Knowledge Discovery Applications in Bioinformatics

Limsoon Wong

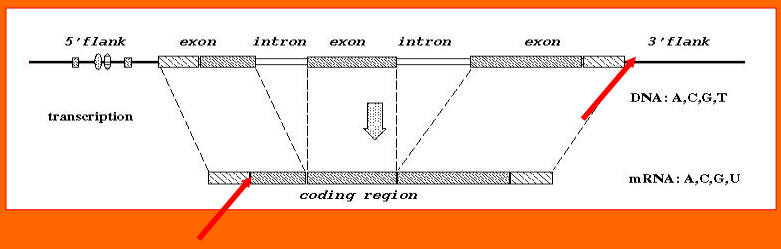
Objectives

- **To give a flavor of bioinformatics problems where data mining and machine learning are applicable**
- **To highlight adaptations that are necessary for data mining and machine learning techniques to be successfully applied in bioinformatics**

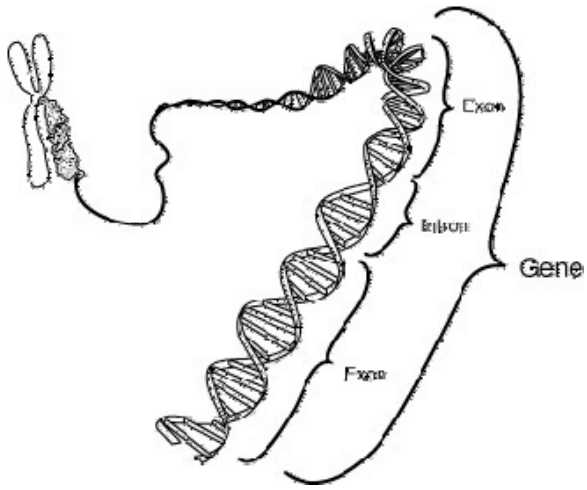
Plan

- Genomic sequence feature discovery
- Protein function predictiom
- Protein complex prediction
- Syntenic gene cluster determination
- Drug pathway inference

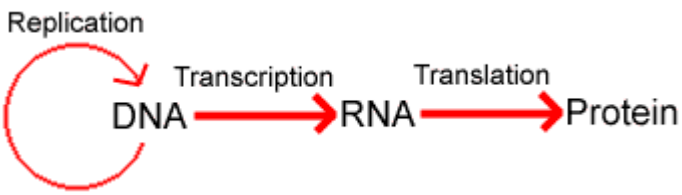
Gene Feature Recognition

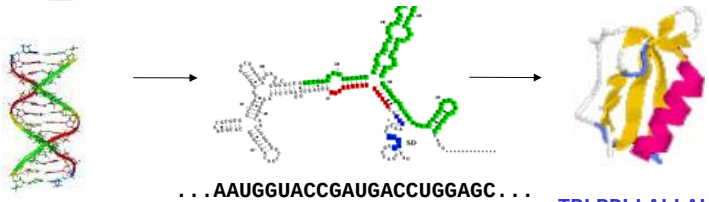


What is a gene?



Central Dogma



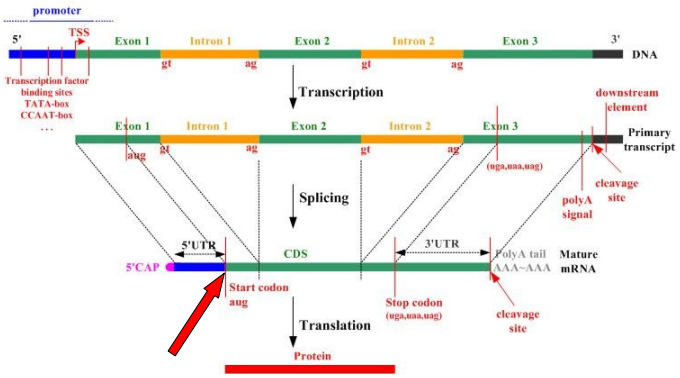


...AATGGTACCGATGACCTG...

...AAUGGUACCGAUGACCUGGAGC...

...TRLRPLLALLALWP...

Translation Initiation Site



A Sample cDNA

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

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

Approach

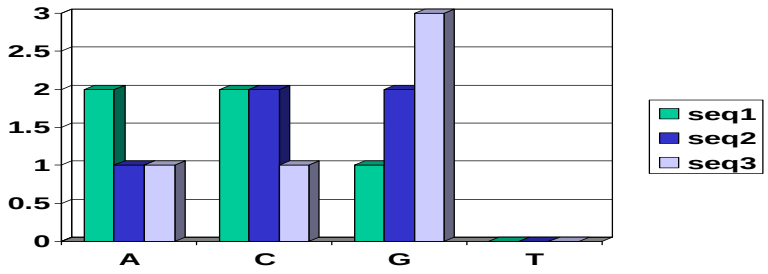
- **Training data gathering**
- **Signal generation**
  - k-grams, distance, domain know-how, ...
- **Signal selection**
  - Entropy, $\chi 2$, CFS, t-test, domain know-how...
- **Signal integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...

