An Introduction to Knowledge Discovery Applications and Challenges in Life Sciences

Limsoon Wong

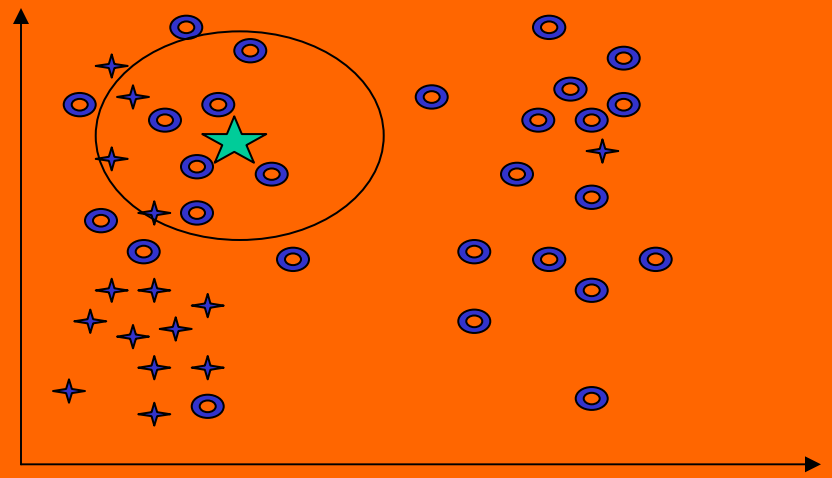
ICDE'07, Istanbul, Turkey, 16-20 April 2007

Plan

- **Quick introduction to knowledge discovery** (10 min)
- **Protein function inference** (40 min)
- **Gene feature recognition** (50 min)
- **Disease diagnosis, treatment, and understanding** (70 min)
- **Advancing knowledge discovery** (10 min)

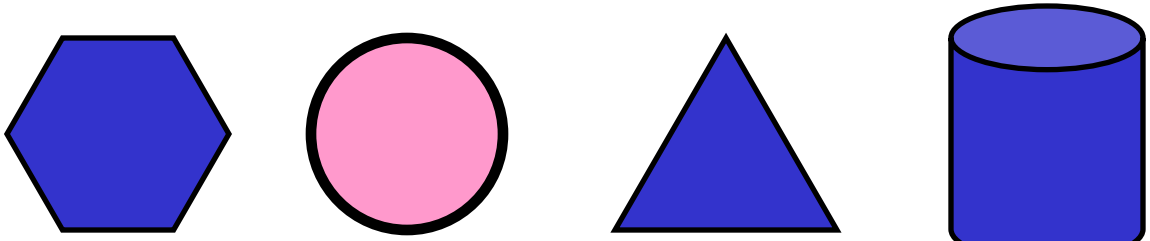
10 min

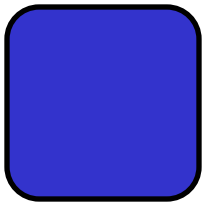
Quick Intro to Knowledge Discovery



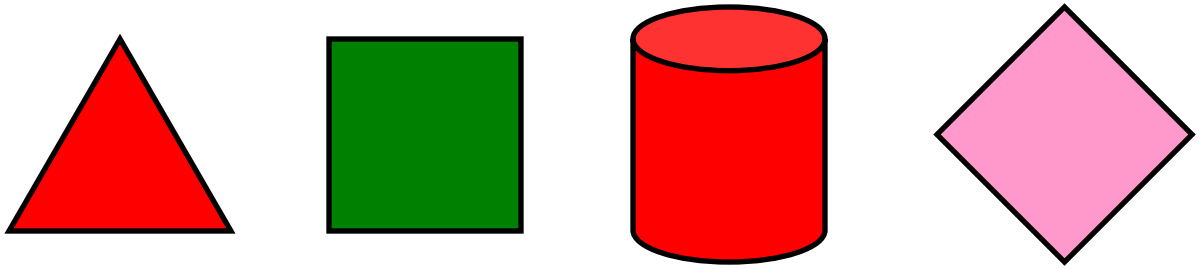
What is Knowledge Discovery?

Jonathan's blocks





Jessica's blocks



Whose block is this?

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

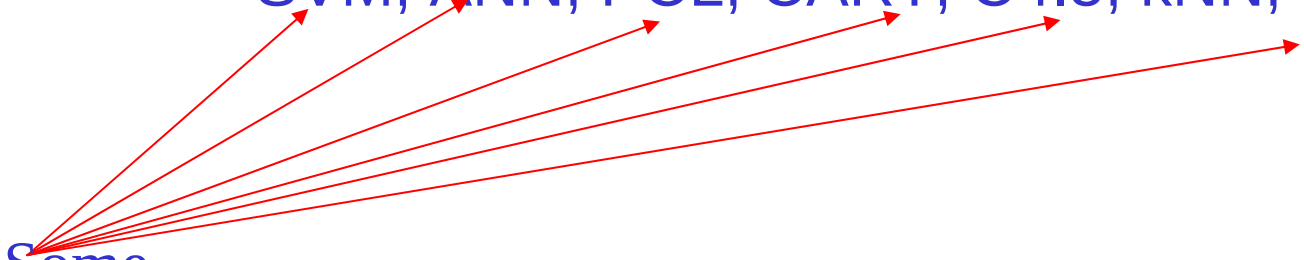
What is Knowledge Discovery?




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

An Example Classifier: kNN (k=8)

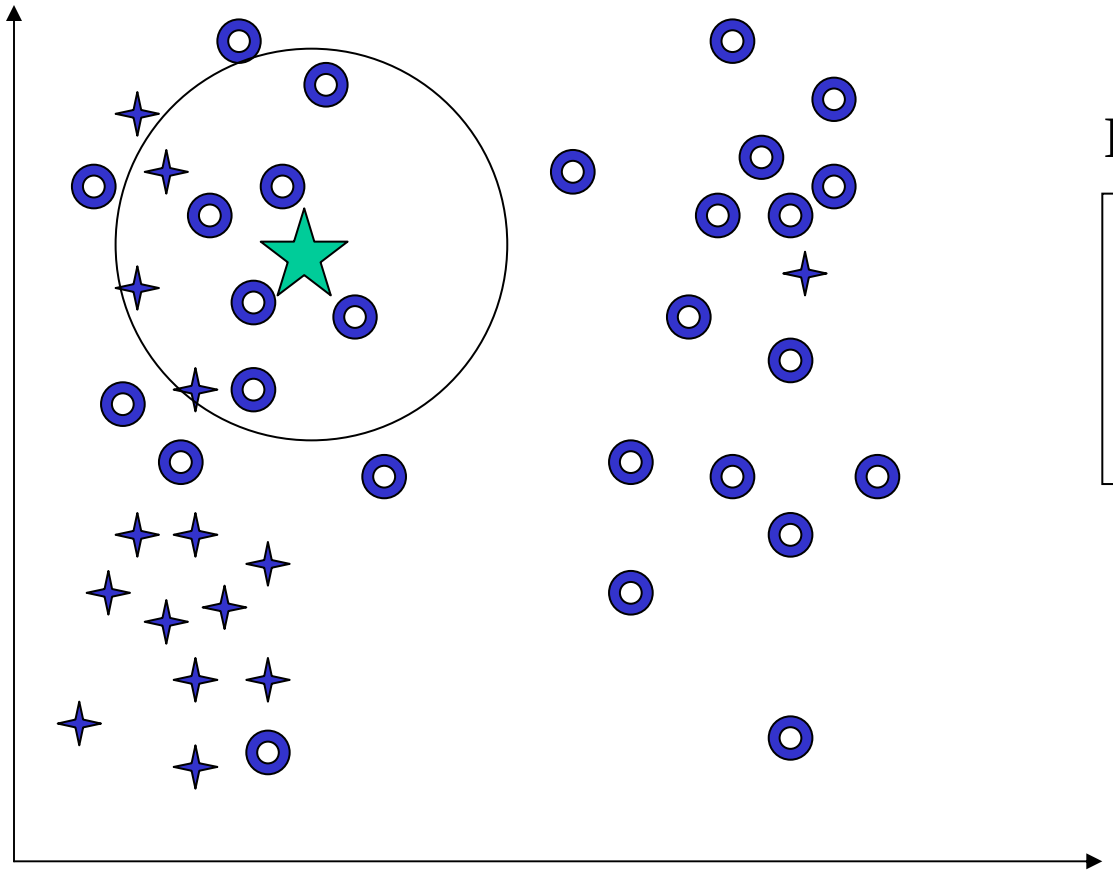
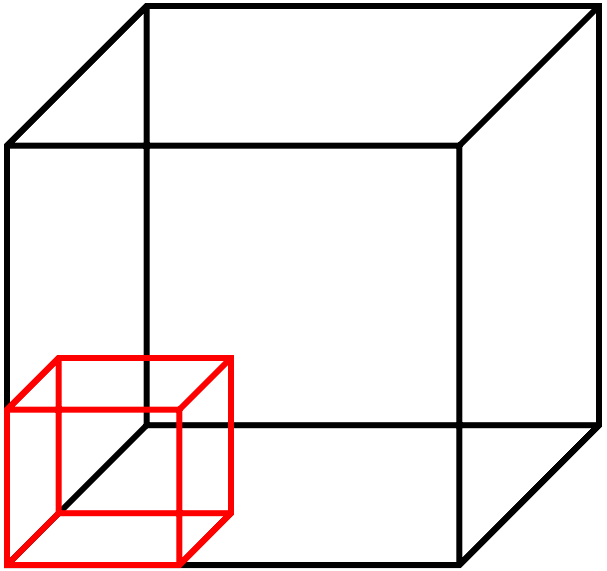


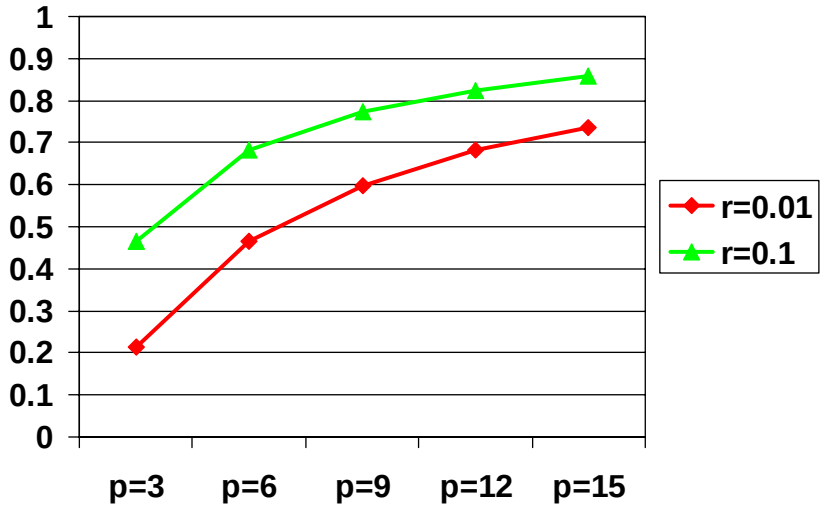
Image credit: Zaki

Curse of Dimensionality

- **How much of each dimension is needed to cover a proportion r of total sample space?**



- **Calculate by $e_p(r) = r^{1/p}$**
- **So, to cover 1% of a 15-D space, need 85% of each dimension!**



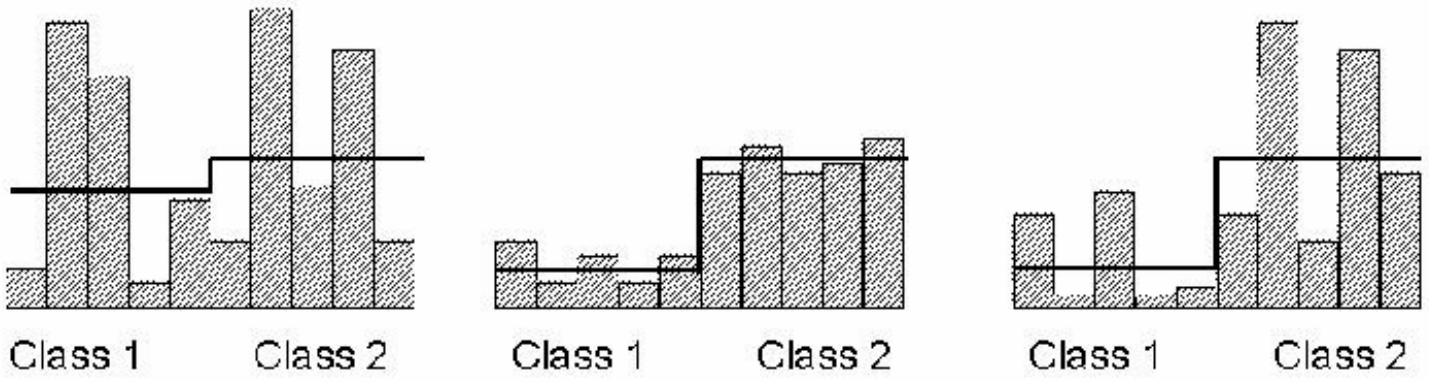
Tackling the Curse: Signal Selection

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

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.



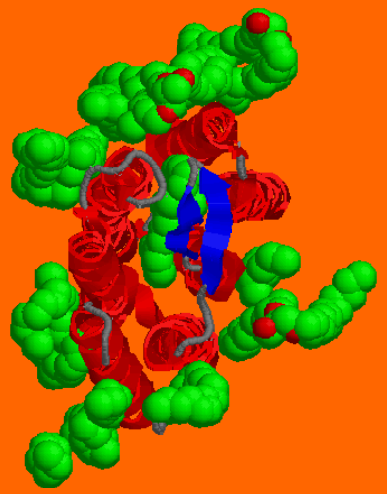
Self-fulfilling Oracle

- Construct artificial dataset with 100 samples, each with 100,000 randomly generated features and randomly assigned class labels
- select 20 features with the best t-statistics (or other methods)
- Evaluate accuracy by cross validation using only the 20 selected features
- The resultant estimated accuracy can be ~90%
- But the true accuracy should be 50%, as the data were derived randomly

What went wrong?

40 min

Protein Function Inference



Earliest Research in Seq Alignment

Source: Ken Sung

- **Doolittle et al. (*Science*, July 1983) searched for platelet-derived growth factor (PDGF) in his own DB. He found that PDGF is similar to v-sis oncogene**

```
PDGF-2  1       SLGSLTIAEPAMIAECKTREEVFCICRRL?DR?? 34
p28sis 61 LARGKRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTN 100
```

- **A seq alignment maximizes the number of positions that are in agreement in two sequences**

Sequence Alignment: Poor Example

- **Poor seq alignment shows few matched positions**

$\Rightarrow$ **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                 MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE
                                                    :..:   .  ::. ::
Ascorbate Oxidase         ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI
                                  70        80        90        100       110       120
```

No obvious match between
Amicyanin and Ascorbate Oxidase

Sequence Alignment: Good Example

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

good match between
Amicyanin and unknown M. loti protein

Function Assignment to Protein Seq

**SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT**

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

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

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

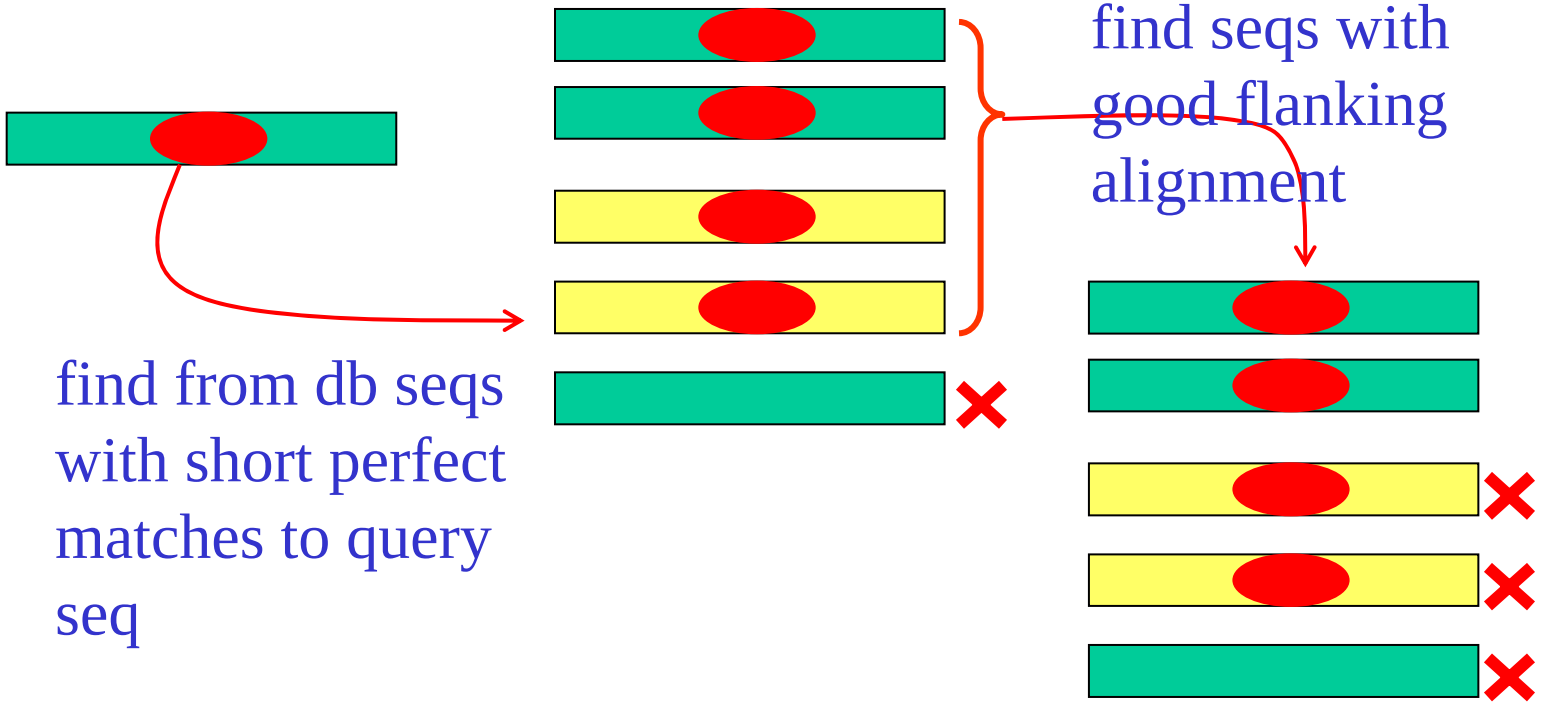
Assign to *T* same function as homologs

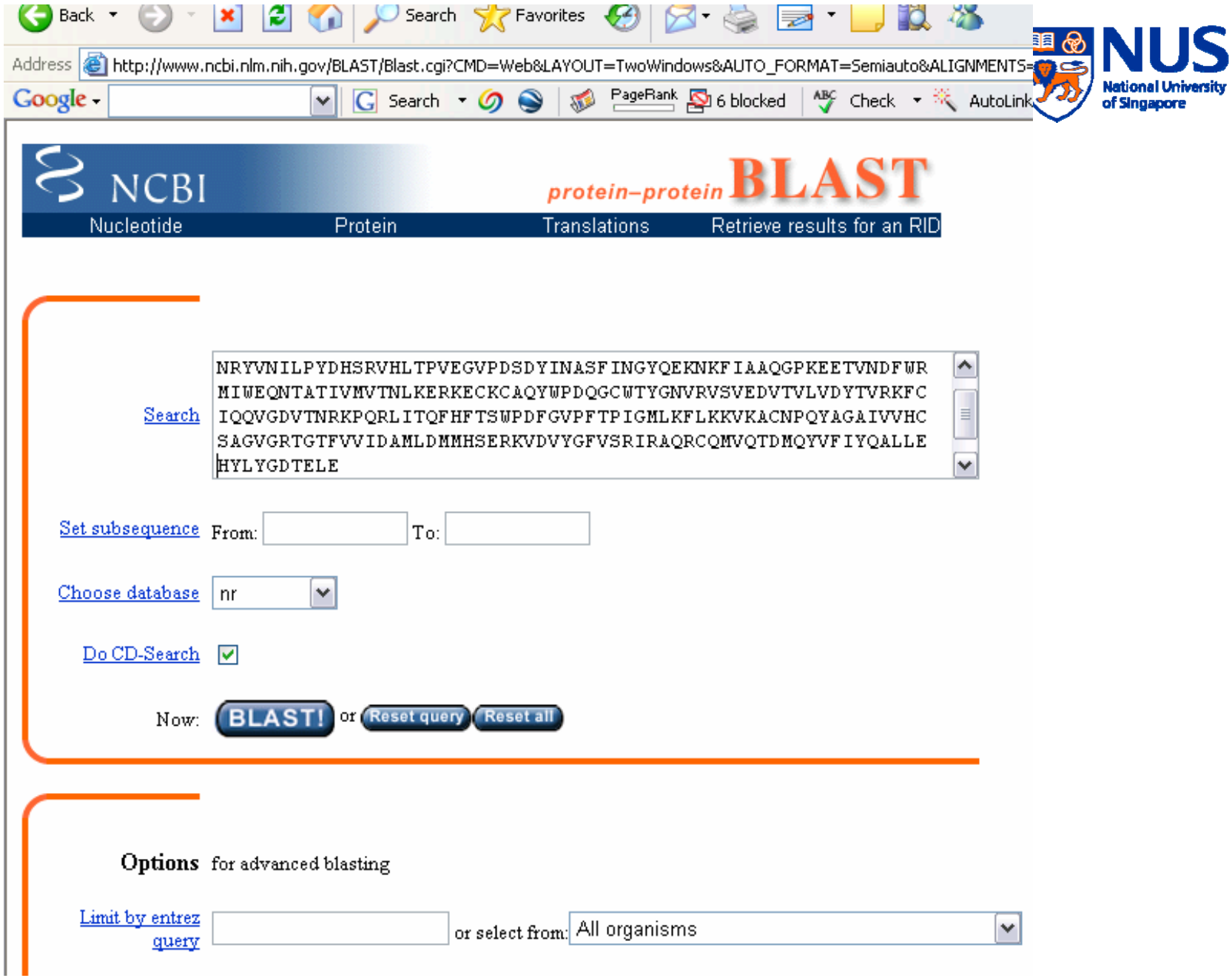
Confirm with suitable wet experiments

BLAST: How It Works

Altschul et al., *JMB*, 215:403--410, 1990

- **BLAST is one of the most popular tool for doing "guilt-by-association" sequence homology search**





Homologs obtained by BLAST

Sequences producing significant alignments:	Score (bits)	E Value
gi|14193729|gb|AAK56109.1|AF332081_1 protein tyrosin phosph...	62 L	e-177
gi|126467|sp|P18433|PTRA_HUMAN Protein-tyrosine phosphatase...	62 L	e-177
gi|4506303|ref|NP_002827.1| protein tyrosine phosphatase, r...	62 L	e-176
gi|227294|prf||1701300A protein Tyr phosphatase	620	e-176
gi|18450369|ref|NP_543030.1| protein tyrosine phosphatase, ...	62 L	e-176
gi|32067|emb|CAA37447.1| tyrosine phosphatase precursor [Ho...	61 L	e-176
gi|285113|pir||JC1285 protein-tyrosine-phosphatase (EC 3.1....	619	e-176
gi|6981446|ref|NP_036895.1| protein tyrosine phosphatase, r...	61 L	e-176
gi|2098414|pdb|1YFO|A Chain A, Receptor Protein Tyrosine Ph...	61 S	e-174
gi|32313|emb|CAA38662.1| protein-tyrosine phosphatase [Homo...	61 L	e-174
gi|450583|gb|AAB04150.1| protein tyrosine phosphatase >gi|4...	605	e-172
gi|6679557|ref|NP_033006.1| protein tyrosine phosphatase, r...	60 L	e-172
gi|483922|gb|AAA17990.1| protein tyrosine phosphatase alpha	599	e-170

- **Thus our example sequence could be a protein tyrosine phosphatase α (PTPα)**

Example Alignment with PTPα


