For written notes on this lecture, please read chapter 14 of *The Practical Bioinformatician*,

CS2220: Introduction to Computational Biology

Lecture 4: Gene Expression Analysis

Limsoon Wong

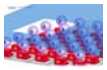
Plan

- **Microarray background**
- **Gene expression profile classification**
- **Gene expression profile clustering**
- **Normalization**
- **Extreme sample selection**
- **Intersection Analysis**

Background on Microarrays

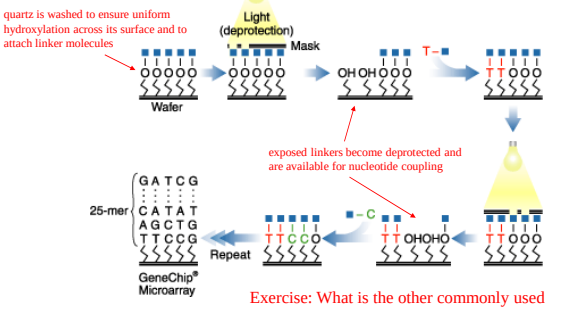
What is a Microarray?

- **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**



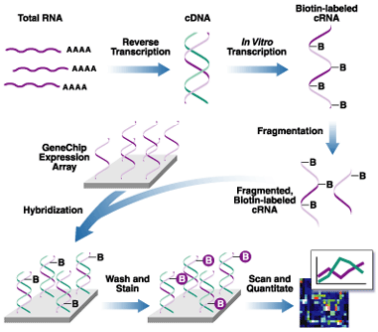
Affymetrix GeneChip Array

Making Affymetrix GeneChip Array



Exercise: What is the other commonly used type of microarray? How is that one different from Affymetrix's?

Gene Expression Measurement by Affymetrix GeneChip Array



Click to watch an interesting movie explaining the working of microarray

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	M25692 Mus musculus interleukin 4 (Il-4) mRNA, comp
AFFX-MurI	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

Some Advice on Affymetrix Gene Chip Data

- **Ignore AFFX genes**
  - These genes are control genes
- **Ignore genes with "Abs Call" equal to "A" or "M"**
  - Measurement quality is suspect
- **Upperbound 40000, lowerbound 100**
  - Accuracy of laser scanner
- **Deal with missing values**

Exercise: Suggest 2 ways to deal with missing value

Type of Gene Expression Datasets

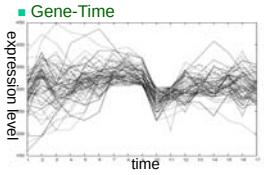
- Gene-Conditions or **Gene-Sample** (**numeric** or discretized)

1000 - 100,000 columns

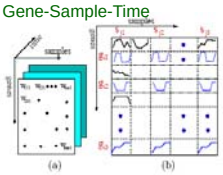
100-500 rows

	Class	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	.....	
Sample1	Cancer	0.12	-1.3	1.7	1.0	-3.2	0.78	-0.12		
Sample2	Cancer							1.3		
.										
	~Cancer									
SampleN	~Cancer									

- Gene-Time



- Gene-Sample-Time



Type of Gene Expression Datasets

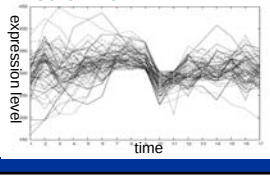
- Gene-Conditions or **Gene-Sample** (numeric or **discretized**)

1000 - 100,000 columns

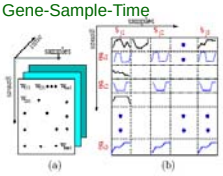
100-500 rows

	Class	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	.....	
Sample1	Cancer	1	0	1	1	1	0	0		
Sample2	Cancer							1		
.										
	~Cancer									
SampleN	~Cancer									

