An Introduction to Knowledge Discovery Applications in Biomedical Sciences

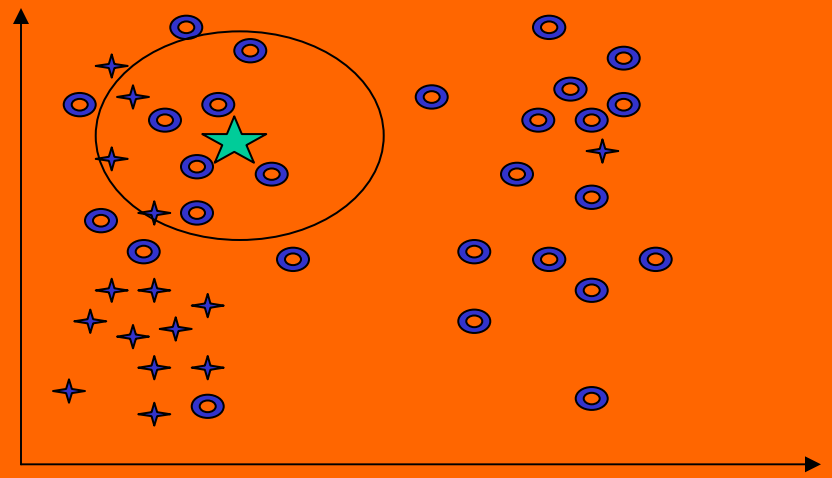
Limsoon Wong

Invited talk at Taipei Medical University, June 2006

Plan

- **Quick introduction to knowledge discovery**
- **Example applications**
  - Translation Initiation Site Recognition
  - Protein subcellular localization prediction
  - Protein function inference
  - Treatment optimization of childhood ALL

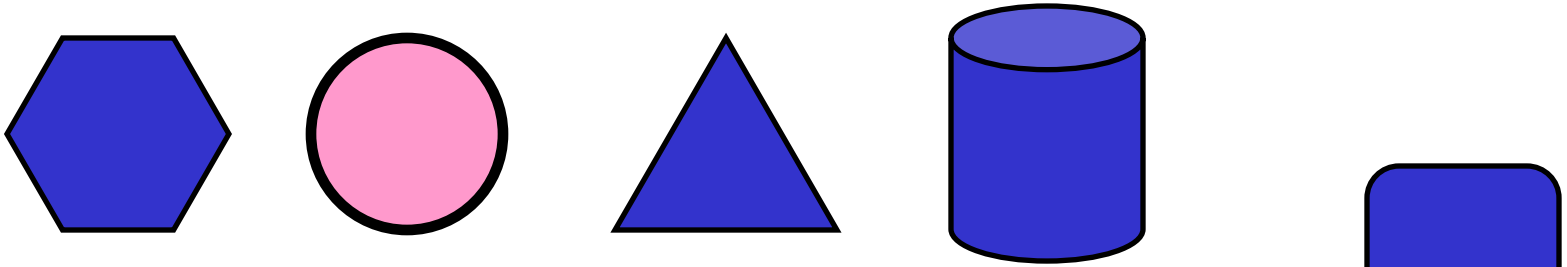
Quick Intro to Knowledge Discovery



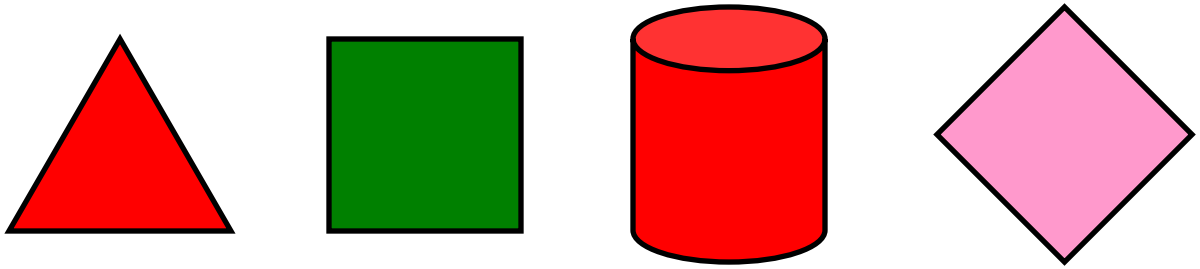
Invited talk at Taipei Medical University, June 2006

What is Knowledge Discovery?

Jonathan's blocks



Jessica's blocks



Whose block is this?

Jonathan's rules : Blue or Circle
Jessica's rules : All the rest

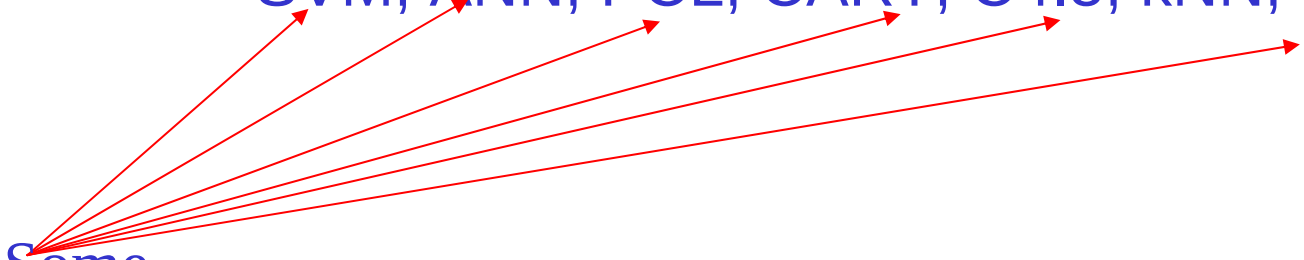
What is Knowledge Discovery?




Question: Can you explain how?

Main Steps of Knowledge Discovery

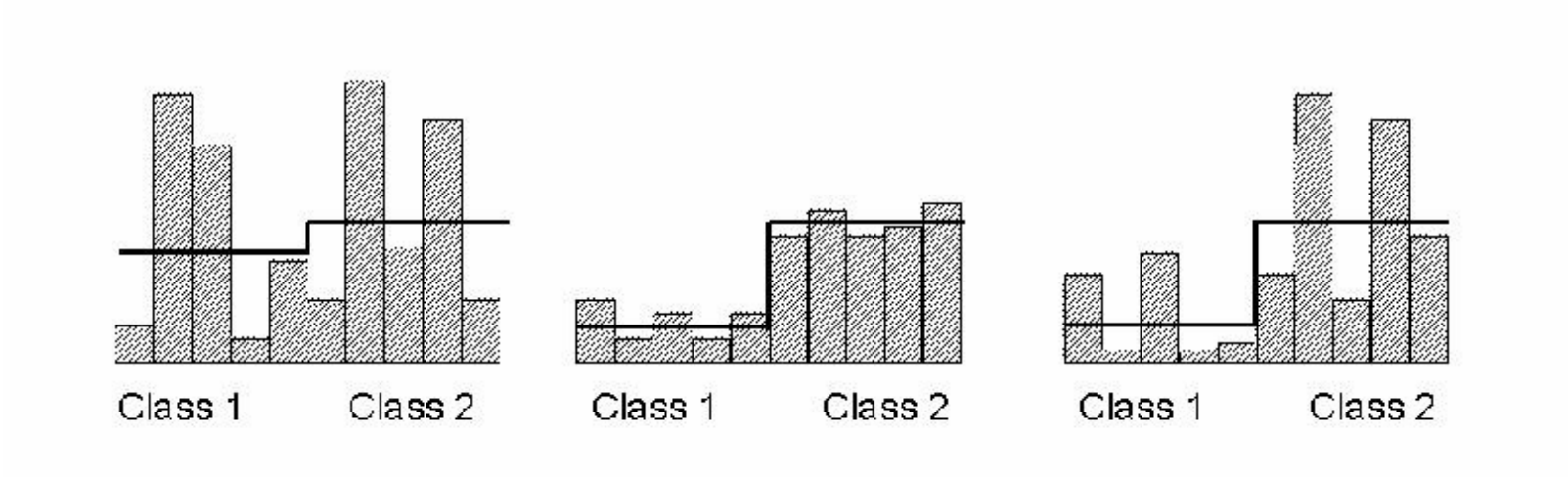
- **Training data gathering**
- **Feature generation**
  - k-grams, colour, texture, domain know-how, ...
- **Feature selection**
  - Entropy, $\chi 2$, CFS, t-test, domain know-how...
- **Feature integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...