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
```

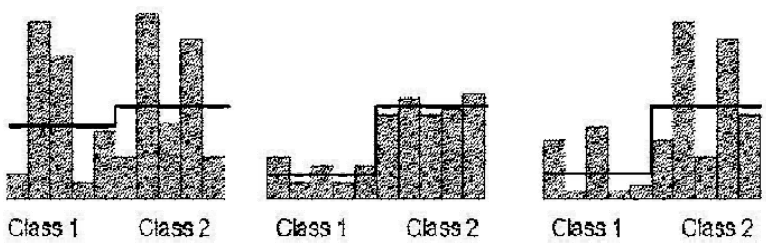
- **Window = ±100 bases**
- **In-frame, downstream**
  - GCT = 1, TTT = 1, ATG = 1…
- **Any-frame, downstream**
  - GCT = 3, TTT = 2, ATG = 2…
- **In-frame, upstream**
  - GCT = 2, TTT = 0, ATG = 0, ...

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., χ2)

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

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

where m is the number of intervals, k the number of classes, A_{ij} the number of samples in the ith interval, jth class, R_i the number of samples in the ith interval, C_j the number of samples in the jth class, N the total number of samples, and E_{ij} the expected frequency of A_{ij} ($E_{ij} = R_i * C_j/N$).

Signal Selection (e.g., CFS)

- **Instead of scoring individual signals, how about scoring a group of signals as a whole?**
- **CFS**
  - Correlation-based Feature Selection
  - A good group contains signals that are highly correlated with the class, and yet uncorrelated with each other

Sample k-grams Selected by CFS for Recognizing TIS

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

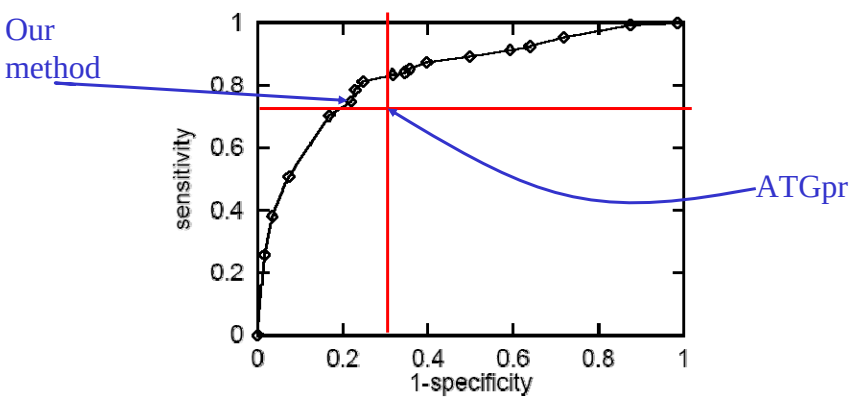
	TP/(TP + FN)	TN/(TN + FP)	TP/(TP + FP)	Accuracy
Naïve Bayes	84.3%	86.1%	66.3%	85.7%
SVM	73.9%	93.2%	77.9%	88.5%
Neural Network	77.6%	93.2%	78.8%	89.4%
Decision Tree	74.0%	94.4%	81.1%	89.4%

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%

Validation Results (on Chr X and Chr 21)

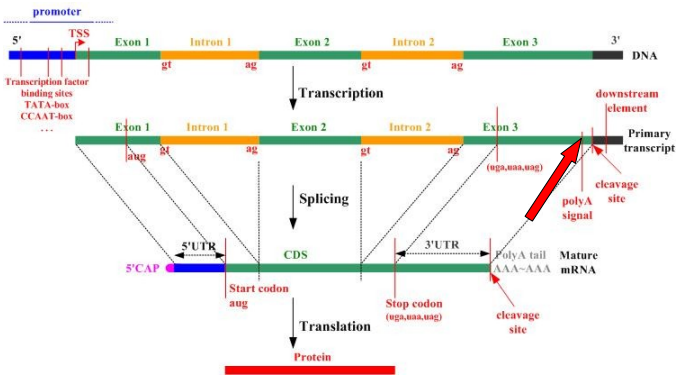


- **Using top 100 features selected by entropy and trained on Pedersen & Nielsen's**

Question

- **What would you do if you have to recognize a gene feature that has no anchor signal like ATG?**

PolyA Signal Site



Eukaryotic Pre-mRNA Processing

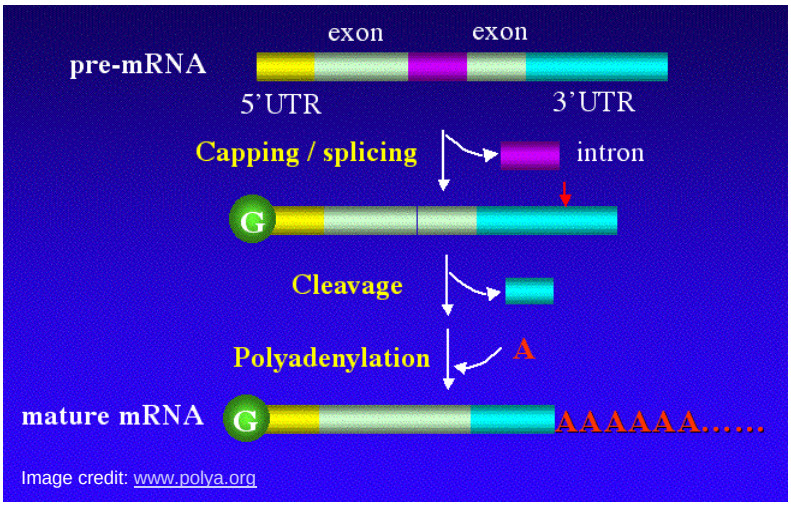
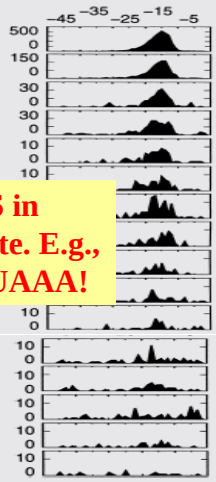


Image credit: www.polya.org

Poly-A Signals in Human

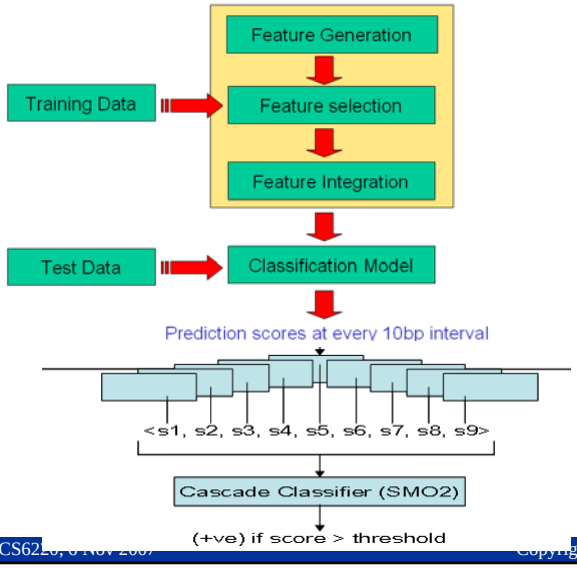
Table 2. Most Significant Hexamers in 3′ Fragments: Clustered Hexamers

Hexamer	Observed (expected)[a]	% sites	p[b]	Position average ± SD	Location[c]
AAUAAA	3286 (317)	58.2	0	−16 ± 4.7	
AUUAAA	843 (112)	14.9	0	−17 ± 5.3	
AGUAAA	156 (32)	2.7	6×10^{-57}	−16 ± 5.9	
UAUAAA	180 (53)	3.2	4×10^{-45}	−18 ± 7.8	
CAUAAA	76 (23)	1.3	1×10^{-16}	−17 ± 5.9	
GAUAAA	72				
AAUAUA	96				
AAUACA	70				
AAUAGA	43				
AAAAAG	49				
ACUAAA	36 (11)	0.6	1×10^{-08}	−17 ± 8.1	
AAGAAA	62 (10)	1.1	9×10^{-28}	−19 ± 11	
AAUGAA	49 (10)	0.8	4×10^{-16}	−20 ± 10	
