

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
6 February 2009



2

Plan



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

Copyright 2009 © Limsoon Wong

Background on Microarrays

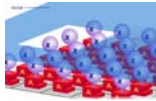


4

What's a Microarray?



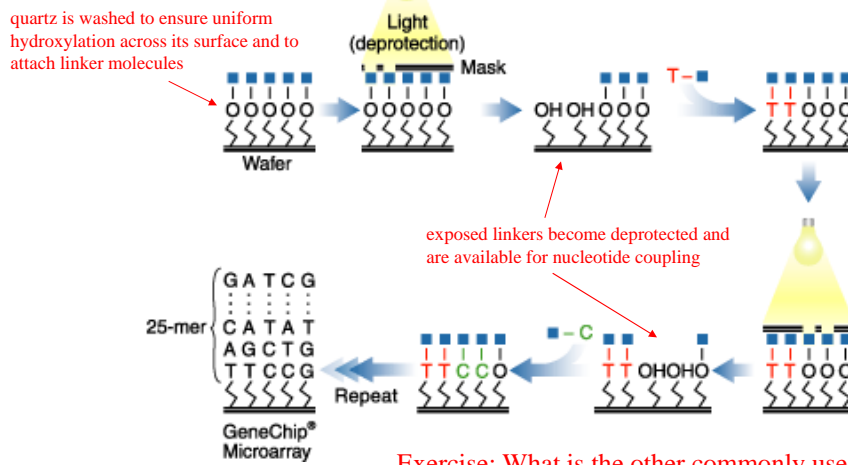
- **Contain large number of DNA molecules spotted on glass slides, nylon membranes, or silicon wafers**
- **Detect what genes are being expressed or found in a cell of a tissue sample**
- **Measure expression of thousands of genes simultaneously**



Affymetrix GeneChip Array

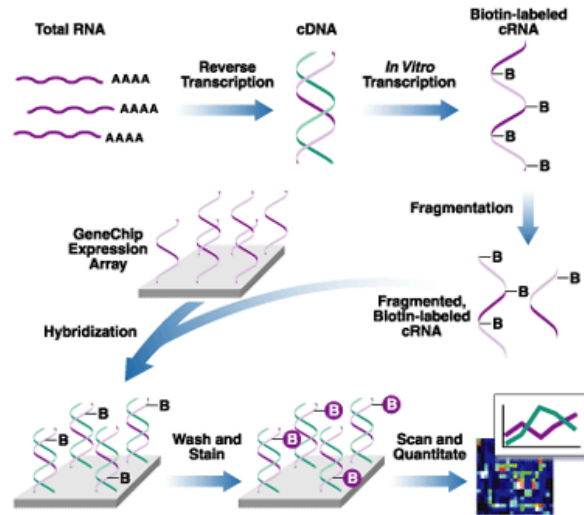


Making Affymetrix GeneChip Array



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



Copyright 2009 © Limsoon Wong

A Sample Affymetrix GeneChip Data File (U95A)



	00-0586-UX	00-0586-UX	00-0586-UX	00-0586-UX	00-0586-UX	Descriptions
	Positive	Negative	Pairs In	Avg 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, complete cds
AFFX-Murf	1	3	19	141	A	M83649 Mus musculus Fas antigen mRNA, complete cds
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

Copyright 2009 © Limsoon Wong



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

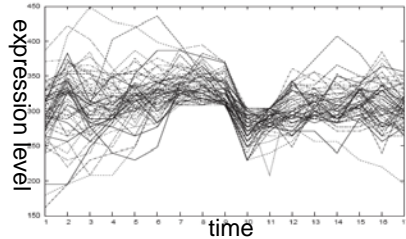
- Gene-Conditions or **Gene-Sample (numeric or discretized)**

← 1000 - 100,000 columns →

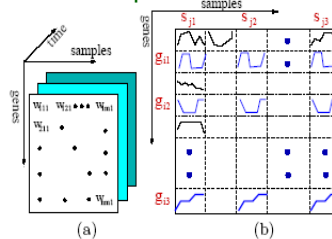
	Class	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7
Sample1	Cancer	0.12	-1.3	1.7	1.0	-3.2	0.78	-0.12	
Sample2	Cancer							1.3	
SampleN	-Cancer								