Some classifier/ methods

Feature Selection Statistics Principle

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



Classifier Learning/Operation Principle

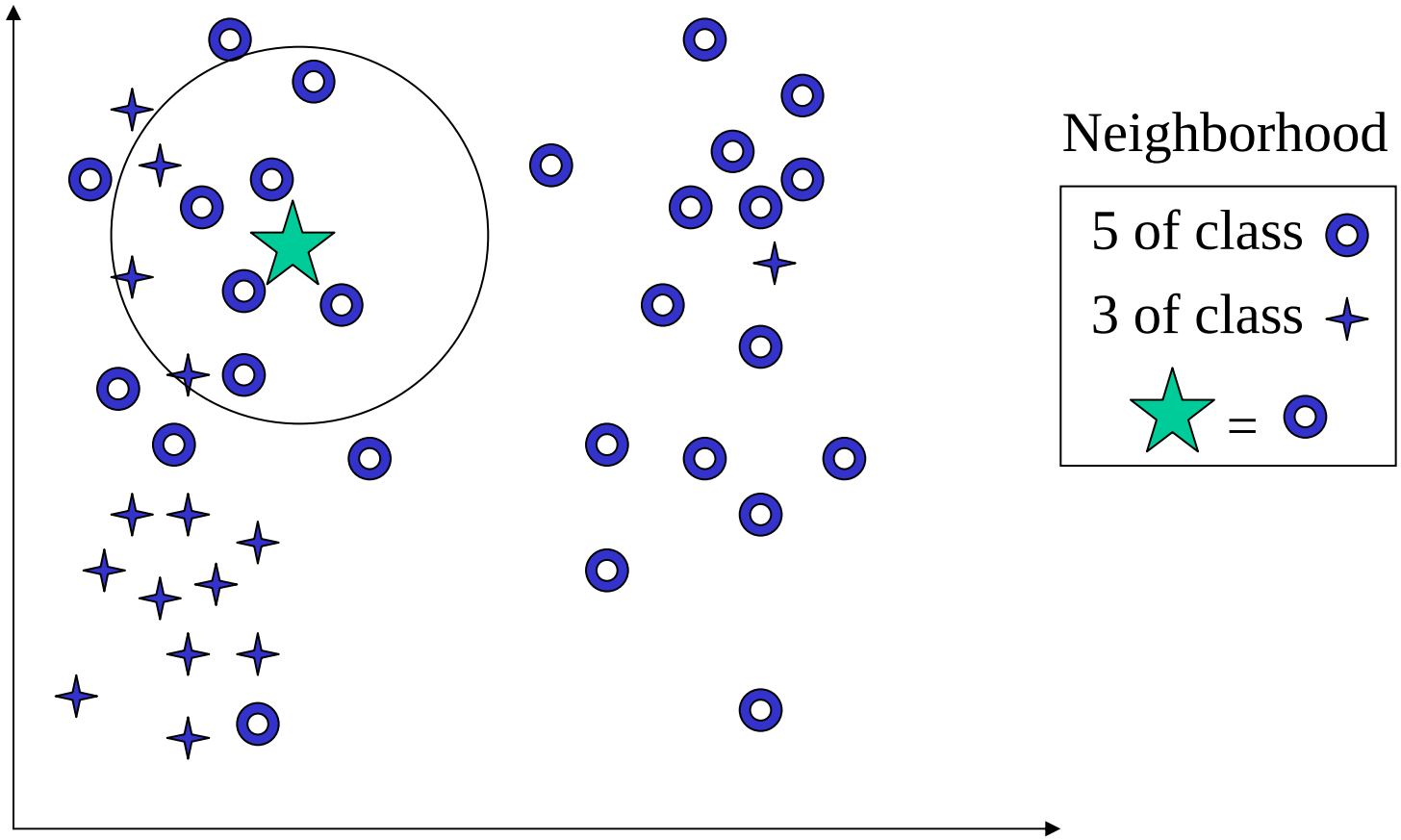
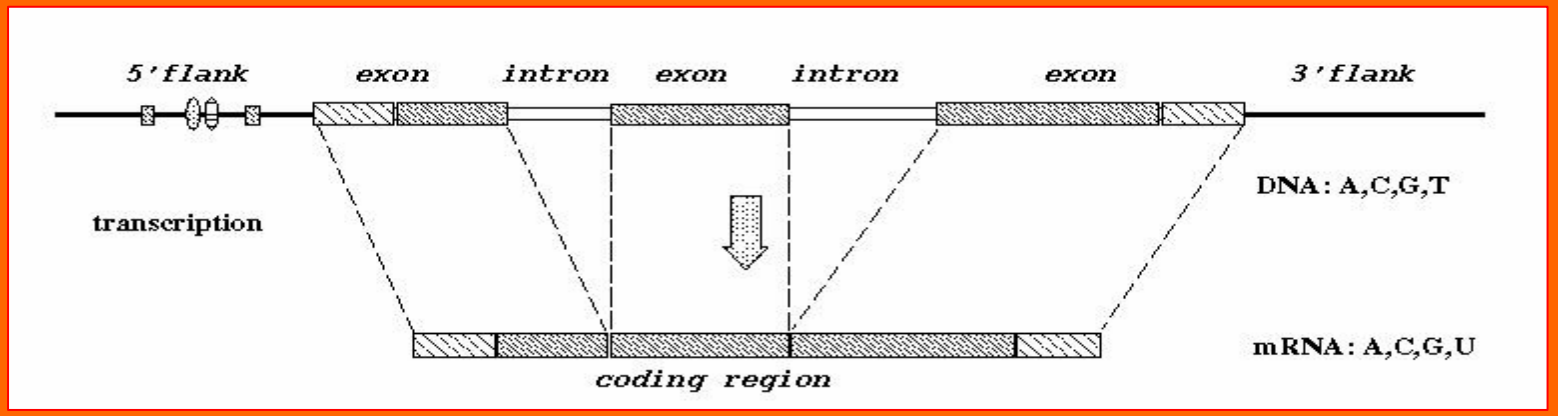


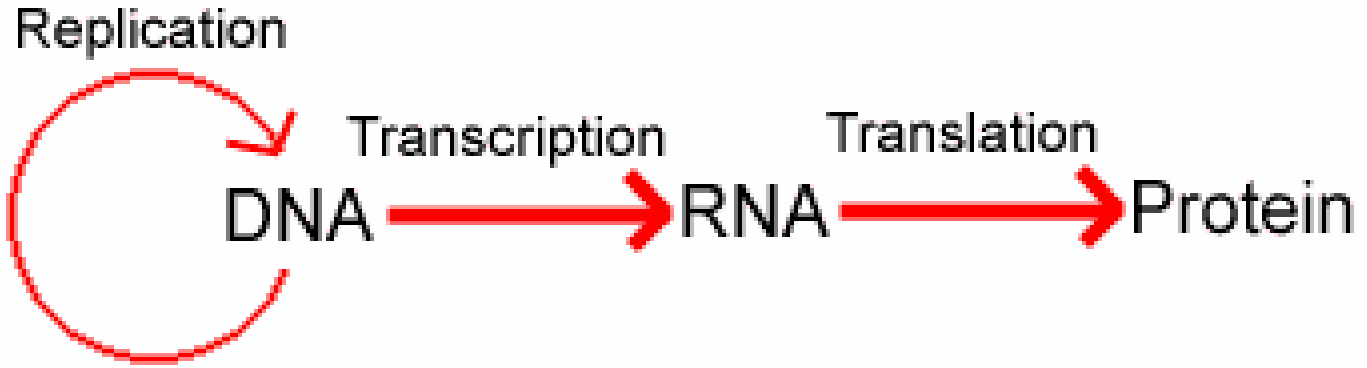
Image credit: Zaki

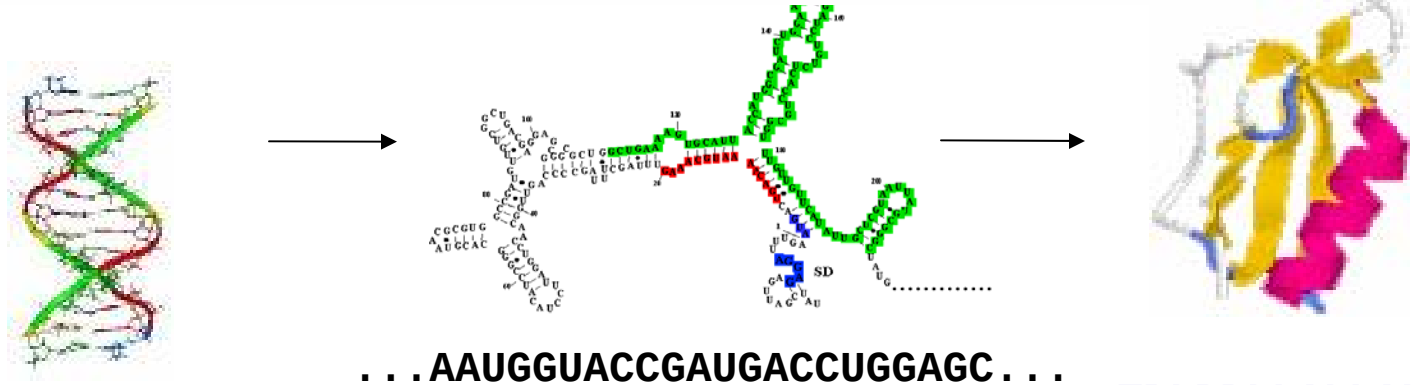
Translation Initiation Site Recognition



NUS National University of Singapore

Central Dogma



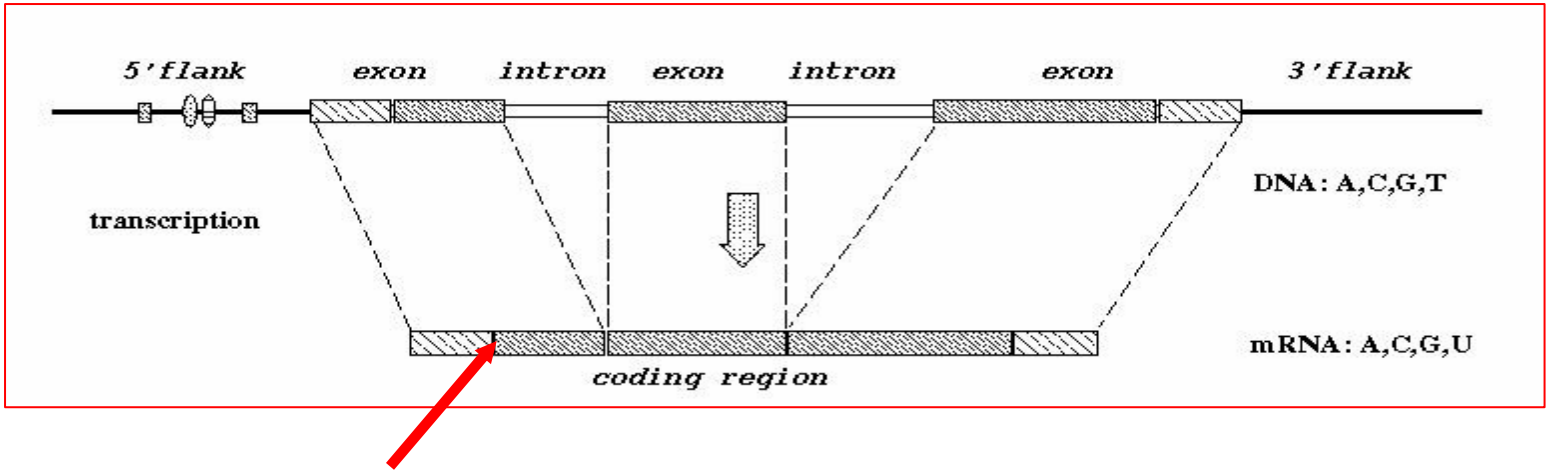


. . .AATGGTACCGATGACCTG. . .

. . .AAUGGUACCGAUGACCUGGAGC. . .

. . .TRLRPLLALLALWP. . .