```
 Score =  632 bits (1629), Expect = e-180
 Identities = 294/302 (97%), Positives = 294/302 (97%)

Query:  1   SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASXXXXXXXXR  60
           SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAAS        R
Sbjct: 202 SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR  261

Query: 61  YVNILPYDHSRVHLTPVEGVFDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE  120
           YVNILPYDHSRVHLTPVEGVFDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
Sbjct: 262 YVNILPYDHSRVHLTPVEGVFDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE  321

Query: 121 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD  180
           QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
Sbjct: 322 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD  381

Query: 181 VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG  240
           VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
Sbjct: 382 VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG  441

Query: 241 TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE  300
           TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
Sbjct: 442 TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE  501
```

Guilt-by-Association: Caveats

- **Ensure that the effects of database size and composition have been accounted for**
- **Ensure that the function of the homology is not derived via invalid "transitive assignment''**
- **Ensure that the target sequence has all the key features associated with the function, e.g., active site and/or domain**

Law of Large Numbers

- **Suppose you are in a room with 365 other people**
- **Q: What is the prob that a specific person in the room has the same birthday as you?**
- **A: 1/365 = 0.3%**
- **Q: What is the prob that there is a person in the room having the same birthday as you?**
- **A: $1 - (364/365)^{365} = 63\%$**
- **Q: What is the prob that there are two persons in the room having the same birthday?**
- **A: 100%**

Interpretation of P-value

- **Seq. comparison progs, e.g. BLAST, often associate a P-value to each hit**
- **P-value is interpreted as prob that a random seq has an equally good alignment**
- **Suppose the P-value of an alignment is 10^{-6}**
- **If database has 10^7 seqs, then you expect $10^7 * 10^{-6} = 10$ seqs in it that give an equally good alignment**

$\Rightarrow$ **Need to correct for database size if your seq comparison prog does not do that!**

Lightning Does Strike Twice!

- **Roy Sullivan, a former park ranger from Virgina, was struck by lightning 7 times**
  - 1942 (lost big-toe nail)
  - 1969 (lost eyebrows)
  - 1970 (left shoulder seared)
  - 1972 (hair set on fire)
  - 1973 (hair set on fire & legs seared)
  - 1976 (ankle injured)
  - 1977 (chest & stomach burned)
- **September 1983, he committed suicide**

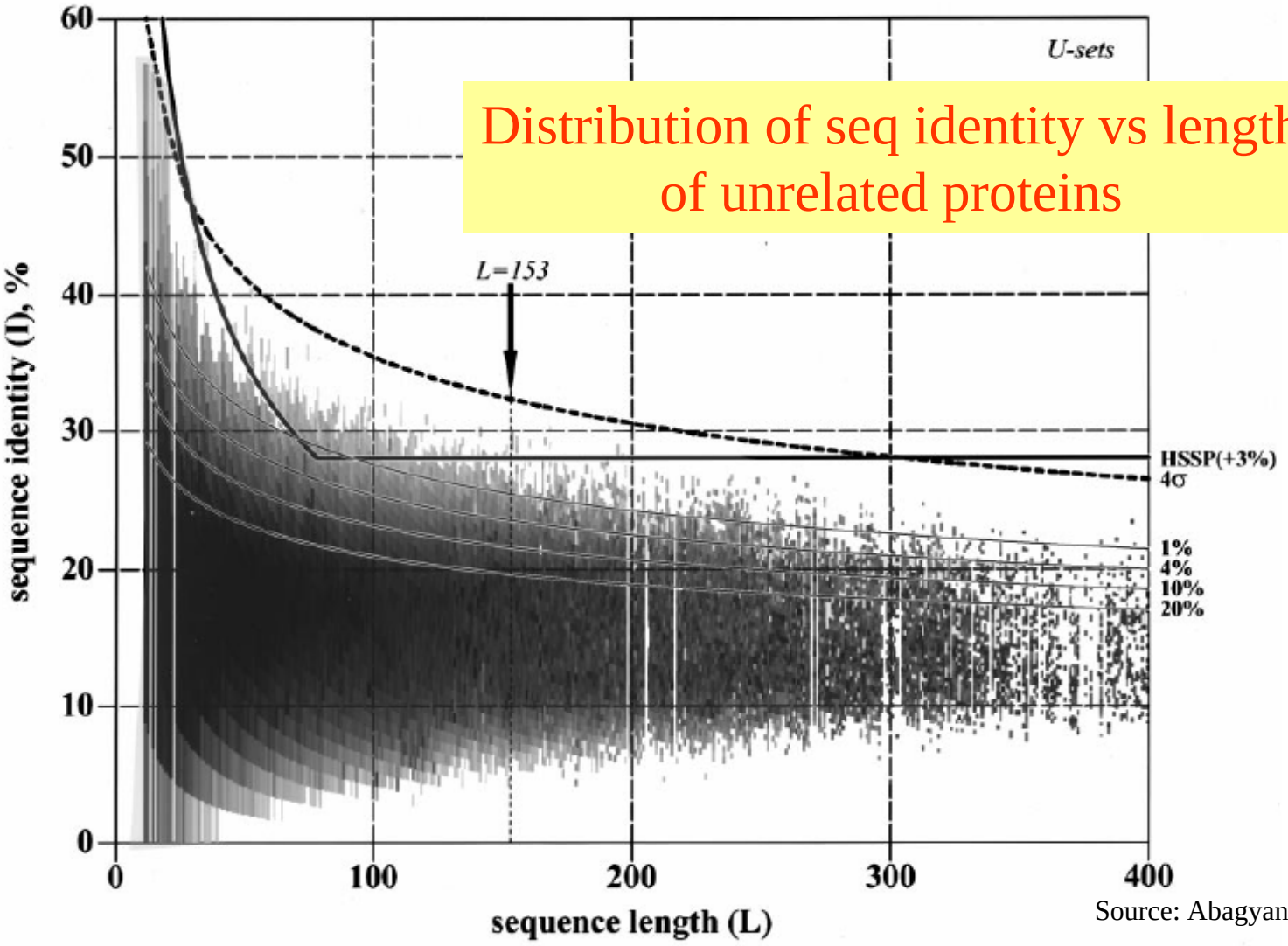
Cartoon: Ron Hipschman
Data: David Hand

Effect of Seq Compositional Bias

- One fourth of all residues in protein seqs occur in regions with biased amino acid composition
- Alignments of two such regions achieves high score purely due to segment composition
- While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments
- BLAST employs the SEG algorithm to filter low complexity regions from proteins before executing a search

Source: NCBI

Effect of Sequence Length



Distribution of seq identity vs length of unrelated proteins

Source: Abagyan & Batalov

Important Unsolved Challenges

- **What if there is no useful sequence homolog?**
- **Guilt by other types of association!**
  - Domain modeling (e.g., HMMPFAM)
  - Similarity of dissimilarities (e.g., SVM-PAIRWISE)
  - Similarity of phylogenetic profiles
