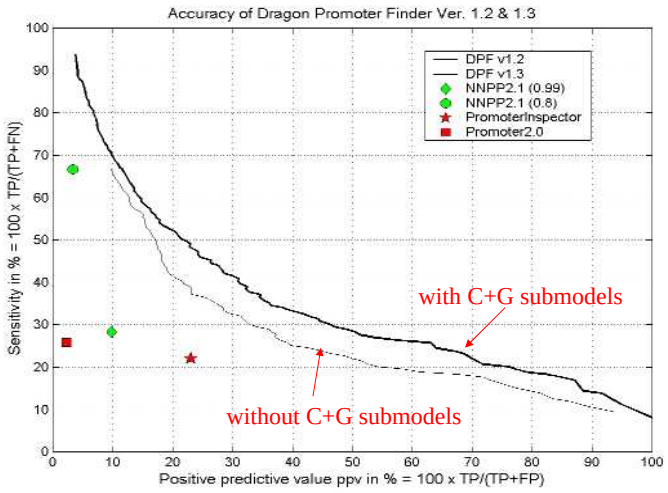
Simple feedforward ANN trained by the Bayesian regularisation method



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

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

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, GeneId, Genie, etc.**

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
- A.G.Pedersen et al., "The biology of eukaryotic promoter prediction---a review", *Computer & Chemistry* 23:191--207, 1999
- 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**

Recognition of Poly-A Signal Sites

Recap & twists to the "feature generation, feature selection, feature integration" approach

Eukaryotic Pre-mRNA Processing

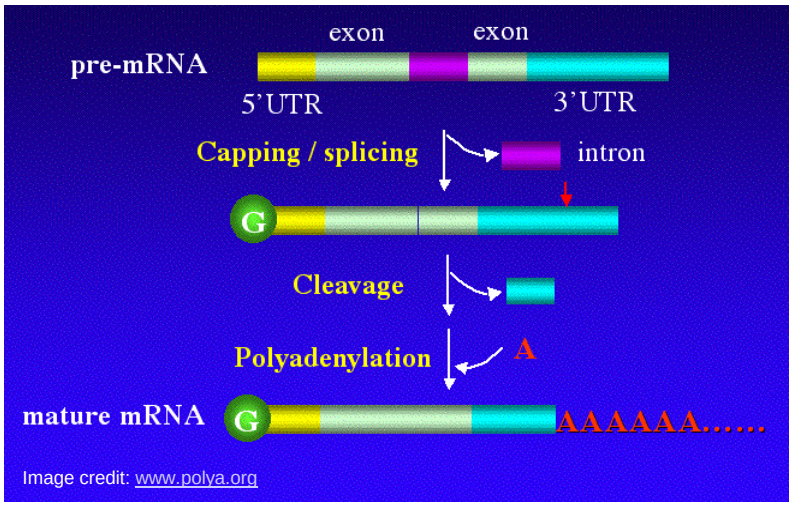
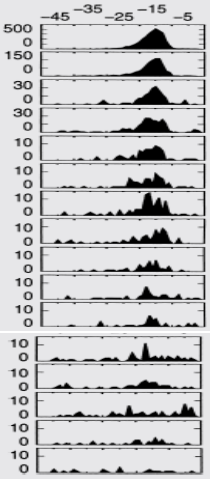


Image credit: www.polya.org

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

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

Hexamer	Observed (expected)[a]	% sites	P[b]	Position average ± 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^{-37}	−16 ± 5.9	
UAUAAA	180 (53)	3.2	4×10^{-45}	−18 ± 7.8	
CAUAAA	76 (23)	1.3	1×10^{-16}	−17 ± 5.9	
GAUAAA	72 (21)	1.3	2×10^{-16}	−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^{-06}	−17 ± 8.1	
AAGAAA	62 (10)	1.1	9×10^{-26}	−19 ± 11	
AAUGAA	49 (10)	0.8	4×10^{-16}	−20 ± 10	
UUUAAA	69 (20)	1.2	3×10^{-16}	−17 ± 12	
AAAACA	29 (5)	0.5	8×10^{-12}	−20 ± 10	
GGGGCU	22 (3)	0.3	9×10^{-12}	−24 ± 13	



Approach on Human PAS Sites

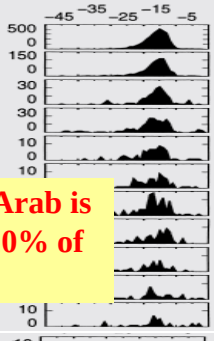
- **Training data collection**
  - 2327 +ve seq (Legendre & Gautheret, 2003)
  - 2300 -ve seq (Liu et al., 2003)
    - **713 CDS seq**
    - **904 intronic seq of first intron**
    - **861 randomized UTR seq of same mono nucleotide composition as human 3' UTRs**
- **Feature generation**
  - 3-grams, compositional features (5U/1N, G/U*7, etc)
  - Freq of features above in 3 diff windows (-60/-1), (-40/+55), (+1/+60) around a candidate AAUAA site
- **Feature selection**
  - χ^2
- **Feature integration**
  - C4.5, SVM