Translation Initiation Site



A Sample cDNA

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

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

Approach

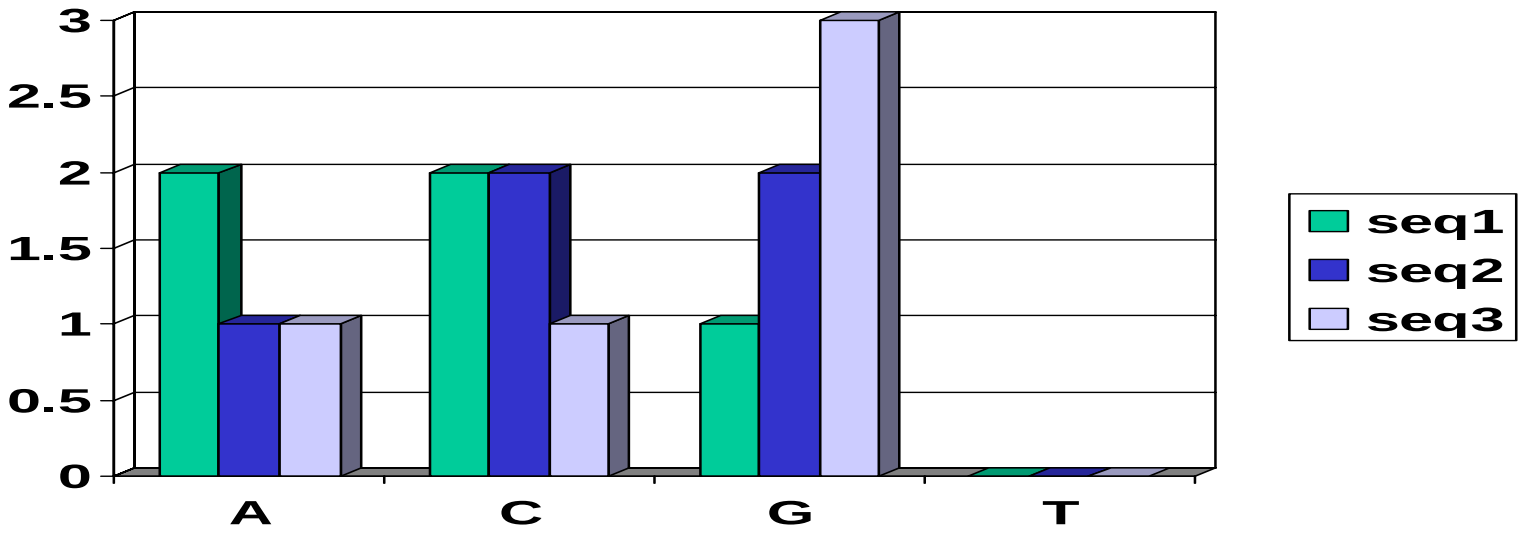
- **Training data gathering**
- **Signal generation**
  - k-grams, distance, domain know-how, ...
- **Signal selection**
  - Entropy, $\chi 2$, CFS, t-test, domain know-how...
- **Signal integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...