  - Similarity of subcellular co-localization & other physico-chemico properties(e.g., PROTFUN)
  - Similarity of gene expression profiles
  - Similarity of protein-protein interaction partners
  - …
  - Fusion of multiple types of info

Similarity of Dissimilarities

	$orange_1$	$banana_1$	…
$apple_1$	Color = red vs orange Skin = smooth vs rough Size = small vs small Shape = round vs round	Color = red vs yellow Skin = smooth vs smooth Size = small vs small Shape = round vs oblong	…
$apple_2$	Color = red vs orange Skin = smooth vs rough Size = small vs small Shape = round vs round	Color = red vs yellow Skin = smooth vs smooth Size = small vs small Shape = round vs oblong	…
$orange_2$	Color = orange vs orange Skin = rough vs rough Size = small vs small Shape = round vs round	Color = orange vs yellow Skin = rough vs smooth Size = small vs small Shape = round vs oblong	..
…	…	…	…

SVM-Pairwise Framework

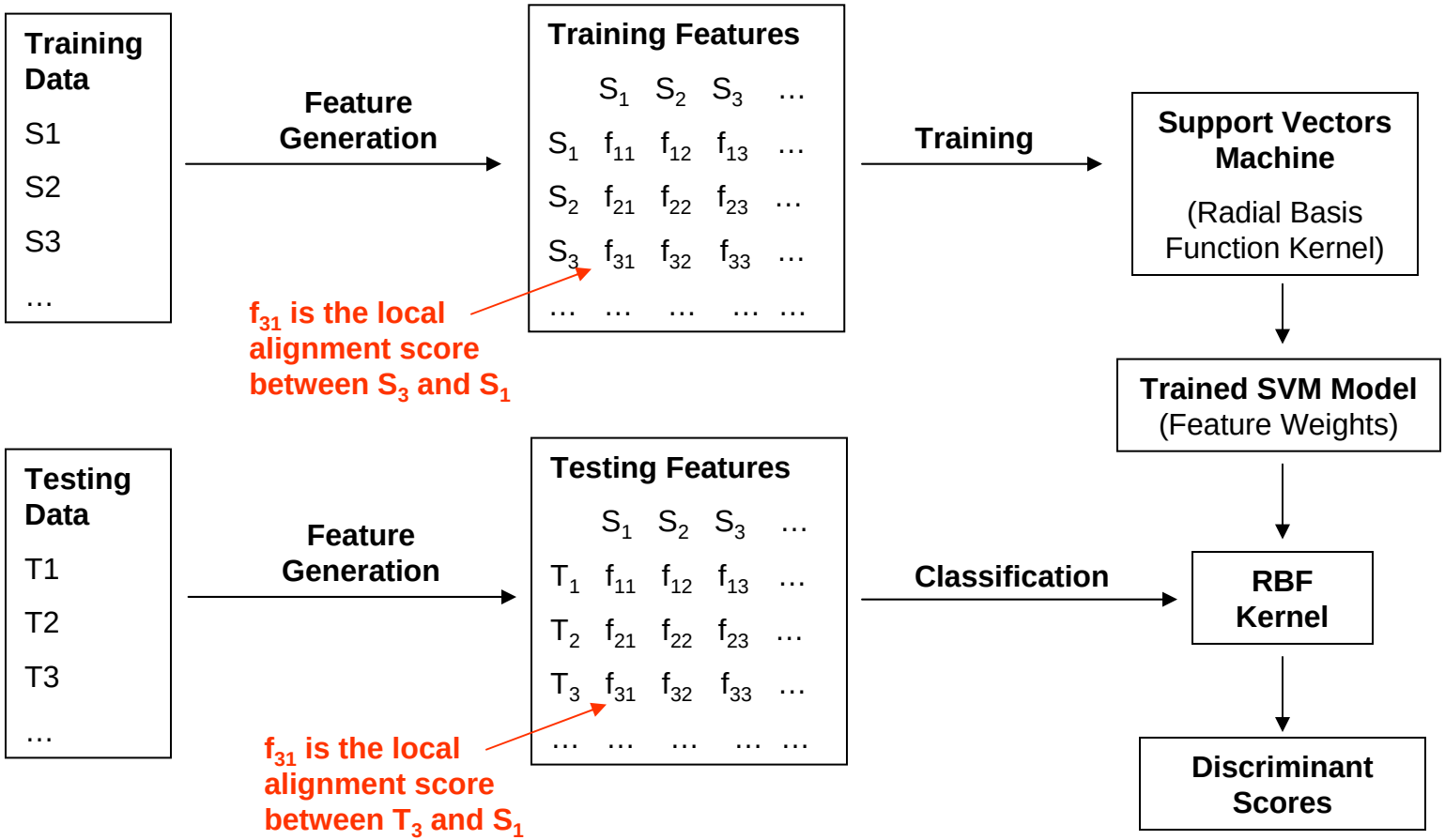
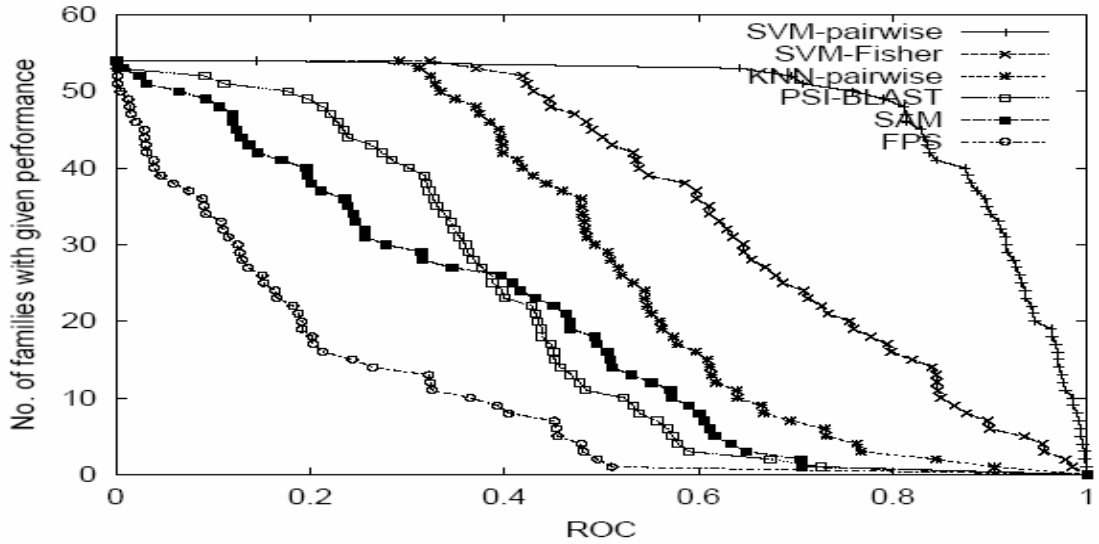
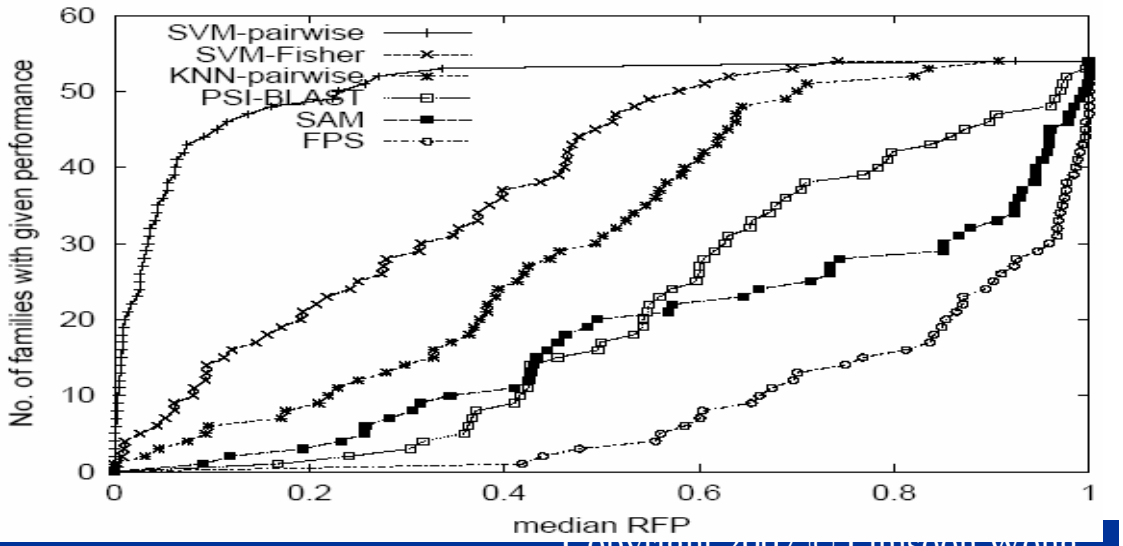


Image credit: Kenny Chua

Performance of SVM-Pairwise

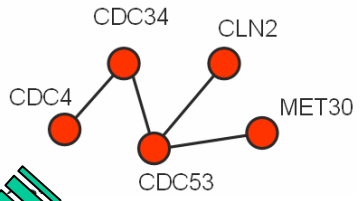
- **Receiver Operating Characteristic (ROC)**
  - The area under the curve derived from plotting true positives as a function of false positives for various thresholds.
- **Rate of median False Positives (RFP)**
  - The fraction of negative test examples with a score better or equals to the median of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

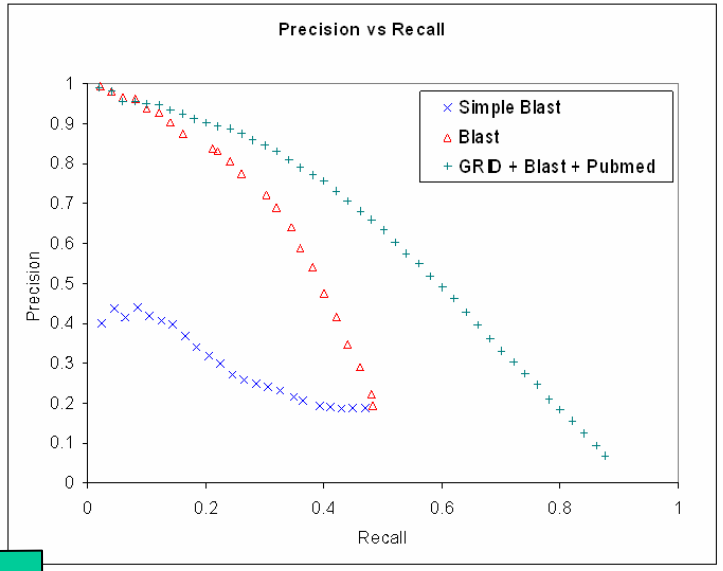




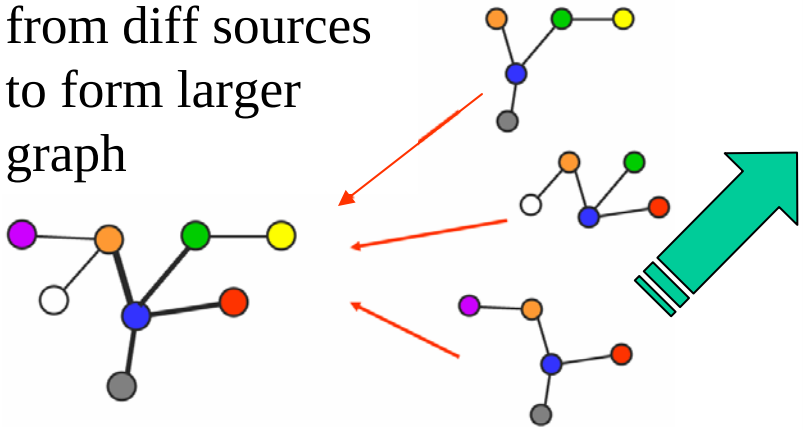
Fusion of Multiple Types of Info

Model each data source into a graph





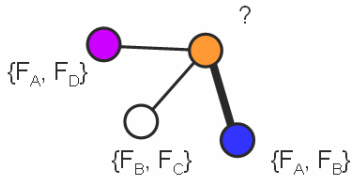
Combine graphs from diff sources to form larger graph



Function Prediction

- Weighted Average

$$S_f(u) = \frac{\sum_{v \in N_u} e_f(v) \times r_{uv}}{1 + \sum_{v \in N_u} r_{uv}}$$



$S_f(u)$ is the score of function f for protein u
$e_f(v)$ is 1 if protein v has function f, 0 otherwise
N_u is the set of neighbours of protein u
r_{uv} is the link confidence between protein u and protein v

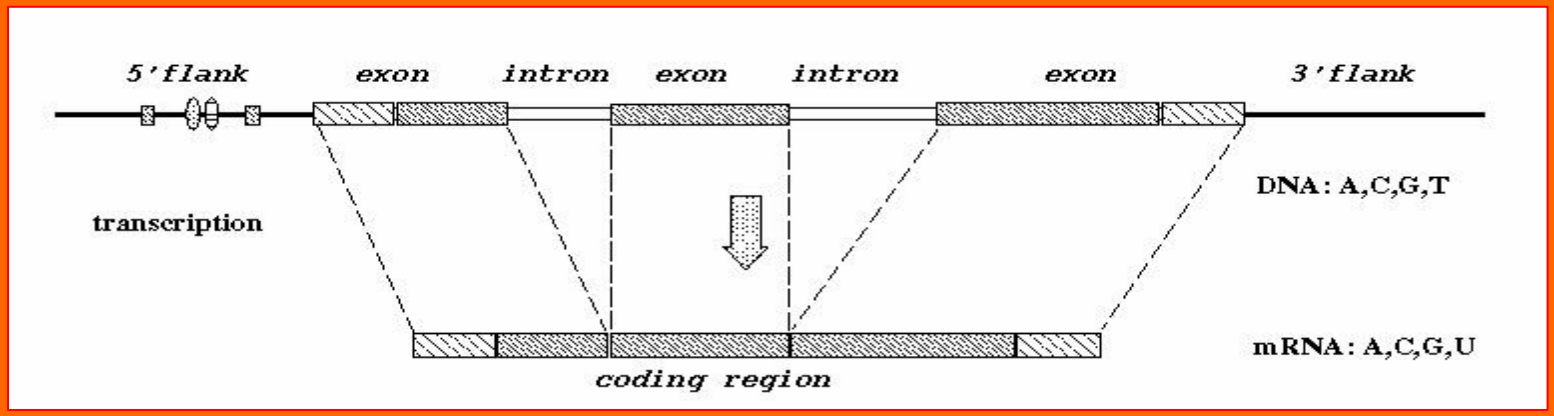
Recommended Readings

- Wong, *The Practical Bioinformatician*, 2004, ICP. Chapters 10 & 19
- Chua et al., Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22:1623-1630, 2006.
- Bateman et al., The Pfam protein families database. *NAR*, 32:D138-D141, 2004
- Jaakkola et al., A discriminative framework for detecting remote protein homologies. *JCB*, 7:95-114, 2000
- S.F.Altschul et al. “Gapped BLAST and PSI-BLAST: A new generation of protein database search programs”, *NAR*, 25(17):3389--3402, 1997
- Pellegrini et al., Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *PNAS*, 96:4285-4288, 1999

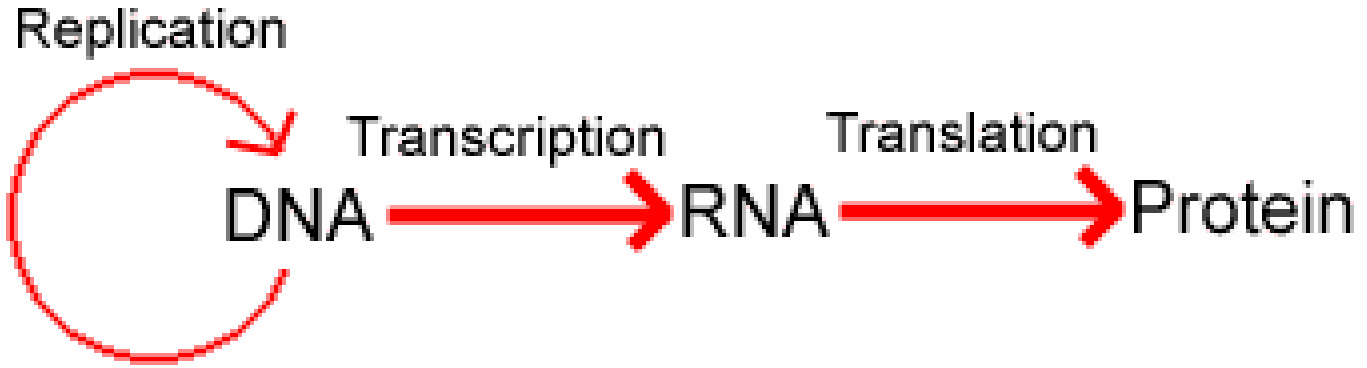
5-Minute Break?

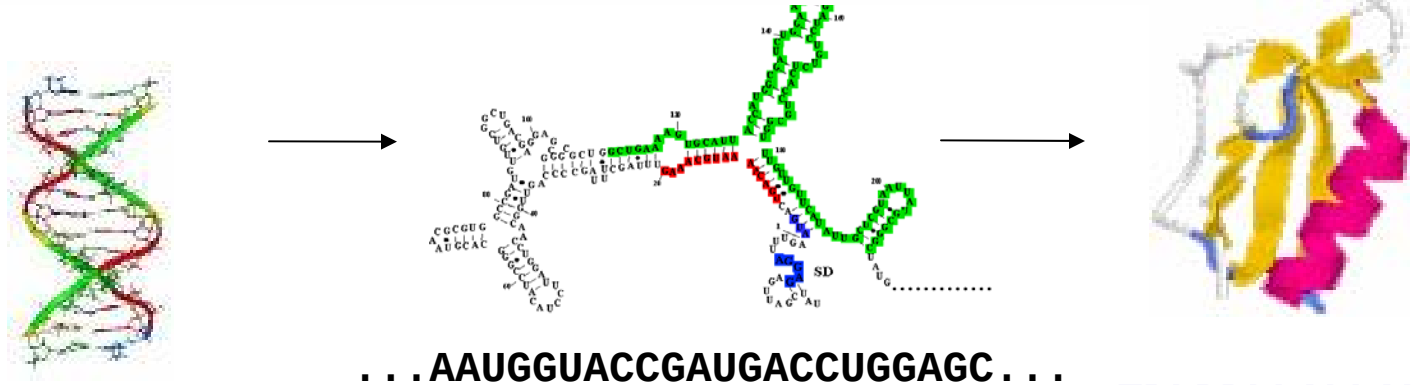
60 min

Gene Feature Recognition



Central Dogma



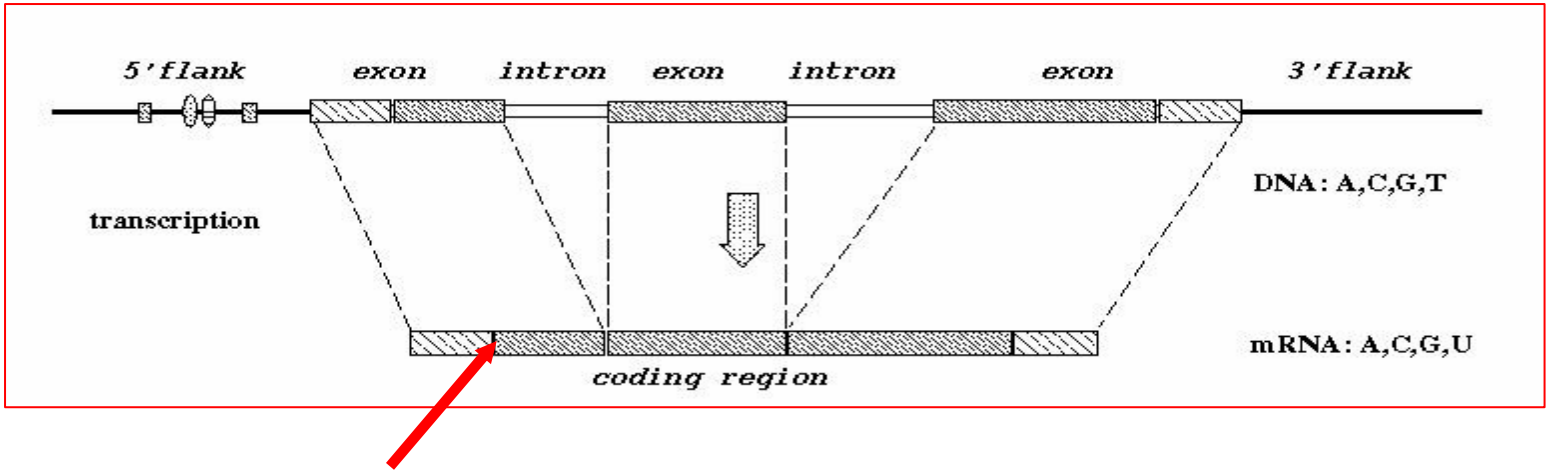


. . .AATGGTACCGATGACCTG. . .

. . .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
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE     240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

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

Approach

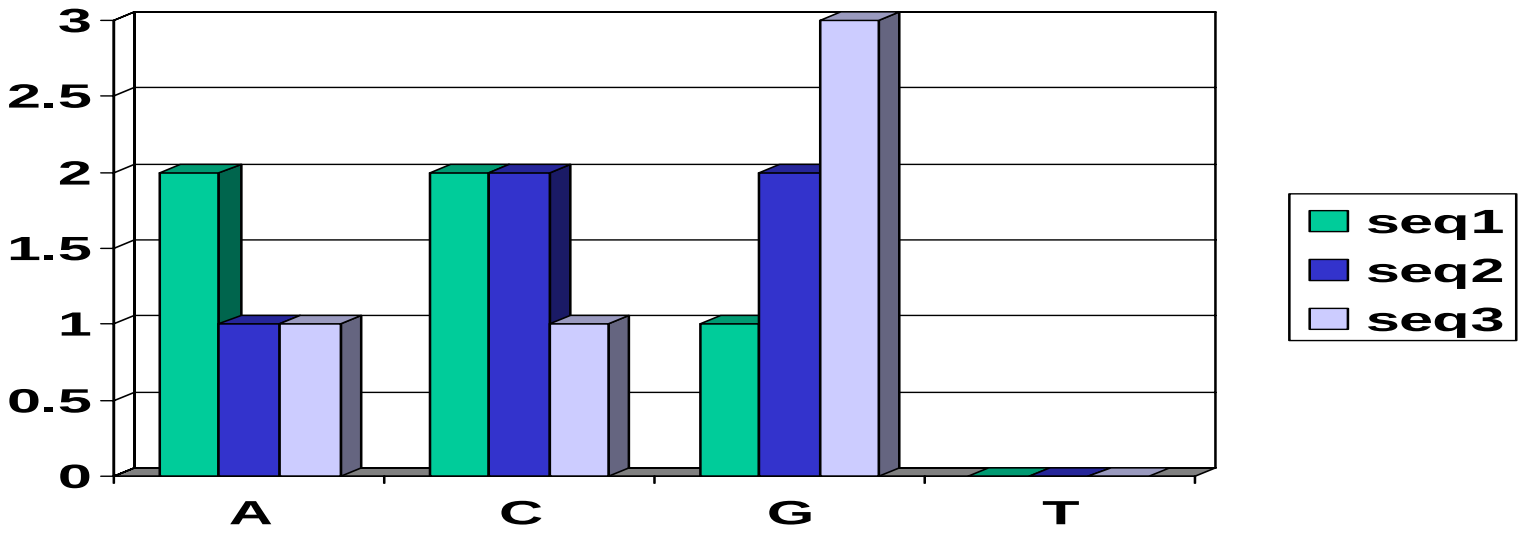
- **Training data gathering**
- **Signal generation**
  - k-grams, distance, domain know-how, ...
- **Signal selection**
  - Entropy, $\chi 2$, CFS, t-test, domain know-how...
- **Signal integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...

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
```

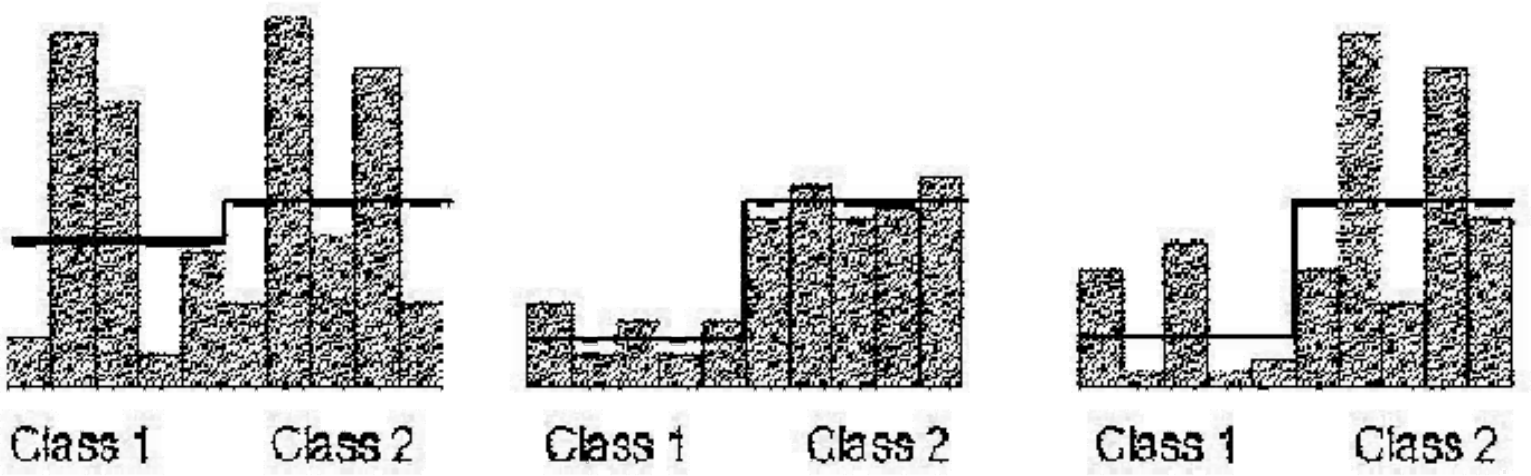
- **Window = ±100 bases**
- **In-frame, downstream**
  - GCT = 1, TTT = 1, ATG = 1…
- **Any-frame, downstream**
  - GCT = 3, TTT = 2, ATG = 2…
- **In-frame, upstream**
  - GCT = 2, TTT = 0, ATG = 0, ...