UUUAAA	69 (20)	1.2	3×10^{-16}	−17 ± 12	
AAAACA	29 (5)	0.5	8×10^{-12}	−20 ± 10	
GGGGCU	22 (3)	0.3	9×10^{-12}	−24 ± 13	



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

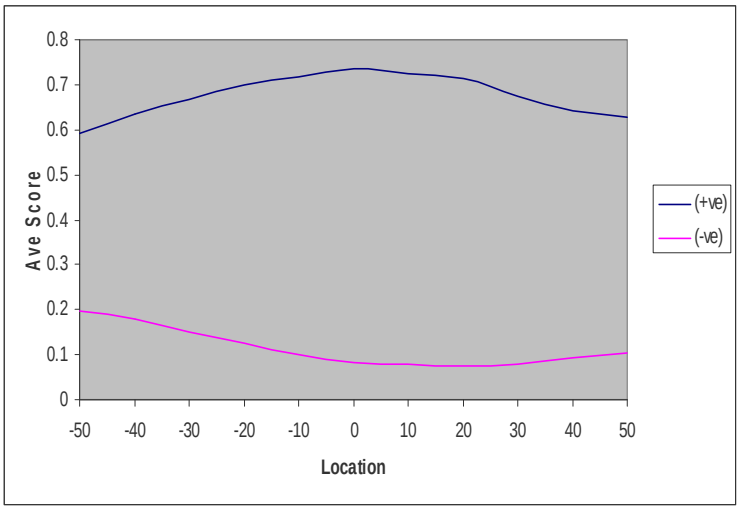
Approach on Arab PAS Sites (I)



Approach on Arab PAS Sites (II)

- **Data from Qingshun Li**
  - 6209 (+ve) seq
  - 1581 (-ve) intron
  - 1501 (-ve) coding
  - 864 (-ve) 5'utr
- **Feature generation**
  - 3-grams, compositional features (4U/1N. G/U*7, etc)
  - Freq of features above in 3 diff windows: (-110/+5), (-35/+15), (-50/+30)
- **Feature selection**
  - $\chi 2$
- **Feature integration & Cascade**
  - SVM

Score Profile Relative to Candidate Sites



Validation Results

SN_0	SMO 1		SMO 2		PASS 1.0	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
CDS	90%	0.26	94%	0.24	95%	3.7
5'UTR	79%	0.42	85%	0.49	78%	5.5
Intron	64%	0.59	71%	0.67	63%	6.3

Table 2. Equal-error-rate points of SMO1, SMO2, and PASS 1.0 for SN_10.

SN_10	SMO 1		SMO 2		PASS 1.0	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
CDS	94%	0.36	96%	0.31	96%	4
5'UTR	86%	0.53	89%	0.6	81%	5.7
Intron	73%	0.68	77%	0.77	67%	6.6

Table 3. Equal-error-rate points of SMO1, SMO2, and PASS 1.0 for SN_30.

SN_30	SMO 1		SMO 2		PASS 1.0	
Control Sequences	SN & SP	Threshold	SN & SP	Threshold	SN & SP	Threshold
CDS	97%	0.44	97%	0.37	97%	4.3
5'UTR	90%	0.62	92%	0.67	84%	6.2
Intron	79%	0.75	83%	0.81	72%	6.8

What have we learned?

- **Gene feature recognition applications**
  - TIS, PAS
- **General methodology**
  - "Feature generation, feature selection, feature integration"
    - **Explicit feature generation**
    - **Explicit feature selection**
    - **Use any machine learning method w/o any form of complicated tuning**
- **Important tactics**
  - Multiple models to optimize overall performance
  - Classifier cascades

References

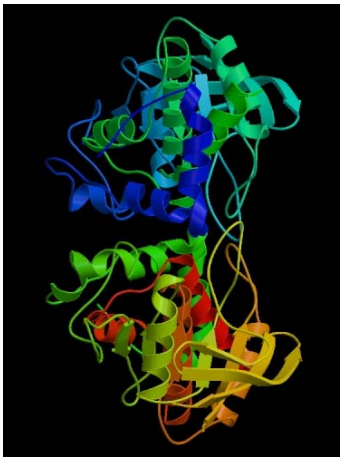
- L. Wong et al., "Using feature generation and feature selection for accurate prediction of translation initiation sites", GIW, 13:192--200, 2002
- J. Li et al., "Techniques for Recognition of Translation Initiation Sites", The Practical Bioinformatician, Chapter 4, pages 71—90, 2004
- C. H. Koh et al., "Recognition of Polyadenylation Sites from Arabidopsis Genomic Sequences". GIW, 18:73-82, 2007

Protein Function Prediction

NUS National University of Singapore

What is a protein?

- **A protein is a large complex molecule made up of one or more chains of amino acids**
- **Protein performs a wide variety of activities in the cell**



Function Assignment to Protein Seq