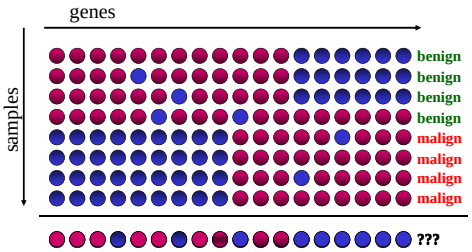
- Gene-Time



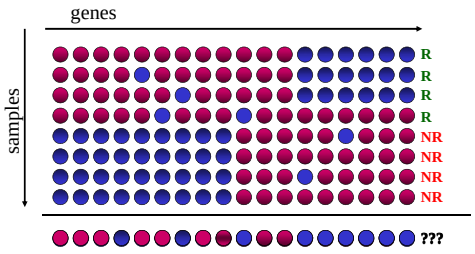
- Gene-Sample-Time



Application: Disease Subtype Diagnosis



Application: Treatment Prognosis



Type of Gene Expression Datasets

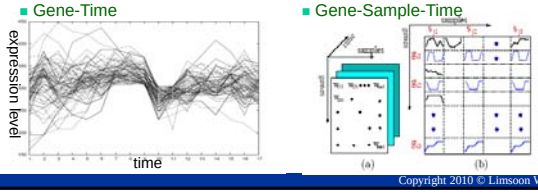
- **Gene-Conditions** or Gene-Sample (**numeric** or discretized)

1000 - 100,000 columns

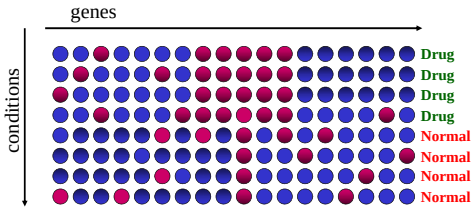
100-500 rows

	Gene1	Gene2	Gene3	Gene 4	Gene5	Gene6	Gene7		
Cond1	0.12	-1.3	1.7	1.0	-3.2	0.78	-0.12		
Cond2							1.3		
.									
CondN									

- Gene-Time
- Gene-Sample-Time



Application: Drug Action Detection



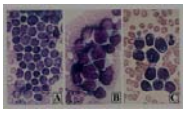
Which group of genes are the drug affecting on?

Gene Expression Profile Classification

Diagnosis of Childhood Acute Lymphoblastic Leukemia and Optimization of Risk-Benefit Ratio of Therapy

Childhood ALL

- **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**



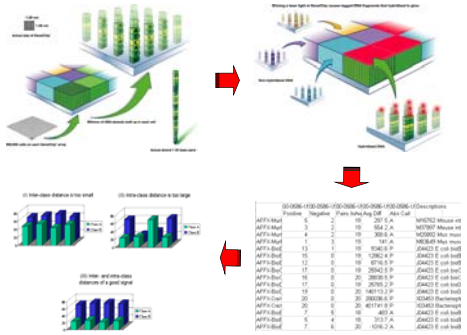
- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- **Unavailable in most ASEAN countries**

Mission

- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists**
- **Generally available only in major advanced hospitals**

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

Single-Test Platform of Microarray & Machine Learning



Overall Strategy

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

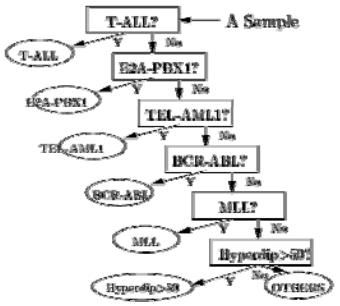
- For each subtype, select genes to develop classification model for diagnosing that subtype
- For each subtype, select genes to develop prediction model for prognosis of that subtype

Subtype Diagnosis by PCL

- Gene expression data collection
- 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

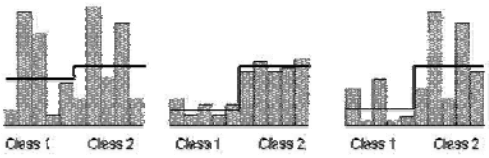


Training and Testing Sets

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 109	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, None}	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 $EP_1^P, \ldots, EP_t^P$ 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):

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

- 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

EP_1^P (90%)
EP_2^P (86%)
.
.
EP_n^P (68%)

Top-Ranked EPs in Negative class

EP_1^N (100%)
EP_2^N (95%)
.
.
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