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**

Signal Selection (Basic Idea)

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

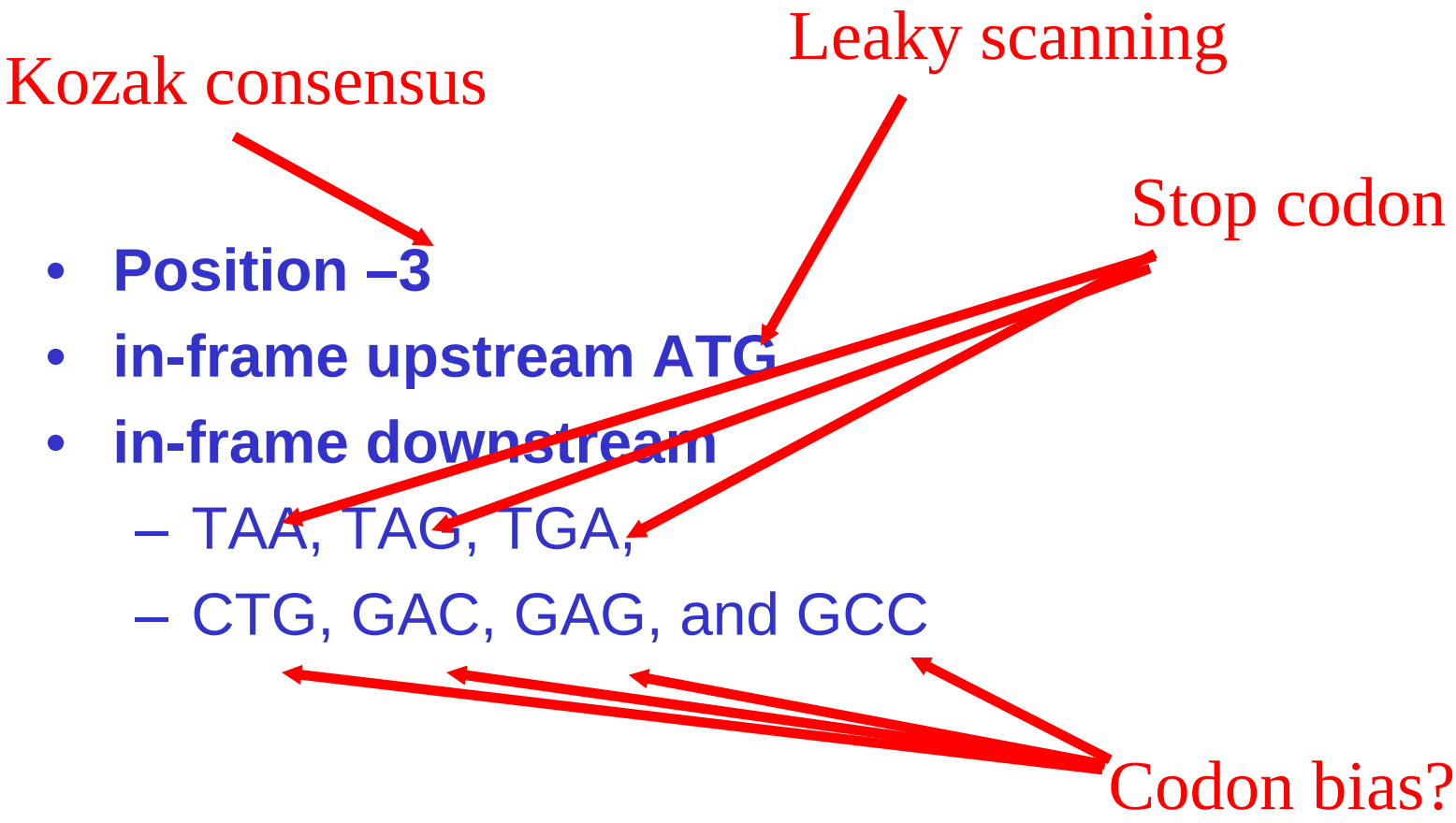


Signal Selection (e.g., CFS)

- **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



- **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

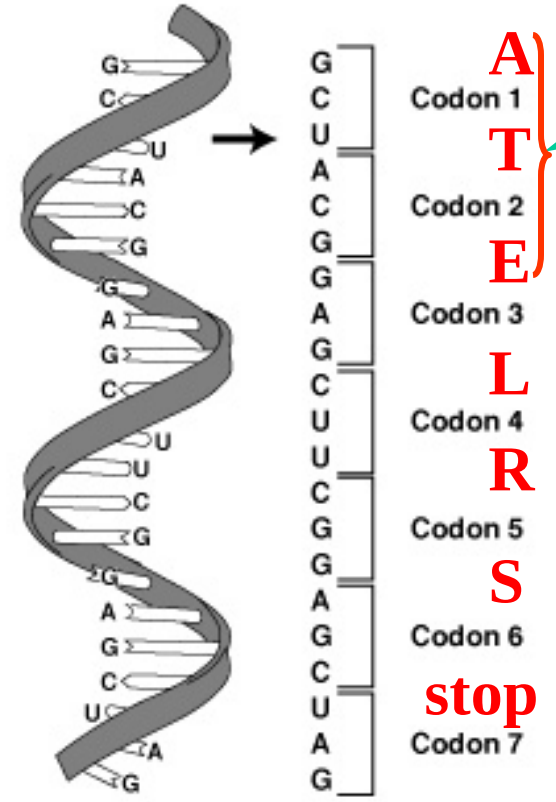
	TP/(TP + FN)	TN/(TN + FP)	TP/(TP + FP)	Accuracy
Naïve Bayes	84.3%	86.1%	66.3%	85.7%
SVM	73.9%	93.2%	77.9%	88.5%
Neural Network	77.6%	93.2%	78.8%	89.4%
Decision Tree	74.0%	94.4%	81.1%	89.4%

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%**

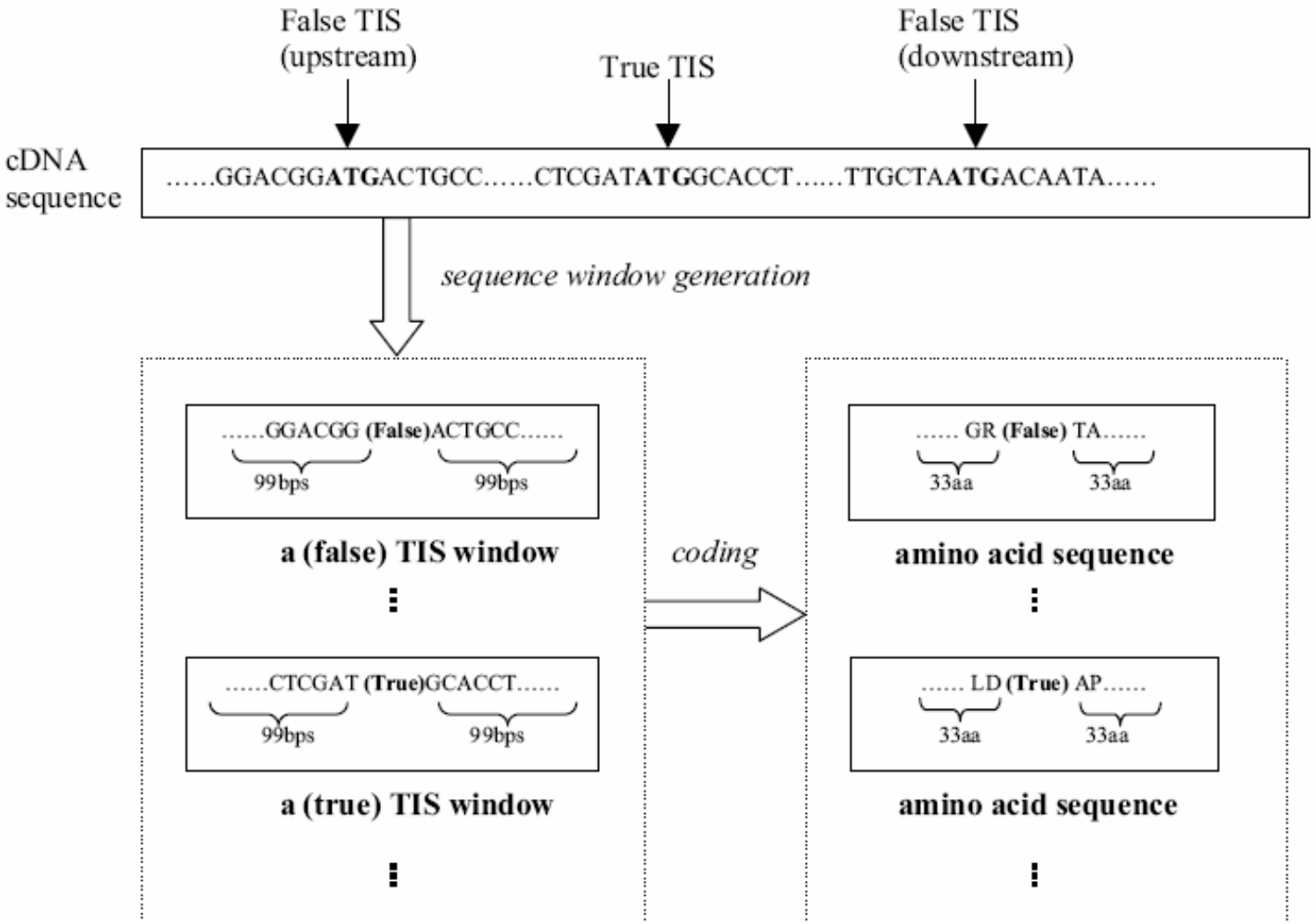
mRNA→Protein



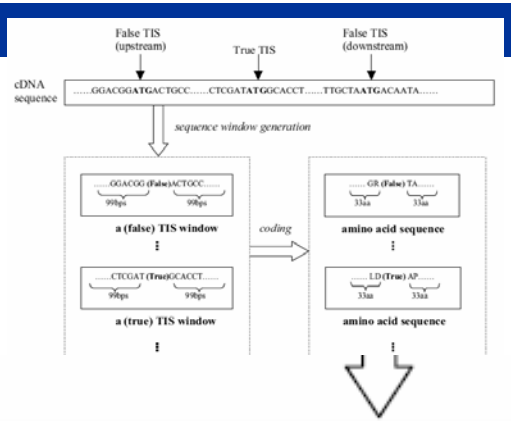
How about using k-grams from the translation?

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

Amino-Acid Features

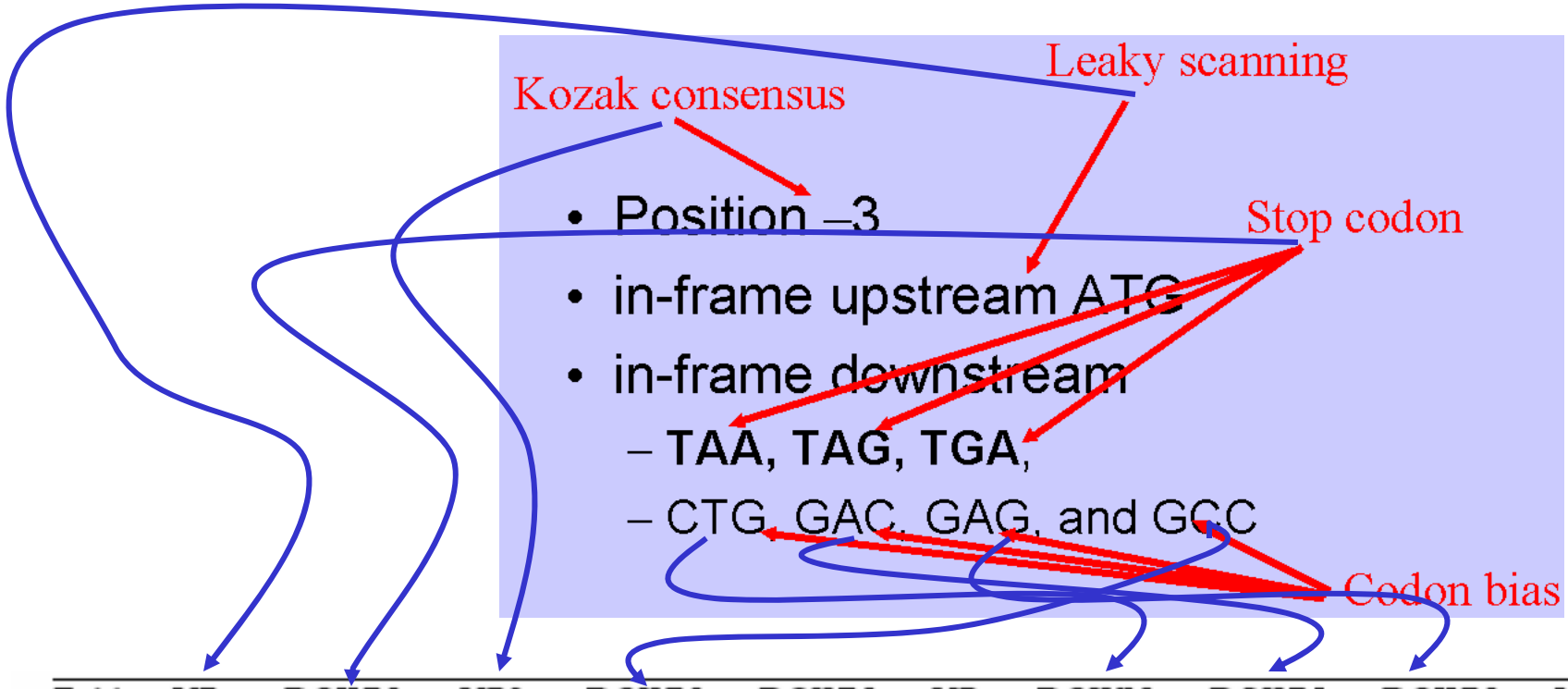


Amino-Acid Features



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, 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)



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

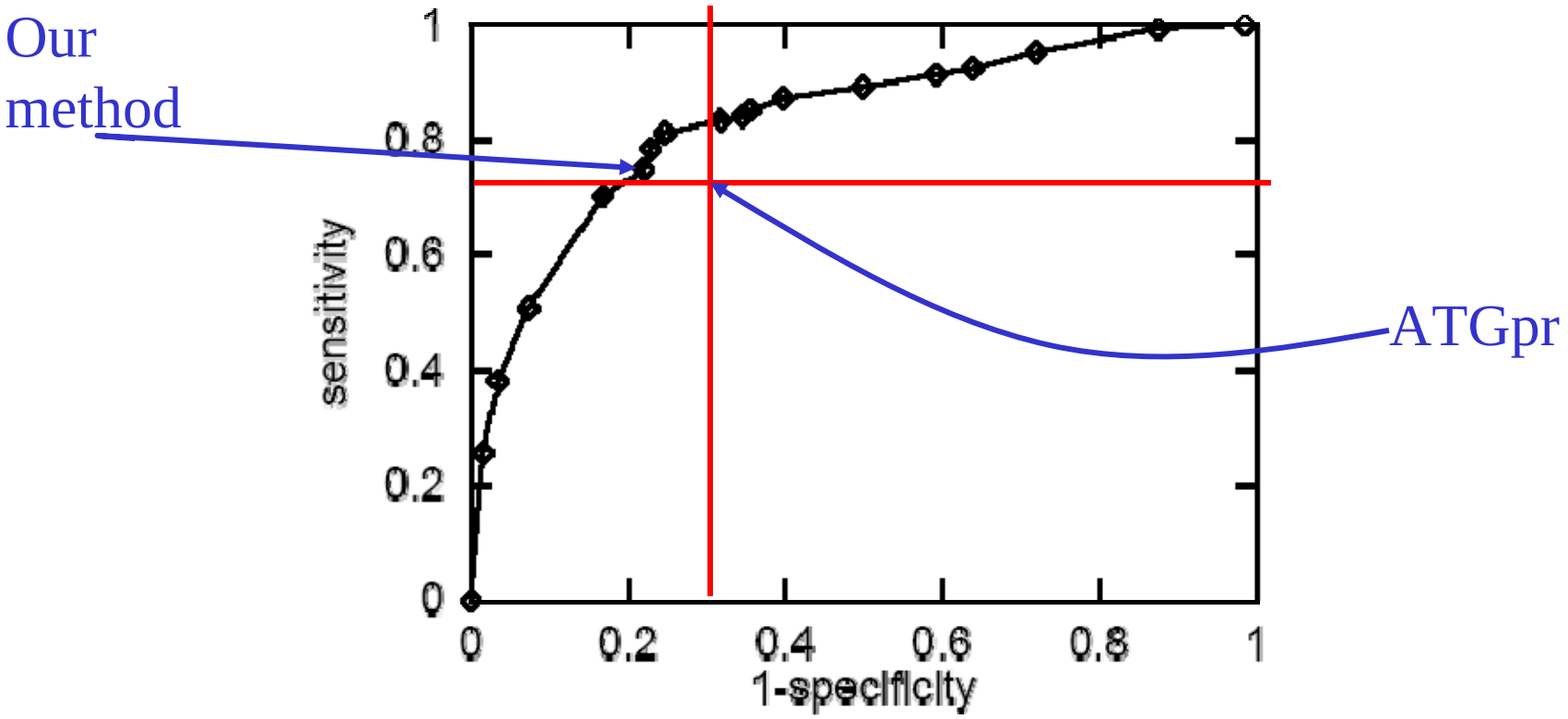
- **A. Hatzigeorgiou:**
  - 480 fully sequenced human cDNAs
  - 188 left after eliminating sequences similar to training set (Pedersen & Nielsen's)
  - 3.42% of ATGs are TIS
- **Our own:**
  - well characterized human gene sequences from chromosome X (565 TIS) and chromosome 21 (180 TIS)

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

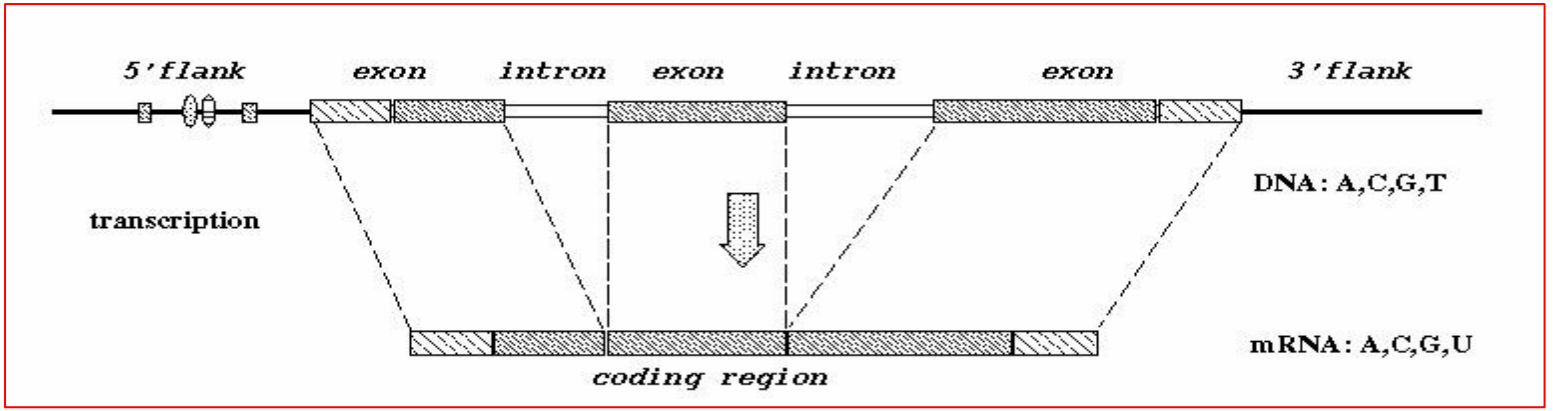
Validation Results (on Chr X and Chr 21)



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

Important Unsolved Challenges

- **Other gene features: TSS, PolyA, Splicing, …**

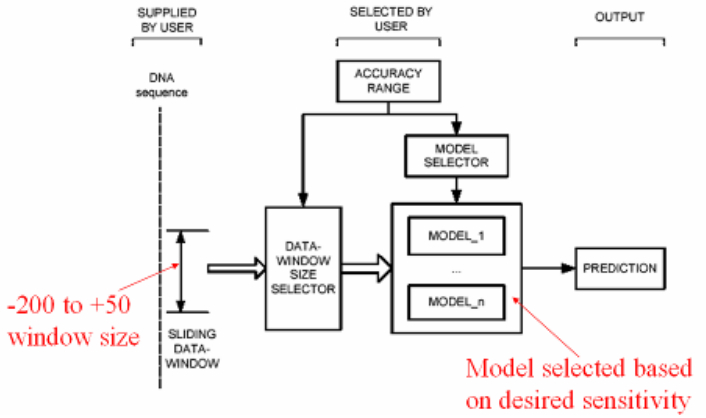


- **Gene regulation: TFBS, EREs, miRNA targets, …**

- **"Dark matters": miRNA & non-coding genes**

TSS: Is State of the Art Good Enough?