```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT
```

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

An Early Example of Seq Analysis

Source: Ken Sung

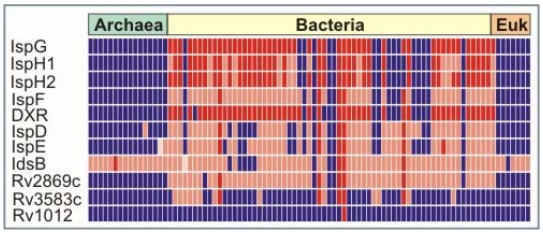
- **Doolittle et al. (*Science*, July 1983) searched for platelet-derived growth factor (PDGF) in his own DB. He found that PDGF is similar to v-sis oncogene**

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

⇒ **"Guilt by association" of sequence similarity!**

Important Unsolved Challenges

- **What if there is no useful seq 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

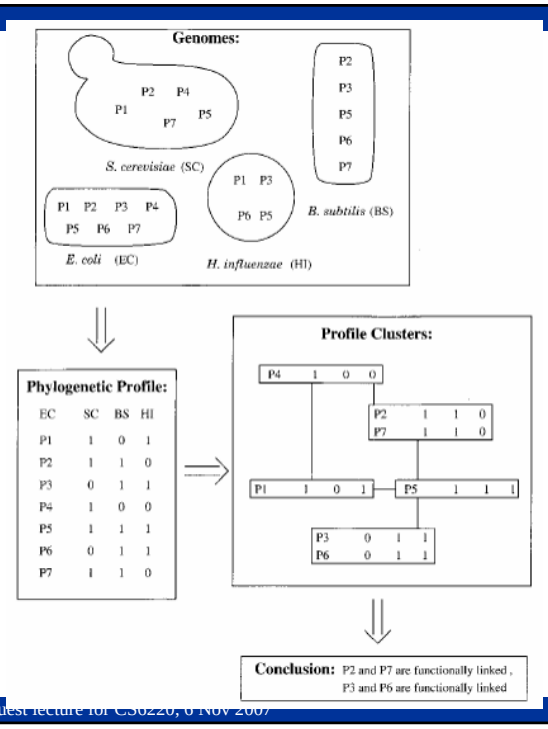


Guilt by Association of Genome Phylogenetic Profiles

- **Gene (and hence proteins) with identical patterns of occurrence across phyla tend to function together**

⇒ **Even if no homolog with known function is available, it is still possible to infer function of a protein**

Image credit: Ed Marcotte, http://apropos.icmb.utexas.edu/plex/tour/isoprenoid.jpg



Phylogenetic Profiling: How It Works

Phylogenetic Profiling: P-value

The probability of observing by chance z occurrences of genes X and Y in a set of N lineages, given that X occurs in x lineages and Y in y lineages is

$$P(z|N,x,y) = \frac{w_z * \overline{w_z}}{W}$$

where

$$w_z = \binom{N}{z}$$

$$\overline{w_z} = \binom{N-z}{x-z} * \binom{N-z}{y-z}$$

$$W = \binom{N}{x} * \binom{N}{y}$$

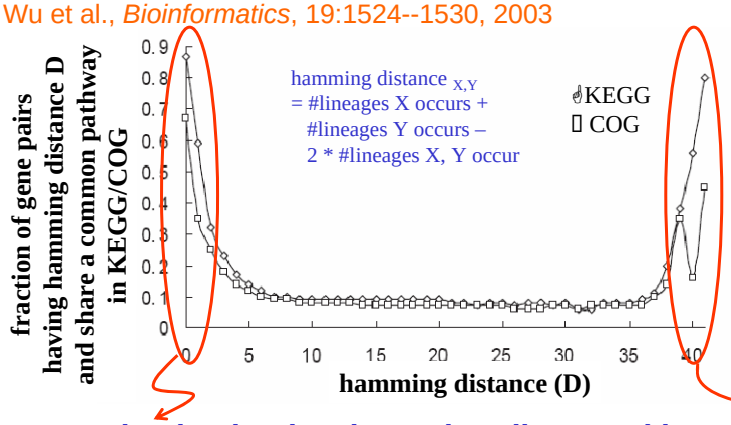
No. of ways to distribute z co-occurrences over N lineage's

No. of ways to distribute the remaining $x - z$ and $y - z$ occurrences over the remaining $N - z$ lineage's

No. of ways of distributing X and Y over N lineage's without restriction

Phylogenetic Profiling: Evidence

Wu et al., *Bioinformatics*, 19:1524--1530, 2003



- **Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways**

Exercise: Why do proteins having high hamming distance also have this behaviour?

Guilt by Association of Similarity of Dissimilarities

- **Suppose the diff betw proteins X and proteins A, B, C, … are same as those betw Y and A, B, C, …**
- **Can we infer X and Y are similar in function?**

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

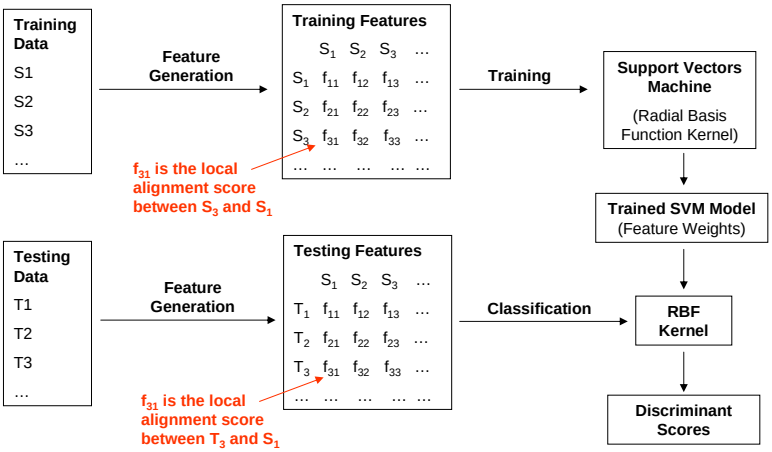
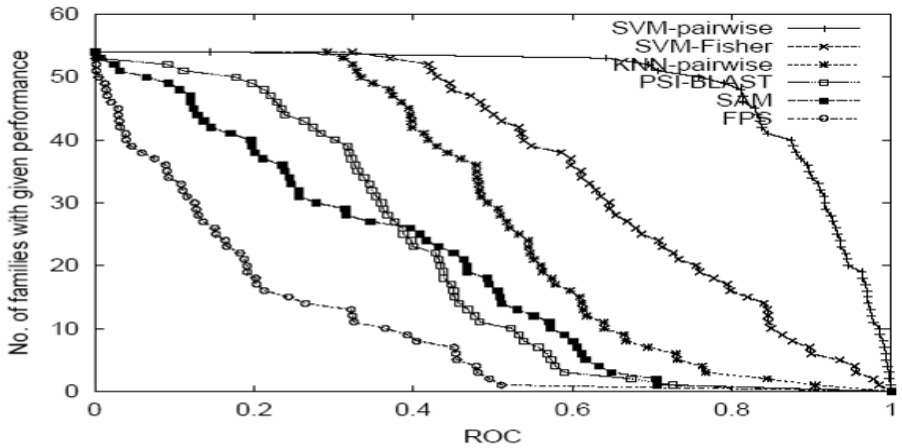


Image credit: Kenny Chua

Performance of SVM-Pairwise



- **ROC: The area under the curve derived from plotting true positives as a function of false positives for various thresholds**

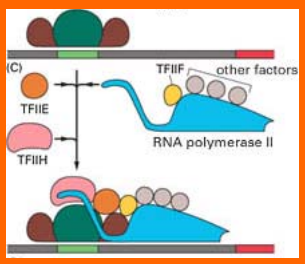
What have we learned?

- **Protein function prediction applications**
  - Genome phylogenetic profile, SVM Pairwise
- **General methodology**
  - "Guilt by association"
    - **Similarity in sequences**
    - **Similarity in genome phylogenetic profiles**
    - **Similarity in dissimilarities of sequences**
- **Important tactics**
  - Clustering of genome phylogenetic profiles
  - Clustering of dissimilarity profiles
  - Discriminative learning on dissimilarity profiles

References

- M. Pellegrini et al. “Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles”, *PNAS*, 96:4285--4288, 1999
- J. Wu et al. “Identification of functional links between genes using phylogenetic profiles”, *Bioinformatics*, 19:1524--1530, 2003
- T. Jaakkola, M. Diekhans, & D. Haussler. “A discriminative framework for detecting remote homologies”. *JCB*, 7(1-2):95-11, 2000
- L. Liao & W.S. Noble. “Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships”. *JCB*, 10(6):857-868, 2003

Protein Complex Prediction



Protein Complexes and Modules

- Protein complexes and functional modules are impt for understanding principles of cellular organizations
- Protein complex = a group of proteins that interact with each other at the same time and place
