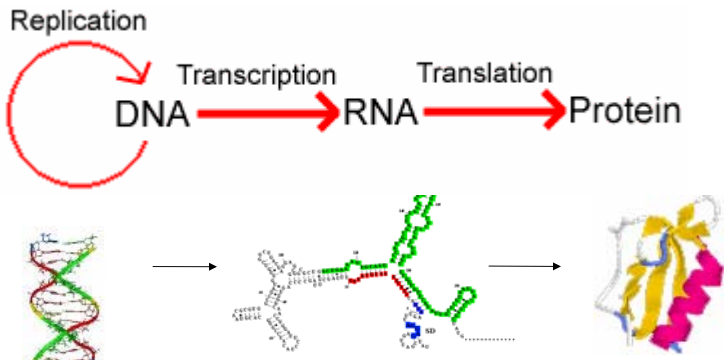
For written notes on this lecture, please read Chapters 4 and 7 of *The Practical Bioinformatician*

Knowledge Discovery Techniques for Bioinformatics, Part III: Applications to Gene Feature Recognition

Limsoon Wong

Central Dogma



...AATGGTACCGATGACCTG...

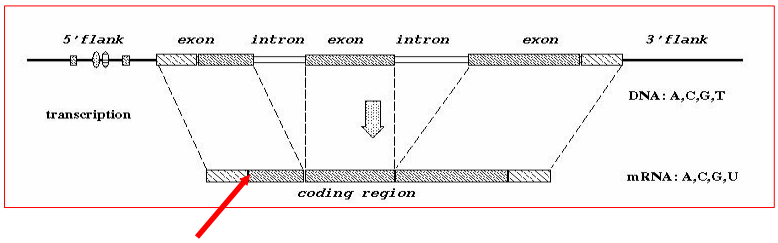
...AAUGGUACCGAUGACCUGGAGC...

...TRLRPLLALLALWP...

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
................................iEEEEEEEEEEEEEEEEEEEEEEEEEEE      160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE      240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

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

Approach

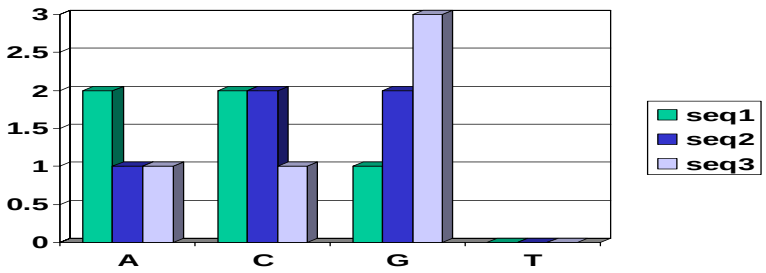
- **Training data gathering**
- **Signal generation**
  - – k-grams, distance, domain know-how, ...
- **Signal selection**
  - – Entropy, $\chi 2$, CFS, t-test, domain know-how...
- **Signal integration**
  - – SVM, ANN, PCL, CART, C4.5, kNN, ...

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, ...

Exercise: Find the in-frame downstream ATG

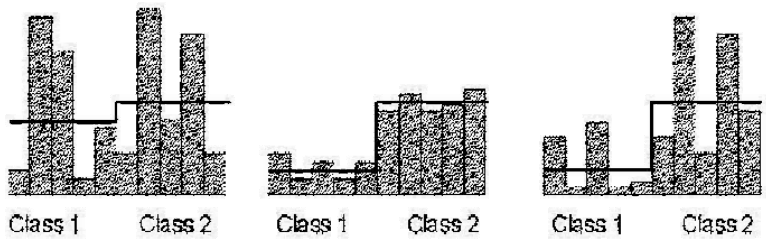
Exercise: What are the possible k-grams (k=3) in this sequence?

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!**
- **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 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., MIT-correlation)

The MIT-correlation value of a signal is defined as

$$MIT = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2}$$

where σ_i is the standard deviation of that signal in class i and μ_i is the mean of that signal in 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

Exercise: What is the main challenge in implementing CFS?