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

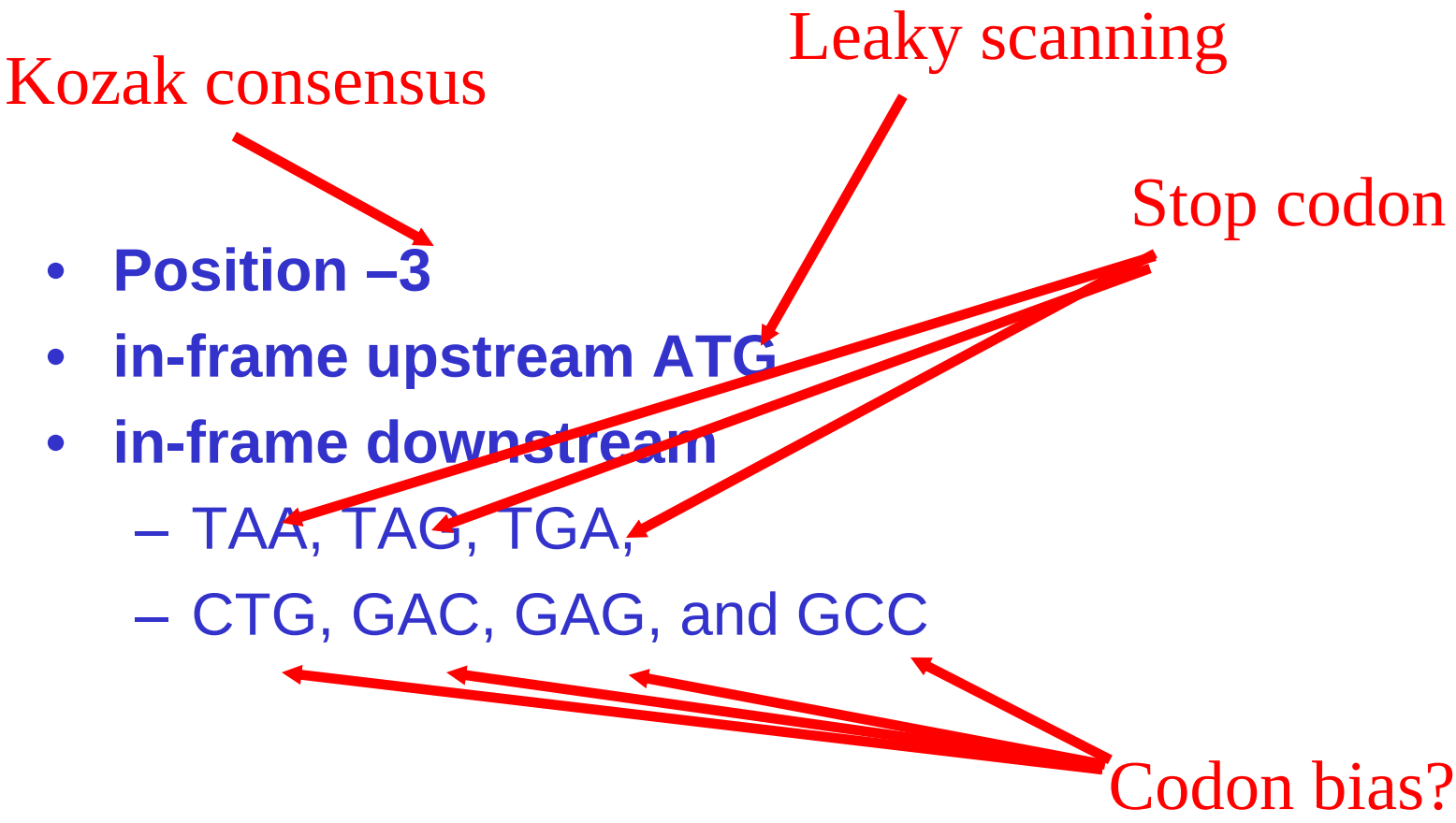
```
 299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA     160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA     240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT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

- **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!**
- **This is too many for most machine learning algorithms**

Sample K-grams Selected by CFS



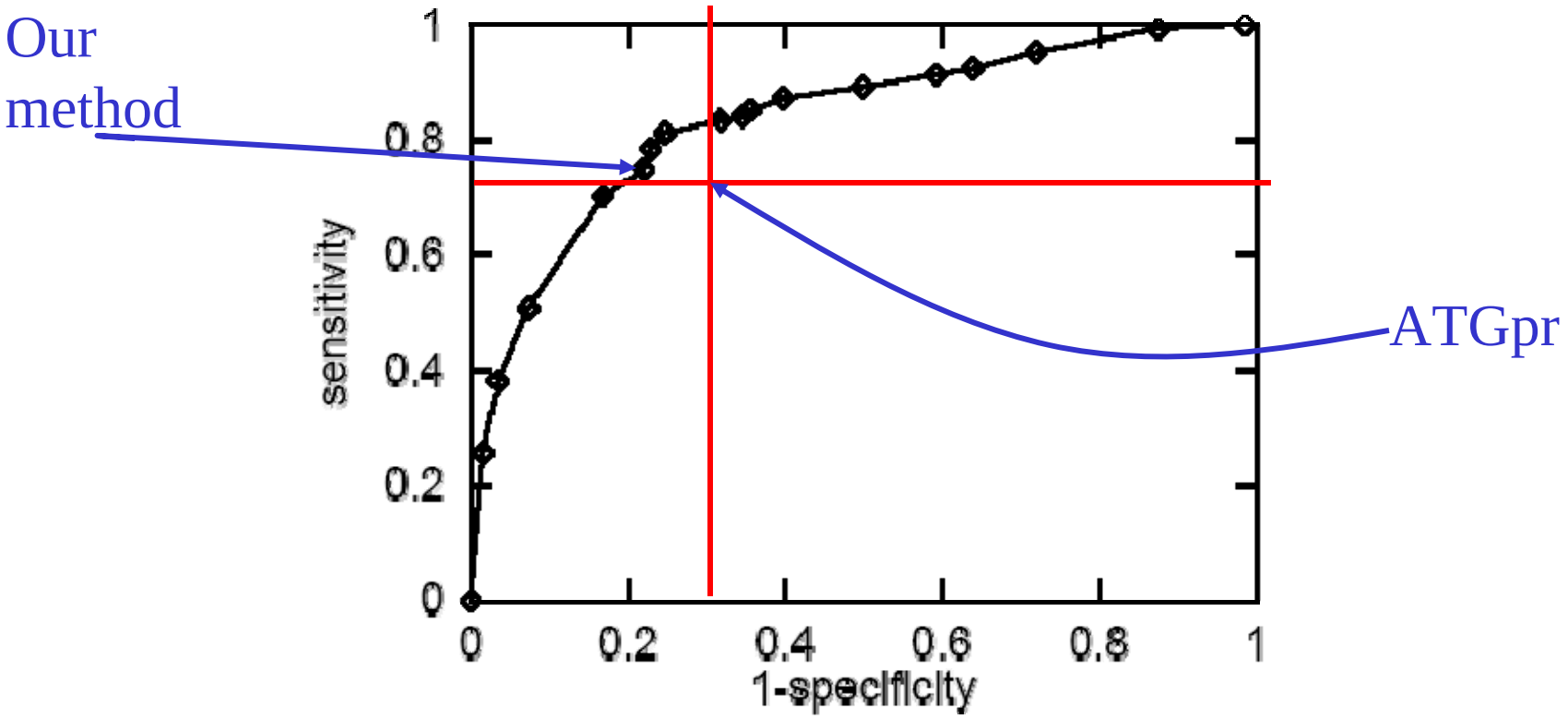
- **Position –3**
- **in-frame upstream ATG**
- **in-frame downstream**
  - TAA, TAG, TGA,
  - CTG, GAC, GAG, and GCC

Results (3-fold x-validation)

	predicted as positive	predicted as negative
positive	TP	FN
negative	FP	TN

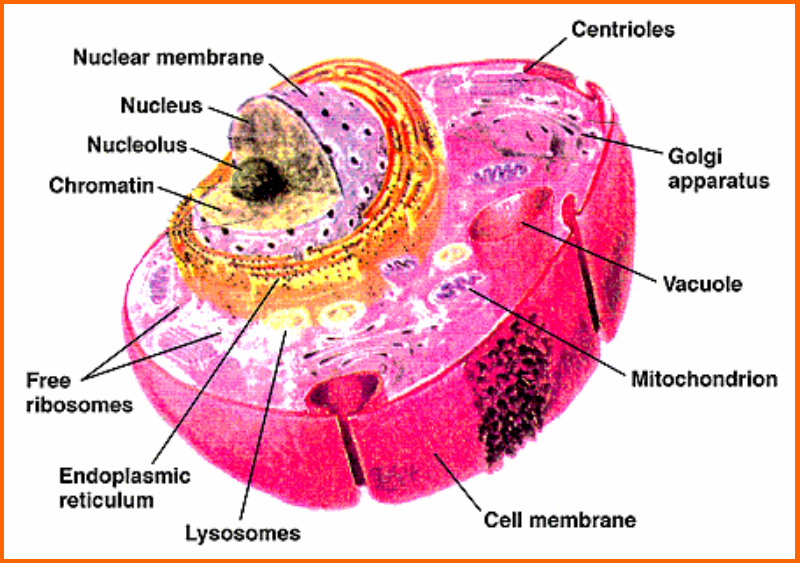
	TP/(TP + FN)	TN/(TN + FP)	TP/(TP + FP)	Accuracy
Naïve Bayes	84.3%	86.1%	66.3%	85.7%
SVM	73.9%	93.2%	77.9%	88.5%
Neural Network	77.6%	93.2%	78.8%	89.4%
Decision Tree	74.0%	94.4%	81.1%	89.4%

Validation Results (on Chr X and Chr 21)



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

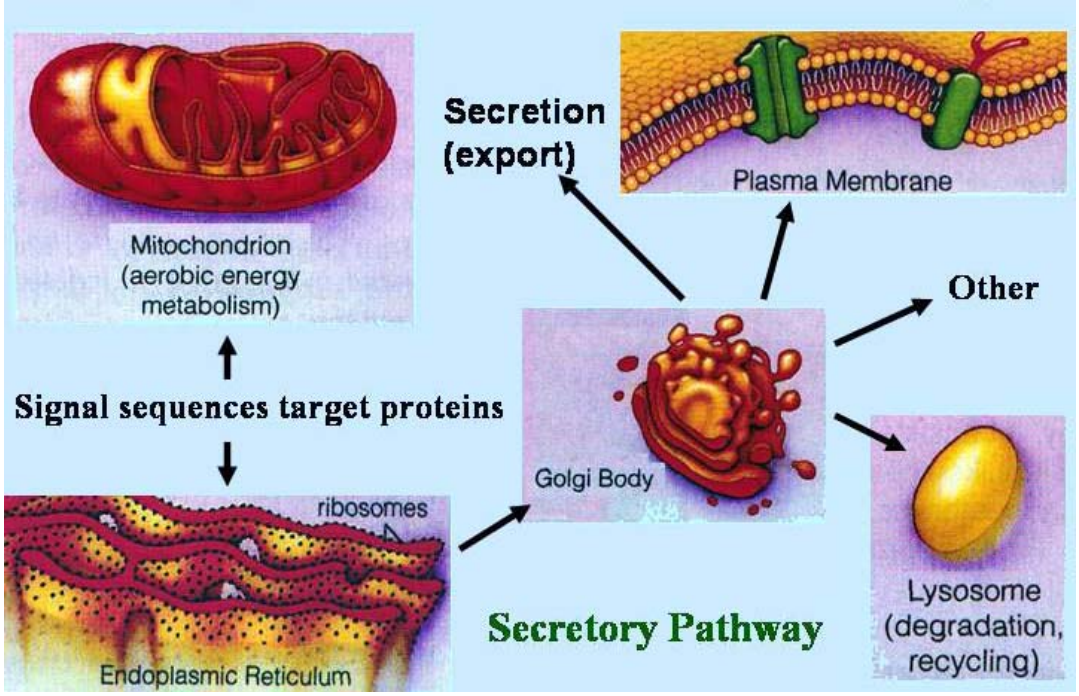
Protein Subcellular Localization Prediction



NUS National University of Singapore

Invited talk at Taipei Medical University, June 2006

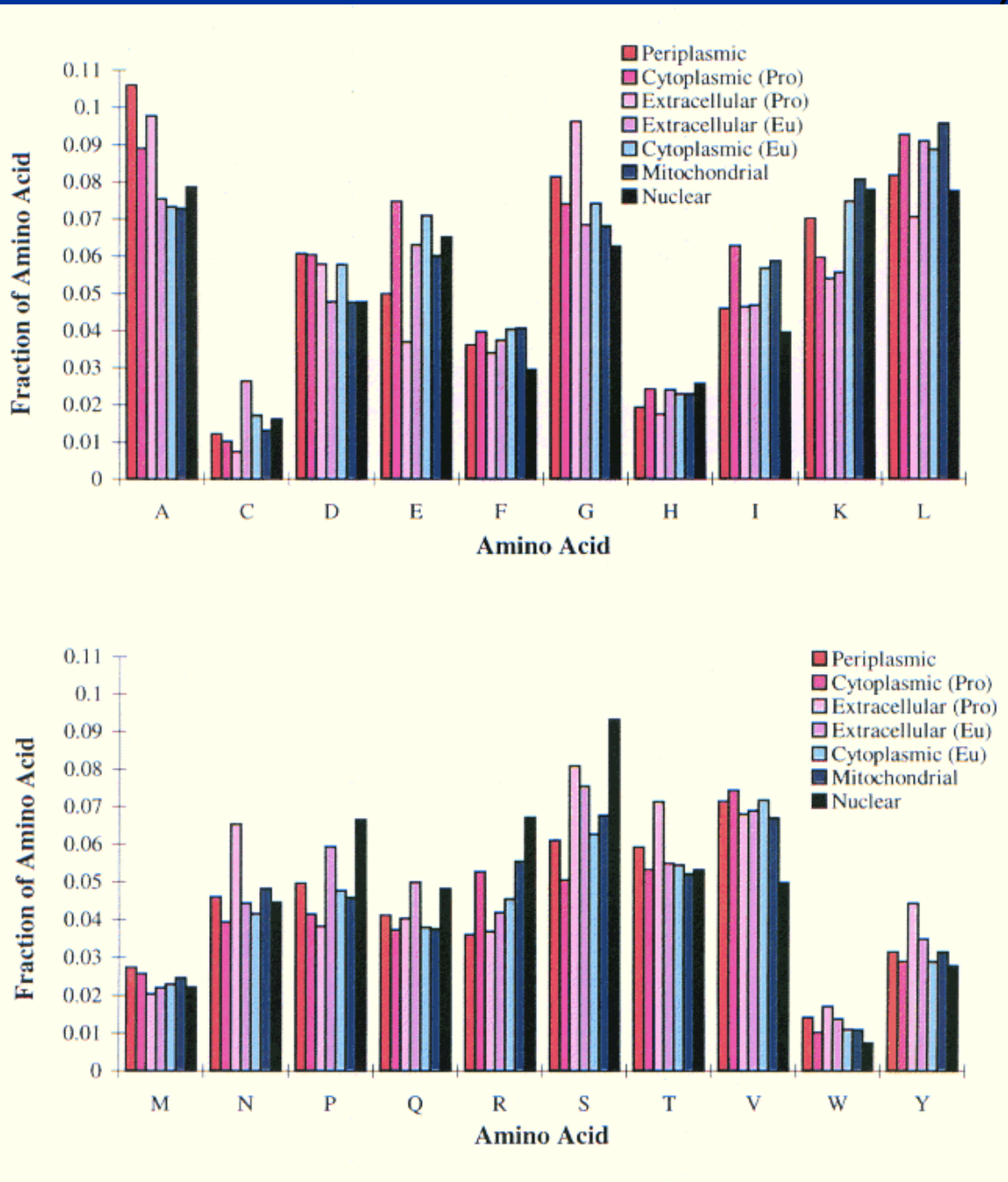
Compartments and Sorting



- **Eukaryotic cells requires proteins be targeted to their subcellular destinations**

- **Protein sorting is determined by specific amino acid sequences, or "signals", within the protein**
- **Secretory pathway targets proteins to plasma membrane, some membrane-bound organelles such as lysosomes, or to export proteins from the cell**

Amino acid composition of proteins residing in different sites are different



Amino Acid Composition Differences

- each cellular location has own characteristic physio-chemical environment
- proteins in each location have adapted thru evolution to that environment
- thus reflected in the protein structure and amino acid composition

- If the above is true, the amino acid composition differences wrt cellular location sites should be more pronounced on protein surfaces than protein interior
- Exercise: Why?

Adaptation of Protein Surfaces

Andrade et al., *JMB*, 1998

- **To test the theory of adaptation of protein surfaces to subcellular localization, we do a plot of 3 types of composition vectors along their first two principal components**