  - **E.g., transcription factor complexes**
- Functional module = proteins that participate in some cellular process, and bind to each other at diff time and place, possibly under diff condition
  - **E.g., MAP signal cascades**

Clique finding as a solution

- **Idea:**
  - Find cliques in the PPI graph
  - Proteins in a large clique for a complex/module
- **Issue:**
  - Genes in a complex/module often don't have full mutual interactions

PPI-Based Complex Prediction Algo

	RNSC	MCODE	MCL
Type	Clustering, local search cost based	Local neighborhood density search	Flow simulation
Multiple assignment of protein	No	Yes	No
Weighted edge	No	No	Yes

- **Issues**
  - Recall vs precision has to be improved
  - Does a "cleaner" PPI network help?

Cleansing PPI Graph

Source	Reliability
Affinity Chromatography	0.823077
Affinity Precipitation	0.455904
Biochemical Assay	0.666667
Dosage Lethality	0.5
Purified Complex	0.891473
Reconstituted Complex	0.5
Synthetic Lethality	0.37386
Synthetic Rescue	1
Two Hybrid	0.265407

- **PPI expt has lots of errors**
- **Real PPI is usually functionally linked**
- **Proteins that are functionally linked have many common partners**

⇒ **Clean up the input PPI network by removing PPI lacking sufficient common partners … how?**

Czekanowski-Dice Distance

- **Functional distance between two proteins** (Brun et al, 2003)

$$D(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}$$

- N_k is the set of interacting partners of k
- X Δ Y is symmetric diff betw two sets X and Y
- Greater weight given to similarity

Is this a good measure if u and v have very diff number of neighbours?

⇒ **Similarity can be defined as**

$$S(u,v) = 1 - D(u,v) = \frac{2X}{2X + (Y+Z)}$$

FS-Weighted Measure

- **FS-weighted measure**

$$S(u,v) = \frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v|} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v|}$$

- N_k is the set of interacting partners of k
- Greater weight given to similarity

⇒ **Intuitively:**

$$S(u,v) = \frac{2X}{2X+Y} \times \frac{2X}{2X+Z}$$

FS-Weighted Measure with Reliability

- **Take reliability into consideration when computing FS-weighted measure:**

$$S_R(u,v)=\frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_u} r_{u,w}+\sum_{w\in(N_u\cap N_v)} r_{u,w}(1-r_{v,w})\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}\times\frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_v} r_{v,w}+\sum_{w\in(N_u\cap N_v)} r_{v,w}(1-r_{u,w})\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}$$

  - **N_k is the set of interacting partners of k**
  - **$r_{u,w}$ is reliability weight of interaction betw u and v**

$\Rightarrow$ **Intuitively**

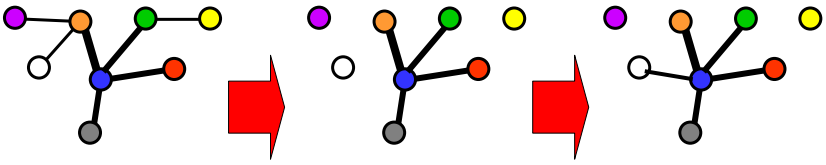
$$S(u,v)=\frac{2X}{2X+Y}\times\frac{2X}{2X+Z}$$

Integrating Reliability

- **Equiv measure shows improved correlation w/ functional similarity when reliability of interactions is considered:**

Neighbours	CD-Distance	FS-Weight	FS-Weight R
S_1	0.471810	0.498745	0.532596
S_2	0.224705	0.298843	0.375317
$S_1 \cup S_2$	0.224581	0.29629	0.363025

Cleaning PPI Network by FS-Weight



- **Modify existing PPI network as follow**
  - Remove level-1 interactions with low FS-weight
  - Add level-2 interactions with high FS-weight
- **Then run RNSC, MCODE, MCL, PCP, etc**

Markov Clustering, MCL

- **$M_{i,j}$: prob of going from i to j**
  - Initialized as adjacency matrix of PPI graph
- **Expansion $M_{i,j,E} = \Sigma_k M_{i,k} * M_{k,j}$**
  - Prob of going from i to j thru k
- **Inflation $M_{i,j,I} = (M_{i,j})^P / \Sigma_k (M_{k,j})^P$**
  - Favour more probably random walks
- **MCL performs random walk by iterating expansion and inflation until PPI is segmented into unconnected subgraphs (the resulting clusters)**