↑ 100-500 rows ↓

- Gene-Time



- Gene-Sample-Time



Type of Gene Expression Datasets

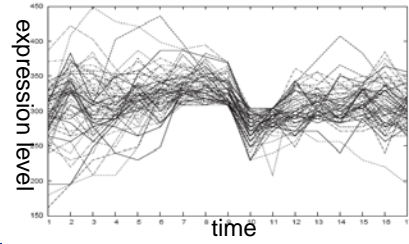
- Gene-Conditions or **Gene-Sample** (numeric or discretized)

← 1000 - 100,000 columns →

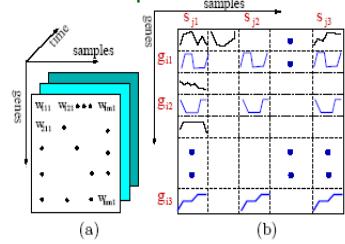
↑ 100-500 rows ↓

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

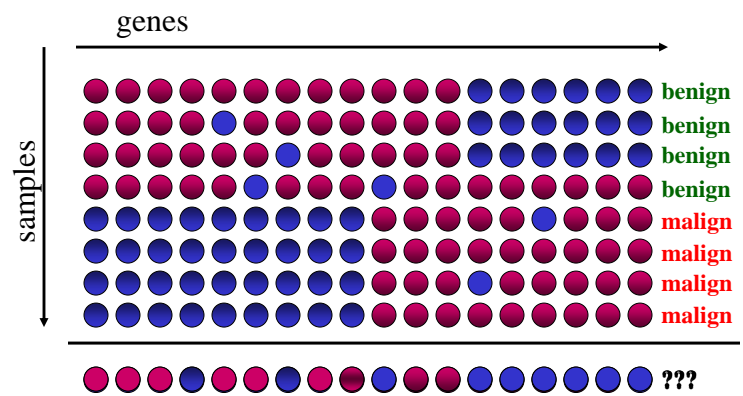
- Gene-Time



- Gene-Sample-Time

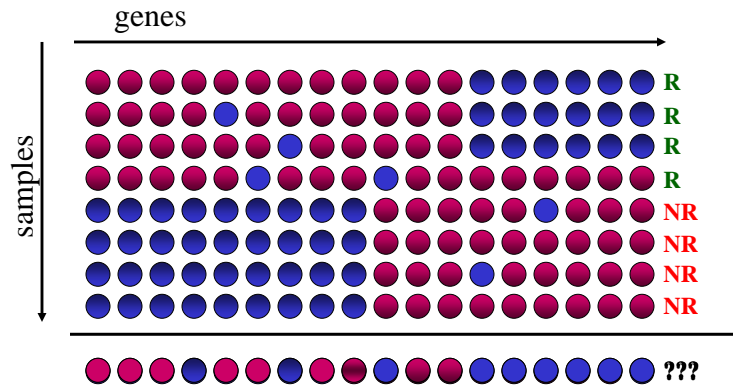


Application: Disease Subtype Diagnosis





Application: Treatment Prognosis



Type of Gene Expression Datasets

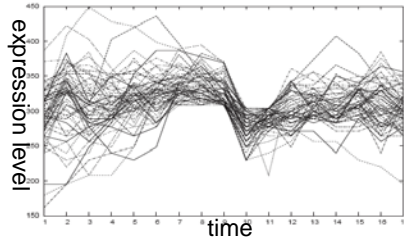
- Gene-Conditions or Gene-Sample (numeric or discretized)

1000 - 100,000 columns

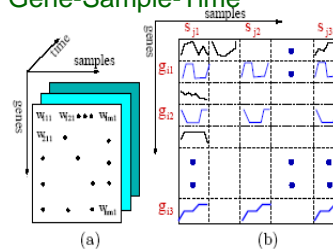
	Gene1	Gene2	Gene3	Gene 4	Gene5	Gene6	Gene7		
Cond1	0.12	-1.3	1.7	1.0	-3.2	0.78	-0.12		
Cond2							1.3		
...									
CondN									

100-500 rows

- Gene-Time

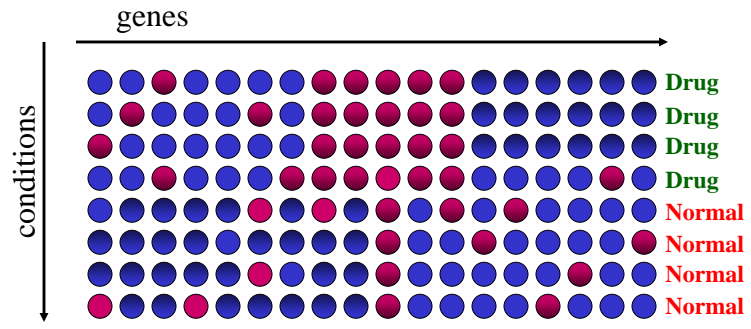


- Gene-Sample-Time





Application: Drug Action Detection



Which group of genes are the drug affecting on?