10-CV Results

		C4.5		SVM	
Chi Squared		**Sensitivity (%)**	**Precision (%)**	**Sensitivity (%)**	**Precision (%)**
Liu et al.	+ve	71.1	71.9	82.5	80.0
	-ve	73.8	73.1	80.6	83.1
This talk	+ve	71.8 (+0.7)	72.1 (+0.2)	83.6 (+1.1)	82.4 (+2.4)
	-ve	74.0 (+0.2)	73.6 (+0.5)	83.2 (+2.6)	84.4 (+1.3)
CFS					
Liu et al.	+ve	73.1	72.7	82.9	79.7
	-ve	74.2	74.6	80.2	83.3
This talk	+ve	72.6 (-0.5)	74.0 (+1.3)	81.9 (-1.0)	80.7 (+1.0)
	-ve	76.1 (+1.9)	74.7 (-0.1)	81.6 (+1.4)	82.7 (-0.5)

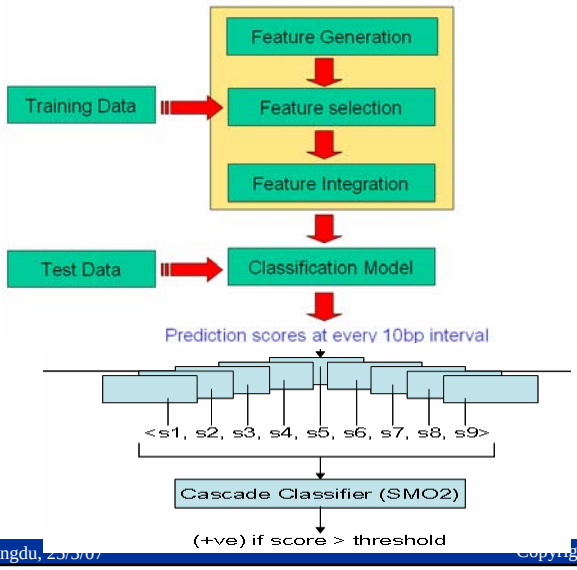
Poly-A Signals in Arabidopsis

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

Hexamer	Observed (expected)[a]	% sites	p[b]	Position average ± SD	Location[c]
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^{-37}	−16 ± 5.9	
UAUAAA	180 (53)	3.2	4×10^{-45}	−18 ± 7.8	
CAUAAA	76 (23)	1.3	1×10^{-16}	−17 ± 5.9	
GAUAAA	72				
AAUAUA	96				
AAUACA	70				
AAUAGA	43				
AAAAAG	49				
ACUAAA	36 (11)	0.6	1×10^{-06}	−17 ± 8.1	
AAGAAA	62 (10)	1.1	9×10^{-26}	−19 ± 11	
AAUGAA	49 (10)	0.8	4×10^{-16}	−20 ± 10	
UUUAAA	69 (20)	1.2	3×10^{-16}	−17 ± 12	
AAAACA	29 (5)	0.5	8×10^{-12}	−20 ± 10	
GGGGCU	22 (3)	0.3	9×10^{-12}	−24 ± 13	

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

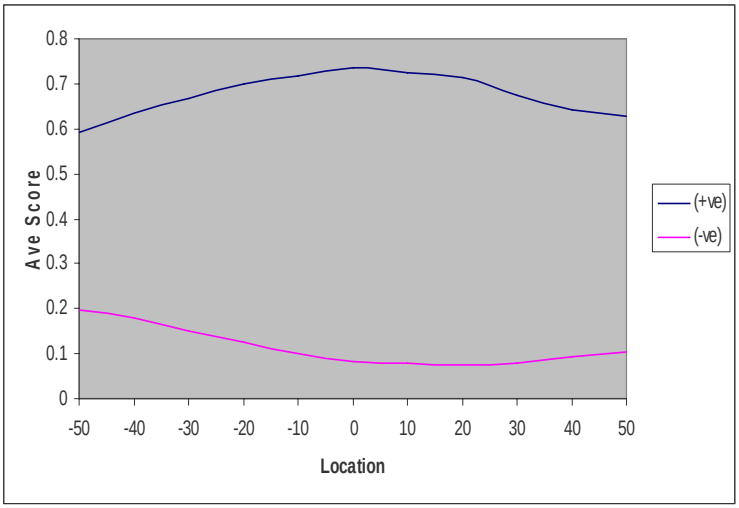
Approach on Arab PAS Sites (I)



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	
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

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
- **Huiqing Liu, Hao Han, Jinyan Li, Limsoon Wong, "An In-Silico Method for Prediction of Polyadenylation Signals in Human Sequences". *Genome Informatics*, 14:84-93, 2003**
- B. Tian et al., "Prediction of mRNA polyadenylation sites by support vector machine". *Bioinformatics*, 22:2320-2325, 2006
- D. Gautheret et al., "Patterns of variant polyadenylation signal usage in human genes". *Genome Res*., 10:1001-1010, 2000

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