Clique Merging, PCP

- **Find all max cliques in PPI graph**
  - If two cliques overlap, distribute the overlapped nodes such that both cliques have larger average FS-weight
- **Merge resulting (partial) cliques with good inter-cluster density**

$$ICD(S_a, S_b) = \frac{\sum S_{FS}(i,j) \mid i \in (V_a - V_b), j \in (V_b - V_a), (i,j) \in E}{|V_a - V_b| \cdot |V_b - V_a|}$$

- **Modify the PPI network by treating the merged partial cliques as vertices**
- **Iterate the steps above**

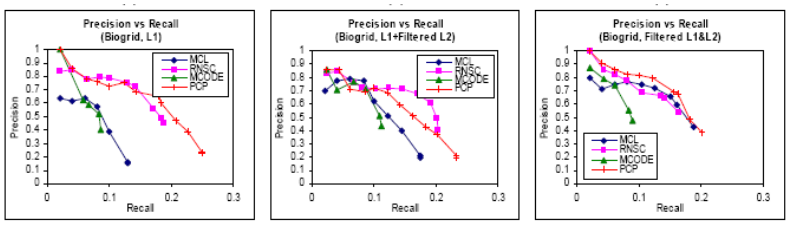
Experiments

- **PPI datasets**
  - PPI[BioGRID], BioGRID db from Stark et al., 2006
- **Gold standards**
  - PC_{2004}, Protein complexes from MIPS 03/30/2004
  - PC_{2006}, Protein complexes from MIPS 05/18/2006
- **Validation criteria**

$$overlap(S,C) = \frac{|V_S \cap V_C|^2}{|V_S| \cdot |V_C|}$$

where
  - S = predicted cluster
  - C = true complex
  - V_x = vertices of subgraph defined by X
- **Overlap(S,C) ≥ 0.25 is considered a correct prediction**

Validation on PC_{2004}

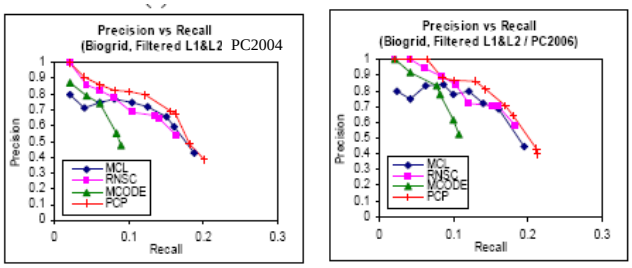


(d) (e) (f)

d) Original level-1 PPI
e) Original level-1 PPI and filtered level-2 PPI
f) Filtered level-1 and level-2 PPI

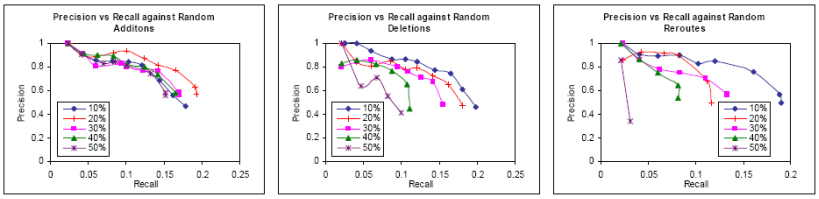
- **Precision improved after cleaning in all methods**
- **PCP performs best**

Validation on PC_{2006}

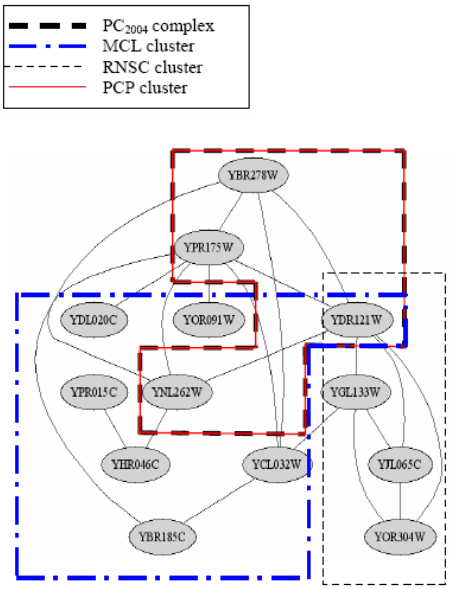


- **When predictions are validated against PC_{2006}, precision of all algo improved**
- **Many "false positives" wrt PC_{2004} are actually real**
- **PCP again performs best**

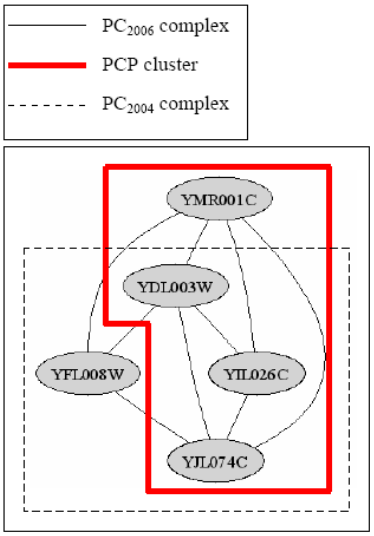
Robustness of PCP Against Noise



- **PCP is robust against 10-50% random additions**
  - FW-weight is able to remove spurious interactions
- **Random deletions negatively impacts recall**
  - Increased sparseness caused edges to received smaller FS-weight; more interactions got filtered
  - Led to insufficient info to form good cliques



PCP Prediction Example 1



PCP Prediction Example 2

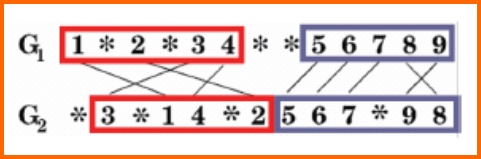
What have we learned?

- **Protein complex/module prediction applications**
  - MCL, PCP
- **General approaches to clustering (of graph nodes)**
  - Random walks, clique merging
- **Important tactics**
  - Cleansing of input
    - **Precision of protein complex prediction can be improved by PPI network augmented with level-2 interactions & cleansed by FS-weight**

References

- J. Chen et al, “Increasing confidence of protein-protein interactomes”, *GIW*, 17:284-297, 2006
- H.N. Chua et al. “Using indirect protein-protein interactions for protein complex prediction”, *Proc. CSB ‘07*, pp. 97-110

Syntenic Gene Cluster Determination



Motivation for Syntenic Gene Clusters

- **Gene order may be altered due to rearrangements**
- **Rearrangements fatal to organism are discarded**
- **Non-functionally related regions accumulate more rearrangements**

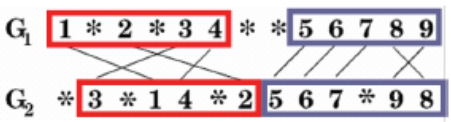
⇒ **By identifying syntenic gene clusters (regions of similarity between multiple genomes), we can discover functional modules**

- **Challenges**
  - Order of genes not preserved
  - Clusters may not be contiguous

A Definition of Syntenic Gene Clusters

- **Let S be a set of gene labels, then a gene order G of length n is a sequence $g_1 g_2 \ldots g_n$ over the the set $S \cup \{*\}$.**

- **Given a set of k gene orders and a parameter d, a gene team is a set of genes in which every adjacent pair of genes is separated by a gap of length at most d in every one of the k gene orders**



Some Possible Solutions

- **Frequent sequential itemset mining**
- **Clustering by gene neighbourhood profile similarity**
  - Gene neighbourhood profile = genes within d positions on either side of the reference gene