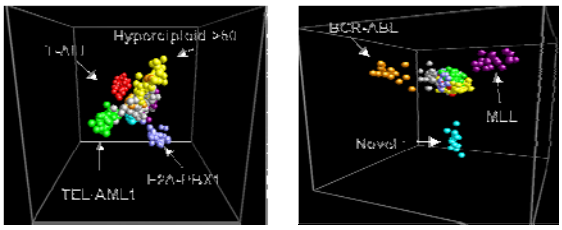
- **E.g., for T-ALL vs. OTHERS, one ideally discriminatory gene 38319_at was found, inducing these 2 EPs**

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

- **These give us the diagnostic rule**

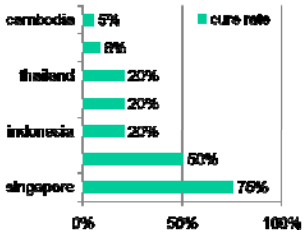
If the expression of 38 319_at is less than 15 975.6, then this ALL sample must be a T-ALL. Otherwise it must be a subtype in OTHERS1.

Multidimensional Scaling Plot for Subtype Diagnosis



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

Childhood ALL Cure Rates



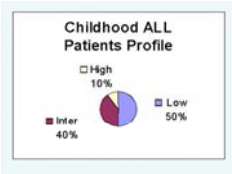
- **Conventional risk assignment procedure requires difficult expensive tests and collective judgement of multiple specialists**

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

Childhood ALL Treatment Cost

- **Treatment for childhood ALL over 2 yrs**
  - Intermediate intensity: US$60k
  - Low intensity: US$36k
  - High intensity: US$72k
- **Treatment for relapse: US$150k**
- **Cost for side-effects: Unquantified**

Current Situation (2000 new cases/yr in ASEAN)



- **Intermediate intensity conventionally applied in less advanced ASEAN countries**

- **Over intensive for 50% of patients, thus more side effects**
- **Under intensive for 10% of patients, thus more relapse**
- **US$120m (US$60k * 2000) for intermediate intensity tx**
- **US$30m (US$150k * 2000 * 10%) for relapse tx**
- **Total US$150m/yr plus un-quantified costs for dealing with side effects**

Using Our Platform

- Low intensity applied to 50% of patients
- Intermediate intensity to 40% of patients
- High intensity to 10% of patients

⇒ Reduced side effects
⇒ Reduced relapse
⇒ 75-80% cure rates

- US$36m (US$36k * 2000 * 50%) for low intensity
- US$48m (US$60k * 2000 * 40%) for intermediate intensity
- US$14.4m (US$72k * 2000 * 10%) for high intensity

- Total US$98.4m/yr

⇒ Save US$51.6m/yr

A Nice Ending…

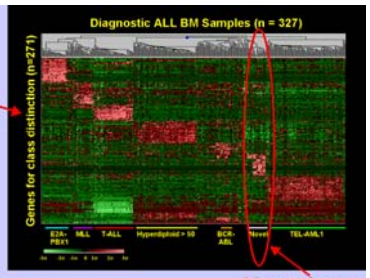
- Asian Innovation Gold Award 2003

Gene Expression Profile Clustering

Novel Disease Subtype Discovery

Is there a new subtype?

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

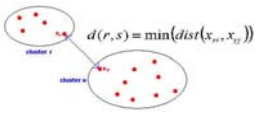


Exercise: Name and describe one bi-clustering method

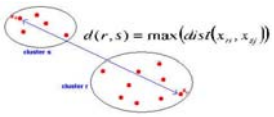
Hierarchical Clustering

- 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



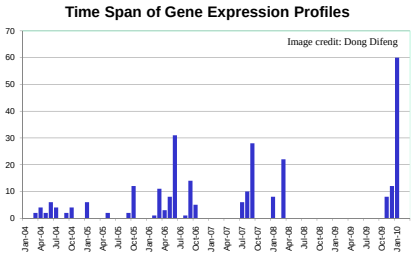
Complete linkage defines distance betw two clusters as max distance betw them

Exercise: Give definition of "average linkage"