composition vectors were calculated for all proteins; these were then used to define a sample variance–co-variance matrix, **S**, as follows:

$$\mathbf{S} = \{s_{jk}\} = \left\{ \sum_{i=1}^{n} (c_{ij} - \bar{c}_j)(c_{ij} - \bar{c}_k)/n \right\} \qquad (2)$$

where:

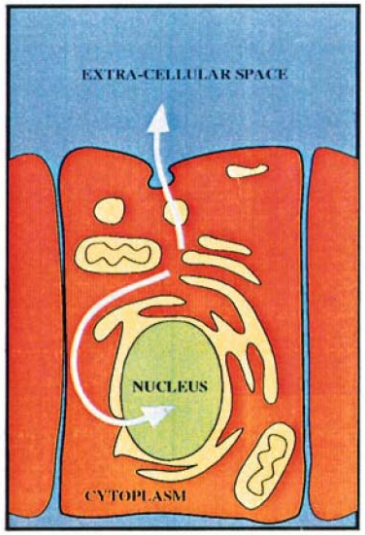
$$\bar{c}_j = \frac{1}{n} \sum_{i=1}^{n} c_{ij} \qquad (3)$$

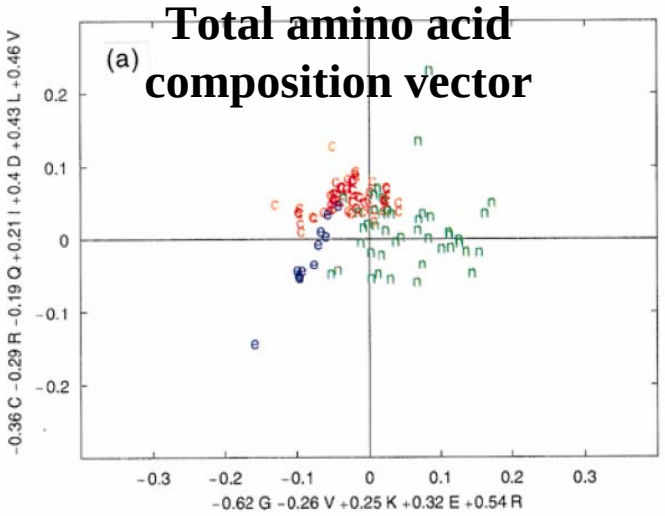
Proportion of j^{th} amino acid type in i^{th} protein

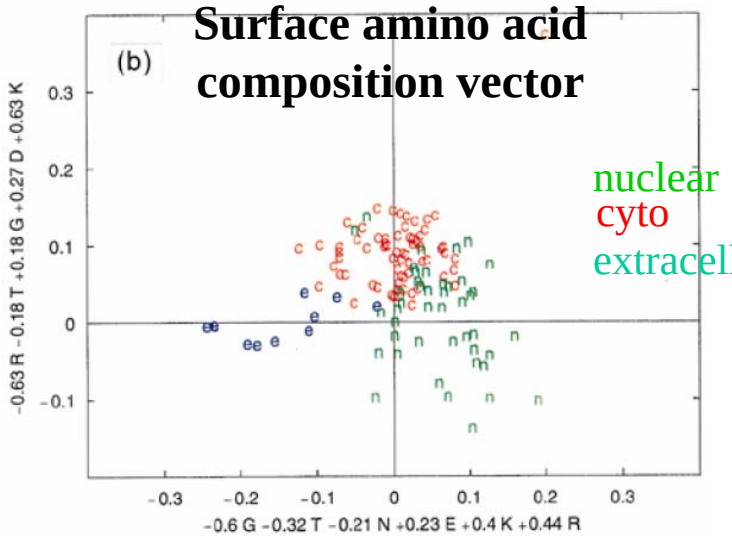
is the average composition of the *j*th amino acid type over the *n* proteins in the data set. The principal components of the set of composition vectors are then the Eigenvectors of **S**

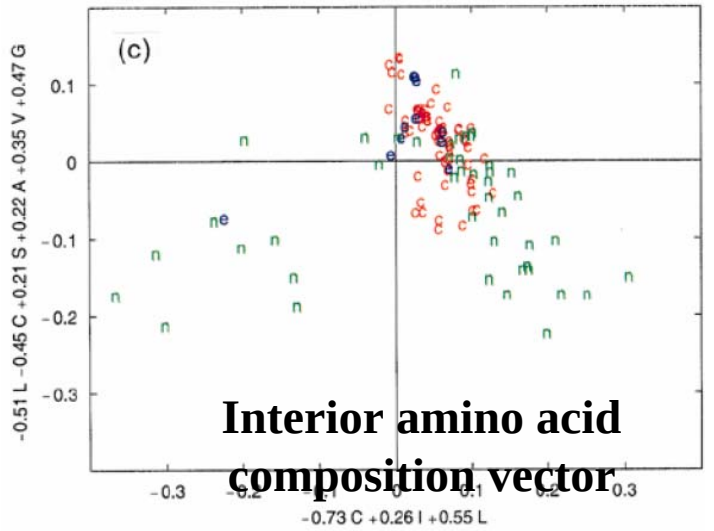
Adaptation of Protein Surfaces

Andrade et al., *JMB*, 1998









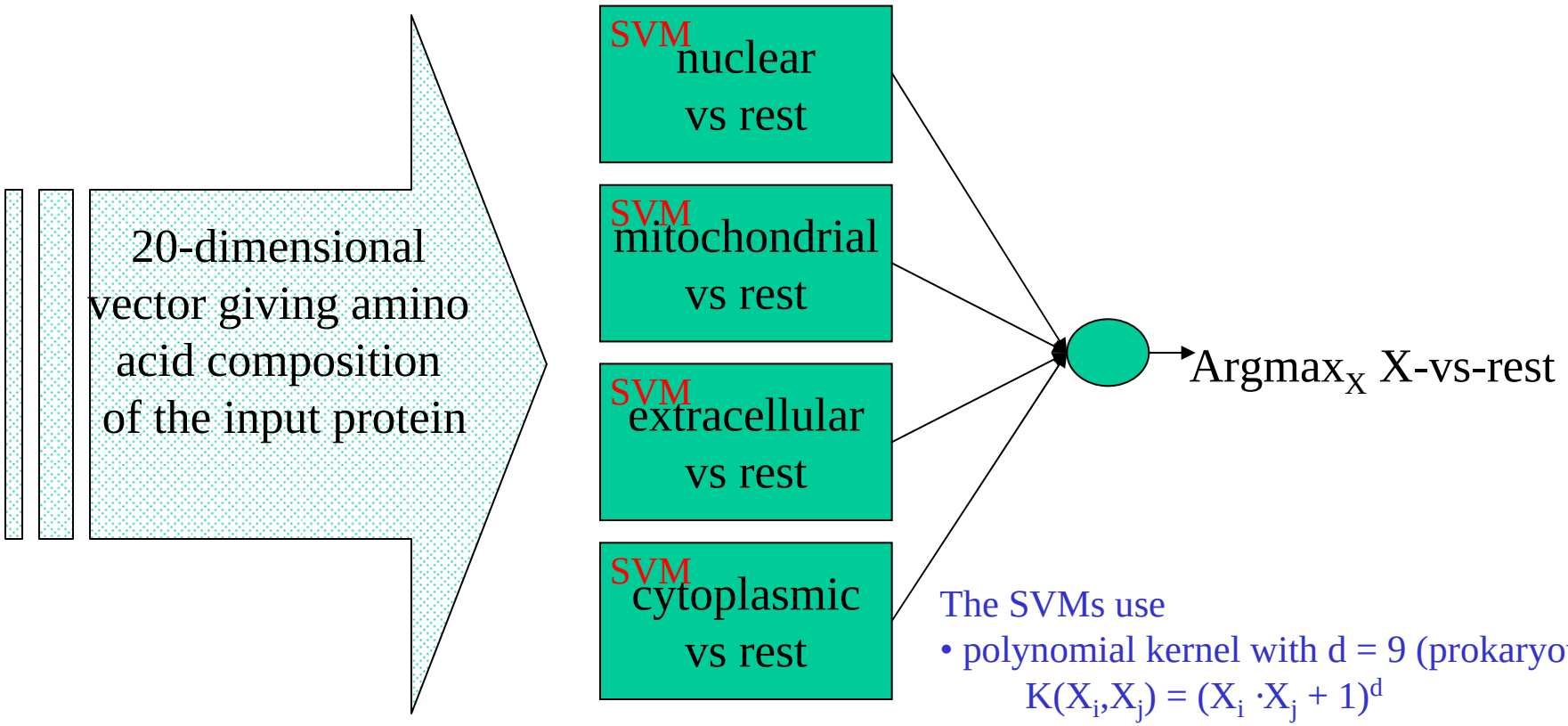
- **Clearly total & surface composition vectors show better separation than interior composition vectors**

Amino Acid Composition

- **This means can use amino acid composition vectors, especially those from protein surfaces, to predict subcellular localization!**
- **Let's see how this turn out….**

Support Vector Machines: SubLoc

Hua & Sun, *Bioinformatics*, 17:721--728, 2001



The SVMs use

- polynomial kernel with d = 9 (prokaryotic),
  $K(X_i,X_j) = (X_i \cdot X_j + 1)^d$
- RBF kernel with γ=16 (eukaryotic),
  $K(X_i, X_j) = \exp(-\gamma |X_i - X_j|^2)$

SubLoc:
Performance

Location (Eukaryotic)	NNPSL	Markov model	SubLoc
	Accuracy (%)	Accuracy (%)	Accuracy (%)
Cytoplasmic	55	78.1	76.9
Extracellular	75	62.2	80.0
Mitochondrial	61	69.2	56.7
Nuclear	72	74.1	87.4
Total accuracy	66	73.0	79.4

Dataset: Reinhardt & Hubbard, *NAR*, 1998

SubLoc: Robustness of Amino Acid Composition Approach

	Accuracy (%)				
	Total	Cyto	Extra	Mito	Nuclear
COMPLETE	78.3	76.7	77.2	56.4	86.0
CUT-10	77.2	74.0	77.8	52.7	86.1
CUT-20	76.3	73.2	78.5	51.4	84.8
CUT-30	76.1	72.5	76.3	50.5	85.8
CUT-40	75.3	71.5	74.2	46.7	86.3

- **Amazingly, accuracy of SubLoc is virtually unaffected when the first 10, 20, 30, & 40 amino acids in a protein are deleted**
- **Amino acid composition is a robust indicator of subcellular localization, and is insensitive to errors in N-terminal sequences**

Amino Acid Composition: Taking it Further

- **How about pairs of consecutive amino acids? (a.k.a 2-grams) How about 3-grams, …, k-grams?**