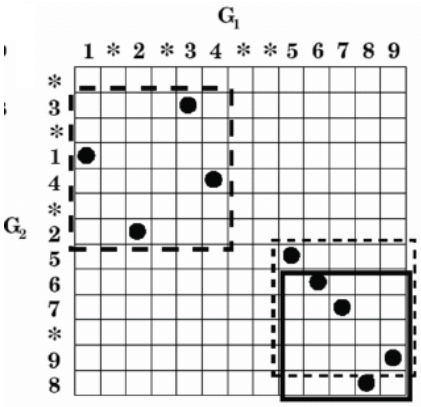
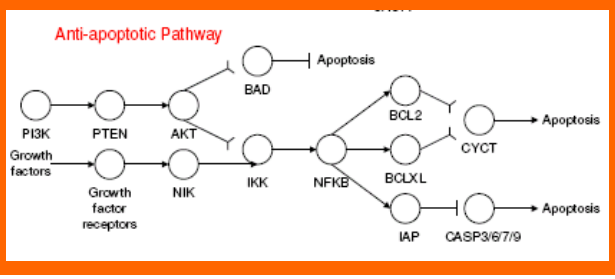


Image credit: Melvin Zhang

References

- A. Bergeron et al., “The algorithmic of gene teams”. *Proc. WABI ’02*, pp. 464–476
- S. Kim et al., “Gene teams with relaxed proximity constraint”. *Proc CSB ’05*, pp. 44–55
- H. Wu et al., “Prediction of functional modules based on gene distributions in microbial genomes”. *GIW*, 16 (2), 247–59, 2005.

Drug Pathway Inference



What's a Microarray?

- **Contain large number of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers**
- **Detect what genes are being expressed or found in a cell of a tissue sample**
- **Measure expression of thousands of genes simultaneously**

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-Murl	5	2	19	297.5	A	M16762 Mouse interleukin 2 (IL-2) gene, exon 4
AFFX-Murl	3	2	19	554.2	A	M37897 Mouse interleukin 10 mRNA, complete cds
AFFX-Murl	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-BioE	13	1	19	9340.6	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioE	15	0	19	12862.4	P	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioE	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-BioE	7	5	18	-483	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioE	5	4	18	313.7	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r
AFFX-BioE	7	6	20	-1016.2	A	J04423 E coli bioB gene biotin synthetase (-5, -M, -3 r

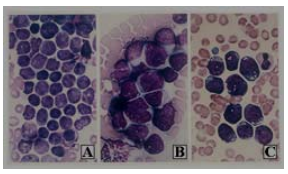
Childhood ALL

- **Major subtypes are:**
  - T-ALL, E2A-PBX, TEL-AML, MLL genome arrangements, BCR-ABL, Hyperdiploid>50
- **Diff subtypes respond differently to same Tx**

⇒ **Over-intensive Tx**
  - Development of sec cancers
  - Reduction of IQ

⇒ **Under-intensiveTx**
  - Relapse

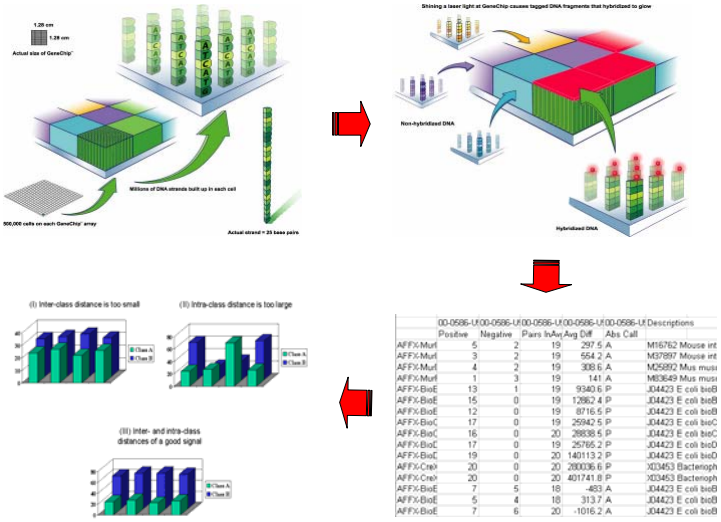
- **The subtypes look similar**



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

⇒ **Unavailable in most ASEAN countries**

Single-Test Platform of Microarray & Machine Learning



Overall Strategy

Diagnosis of subtype → **Subtype-dependent prognosis** → **Risk-stratified treatment intensity**

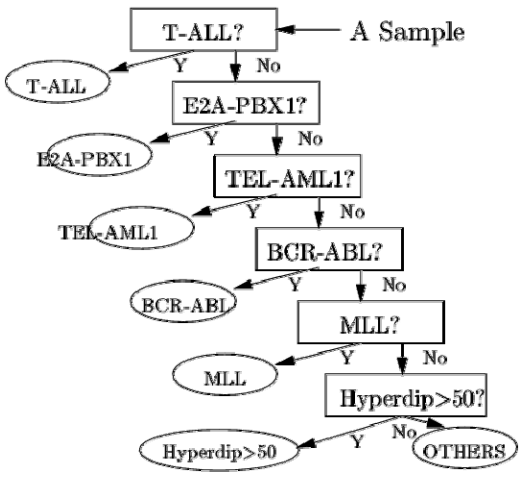
- For each subtype, select genes to develop classification model for diagnosing that subtype
- For each subtype, select genes to develop prediction model for prognosis of that subtype

Subtype Diagnosis by PCL

- **Gene expression data collection**
- ~~**Feature expression feature generation**~~
- **Gene selection (e.g., by $\chi 2$)**
- **Feature integration using classifier (e.g., SVM)**

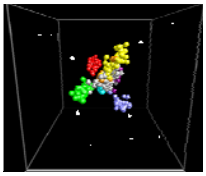
Childhood ALL Subtype Diagnosis Workflow

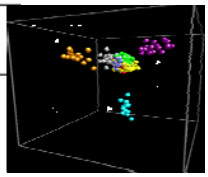
A tree-structured diagnostic workflow was recommended by our doctor collaborator



Accuracy of Various Classifiers

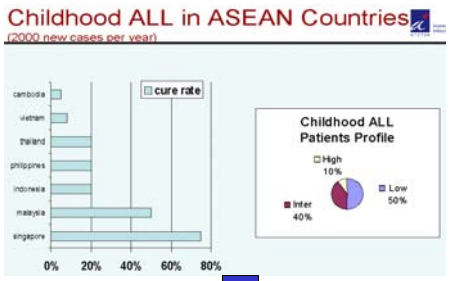
Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS1	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS2	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS3	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS4	2:0	3:0	1:4	2:0
MLL vs OTHERS5	0:1	0:0	0:0	0:0
Hyperdiploid>50 vs OTHERS	2:6	0:2	0:2	0:1
Total Errors	14	6	8	4





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

Practical Outcome



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

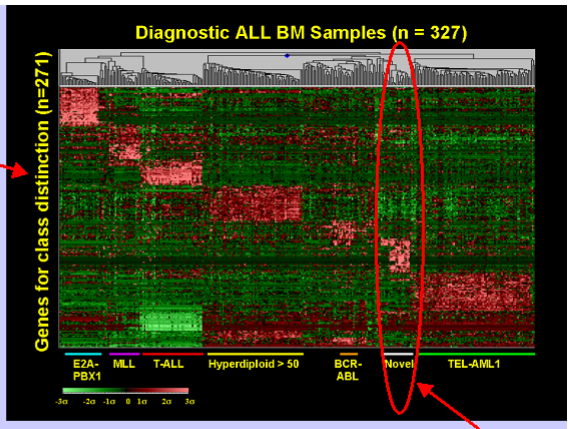
Questions

- **How do you know these are all the ALL subtypes? How do you detect the emergence of a new subtype?**
- **When difference test statistics are used, different genes are selected. In general, they lead to similar level of classification accuracy. So which ones are really relevant?**

Is there a new subtype?

Genes selected by $\chi 2$

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



New subtype discovered

Which genes are really relevant?