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 Comparisons

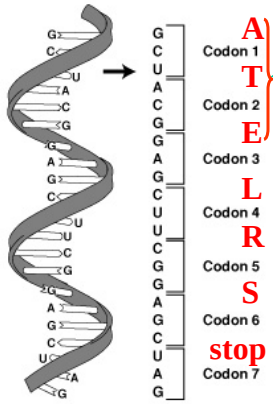
	TP/(TP + FN)	TN/(TN + FP)	TP/(TP + FP)	Accuracy
NB	84.3%	86.1%	66.3%	85.7%
Decision Tree	74.0%	94.4%	81.1%	89.4%
NB+NN+Tree	77.6%	94.5%	82.1%	90.4%
SVM+Scanning	88.5%	96.3%	88.6%	94.4%*
Pedersen&Nielsen	78%	87%	-	85%
Zien	69.9%	94.1%	-	88.1%
Hatzigeorgiou	-	-	-	94%*

* 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

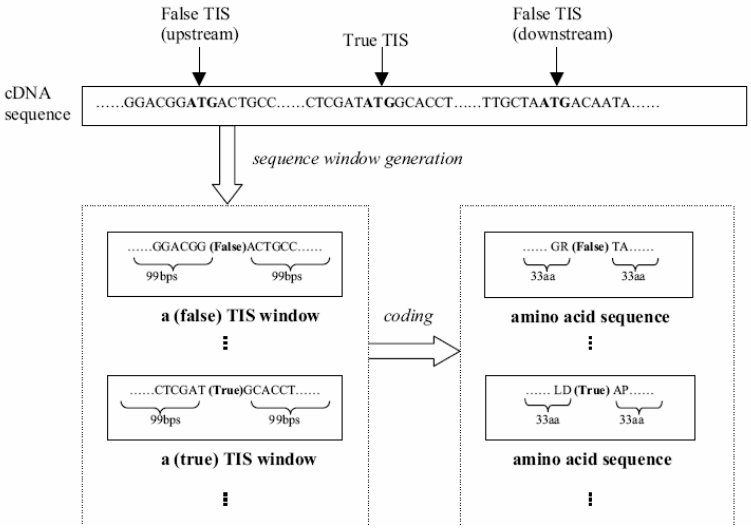


How about using k-grams from the translation?

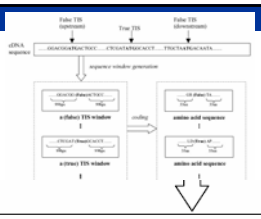
First	U	C	A	G	Last
U	Phe F	Ser S	Tyr Y	Cys C	U
	Phe	Ser	Tyr	Cys	C
	Leu L	Ser	**Stop** (Ochre)	**Stop** (Umber)	A
	Leu	Ser	**Stop** (Amber)	Trp W	G
C	Leu	Pro P	His H	Arg R	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln Q	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile I	Thr T	Asn N	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys K	Arg	A
	Met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