- **How about pseudo amino acid composition?**
- **How about presence of entire functional domains? (I.e. think of the presence/absence of a functional domain as a summary of amino acid sequence info...)**

Functional Domain Composition

Cai & Chou, *BBRC*, 305:407--411, 2003

If a protein got a hit in Interpro,
use NN-5875D; else use NN-40D

Training seqs of various localization sites

BLAST against db of known functional domains (Interpro)

NN-5875D:
Train k-NN (k=1) using these vectors

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_{5875} \end{bmatrix},$$

where

$$x_i = \begin{cases} 1 & \text{hit,} \\ 0 & \text{otherwise} \end{cases}$$

or, if no hit found

NN-40D:
Train k-NN (k=1) using these vectors

$$\mathbf{X} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix},$$

Amino acid composition

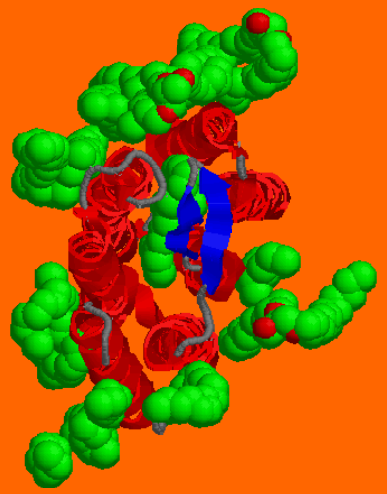
Pseudo amino acid composition

Functional Domain Composition:
Performance

Investigators	Prokaryotic set[b]		Eukaryotic set[c]	
	Re-substitution (%)	Jackknife (%)	Re-substitution (%)	Jackknife (%)
Chou and Elrod [6]	90.4	86.5	N/A	N/A
Yuan [22]	N/A	89.1	N/A	73.0
Cai and Chou [23]	96.1	84.4	95.6	70.6
Feng [24]	93.5	89.2	N/A	N/A
Feng and Zhang [25]	97.7	90.4	N/A	N/A
Hua and Sun [26]	N/A	91.4	N/A	79.4
Authors of this paper	100	**89.3**	100	**90.4**

Dataset: Reinhardt & Hubbard, *NAR*, 1998

Protein Function Prediction



National University of Singapore

Invited talk at Taipei Medical University, June 2006

Function Assignment to Protein Sequence

**SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT**

- **How do we attempt to assign a function to a new protein sequence?**

Guilt-by-Association

- **Compare the target sequence T with sequences S_1, …, S_n of known function in a database**
- **Determine which ones amongst S_1, …, S_n are the mostly likely homologs of T**
- **Then assign to T the same function as these homologs**
- **Finally, confirm with suitable wet experiments**

Guilt-by-Association

Compare *T* with seqs of known function in a db

Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                                  60        70        80        90       100
Amicyanin                MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVV
                                                    :..:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                             70        80        90       100       110
```