- **Intersection Analysis**
  - List A: Genes selected by a feature selection technique
  - List B: Genes from a known pathway P
  - Is the intersection of A and B significant?
- **Do intersection analysis above for as many pathways P_1, P_2, … as possible. Suggest those that are significant to be relevant**

- **How do you test if the intersection of A and B is significant?**

$$pval = \sum_{z \geq z_0} \frac{\binom{n}{z}\binom{n-z}{a-z}\binom{n-a}{b-z}}{\binom{n}{a}\binom{n}{b}}$$

- **n = # genes in P on array**
- **a = |A|**
- **b = |B|**
- **z_0 = |A|$\cap$|B|**

Criticisms on Intersection Analysis

- **List of diff expressed genes (A) defined using test statistics such as χ2 with arbitrary thresholds**
  - Diff test statistics and diff thresholds often result in diff list of diff expressed genes
- **Outcome of whole procedure (the biological pathways that are significant) is not stable wrt variations in test statistics and thresholds used**

Which genes are really relevant?

- **Direct Pathway Analysis**
  - Take a known pathway P, compute

$$z(P) = \frac{1}{\sqrt{k}}\left(\sum_{q:x\to y\in P} z(q) - \sum_{q:x-|y\in P} z(q)\right)$$

  - Estimate pvalue for z(P)

- **Do direct pathway analysis above for as many pathways P_1, P_2, … as possible. Suggest those that are significant to be relevant**

- **How do you compute z(q)?**

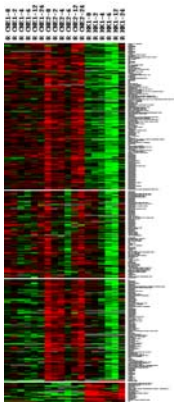
$$z(x \to y) = \frac{corr(x,y) - \mu}{\sigma}$$

where μ, σ are estimated corr(x',y') of all possible pairs of genes x', y'

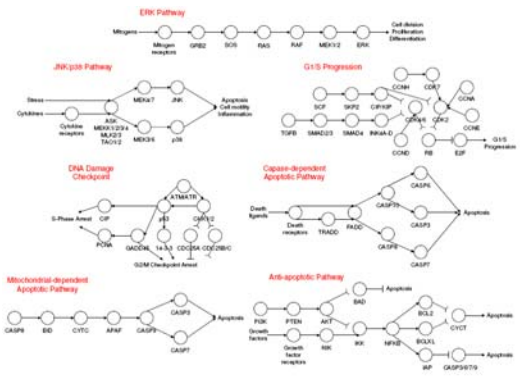
- **Exercise: How do you compute pvalue for z(P)?**

CYC202 Drug Pathway in NPC

- **CNE1 responds poorly to CYC202, CNE2 responds limitedly,HK1 responds fully**

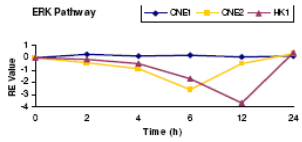


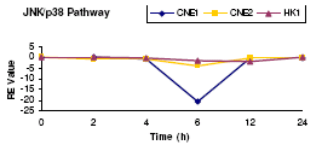
- **How does CYC202 act on HK1? How do CNE1/2 escape?**



Example Significant Pathways Detected

Signaling Pathway	Genetic Pathway	Confidence
CNE1		
ERK	GRB2→SOS2→HRAS→RAF1→MAP2K1→ MAPK1/MAPK3	> 0.9986
JNK/p38	TRADD→RIPK1→TRAF2→MAP3K3/MAP3K11→MAP2K7→MAPK9	> 0.9993
Apoptosis	PIK3CB→PTEN→AKT1→IKBKB→NFKB2→BIRC2/BIRC5→CASP3/CASP7	> 0.9997
G1/S Transition	RBX1/SKP1A→SKP2→CDKN1B→CDK6→RB1→E2F1	0.9779
DNA Damage	ATM→CHEK2→TP53→GADD45B/GADD45G	0.9945
CNE2		
ERK	GRB2→SOS1→MRAS/KRAS/NRAS/RRAS→BRAF→MAP2K1→MAPK1	0.9947
JNK/p38	IRAK1→TRAF6→MAP3K1→MAP2K4→MAPK8	0.9844
Apoptosis	PIK3CA→PTEN→AKT1→IKBKB/IKBKG→NFKB2→BIRC3→CASP6	> 0.9967
HK1		
ERK	GRB2→SOS1→HRAS→BRAF→MAP2K1/MAP2K2→MAPK1	> 0.9833
JNK/p38	IRAK1→TRAF6→MAP3K4→MAP2K4→MAPK8	0.9891
Apoptosis	TRADD→FADD→CASP8→BID→CYCS→APAF1→CASP9→CASP7	0.9593
	TRADD→FADD→CASP8→CASP7	0.9507





- **ERK regulates cell survival & proliferation**
- **ERK suppressed in HK1**

⇒ **CYC202 suppresses ERK to inhibit cancer cell growth**

- **JNK activation suppressed cell viability**
- **JNK suppressed in CNE1**

⇒ **CNE1 escapes CYC202 thru JNK inactivation**

Analysis by Difeng Dong

What have we learned?

- **Gene expression analysis applications**
  - Disease subtype diagnosis
  - Disease subtype discovery
  - Drug response pathway understanding
- **General approaches**
  - "Feature selection, feature integration"
- **Important tactics**
  - Intersection analysis
  - Direct pathway analysis

References

- E.-J. Yeoh et al., "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling", *Cancer Cell*, 1:133--143, 2002
- J.Li, L. Wong, "Identifying good diagnostic genes or gene groups from gene expression data by using the concept of emerging patterns", *Bioinformatics*, 18:725--734, 2002
- D. Soh et al., "Enabling More Sophisticated Gene Expression Analysis for Understanding Diseases and Optimizing Treatments". *ACM SIGKDD Explorations*, 9(1):3--14, 2007

Any Question?