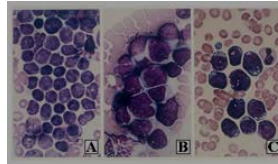
Gene Expression Profile Classification

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-intensive Tx
 - Relapse
- The subtypes look similar



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

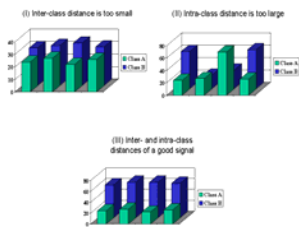
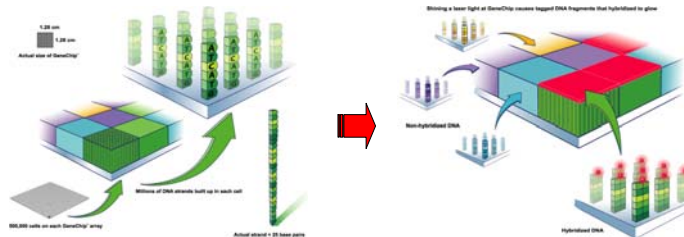
Copyright 2009 © Limsoon Wong

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?**

Copyright 2009 © Limsoon Wong

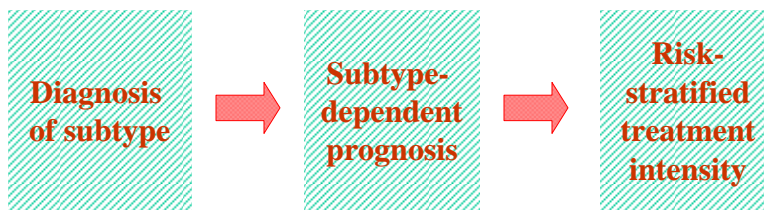
Single-Test Platform of Microarray & Machine Learning



	Positive	Negative	Pairs	In	Avg	Diff	Abs	Call	Descriptions
AFFX:Murf	5	2	19	207.5	A	M16762	Mouse int		
AFFX:Murf	3	2	19	554.2	A	M37997	Mouse int		
AFFX:Murf	4	2	19	303.6	A	M25992	Mus mus		
AFFX:Murf	1	3	19	141	A	M23649	Mus mus		
AFFX:BioE	13	1	19	9340.6	P	J04423	E coli bioB		
AFFX:BioE	15	0	19	12602.4	P	J04423	E coli bioB		
AFFX:BioE	12	0	19	8716.5	P	J04423	E coli bioB		
AFFX:BioC	17	0	19	25642.5	P	J04423	E coli bioC		
AFFX:BioC	16	0	20	28838.5	P	J04423	E coli bioC		
AFFX:BioC	17	0	19	25765.2	P	J04423	E coli bioC		
AFFX:BioC	19	0	20	140113.2	P	J04423	E coli bioC		
AFFX:Creb	20	0	20	280036.6	P	J03463	Bacterioph		
AFFX:Creb	20	0	20	401741.8	P	J03463	Bacterioph		
AFFX:BioE	7	5	18	-483	A	J04423	E coli bioB		
AFFX:BioE	5	4	18	313.7	A	J04423	E coli bioB		
AFFX:BioE	7	6	20	-1016.2	A	J04423	E coli bioB		

Copyright 2009 © Limsoon Wong

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

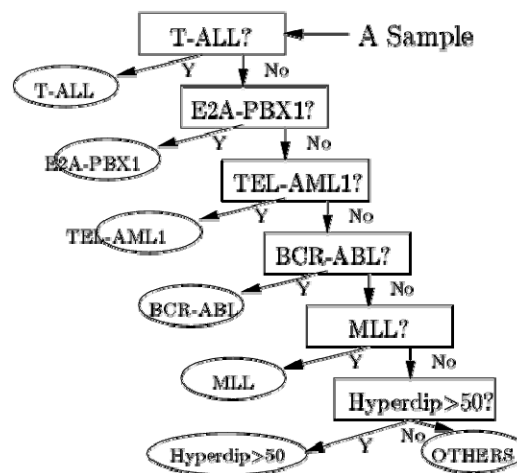
Copyright 2009 © Limsoon Wong

Subtype Diagnosis by PCL

- Gene expression data collection
- Gene selection by χ^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