No obvious match between Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

```
>gi|13476732|ref|NP_108301.1|  unknown protein [Mesorhizobium loti]
gi|14027493|dbj|BAB53762.1|  unknown protein [Mesorhizobium loti]
          Length = 105

 Score =  105 bits (262), Expect = 1e-22
 Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALALIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
           MK G L  ++       MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between Amicyanin and u[illegible]wn M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

NUS National University of Singapore

What if no homolog of known function is found?

SVM-Pairwise Framework

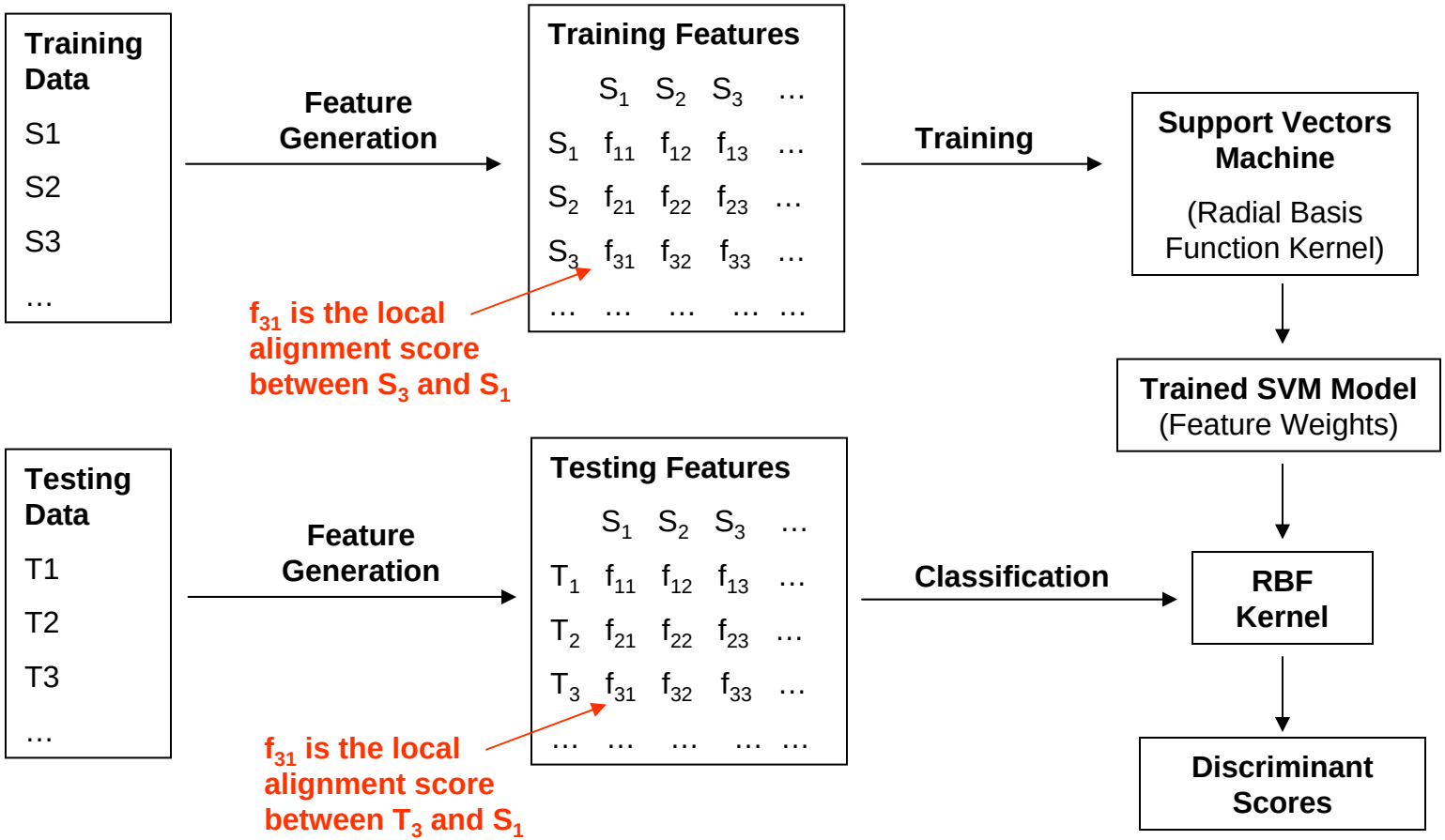
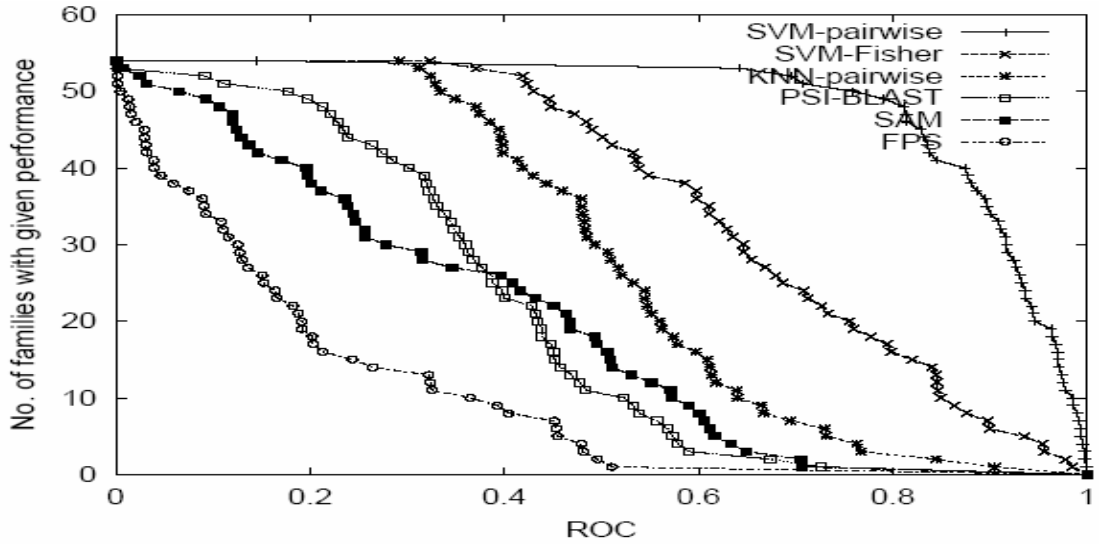
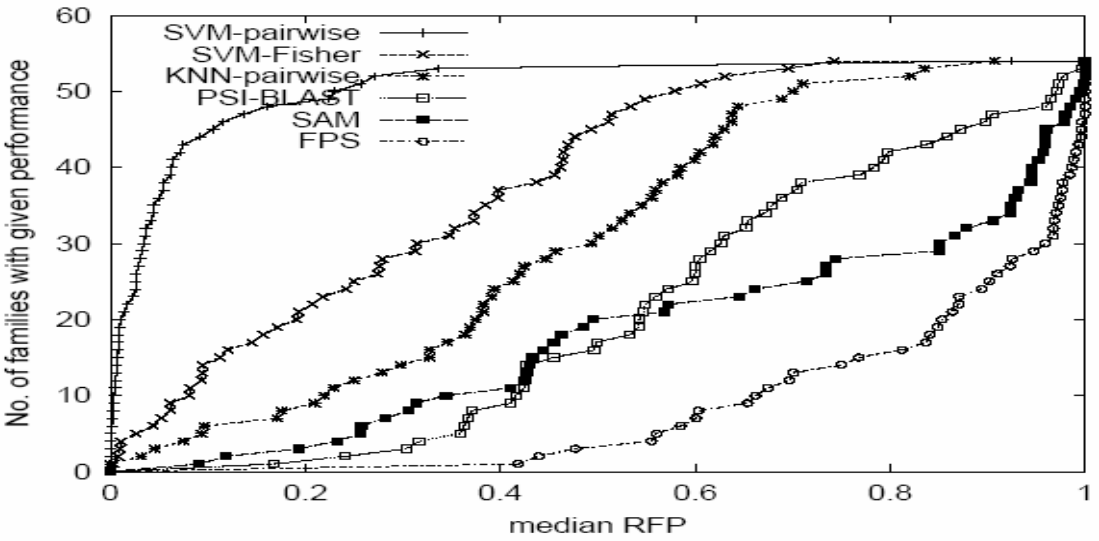


Image credit: Kenny Chua

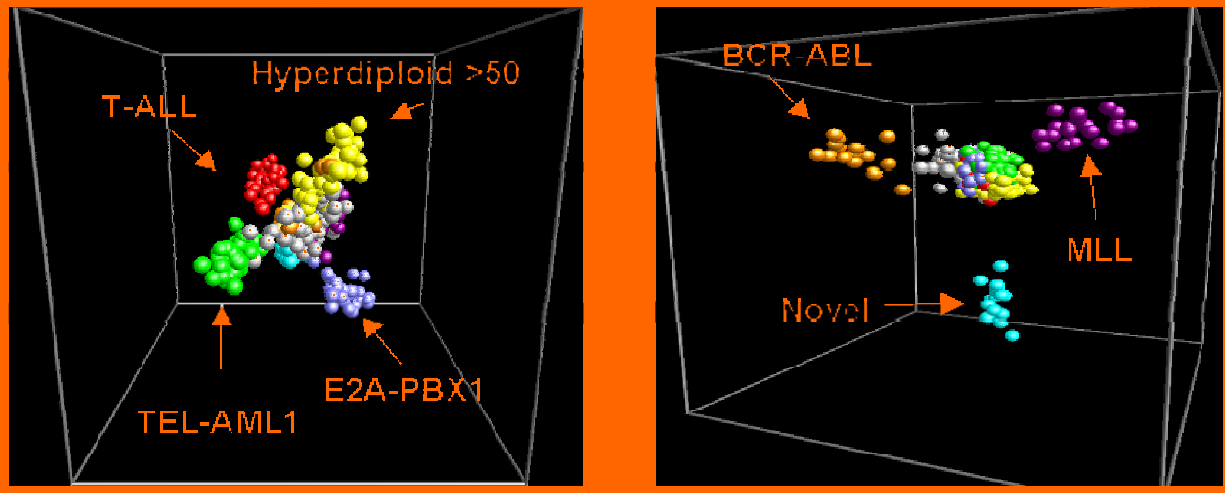
Performance of SVM-Pairwise

- **Receiver Operating Characteristic (ROC)**
  - The area under the curve derived from plotting true positives as a function of false positives for various thresholds.
- **Rate of median False Positives (RFP)**
  - The fraction of negative test examples with a score better or equals to the median of the scores of positive test examples.





Treatment Optimization of Childhood Leukemia



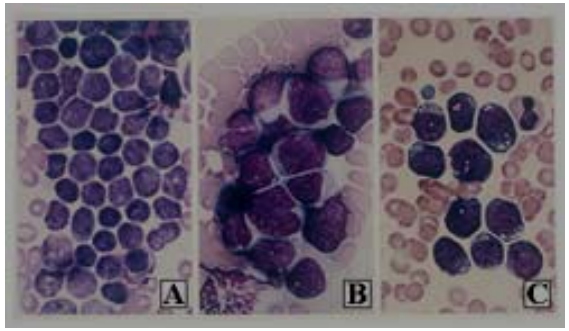
NUS National University of Singapore

Invited talk at Taipei Medical University, June 2006

Childhood ALL

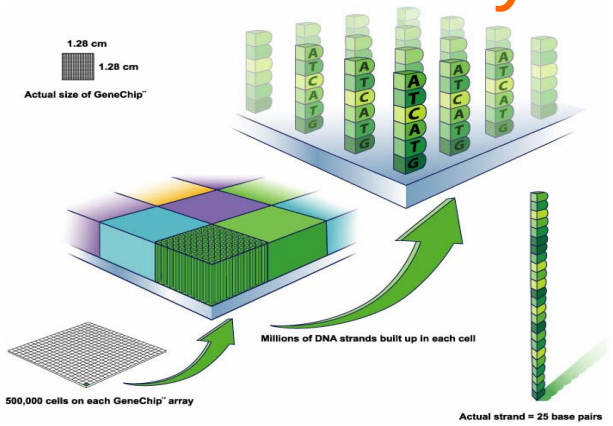
- **Major subtypes are: T-ALL, E2A-PBX, TEL-AML, MLL genome rearrangements, Hyperdiploid>50, BCR-ABL**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse

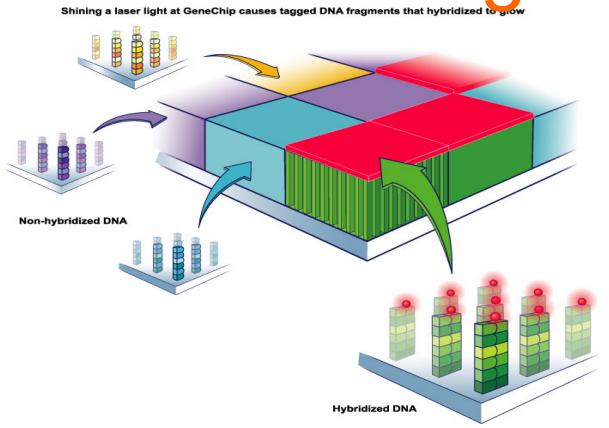
- **The subtypes look similar**

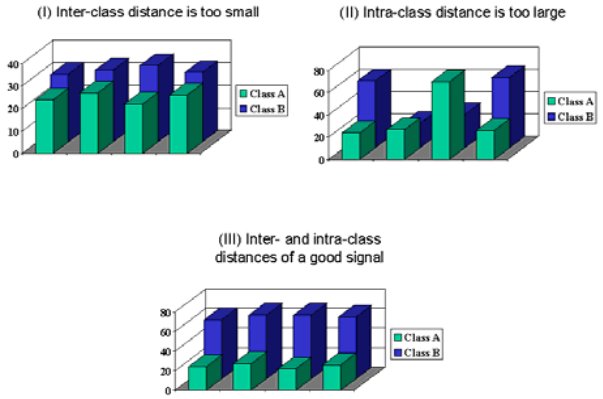


- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- **Unavailable in most ASEAN countries**

Single-Test Platform of Microarray & Machine Learning





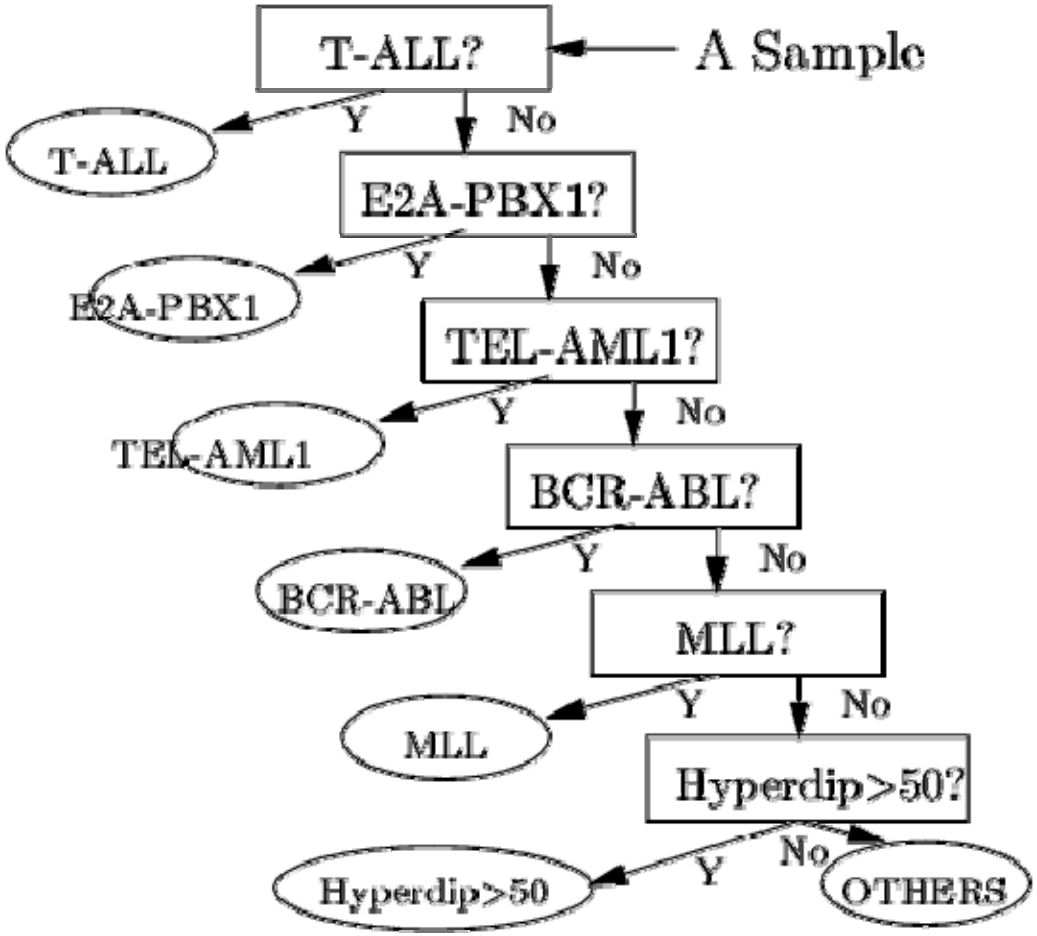


	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs In	Avg Diff	Abs Call	