Image source: UCL Microcore Website

Normalization

Sometimes, a gene expression study may involve batches of data collected over a long period of time…



In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

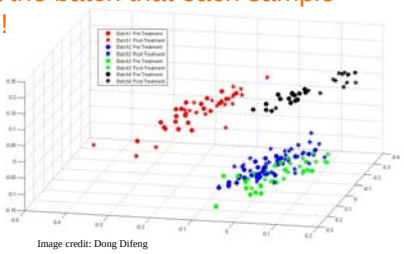


Image credit: Dong Difeng

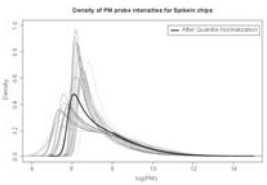
⇒ **Need normalization to correct for batch effect**

Approaches to Normalization

- **Aim of normalization: Reduce variance w/o increasing bias**
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's
- **Xform data so that distribution of probe intensities is same on all arrays**
  - E.g., $(x - \mu) / \sigma$
- **Quantile normalization**

Quantite Normalization

- **Given *n arrays of length p, form X of size p × n where each array is a column***
- **Sort each column of *X to give* X_{sort}**
- **Take means across rows of X_{sort} *and assign this* mean to each elem in the row to get X'_{sort}**
- **Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as *X***



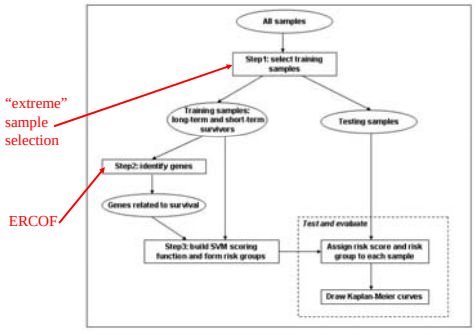
- Implemented in some microarray s/w, e.g., EXPANDER

Selection of Patient Samples and Genes for Disease Prognosis

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)
- **……**

Our Approach

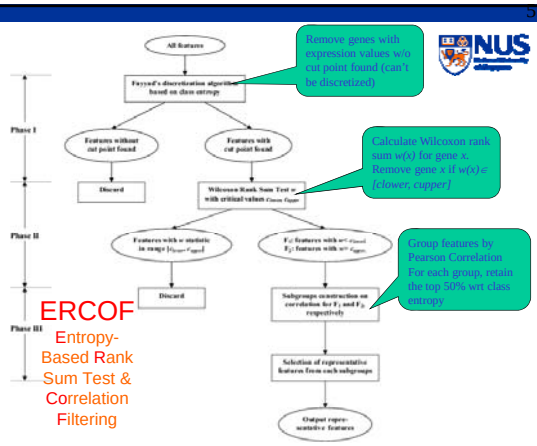


Extreme Sample Selection

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

Short-term survivors who died within a *short* period	*Long-term survivors* who were alive after a *long* follow-up time
⇓	⇓
$F(T) < c_1$ and $E(T) = 1$	$F(T) > c_2$

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**

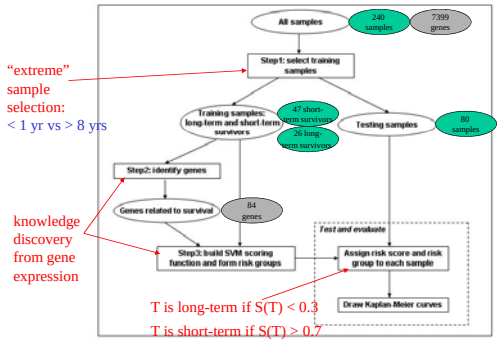
- **Intl Prognostic Index (IPI)**
  - age, "Eastern Cooperative Oncology Group" Performance status, tumor stage, lactate dehydrogenase level, sites of extranodal disease, ...
- **Not very good for stratifying DLBC lymphoma patients for therapeutic trials**

⇒ **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



Discussions: Sample Selection

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

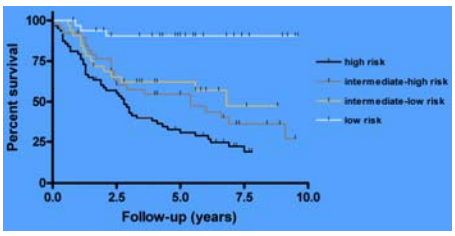
Number of samples in original data and selected informative training set.
(*): Number of samples whose corresponding patient was dead at the end of follow-up time, but selected as a long-term survivor.