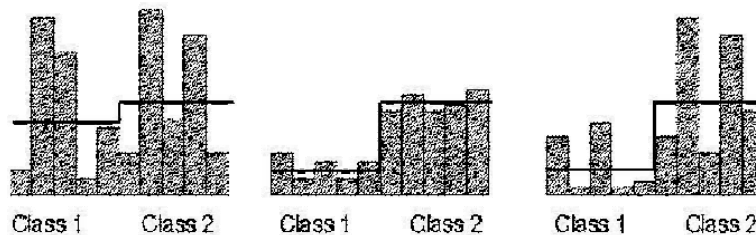


Training and Testing Sets

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 χ^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 i th interval, j th class, R_i the number of samples in the i th interval, C_j the number of samples in the j th 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, \dots, 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):

$$EP_{t_1}^P, EP_{t_2}^P, \dots, EP_{t_k}^P$$

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

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

- Ditto for $\text{score}(T, D^N)$.
- If $\text{score}(T, D^P) > \text{score}(T, D^N)$, 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

$$\text{Score}^P = \frac{EP_1^{P'}}{EP_1^P} + \dots + \frac{EP_k^{P'}}{EP_k^P}$$

Most freq EP of pos class

Similarly,

$$\text{Score}^N = \frac{EP_1^{N'}}{EP_1^N} + \dots + \frac{EP_k^{N'}}{EP_k^N}$$

**If $\text{Score}^P > \text{Score}^N$, 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 χ^2 at each level of the tree

Understandability of PCL

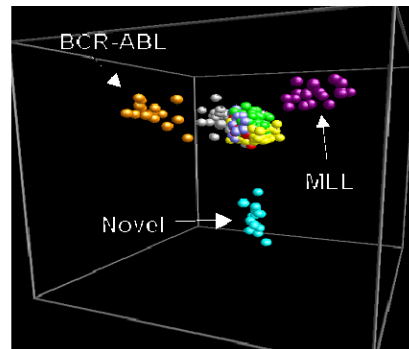
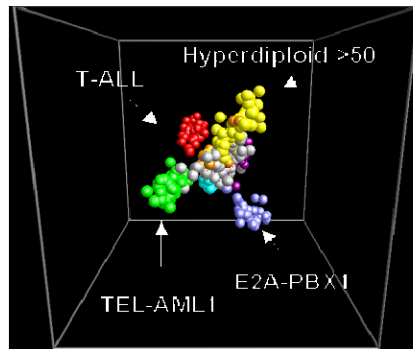
- E.g., for T-ALL vs. OTHERS, one ideally discriminatory gene 38319_at was found, inducing these 2 EPs

$$\{gene_{-(38319_at)} @ (-\infty, 15975.6)\} \text{ and } \{gene_{-(38319_at)} @[15975.6, +\infty)\}.$$

- These give us the diagnostic rule

If the expression of 38319_at is less than 15975.6, then this ALL sample must be a T-ALL.
 Otherwise it must be a subtype in OTHERS1.

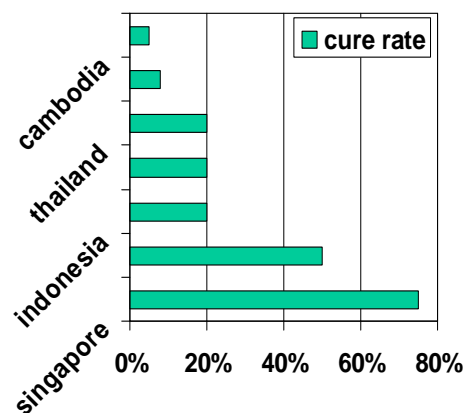
Multidimensional Scaling Plot for Subtype Diagnosis



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

Copyright 2009 © Limsoon Wong

Childhood ALL Cure Rates



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

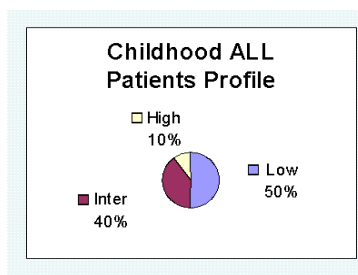
⇒ Not available in less advanced ASEAN countries