Structure of Dragon Promoter Finder



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

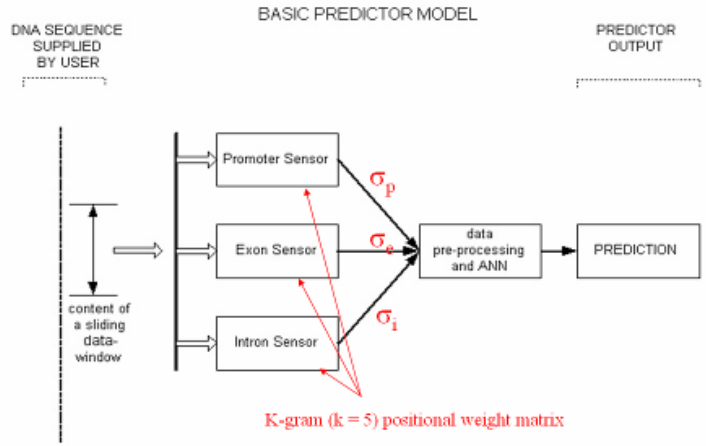
$$\sigma = \frac{\left(\sum_{i=1}^{L-4} p_j^i \otimes f_{j,i}\right)}{\left(\sum_{i=1}^{L-4} \max_j f_{j,i}\right)}, \qquad p_j^i \otimes f_{j,i} = \begin{cases} f_{j,i}, \text{ if } p_i = p_j^i \\ 0, \text{ if } p_i \neq p_j^i \end{cases},$$

Pentamer at i^{th} position in input

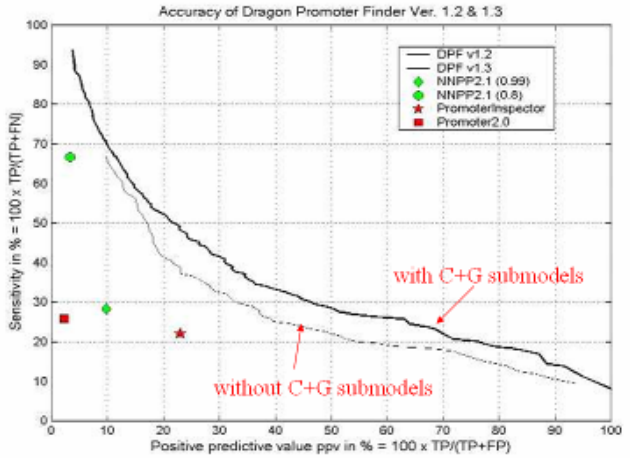
Frequency of jth pentamer at ith position in training window

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

Data Analysis Within Submodel



Accuracy Comparisons



TFs and TFBSs

- **Recognize TFBS**
  - Short & degenerate
  - Incomplete erroneous data
- **Characterize type of regulation, e.g., suppressor vs enhancer**
- **Identify synergistic TF pairs**
  - Genes that are coordinately bound are coordinately expressed
  - Requires data over many time points
  - TFs may interact only under certain conditions

- **A method to identify synergistic TF pairs in cell cycle**
  - Study if a pair of TFs are assoc with the same genes more often than random
  - For each significant pair, test whether there is a phase in cell cycle where the common associated genes are significantly diff in other cell cycle phases

$\Rightarrow$ **Quite primitive. Lots more work needed!**

miRNA: New Frontier in Molecular Biology

- **Impt functions of miRNA:**
  - Control of cell fate and fat metabolism in flies
  - Neuronal patterning in nematodes
  - Modulation of hematopoietic lineage differentiation in mammals
  - control of leaf and flower development in plants
- **By regulating gene expr in binding targets:**
  - Repression/inhibition of translation
  - Degradation of mRNA
- **miRNA identification thru direct cloning has limitations:**
  - miRNAs may be expressed only in certain tissues or at certain times
  - Expr levels vary greatly
  - Degradation products from mRNAs and other endogenous non-coding RNAs coexist w/ miRNAs and are sometimes dominant in small RNA molecule samples extracted from cells

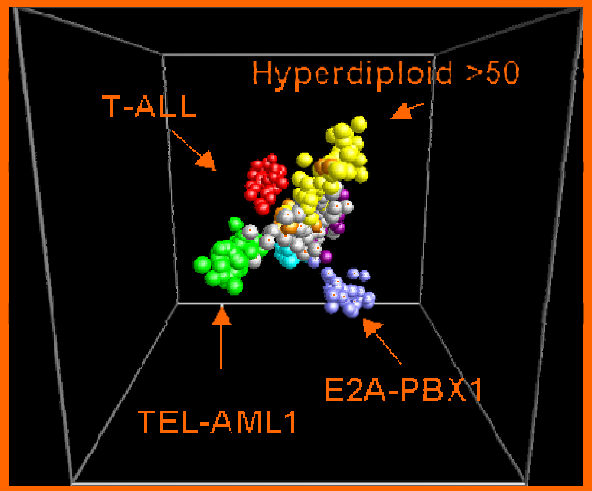
Recommended Readings

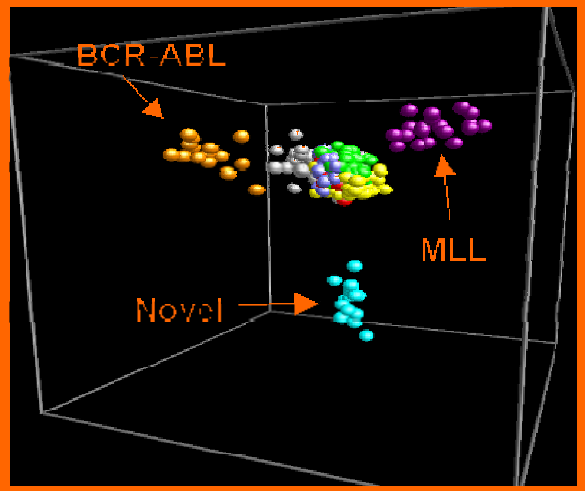
- Wong, *The practical Bioinformatician*, 2004, ICP. Chapters 4, 5, 6, & 7
- Liu & Wong, Data mining tools for biological sequences. *JBCB*, 1:139-168, 2003
- Tsai et al., Statistical methods for identifying yeast cell cycle transcription factors. *PNAS*, 102:13532-13537, 2005
- Yang et al., Identification of microRNA precursors via SVM. *APBC* 2006, pages 267-276
- Zheng et al., Exploring Essential Attributes for Detecting MicroRNA Precursors from Background Sequences. *VLDB 2006 Workshop on Data Mining in Bioinformatics*, pages 131--145.

5-Minute Break?

70 min

Disease Diagnosis, Treatment, & Understanding





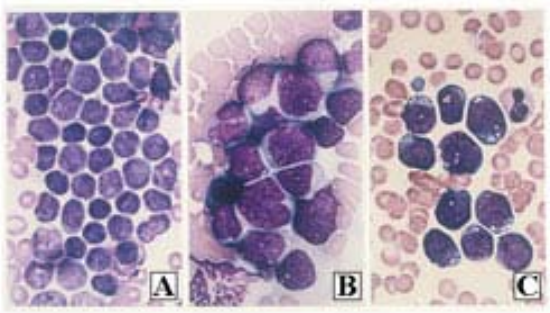
NUS National University of Singapore

Childhood Acute Lymphoblastic Leukemia

- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50**

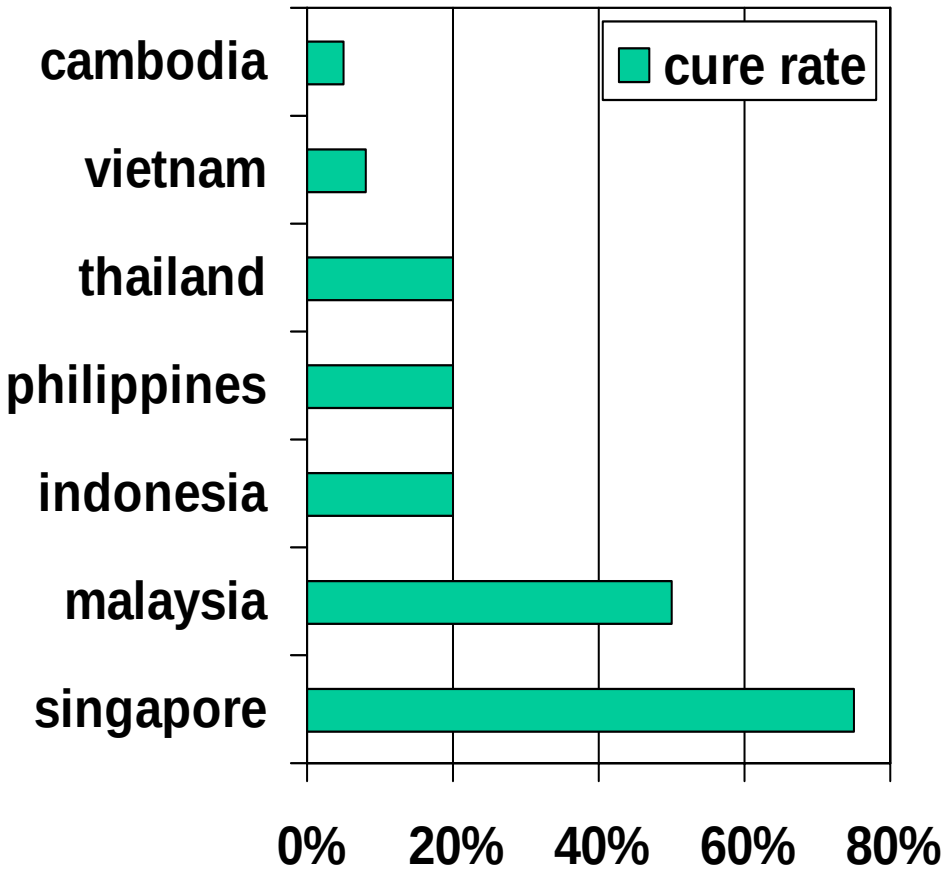
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse

- **The subtypes look similar**



- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics

Childhood ALL Cure Rates

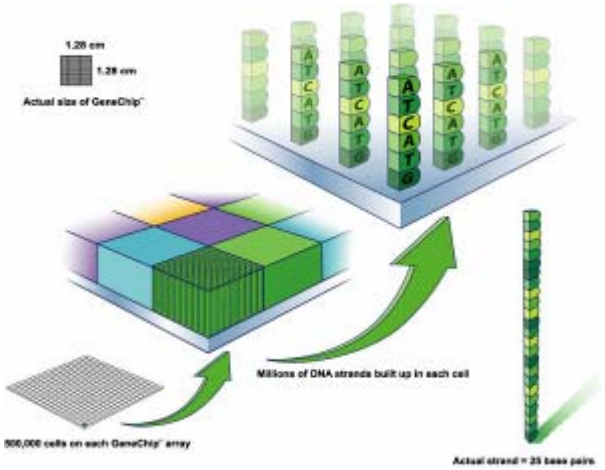


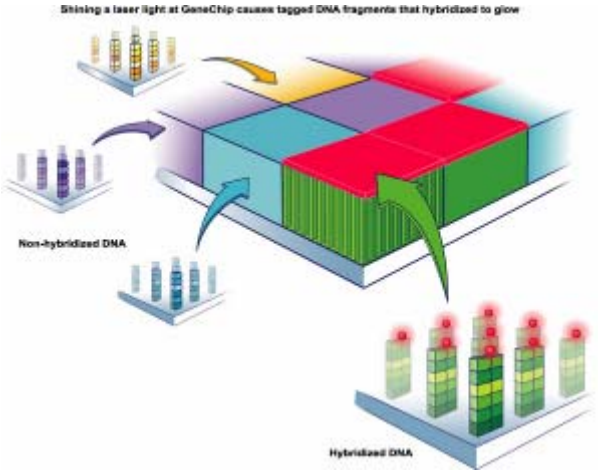
- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists**

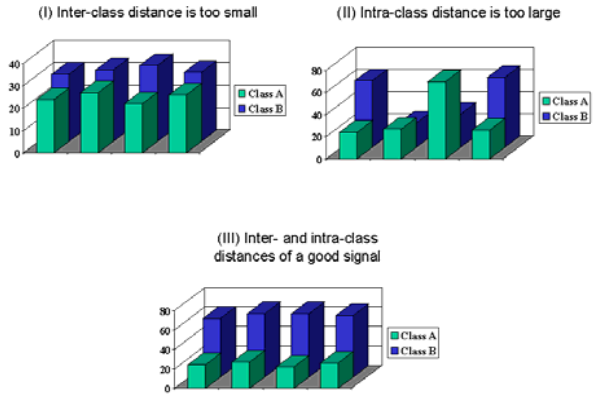
⇒ **Not available in less advanced ASEAN countries**

⇒ **Can we have a single-test easy-to-use platform instead?**

Single-Test Easy-to-Use Platform





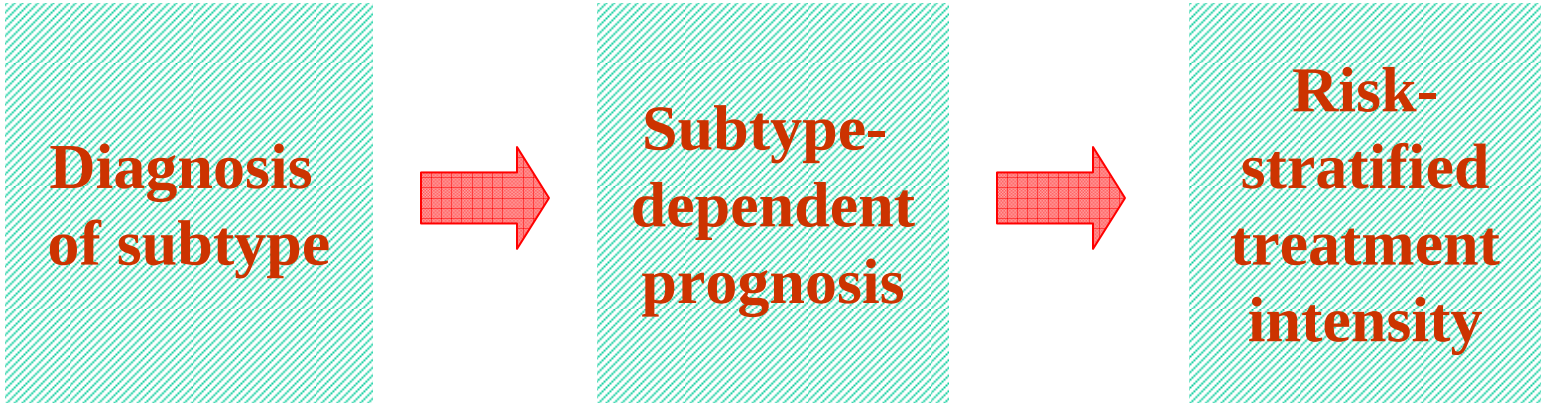


	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAv	Avg Diff	Abs Call	
AFFX-MurI	5	2	19	297.5	A	M16762 Mouse int
AFFX-MurI	3	2	19	554.2	A	M37897 Mouse int
AFFX-MurI	4	2	19	308.6	A	M25892 Mus musc
AFFX-MurI	1	3	19	141	A	M83649 Mus musc
AFFX-BioB	13	1	19	9340.6	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	25838.5	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 Bacteriopl
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacteriopl
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

A Sample Affymetrix GeneChip Data File (U95A)

	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAv	Avg Diff	Abs Call	
AFFX-MurI	5	2	19	297.5	A	M16762 Mouse interleukin 2 (IL-2) gene, exon 4
AFFX-MurI	3	2	19	554.2	A	M37897 Mouse interleukin 10 mRNA, complete cds
AFFX-MurI	4	2	19	308.6	A	M25892 Mus musculus interleukin 4 (Il-4) mRNA, comp
AFFX-MurF	1	3	19	141	A	M83649 Mus musculus Fas antigen mRNA, complete
AFFX-BioB	13	1	19	9340.6	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	15	0	19	12862.4	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	12	0	19	8716.5	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC protein (-5 and -3 represent transcr
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC protein (-5 and -3 represent transcr
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD gene dethiobiotin synthetase (-5 ar
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD gene dethiobiotin synthetase (-5 ar
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacteriophage P1 cre recombinase protein (-5
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacteriophage P1 cre recombinase protein (-5
AFFX-BioB	7	5	18	-483	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	5	4	18	313.7	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioB	7	6	20	-1016.2	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r

Overall Strategy



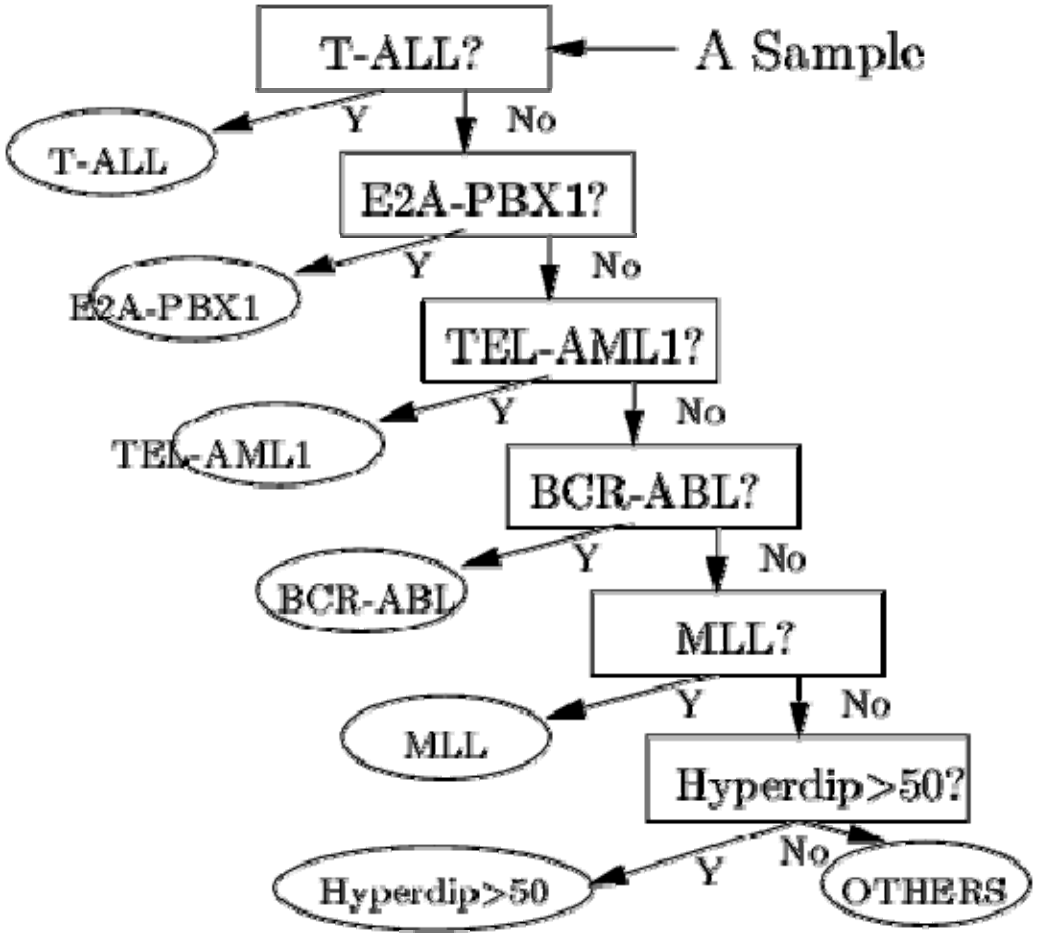
- For each subtype, select genes to develop classification model for diagnosing that subtype
- For each subtype, select genes to develop prediction model for prognosis of that subtype

Subtype Diagnosis by PCL

- **Gene expression data collection**
- **Gene selection by $\chi 2$**
- **Classifier training by emerging pattern**
- ~~**Classifier tuning (optional for some machine learning methods)**~~
- **Apply classifier for diagnosis of future cases by PCL**