Discussions: Gene Identification

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

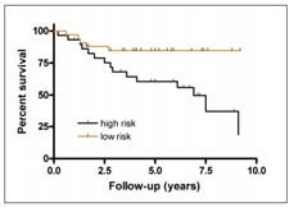
Number of genes left after feature filtering for each phase.
(*): number of genes after removing those genes who were absent in more than 10% of the experiments.

Kaplan-Meier Plot for 80 Test Cases

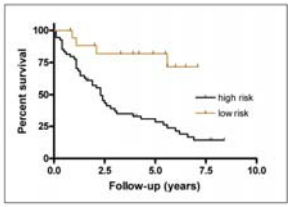


p-value of log-rank test: < 0.0001
Risk score thresholds: 0.7, 0.3

Improvement Over IPI

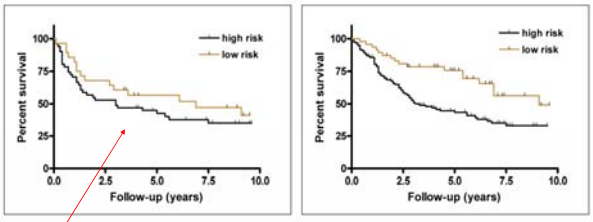


(A) IPI low,
p-value = 0.0063



(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**

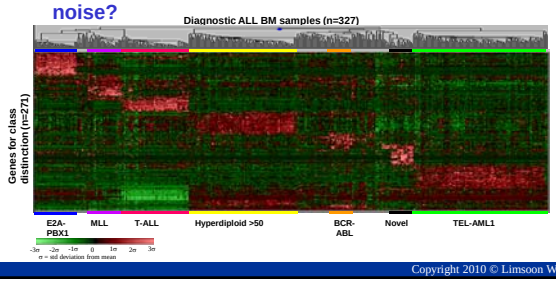
About the Inventor: Huiqing Liu

- **Huiqing Liu**
  - PhD, NUS, 2004
  - Currently Senior Scientist at Centocor
  - Asian Innovation Gold Award 2003
  - New Jersey Cancer Research Award for Scientific Excellence 2008
  - Gallo Prize 2008

Beyond Disease Diagnosis & Prognosis

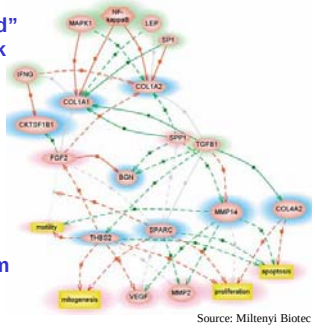
Beyond Classification of Gene Expression Profiles

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



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

Hints to extend reach of prediction

- **Each disease subtype has underlying cause**

⇒ **There is a unifying biological theme for genes that are truly associated with a disease subtype.**

- **Uncertainty in reliability of selected genes can be reduced by considering molecular functions and biological processes associated with the genes**
- **The unifying biological theme is basis for inferring the underlying cause of disease subtype**

Intersection Analysis

- **Intersect the list of differentially expressed genes with a list of genes on a pathway**
- **If intersection is significant, the pathway is postulated as basis of disease subtype or treatment response**

Exercise: What is a good test statistics to determine if the intersection is significant?