Copyright 2009 © Limsoon Wong

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)



- **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 unquantified costs for dealing with side effects**
- **Intermediate intensity conventionally applied in less advanced ASEAN countries**



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**

Copyright 2009 © Limsoon Wong

Gene Expression Profile Clustering

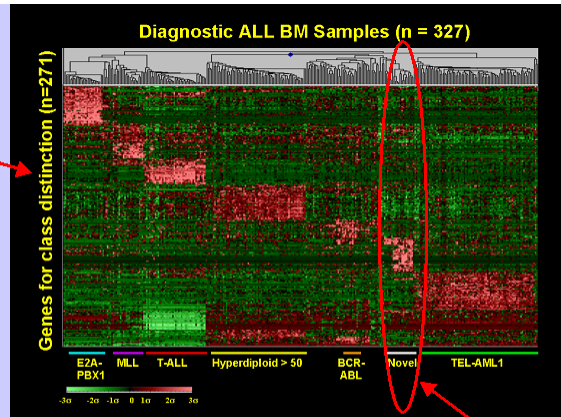
Novel Disease Subtype Discovery



Is there a new subtype?

Genes selected by χ^2

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



New subtype discovered

Exercise: Name and describe one bi-clustering method

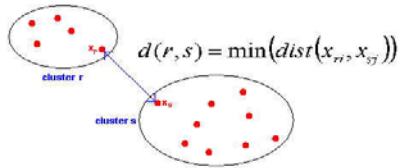
Copyright 2009 © Limsoon Wong

Hierarchical Clustering

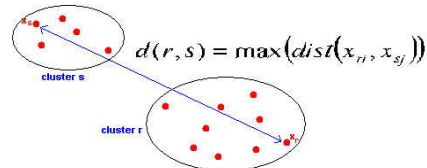
- Assign each item to its own cluster
 - If there are N items initially, we get N clusters, each containing just one item
- Find the “most similar” pair of clusters, merge them into a single cluster, so we now have one less cluster
 - “Similarity” is often defined using
 - Single linkage
 - Complete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

Copyright 2009 © Limsoon Wong

Single, Complete, & Average Linkage



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



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

Exercise: Give definition of “average linkage”