Childhood ALL Subtype Diagnosis Workflow

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

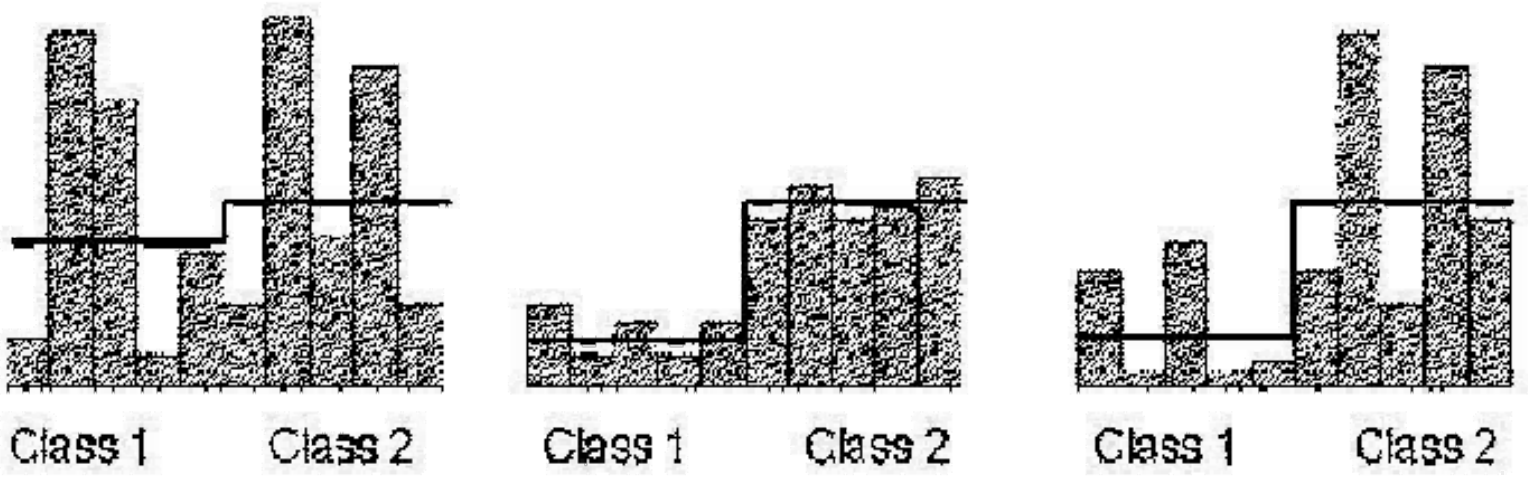


Training and Testing Sets

Paired datasets	Ingredients	Training	Testing
T-ALL vs OTHERS1	OTHERS1 ={E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}	28 vs 187	15 vs 97
E2A-PBX1 vs OTHERS2	OTHERS2 = {TEL-AML1, BCR-ABL Hyperdip>50, MLL, OTHERS}	18 vs 169	9 vs 88
TEL-AML1 vs OTHERS3	OTHERS3 = {BCR-ABL Hyperdip>50, MLL, OTHERS}	52 vs 117	27 vs 61
BCR-ABL vs OTHERS4	OTHERS4 = {Hyperdip>50, MLL, OTHERS}	9 vs 108	6 vs 55
MLL vs OTHERS5	OTHERS5 = {Hyperdip>50, OTHERS}	14 vs 94	6 vs 49
Hyperdip>50 vs OTHERS	OTHERS = {Hyperdip47-50, Pseudodip, Hypodip, Normo}	42 vs 52	22 vs 27

Signal Selection Basic Idea

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



Signal Selection by $\chi 2$

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} ($E_{ij} = R_i * C_j / N$).

Emerging Patterns

- **An emerging pattern is a set of conditions**
  - usually involving several features
  - that most members of a class satisfy
  - but none or few of the other class satisfy
- **A jumping emerging pattern is an emerging pattern that**
  - some members of a class satisfy
  - but no members of the other class satisfy
- **We use only jumping emerging patterns**

Examples

Patterns	Frequency (P)	Frequency(N)
{9, 36}	38 instances	0
{9, 23}	38	0
{4, 9}	38	0
{9, 14}	38	0
{6, 9}	38	0
{7, 21}	0	36
{7, 11}	0	35
{7, 43}	0	35
{7, 39}	0	34
{24, 29}	0	34

Easy interpretation

Reference number 9: the expression of gene 37720_at > 215
Reference number 36: the expression of gene 38028_at <= 12

PCL: Prediction by Collective Likelihood

- Let $EP_1^P, \ldots, EP_i^P$ be the most general EPs of D^P in descending order of support.
- Suppose the test sample T contains these most general EPs of D^P (in descending order of support):

$$EP_{i_1}^P, EP_{i_2}^P, \cdots, EP_{i_x}^P$$

- Use k top-ranked most general EPs of D^P and D^N. Define the score of T in the D^P class as

$$score(T, D^P) = \sum_{m=1}^{k} \frac{frequency(EP_{i_m}^P)}{frequency(EP_m^P)}$$

- Ditto for $score(T, D^N)$.
- If $score(T, D^P) > score(T, D^N)$, then T is class P. Otherwise it is class N.

PCL Learning

Top-Ranked EPs in Positive class	Top-Ranked EPs in Negative class
EP_1^P (90%)	EP_1^N (100%)
EP_2^P (86%)	EP_2^N (95%)
.	.
.	.
EP_n^P (68%)	EP_n^N (80%)

The idea of summarizing multiple top-ranked EPs is intended to avoid some rare tie cases

PCL Testing

Most freq EP of pos class in the test sample

$$Score^P = EP_1^{P'} / EP_1^{P} + \ldots + EP_k^{P'} / EP_k^{P}$$

Most freq EP of pos class

Similarly,

$$Score^N = EP_1^{N'} / EP_1^{N} + \ldots + EP_k^{N'} / EP_k^{N}$$

If $Score^P > Score^N$, then positive class, Otherwise negative class

Accuracy of PCL (vs. other 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

Understandability of PCL

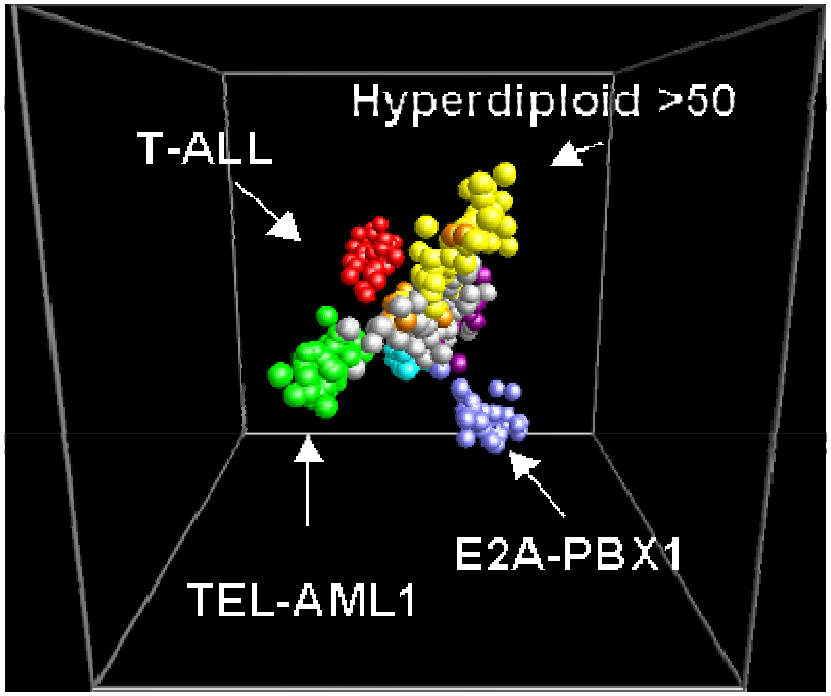
- **E.g., for T-ALL vs. OTHERS, one ideally discriminatory gene 38319_at was found, inducing these 2 EPs**

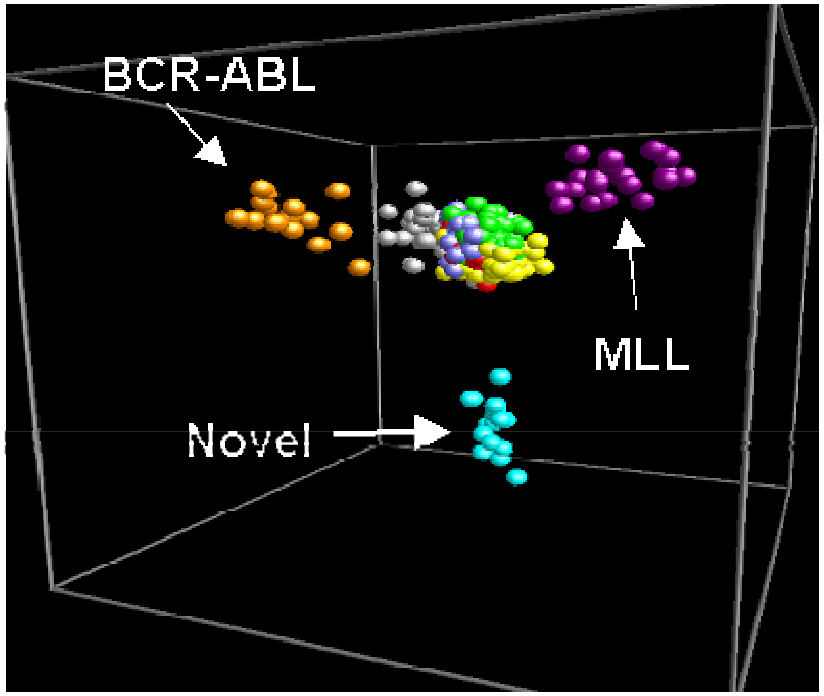
$$\{gene\_(38\,319\_at)\,@\,(-\infty,\ 15\,975.6)\} \text{ and}$$
$$\{gene\_(38\,319\_at)\,@\,[15\,975.6,\ +\infty)\}.$$

- **These give us the diagnostic rule**

If the expression of 38 319_*at* is less than 15 975.6, then
this ALL sample must be a T-ALL.
Otherwise it must be a subtype in OTHERS1.

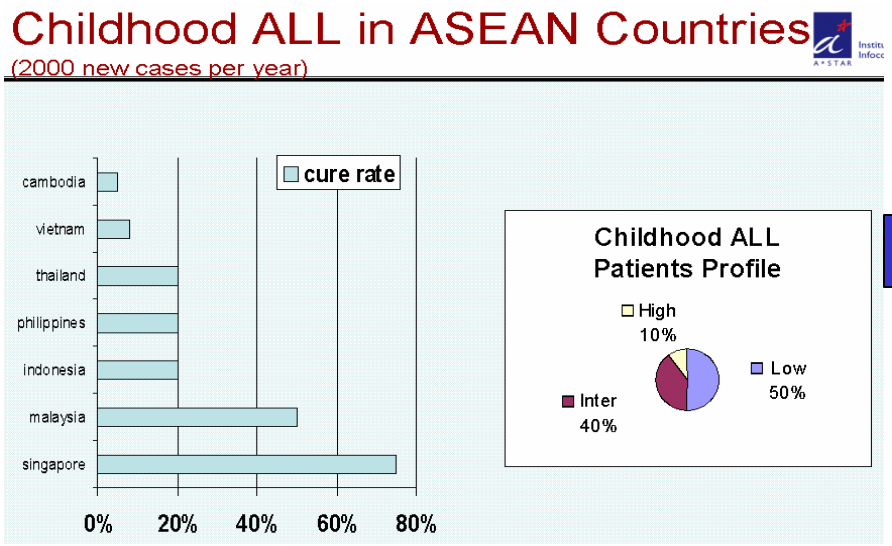
Multidimensional Scaling Plot for Subtype Diagnosis





Obtained by performing PCA on the 20 genes chosen for each level

Impact



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%

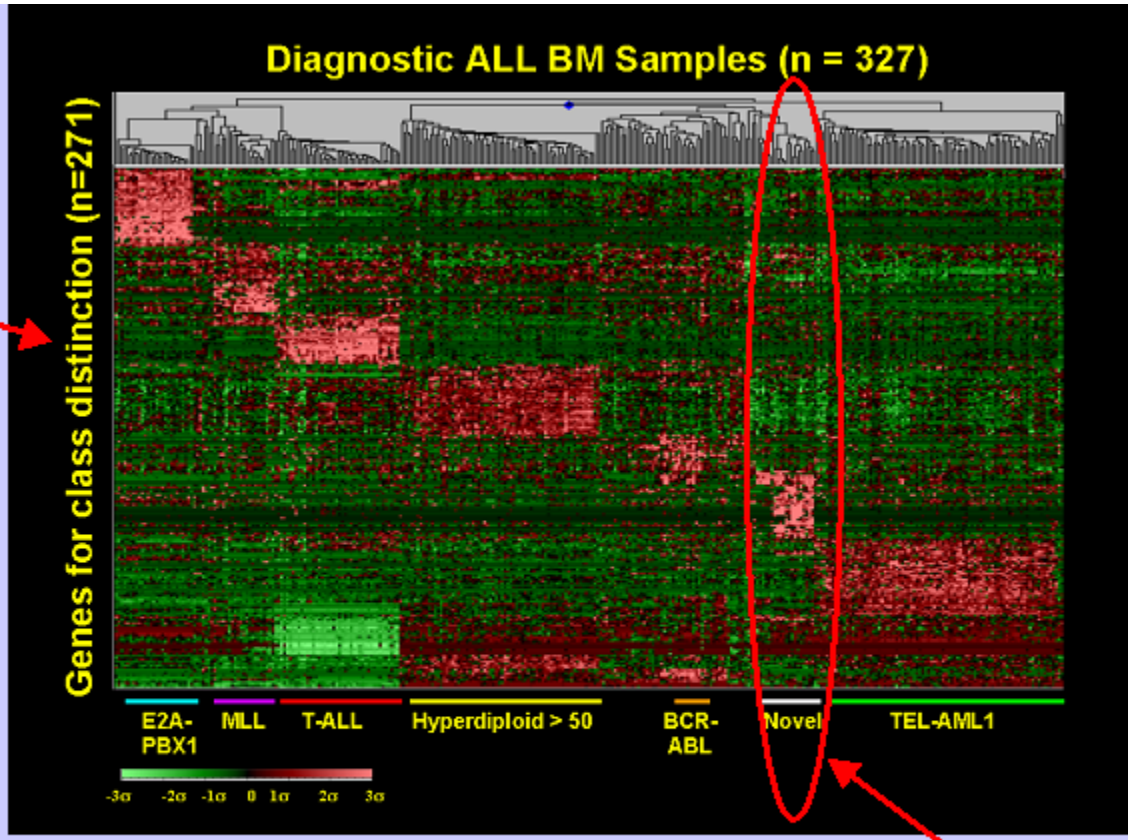
• Less relapse

• Less side effects

• Save US$51.6m/yr

Is there a new subtype?

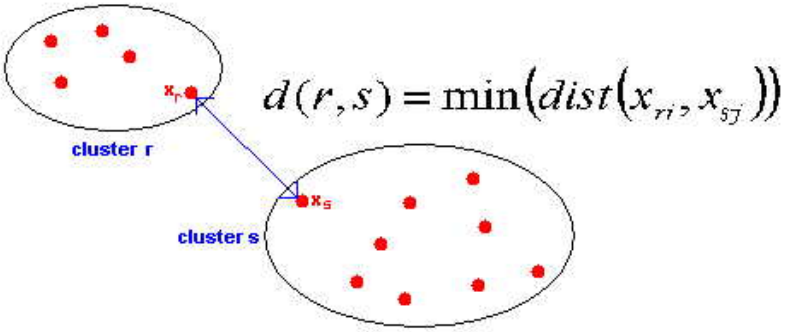
- **Hierarchical clustering of gene expression profiles reveals a novel subtype of childhood ALL**

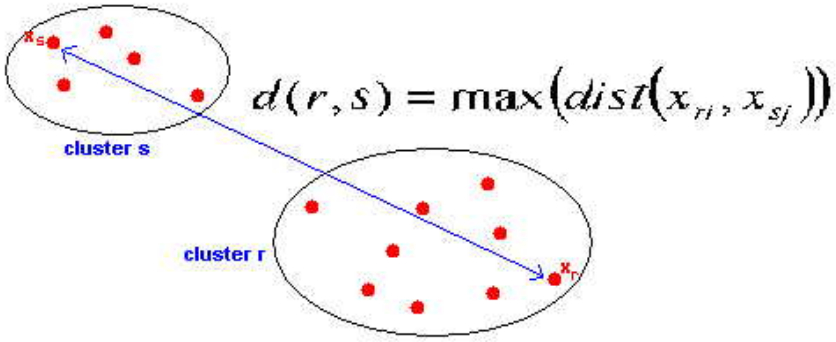


Hierarchical Clustering

- **Assign each item to its own cluster**
  - If there are N items initially, we get N clusters, each containing just one item
- **Find the "most similar" pair of clusters, merge them into a single cluster, so we now have one less cluster**
  - "Similarity" is often defined using
    - **Single linkage**
    - **Complete linkage**
    - **Average linkage**
- **Repeat previous step until all items are clustered into a single cluster of size N**

Single, Complete, & Average Linkage





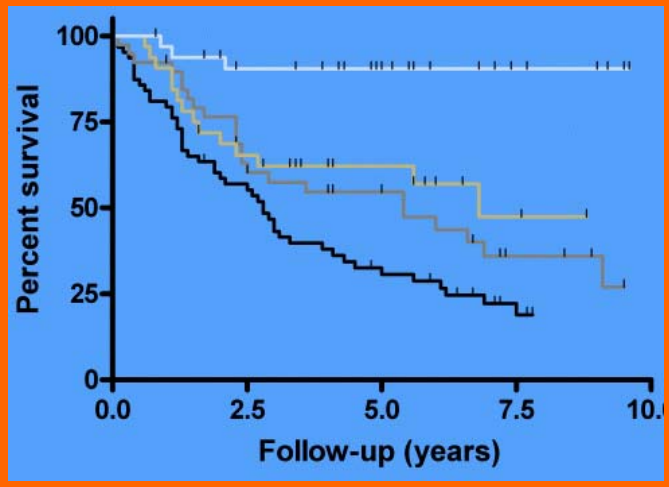
Single linkage defines distance betw two clusters as min distance betw them

Complete linkage defines distance betw two clusters as max distance betw them

Exercise: Give definition of "average linkage"