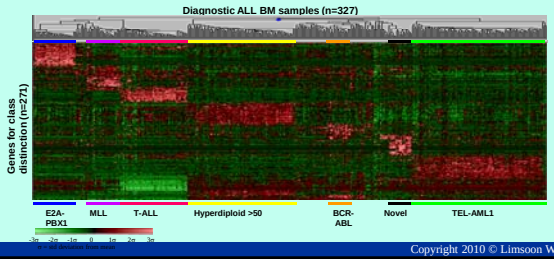
Caution:
- **Initial list of differentially expressed genes is defined using test statistics with arbitrary thresholds**
- **Diff test statistics and diff thresholds result in a diff list of differentially expressed genes**

⇒ **Outcome may be unstable**

Gene Interaction Prediction

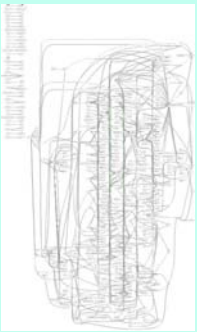
Beyond Classification of Gene Expression Profiles

- **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**
- **Expression of a gene in a network depends on expression of some other genes in the network**
- **Can we reconstruct the gene network from gene expression data?**



Key Questions

- **For each gene in the network:**
- **Which genes affect it?**
- **How they affect it?**
  - Positively?
  - Negatively?
  - More complicated ways?

Some Techniques

- **Bayesian Networks**
  - Friedman et al., *JCB* 7:601--620, 2000
- **Boolean Networks**
  - Akutsu et al., *PSB* 2000, pages 293--304
- **Differential equations**
  - Chen et al., *PSB* 1999, pages 29--40
- **Classification-based method**
  - Soinov et al., "Towards reconstruction of gene network from expression data by supervised learning", *Genome Biology* 4:R6.1--9, 2003

A Classification-Based Technique

Soinov et al., *Genome Biology* 4:R6.1-9, Jan 2003

- **Given a gene expression matrix X**
  - each row is a gene
  - each column is a sample
  - each element x_{ij} is expression of gene i in sample j
- **Find the average value a_i of each gene i**
- **Denote s_{ij} as state of gene i in sample j,**
  - s_{ij} = up if $x_{ij} > a_i$
  - s_{ij} = down if $x_{ij} \leq a_i$

A Classification-Based Technique

Soinov et al., *Genome Biology* 4:R6.1-9, Jan 2003

- **To see whether the state of gene g is determined by the state of other genes**
  - see whether $\langle s_{ij} \mid i \neq g \rangle$ can predict s_{gj}
  - if can predict with high accuracy, then "yes"
  - Any classifier can be used, such as C4.5, PCL, SVM, etc.
- **To see how the state of gene g is determined by the state of other genes**
  - apply C4.5 (or PCL or other "rule-based" classifiers) to predict s_{gj} from $\langle s_{ij} \mid i \neq g \rangle$
  - and extract the decision tree or rules used

Advantages of this method

- **Can identify genes affecting a target gene**
- **Don't need discretization thresholds**
- **Each data sample is treated as an example**
- **Explicit rules can be extracted from the classifier (assuming C4.5 or PCL)**
- **Generalizable to time series**

Concluding Remarks

Bcr-Abl

- **Targeted drug dev**
  - Know what molecular effect you want to achieve
    - **E.g., inhibit a mutated form of a protein**
  - Engineer a compound that directly binds and causes the desired effect
- **Gleevec (imatinib)**
  - 1st success for real drug
  - Targets Bcr-Abl fusion protein (ie, Philadelphia chromosome, Ph)
  - NCI summary of clinical trial of imatinib for ALL at

http://www.cancer.gov/clinicaltrials/results/ALLimatinib1109/print

What have we learned?

- **Technologies**
  - Microarray
  - PCL, ERCOF
- **Microarray applications**
  - Disease diagnosis by supervised learning
  - Subtype discovery by unsupervised learning
- **Important tactics**
  - Extreme sample selection
  - Intersection analysis, Gene network reconstruction

Any Question?

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
- H. Liu, J. Li, L. Wong. Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data. *Bioinformatics*, 21(16):3377--3384, 2005.
- L.D. Miller et al., "Optimal gene expression analysis by microarrays", *Cancer Cell* 2:353--361, 2002
- J. Li, L. Wong, "Techniques for Analysis of Gene Expression", *The Practical Bioinformatician*, Chapter 14, pages 319—346, WSPC, 2004
- B. Bolstad et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias". *Bioinformatics*, 19:185–193. 2003