Image source: UCL Microcore Website

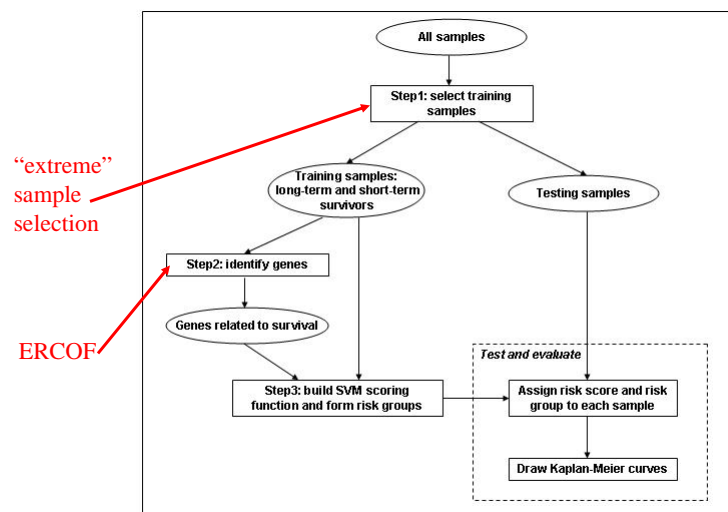
Copyright 2009 © Limsoon Wong

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



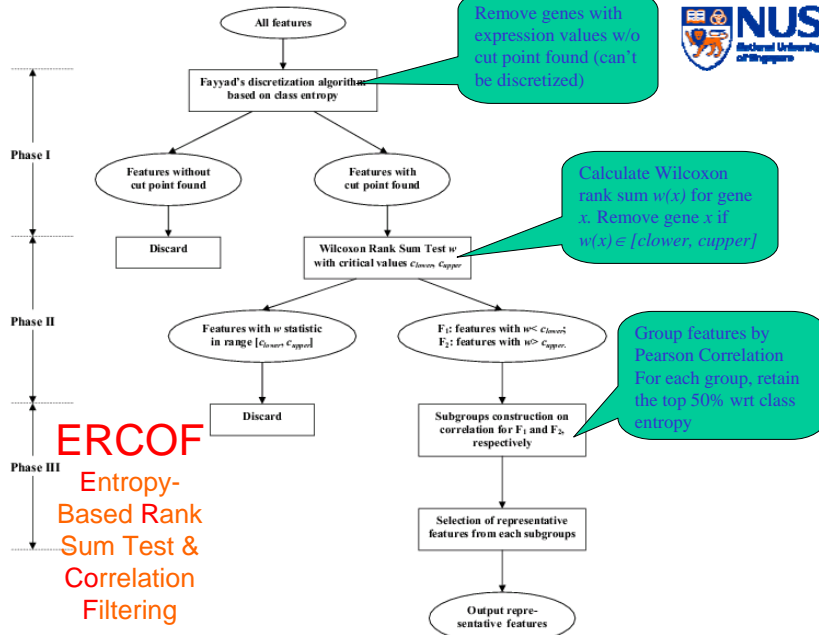
$$F(T) < c_1 \text{ and } E(T) = 1$$

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



$$F(T) > c_2$$

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



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*)

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

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

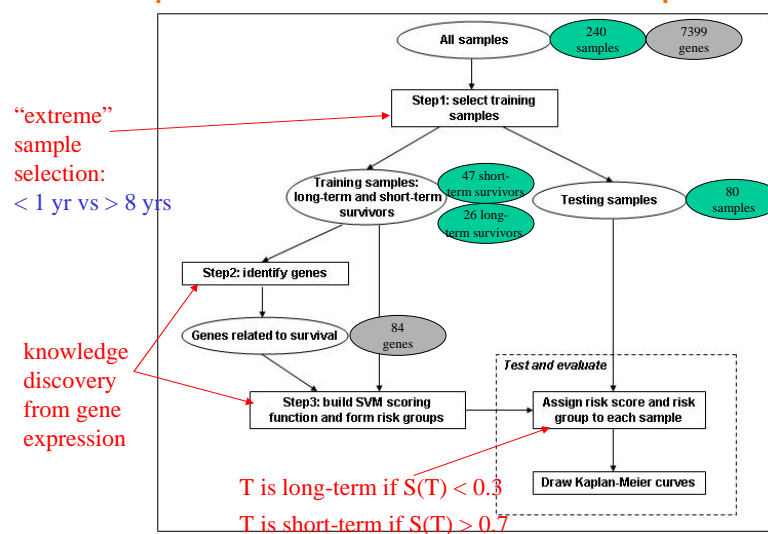
Diffuse Large B-Cell Lymphoma

- 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.

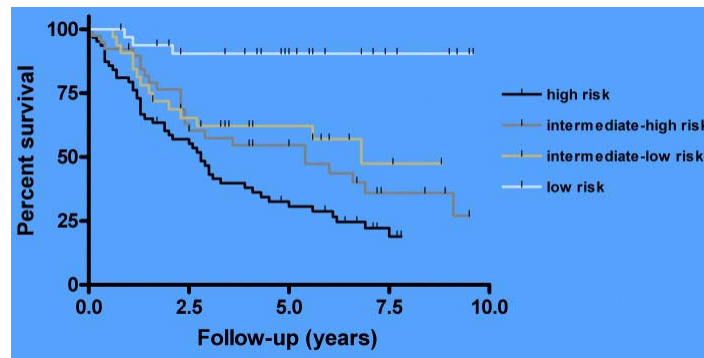
Discussions: Gene Identification



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.

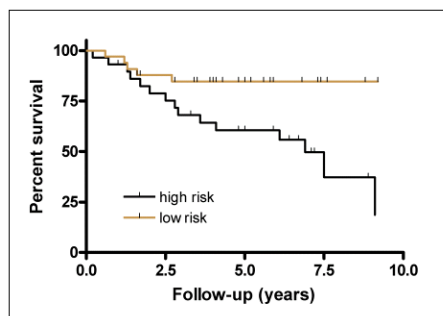
Kaplan-Meier Plot for 80 Test Cases



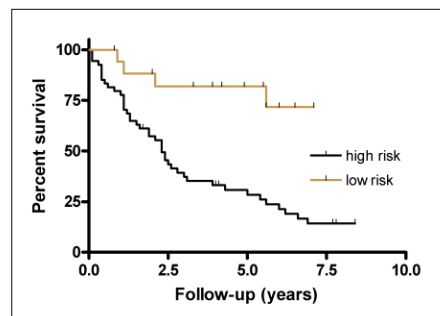
p -value of log-rank test: < 0.0001
Risk score thresholds: 0.7, 0.3