AFFX-MurI	5	2	19	297.5	A	M16762 Mouse int
AFFX-MurI	3	2	19	554.2	A	M37897 Mouse int
AFFX-MurI	4	2	19	308.6	A	M25892 Mus mus
AFFX-MurF	1	3	19	141	A	M83649 Mus mus
AFFX-BioB	13	1	19	9340.6	P	J04423 E coli bioB
AFFX-BioB	15	0	19	12862.4	P	J04423 E coli bioB
AFFX-BioB	12	0	19	8716.5	P	J04423 E coli bioB
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacterioph
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacterioph
AFFX-BioB	7	5	18	-483	A	J04423 E coli bioB
AFFX-BioB	5	4	18	313.7	A	J04423 E coli bioB
AFFX-BioB	7	6	20	-1016.2	A	J04423 E coli bioB

Image credit: Affymetrix

Childhood ALL Subtype Diagnosis Workflow

A tree-structured diagnostic workflow was recommended by our doctor collaborator

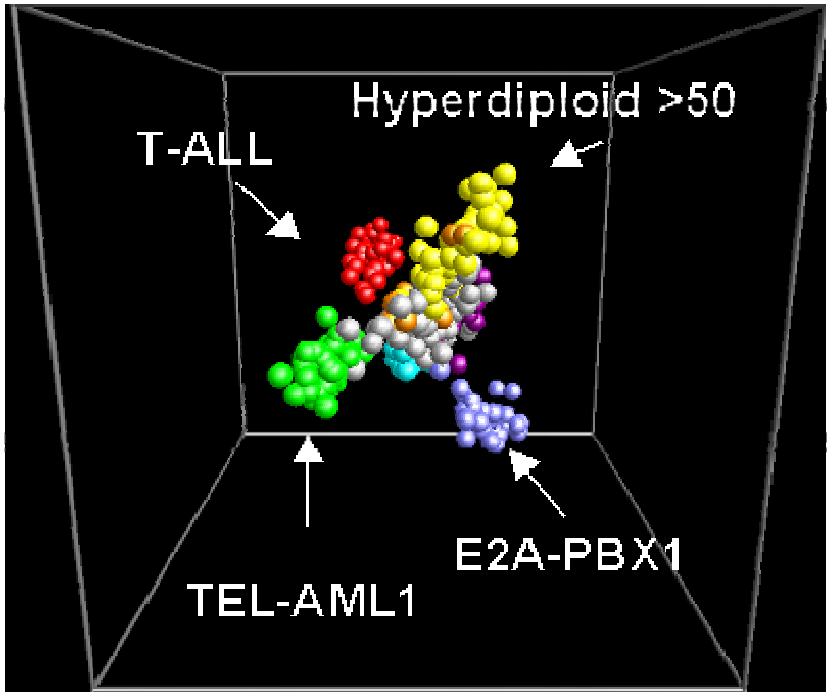


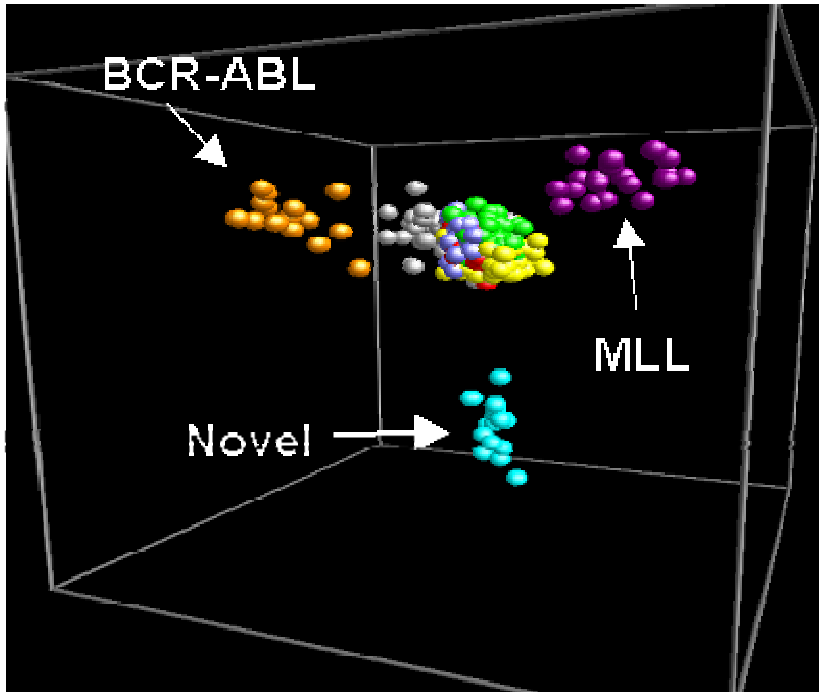
Accuracy Various Classifiers)

Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS1	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS2	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS3	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS4	2:0	3:0	1:4	2:0
MLL vs OTHERS5	0:1	0:0	0:0	0:0
Hyperdiploid>50 vs OTHERS	2:6	0:2	0:2	0:1
Total Errors	14	6	8	4

The classifiers are all applied to the 20 genes selected by $\chi 2$ at each level of the tree

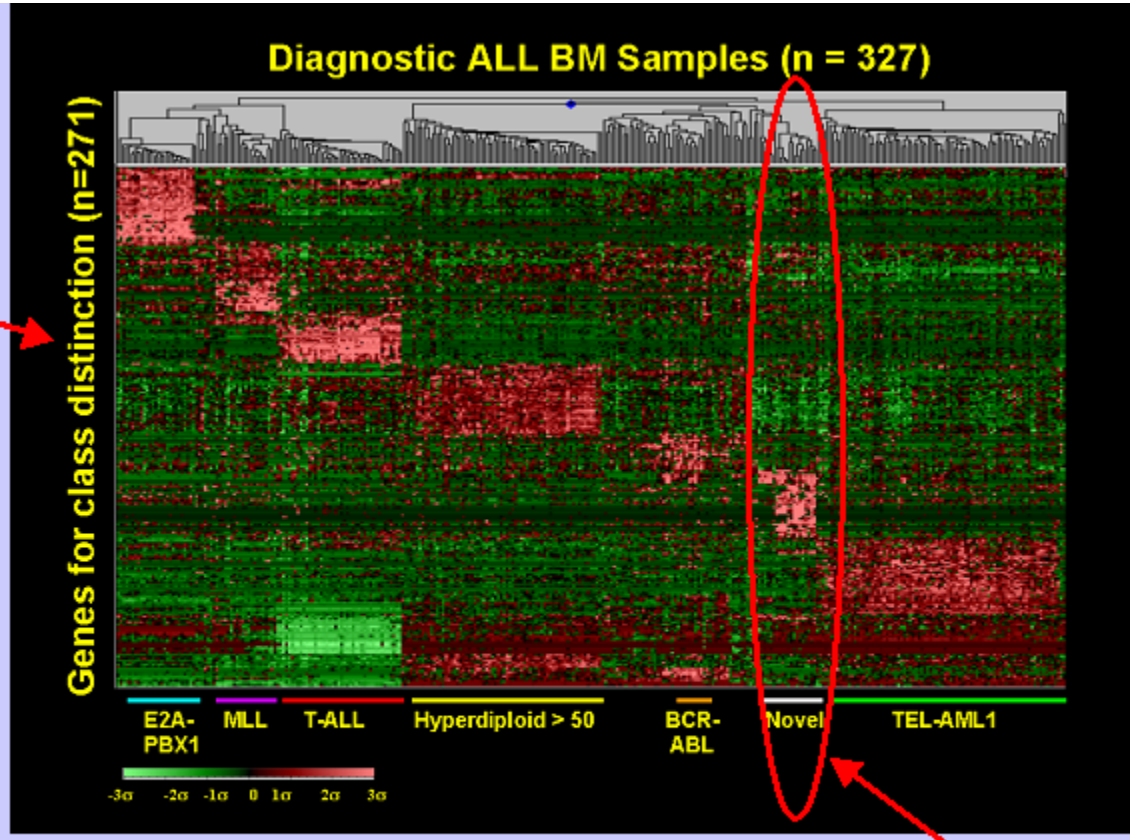
Multidimensional Scaling Plot Subtype Diagnosis





Is there a new subtype?

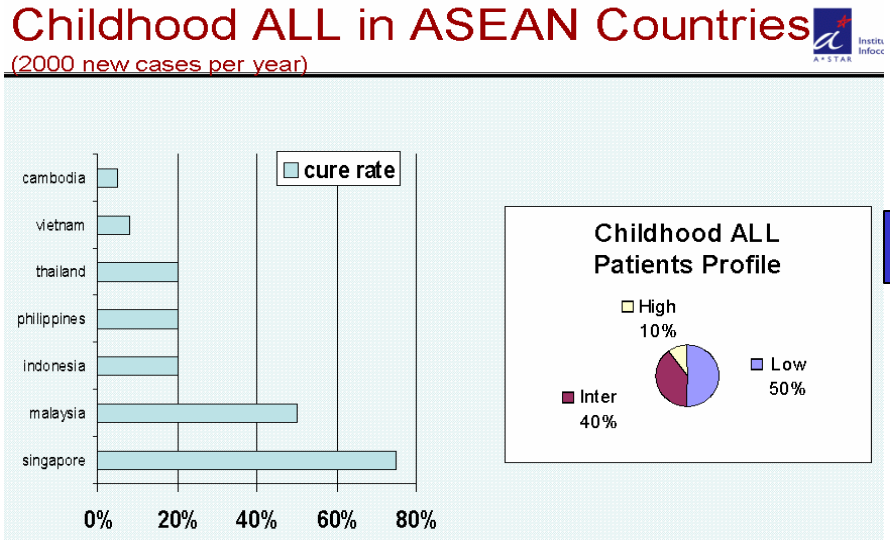
Genes selected by $\chi 2$



New subtype discovered

- **Hierarchical clustering of gene expression profiles reveals a novel subtype of childhood ALL**

Conclusions



Conventional Tx:

- **intermediate intensity to everyone**

⇒ 10% suffers relapse

⇒ 50% suffers side effects

⇒ costs US$150m/yr

Our optimized Tx:

- **high intensity to 10%**
- **intermediate intensity to 40%**
- **low intensity to 50%**
- **costs US$100m/yr**

•High cure rate of 80%

• Less relapse

• Less side effects

• Save US$51.6m/yr

Any Question?

Acknowledgements

- **Huiqing Liu**
- **Fanfan Zeng**
- **Jinyan Li**
- **Allen Yeoh**