Image source: UCL Microcore Website

Selection of Patient Samples and Genes for Disease Prognosis

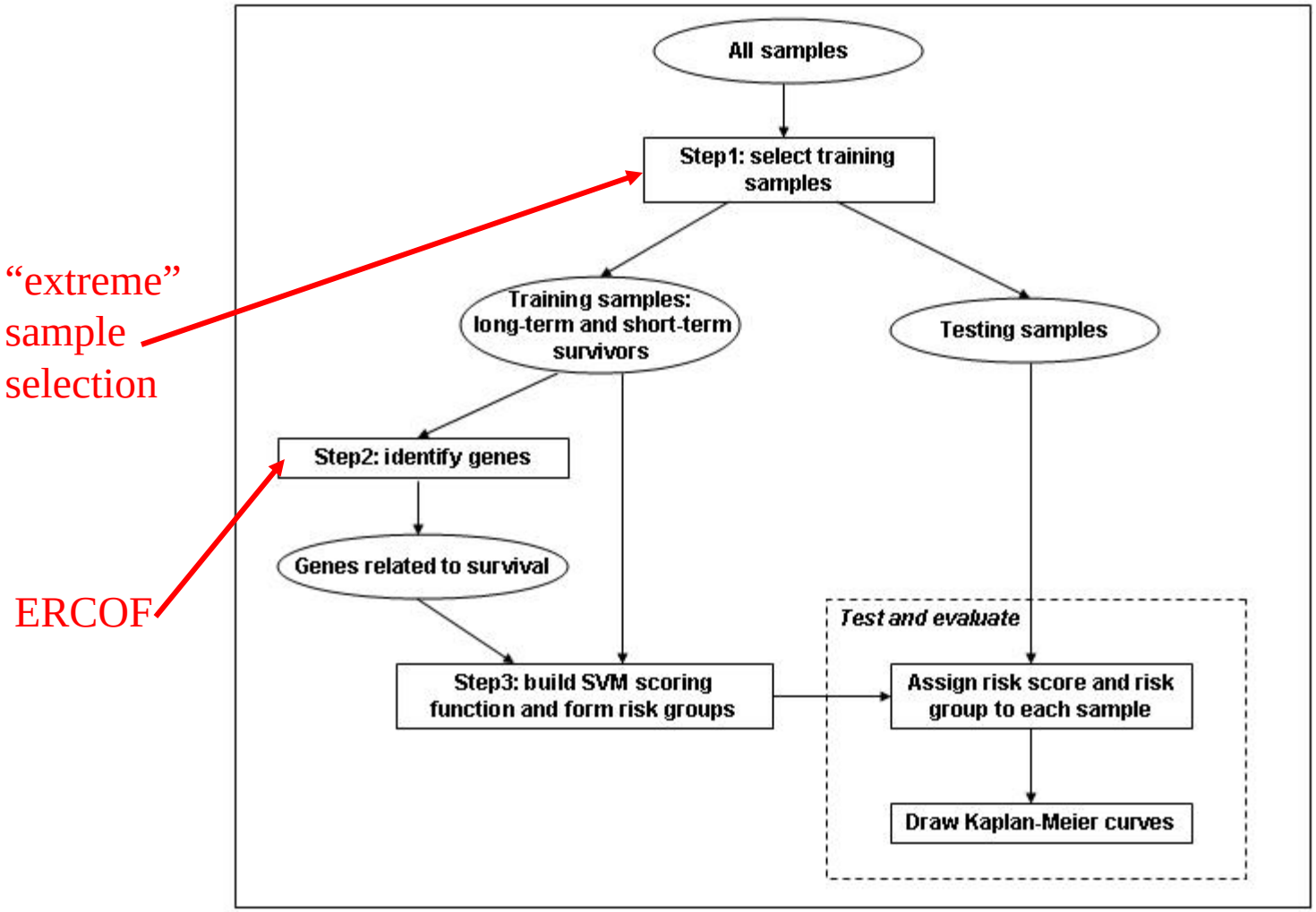


NUS National University of Singapore

Gene Expression Profile + Clinical Data ⇒ Outcome Prediction

- **Univariate & multivariate Cox survival analysis** (Beer et al 2002, Rosenwald et al 2002)
- **Fuzzy neural network** (Ando et al 2002)
- **Partial least squares regression** (Park et al 2002)
- **Weighted voting algorithm** (Shipp et al 2002)
- **Gene index and "reference gene"** (LeBlanc et al 2003)
- **……**

Another Approach …



Extreme Sample Selection

Short-term Survivors *v.s.* Long-term Survivors

Short-term survivors
who died within a *short* period

$\Downarrow$

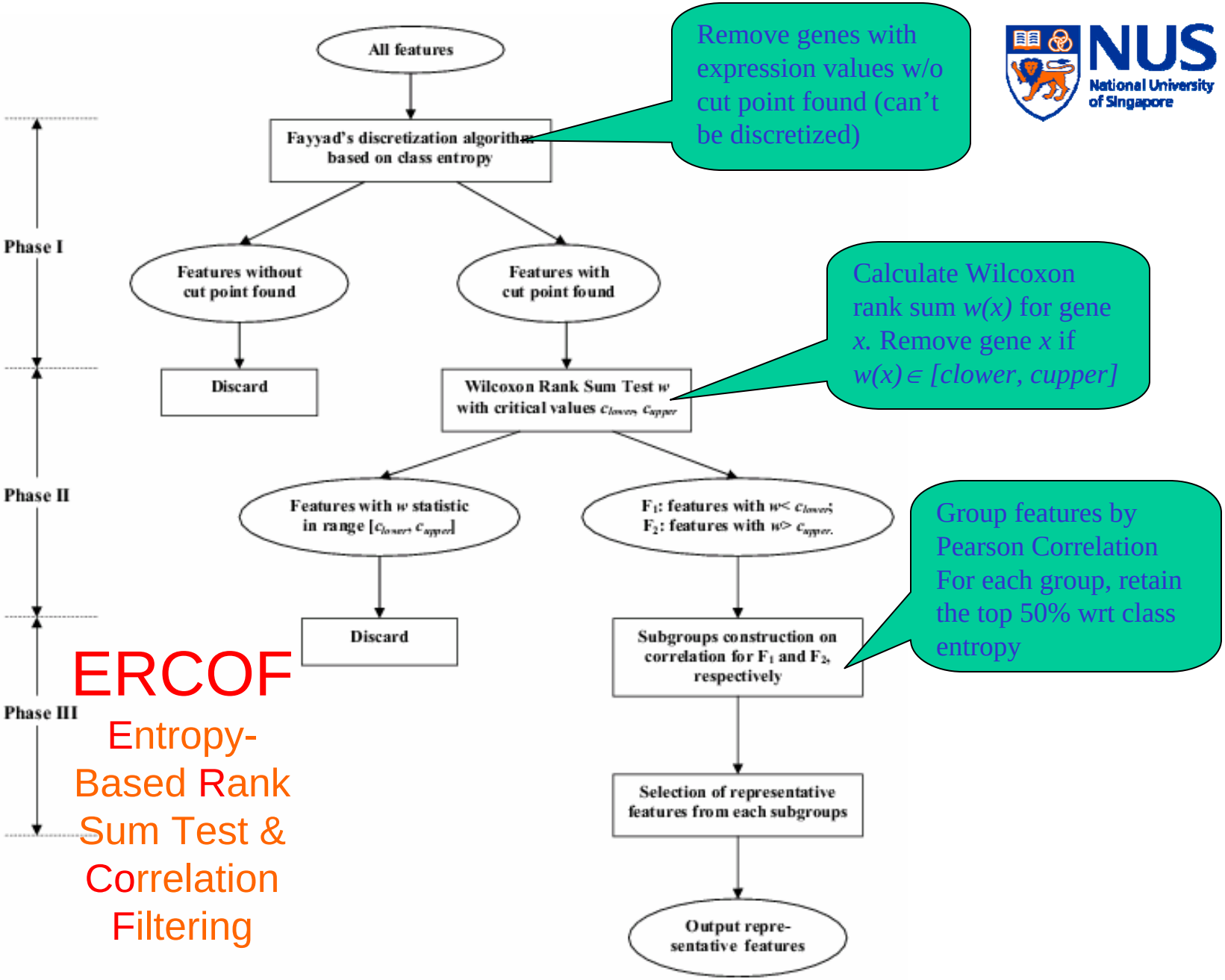
$F(T) < c_1$ and $E(T) = 1$

Long-term survivors
who were alive after a *long* follow-up time

$\Downarrow$

$F(T) > c_2$

T: sample
$F(T)$: follow-up time
$E(T)$: status (1:unfavorable; 0: favorable)
c_1 and c_2: thresholds of survival time



ERCOF

Entropy-Based Rank Sum Test & Correlation Filtering

Risk Score Construction

Linear Kernel SVM regression function

$$G(T) = \sum_i a_i y_i K(T, x(i)) + b$$

T: test sample, $x(i)$: support vector,
y_i: class label (1: short-term survivors; -1: long-term survivors)

Transformation function (*posterior probability*)

$$S(T) = \frac{1}{1 + e^{-G(T)}} \qquad (S(T) \in (0,1))$$

$S(T)$: ***risk score*** of sample T

Diffuse Large B-Cell Lymphoma

- **DLBC lymphoma is the most common type of lymphoma in adults**
- **Can be cured by anthracycline-based chemotherapy in 35 to 40 percent of patients**

⇒ **DLBC lymphoma comprises several diseases that differ in responsiveness to chemotherapy**

- **Intl Prognostic Index (IPI)**
  - age, "Eastern Cooperative Oncology Group" Performance status, tumor stage, lactate dehydrogenase level, sites of extranodal disease, ...
- **Not very good for stratifying DLBC lymphoma patients for therapeutic trials**

⇒ **Use gene-expression profiles to predict outcome of chemotherapy?**

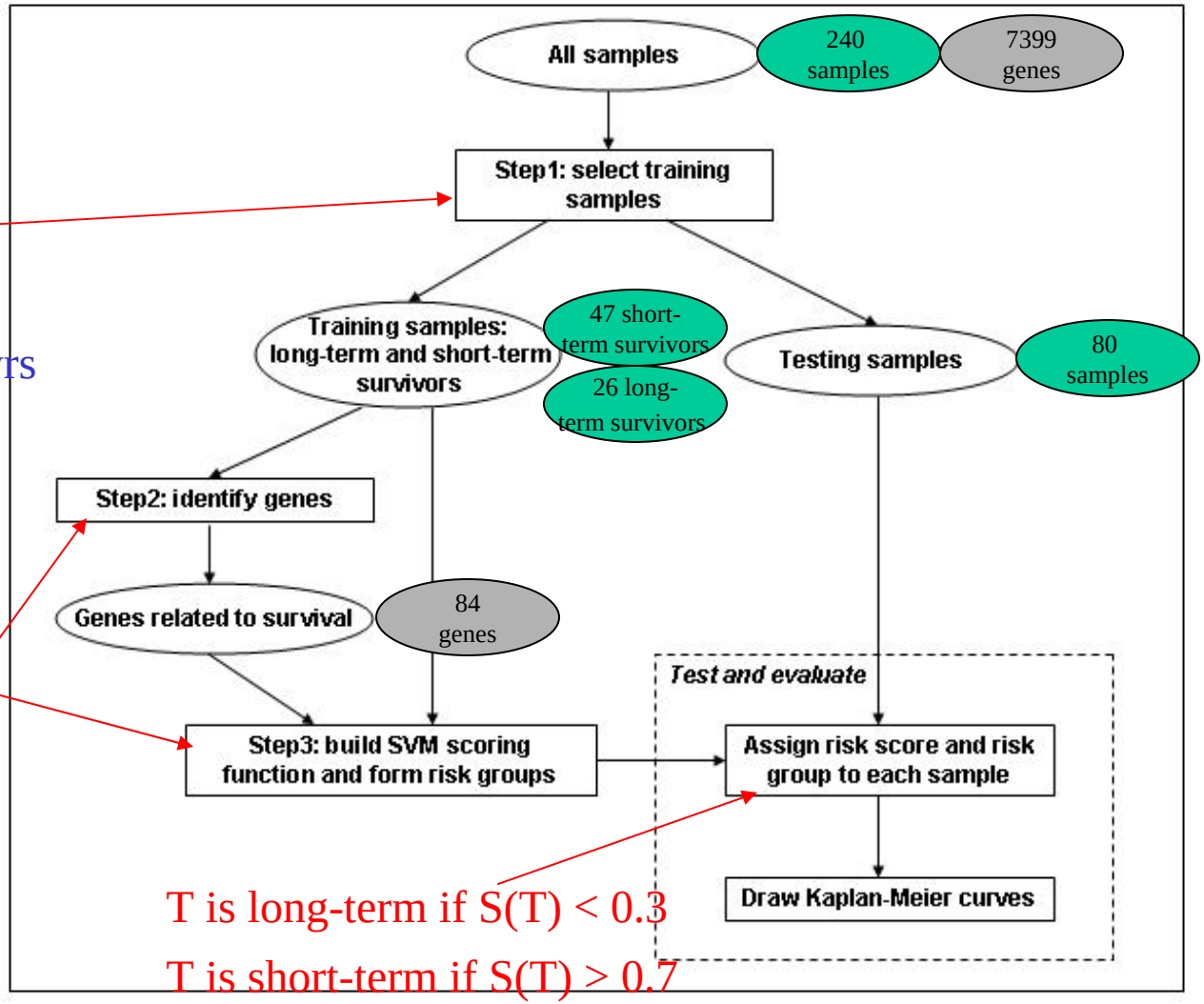
Rosenwald et al., *NEJM* 2002

- **240 data samples**
  - 160 in preliminary group
  - 80 in validation group
  - each sample described by 7399 microarray features
- **Rosenwald et al.'s approach**
  - identify gene: Cox proportional-hazards model
  - cluster identified genes into four gene signatures
  - calculate for each sample an outcome-predictor score
  - divide patients into quartiles according to score

Knowledge Discovery from Gene Expression of "Extreme" Samples

"extreme" sample selection: < 1 yr vs > 8 yrs

knowledge discovery from gene expression



T is long-term if S(T) < 0.3

T is short-term if S(T) > 0.7

Discussions: Sample Selection

Application	Data set	Status		Total
		Dead	Alive	
DLBCL	Original	88	72	160
	Informative	47+1(*)	25	73

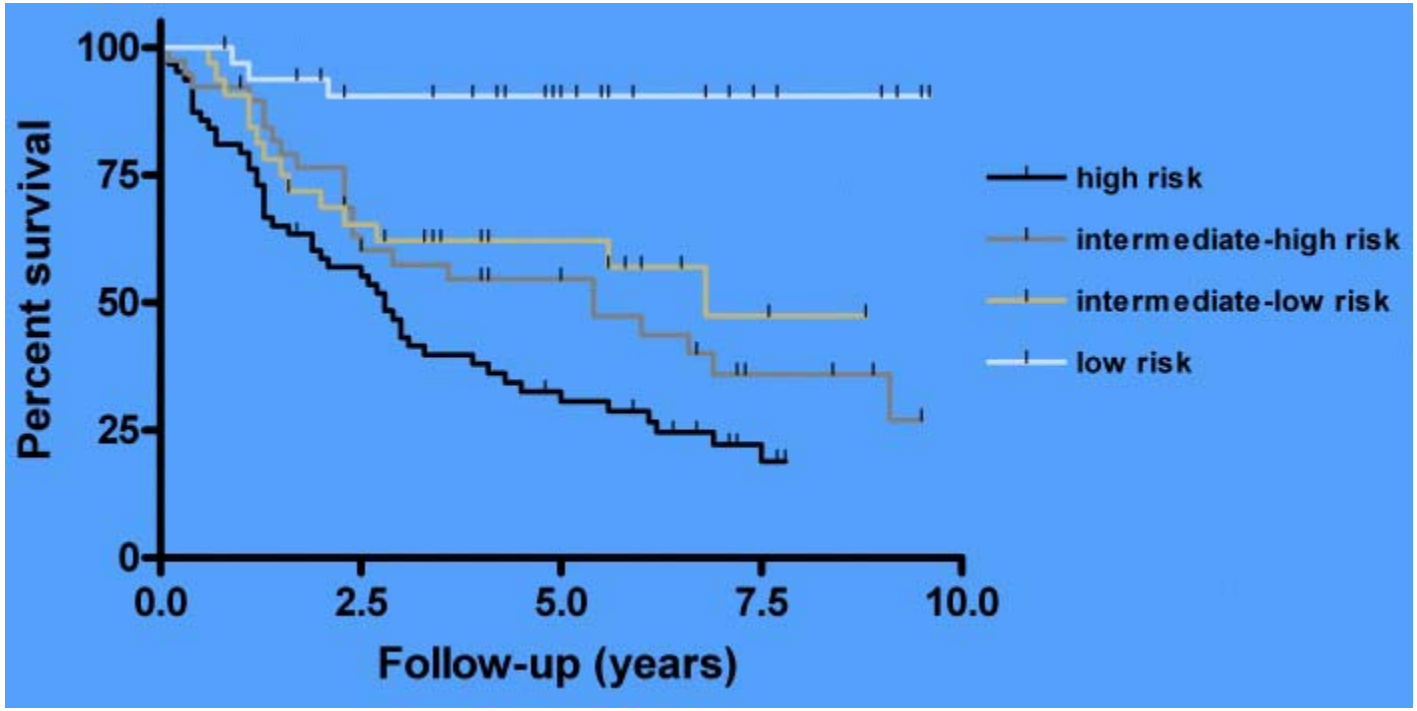
Number of samples in original data and selected informative training set.
(*): Number of samples whose corresponding patient was dead at the end of follow-up time, but selected as a long-term survivor.

Discussions: Gene Identification

Gene selection	DLBCL
Original	4937(*)
Phase I	132(2.7%)
Phase II	84(1.7%)

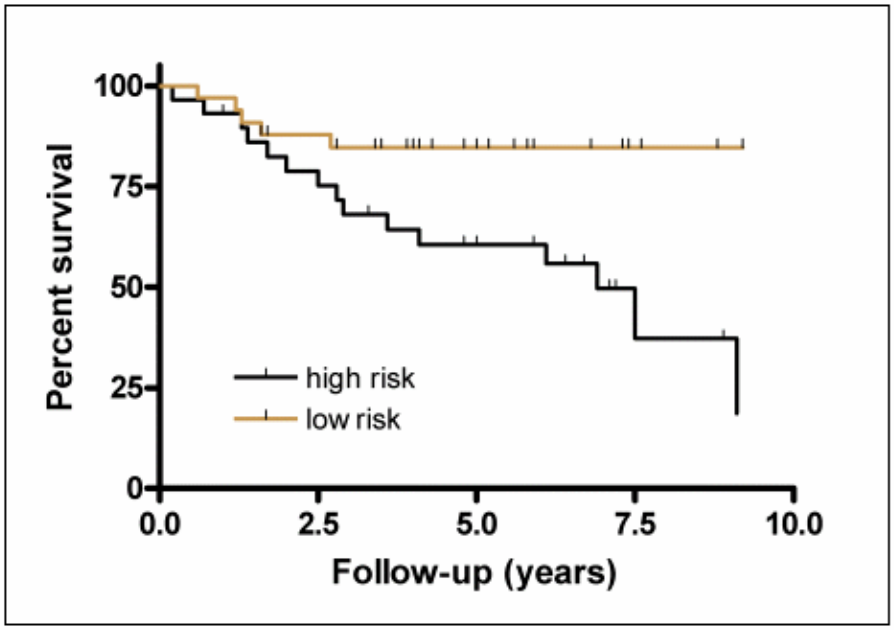
Number of genes left after feature filtering for each phase.
(*): number of genes after removing those genes who were absent in more than 10% of the experiments.