Copyright 2009 © Limsoon Wong

Improvement Over IPI



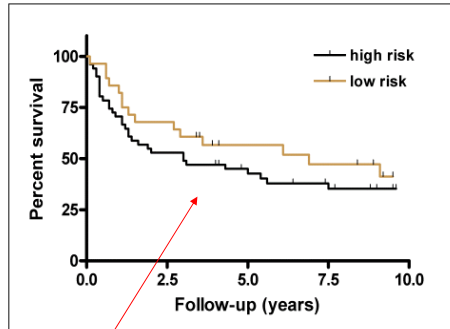
(A) IPI low,
 p -value = 0.0063



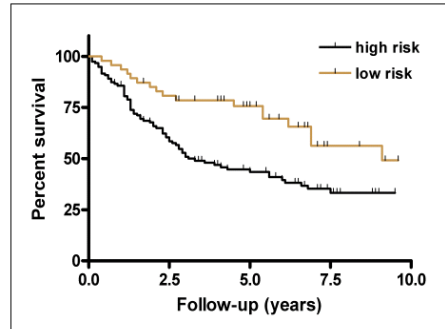
(B) IPI intermediate,
 p -value = 0.0003

Copyright 2009 © Limsoon Wong

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 C:8 Tie:12	A:1 C:12 Tie:7	A:2 C:5 Tie:13	A:0 C:10 Tie:10	A:1 C:14 Tie:5	A:0 C:11 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 C:7 Tie:13	B:3 C:9 Tie:8	B:6 C:4 Tie:10	B:4 C:5 Tie:11	B:4 C:10 Tie:6	B:1 C:7 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		
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

Copyright 2009 © Limsoon Wong

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**
- ⇒ **ERCOF is a systematic feature selection method that is very useful**

Copyright 2009 © Limsoon Wong

Beyond Disease Diagnosis & Prognosis

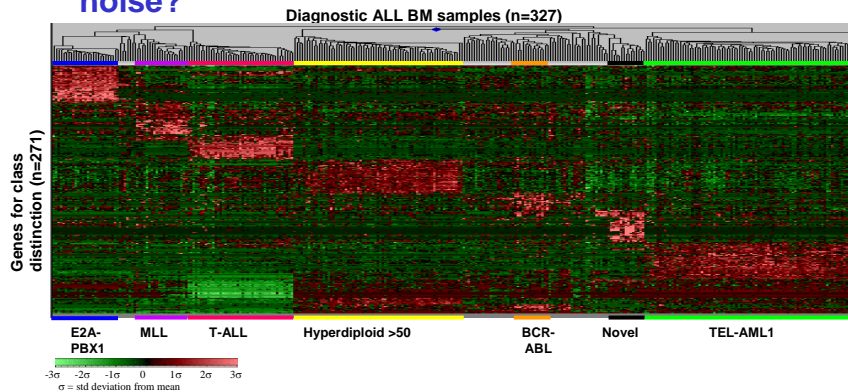


62

Beyond Classification of Gene Expression Profiles



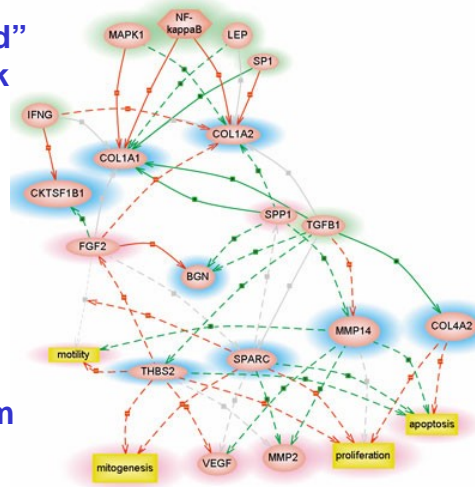
- After identifying the candidate genes by feature selection, do we know which ones are causal genes, which ones are surrogates, and which are noise?



Copyright 2009 © Limsoon Wong

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

Copyright 2009 © Limsoon Wong

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**

Copyright 2009 © Limsoon Wong

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

Concluding Remarks

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

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
- D. Soh, D. Dong, Y. Guo, L. Wong. “Enabling More Sophisticated Gene Expression Analysis for Understanding Diseases and Optimizing Treatments”. *ACM SIGKDD Explorations*, 9(1):3--14, 2007