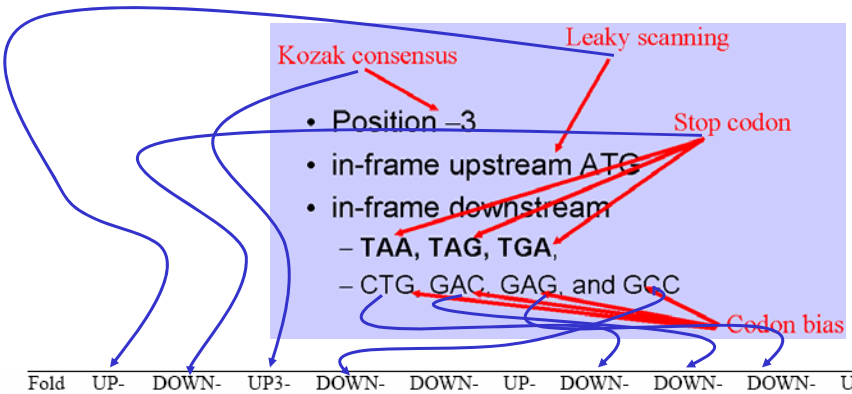


Amino-Acid Features



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

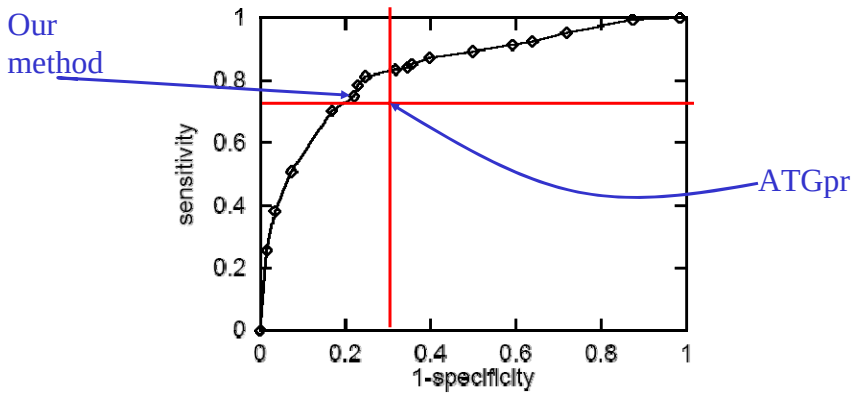
- **A. Hatzigeorgiou:**
  - 480 fully sequenced human cDNAs
  - 188 left after eliminating sequences similar to training set (Pedersen & Nielsen's)
  - 3.42% of ATGs are TIS
- **Our own:**
  - well characterized human gene sequences from chromosome X (565 TIS) and chromosome 21 (180 TIS)

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)



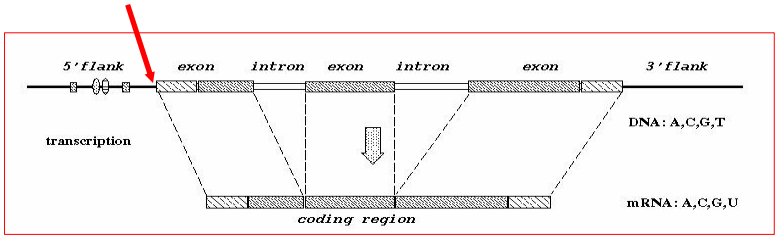
- **Using top 100 features selected by entropy and trained on Pedersen & Nielsen's**

Recognition of Transcription Start Sites

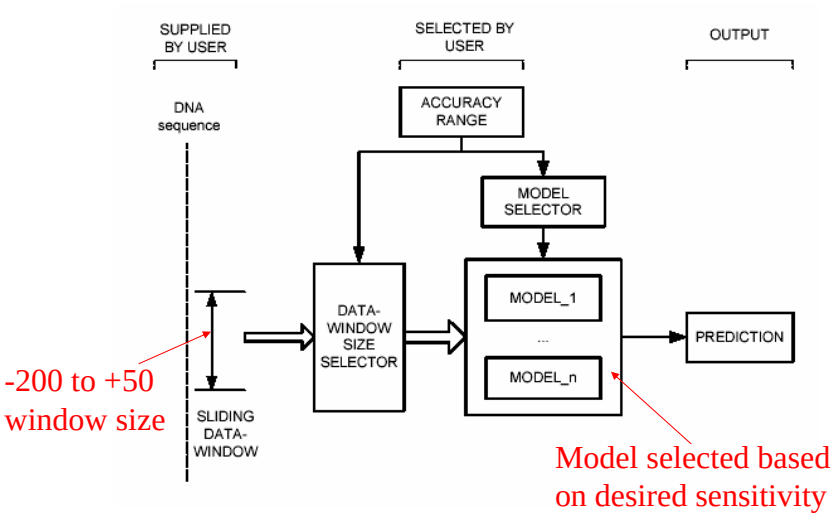
An introduction to the World's best TSS recognition system:

A heavy tuning approach

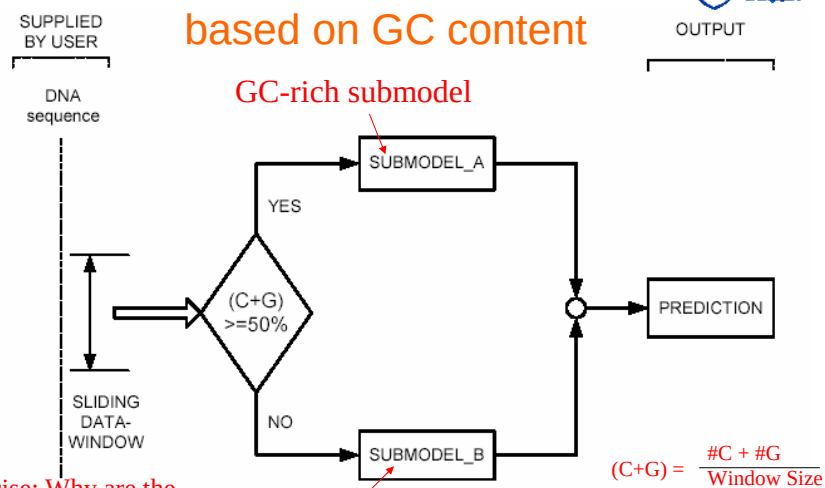
Transcription Start Site



Structure of Dragon Promoter Finder



Each model has two submodels based on GC content



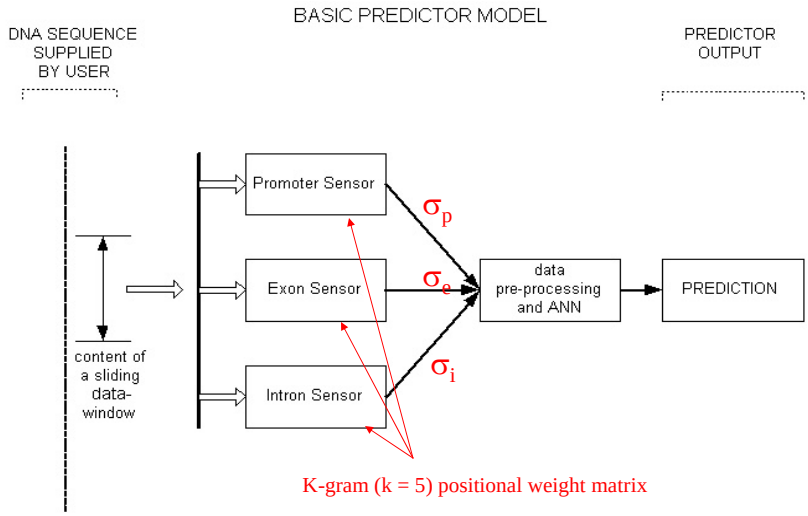
GC-rich submodel

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

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

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							
G	0/3	2/3	0/3							
T	0/3	0/3	2/3							

Exercise: Fill in the rest of the table

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						
…	…	…	…						
TT	0/3	0/3	1/3				1/3		

Exercise: Fill in the rest of the table

Data Preprocessing & ANN

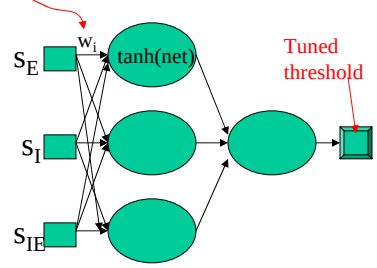
Tuning parameters

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

where the function sat is defined by

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

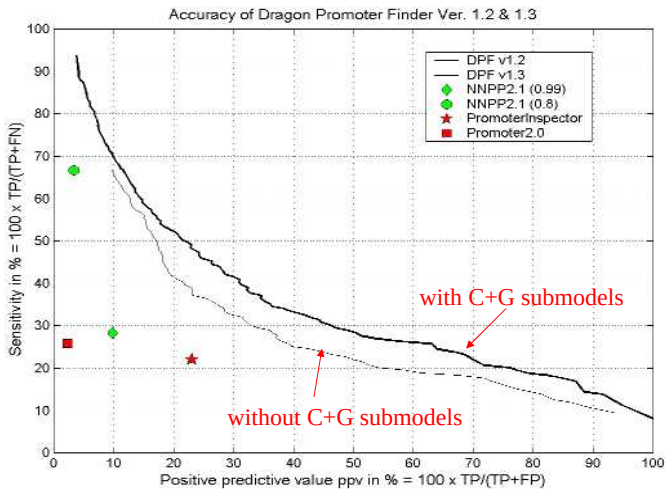
Simple feedforward ANN trained by the Bayesian regularisation method



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

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

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.**

Any Question?

References (TIS Recognition)

- A. G. Pedersen, H. Nielsen, “Neural network prediction of translation initiation sites in eukaryotes”, *ISMB* 5:226--233, 1997
- L. Wong et al., “Using feature generation and feature selection for accurate prediction of translation initiation sites”, *GIW* 13:192--200, 2002
- 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
- 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

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

References (Misc.)

- C. P. Joshi et al., “Context sequences of translation initiation codon in plants”, *PMB* 35:993--1001, 1997
- D. J. States, W. Gish, “Combined use of sequence similarity and codon bias for coding region identification”, *JCB* 1:39--50, 1994
- G. D. Stormo et al., “Use of Perceptron algorithm to distinguish translational initiation sites in E. coli”, *NAR* 10:2997--3011, 1982
- J. E. Tabaska, M. Q. Zhang, “Detection of polyadenylation signals in human DNA sequences”, *Gene* 231:77--86, 1999