Kaplan-Meier Plot for 80 Test Cases

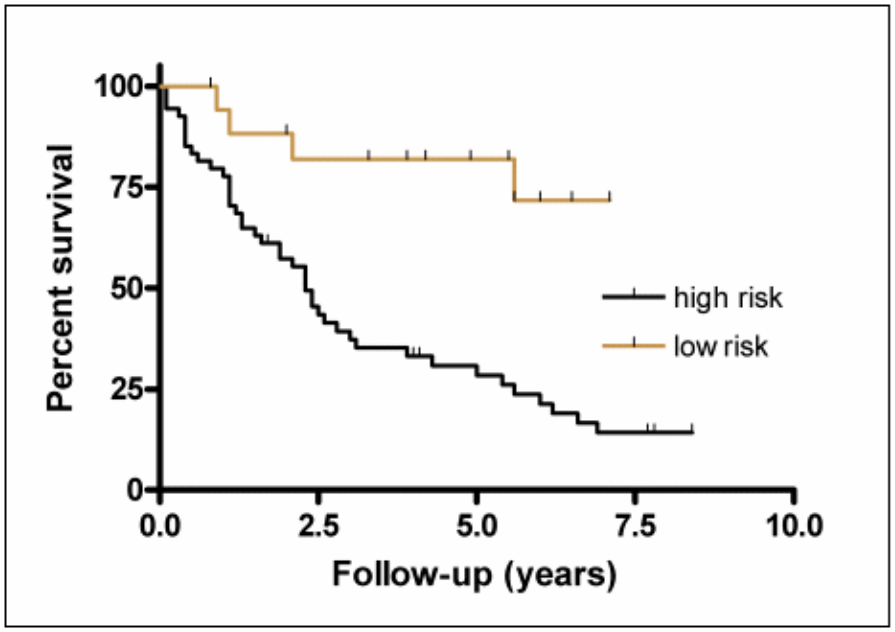


p-value of log-rank test: < 0.0001
Risk score thresholds: 0.7, 0.3

Improvement Over IPI

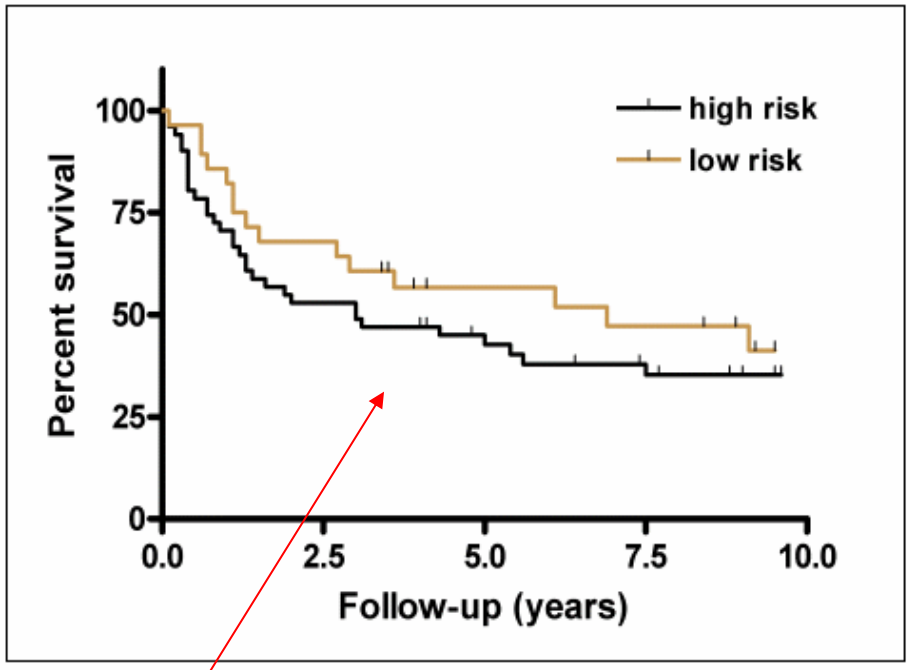


(A) IPI low,
p-value = 0.0063

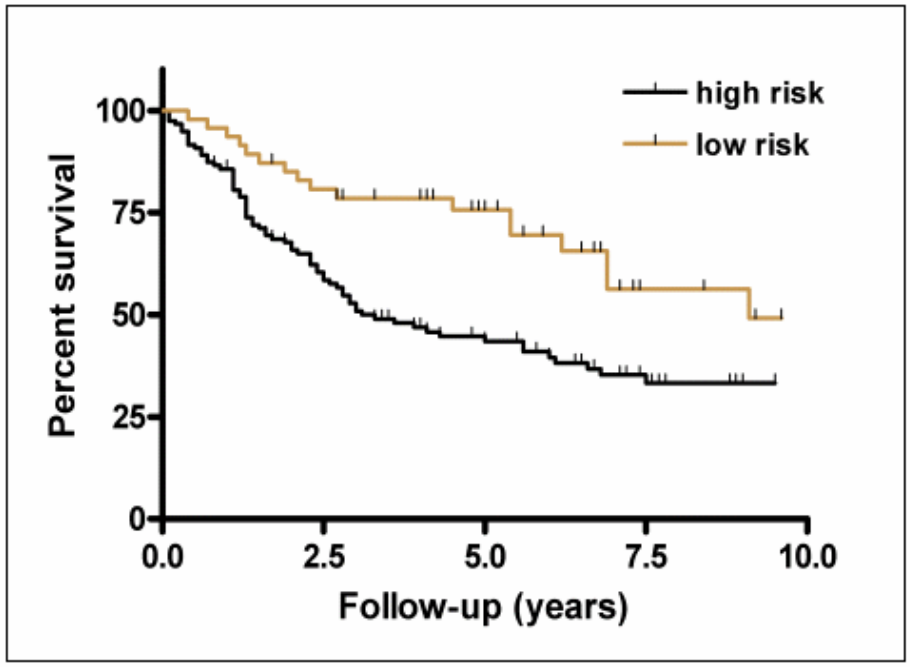


(B) IPI intermediate,
p-value = 0.0003

Merit of "Extreme" Samples



(A) W/o sample selection (p =0.38)



(B) With sample selection (p=0.009)

No clear difference on the overall survival of the 80 samples in the validation group of DLBCL study, if **no training sample selection conducted**

Is ERCOF Useful? Observations from 1000+ Expts

- **Feature selection methods considered**
  - All use all features
  - All-entropy select features whose value range can be partitioned by Fayyad & Irani's entropy method
  - Mean-entropy select features whose entropy is better than the mean entropy
  - Top-number-entropy select the top 20, 50, 100, 200 genes by their entropy
  - ERCOF at 5% significant level for Wilcoxon rank sum test and 0.99 Pearson correlation coeff threshold
- **Data sets considered**
  - Colon tumor
  - Prostate cancer
  - Lung cancer
  - Ovarian cancer
  - DLBC lymphoma
  - ALL-AML
  - Childhood ALL
- **Learning methods considered**
  - C4.5
  - Bagging, Boosting, CS4
  - SVM, 3-NN

ERCOF vs All-Entropy

Experiment	SVM	3-NN	Bagging	AdaBoostM1	RandomForests	CS4
ColonTumor	C	A,C	C	C	C	C
Prostate	C	C	A,C	A,C	C	C
Lung test	C	A,C	A	A,C	C	A,C
Lung	A,C	C	A,C	C	C	C
Ovarian	A,C	C	A,C	C	C	A,C
DLBCL	C	C	A	C	A	A,C
ALLAML test	A,C	C	A,C	A,C	C	C
ALLAML	A,C	A,C	A,C	C	A,C	A,C
Pediatric ALL data — test						
T-ALL	A,C	A,C	A,C	A,C	A,C	A,C
E2A-PBX1	A,C	A,C	A,C	A,C	A,C	A,C
TEL-AML1	A,C	A,C	A,C	A,C	A,C	C
BCR-ABL	A,C	C	A,C	A,C	C	A,C
MLL	A,C	A,C	C	A,C	C	C
Hyperdip>50	A,C	A	A,C	C	C	C
Pediatric ALL data — 10-fold cross validation						
T-ALL	A,C	C	A,C	A,C	A,C	C
E2A-PBX1	C	C	A,C	A,C	C	C
TEL-AML1	C	C	C	C	C	C
BCR-ABL	C	C	C	C	C	A,C
MLL	A,C	C	C	C	C	C
Hyperdip>50	C	C	A,C	C	C	A,C
Sum	A:0	A:1	A:2	A:0	A:1	A:0
	C:8	C:12	C:5	C:10	C:14	C:11
	Tie:12	Tie:7	Tie:13	Tie:10	Tie:5	Tie:9

All-entropy wins 4 times

ERCOF wins 60 times

ERCOF vs Mean-Entropy

Experiment	SVM	3-NN	Bagging	AdaBoostM1	RandomForests	CS4
ColonTumor	C	C	B,C	C	C	C
Prostate	C	B,C	C	B	C	B,C
Lung test	B,C	B,C	B	B,C	C	B,C
Lung	B,C	C	B,C	B	B	B,C
Ovarian	B,C	C	B	C	B,C	C
DLBCL	B,C	C	B	B,C	C	B,C
ALLAML test	B,C	C	B,C	B,C	C	B,C
ALLAML	B,C	B	B	C	B	B,C
Pediatric ALL data — test						
T-ALL	B,C	B,C	B,C	B,C	B,C	B,C
E2A-PBX1	B,C	B,C	B,C	B,C	B,C	B,C
TEL-AML1	B,C	B,C	B,C	B	C	C
BCR-ABL	C	B,C	B	B,C	B,C	B,C
MLL	B,C	B,C	B,C	B,C	B,C	B,C
Hyperdip>50	B,C	B	B	B,C	C	B,C
Pediatric ALL data — 10-fold cross validation						
T-ALL	B,C	B	B,C	B,C	B,C	B,C
E2A-PBX1	C	C	B,C	B,C	C	C
TEL-AML1	C	C	B,C	C	C	C
BCR-ABL	C	C	C	B	B	B
MLL	B,C	B,C	C	C	B	C
Hyperdip>50	C	C	C	B,C	C	C
Sum	B:0	B:3	B:6	B:4	B:4	B:1
	C:7	C:9	C:4	C:5	C:10	C:7
	Tie:13	Tie:8	Tie:10	Tie:11	Tie:6	Tie:12

Mean-entropy wins 18 times

ERCOF wins 42 times

Effectiveness of ERCOF

Table 5.32: A summary of the total winning times (including tie cases) of each classifier (under different feature selection methods) across the 20 validation tests on the six gene expression profiles and one proteomic data set. The number with bold font in each row indicates the feature selection method that owns most winning times for the relevant classifier. In the brackets, there is the total number of misclassified samples across the same 20 validation tests. Similarly, the figure with bold font in the brackets in each row is the minimum number of total misclassified samples among feature selection methods for the classifier.

Classifier	All	All-entropy	Mean-entropy	Top-number-entropy				ERCOF
				20	50	100	200	
SVM	4(100)	9(52)	11(48)	6(76)	6(74)	11(52)	11(59)	**16(38)**
3-NN	1(187)	5(87)	8(77)	6(88)	4(81)	6(77)	5(73)	**12(61)**
Bagging	7(123)	5(117)	8(115)	**11**(123)	**11**(122)	7(122)	9(114)	8(**112**)
AdaBoostM1	5(191)	8(181)	8(166)	**11(138)**	10(144)	10(157)	9(162)	10(154)
RandomForests	0(228)	5(111)	5(93)	6(96)	7(83)	8(96)	5(90)	**9(80)**
CS4	5(87)	6(77)	6(76)	7(101)	10(81)	9(74)	8(74)	**12(66)**
Total wins	22	38	46	47	48	51	47	67

Conclusions

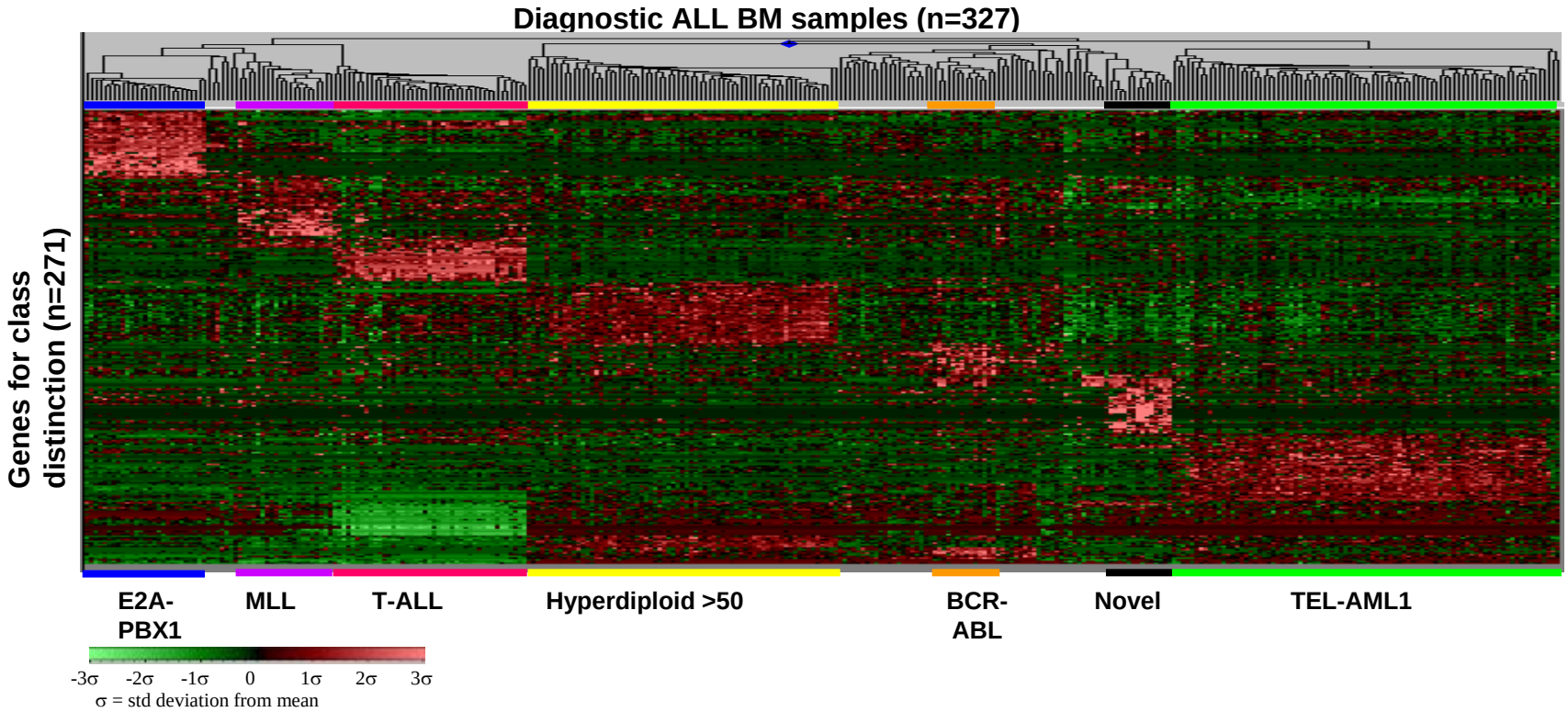
- **Selecting extreme cases as training samples is an effective way to improve patient outcome prediction based on gene expression profiles and clinical information**

- **ERCOF is very suitable for SVM, 3-NN, CS4, Random Forest, as it gives these learning algos highest no. of wins**
- **ERCOF is suitable for Bagging also, as it gives this classifier the lowest no. of errors**

$\Rightarrow$ **ERCOF is a systematic feature selection method that is very useful**

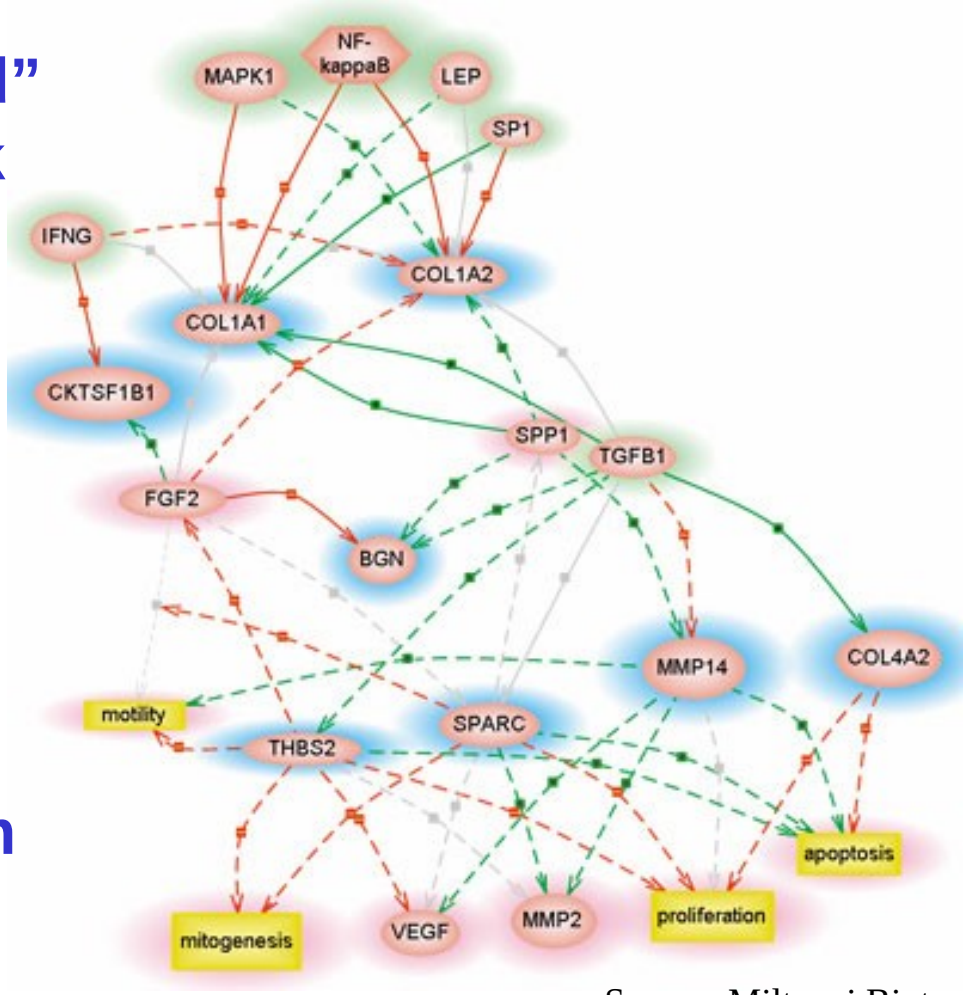
Beyond Classification of Gene Expression Profiles

- **After identifying the candidate genes by feature selection, do we know which ones are causal genes and which ones are surrogates?**



Gene Regulatory Circuits

- **Genes are "connected" in "circuit" or network**
- **Expr of a gene in a network depends on expr of some other genes in the network**
- **Can we "reconstruct" the gene network from gene expression and other data?**



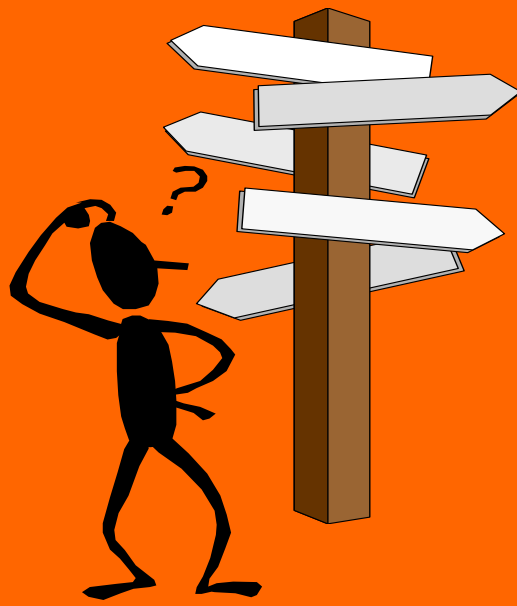
Source: Miltenyi Biotec

Recommended Readings

- Wong, *The Practical Bioinformatician*, 2004, ICP. Chapter 14
- Li & Wong, Identifying good diagnostic genes or gene groups from gene expression data by using the concept of emerging patterns. *Bioinformatics*, 18:725-734, 2002
- Miller et al., Optimal gene expression analysis by microarrays. *Cancer Cell*, 2:353-361, 2002
- Liu et al,. Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data. *Bioinformatics*, 21:3377-3384, 2005
- Eisen et al., Cluster analysis and display of genome-wide expression patterns. *PNAS*, 8:14863-14868, 1998
- Tibshirani et al., Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99:6567-6572, 2002

5 min

Advancing Knowledge Discovery



Some of those "techniques" frequently needed in analysis of biomedical data are insufficiently studied by current data mining researchers

- **Recognizing what samples are relevant and what are not**
- **Recognizing what features are relevant and what are not & handling missing or incorrect values**
- **Recognizing trends, changes, and their causes**

Any Question?

Acknowledgements

- **Disease Treatment Planning & Understanding**
  - Jinyan Li, Huiqing Liu
  - Allen Yeoh

- **Gene Feature Recognition**
  - Huiqing Liu, Rajesh Chowdhary, Vladimir Bajic
  - Fanfan Zeng, Roland Yap

- **Protein Function Inference**
  - Hon Nian Chua, Wing-Kin Sung