



약 력

1. 인적사항



성 명	엄 원 석
소속기관	(주) 이즈텍
직 위	책임연구원
전자메일	wsum@istech21.com

2. 학력/경력

연 도	학교 / 기관	전공 / 직위	학위 / 비고
1994	포항공과대학교	물리학과	학사
1996	포항공과대학교	물리학과(비선형통계)	석사
1995 - 1998	생물학전문연구정보센터	객원연구원	
2003	포항공과대학교	물리학과(생물물리)	박사
2003	포항공과대학교	뇌연구센터	연구원
2004 -	(주) 이즈텍	생물정보학사업부	책임연구원

3. 주요연구실적(개조식, 간단하게)

1. **V. Astakhov**, S. Shabunin, **W. Uhm**, and S. Kim (2002), Multistability formation and synchronization loss in couple Henon maps: Two sides of the single bifurcational mechanism, *Physical Review E*, 63:056212.
2. **W. Uhm**, V. Astakhov, A. Akopov, and S. Kim (2003), Multistability formations and loss of chaos synchronization in coupled period-doubling systems, *International Journal of Modern Physics B*, 17:4013-4022.
3. **W. Uhm** and Seunghwan Kim (2004), Phase synchronization and crisis in coupled periodically driven pendulums, *Physics Letters A*, 327:167-173.

DNA칩을 이용한 위암의 진단 및 예후 측정

(주) 이즈텍 책임연구원 엄원석

바이오칩의 대표 주자인 DNA 칩은 점차 분자생물학의 주요 도구로 인식되고 있다. 쓰임새 또한 다양해져 기초 생물학, 기능 유전체학 연구뿐만 아니라 임상 현장에서의 적용을 위한 연구가 활발히 진행되고 있다. 임상분야에서 최근 주목 받고 있는 분야가 DNA 칩을 이용한 질병 진단 및 예후 측정이다. 개별 환자 세포의 분자유전학적 상태는 DNA 칩의 유전체 프로파일링 (genome-wide profiling)으로 상세히 파악될 수 있으므로, DNA 칩은 질병의 세부아형 진단, 약물에 대한 개인 민감도 측정, 정확한 예후 측정을 통한 환자의 세심한 관리 등 미래 의료의 핵심이라 할 수 있는 개인별 맞춤 치료(personalized medicare)를 가능하게 하는데 지대한 역할을 할 것으로 기대되고 있다. 특히 수많은 질병 중에서 현대인의 난치병으로 손꼽히는 암은 DNA 칩 분석의 주요 적용 대상이다. 암에 연관된 복잡한 메커니즘을 기존의 단일 표지자로 진단하는 데는 한계가 있기 때문에, DNA 칩을 이용해 질병의 특정 phenotype과 관련 있는 암의 특이 패턴을 전사체 수준에서 분석하여 새로운 형태의 분자유전학적 표지자(transcriptional molecular signature)를 발굴하는 것이다. 본 발표에서는 이러한 연구에 쓰이는 DNA 칩 분석 방법들과 실제 위암 데이터에 적용한 사례에 대해 논의하고자 한다. 연세의대 암전이 연구센터의 17K cDNA 칩을 이용하였으며, 진단 및 예후 측정을 위한 여러 분석 방법을 수행하였다.

DNA칩을 이용한 위암의 진단 및 예후 측정

2006. 2. 15
엄원석*, 이성근, 김양석
Bioinformatics Unit
ISTECH Inc.

Objectives

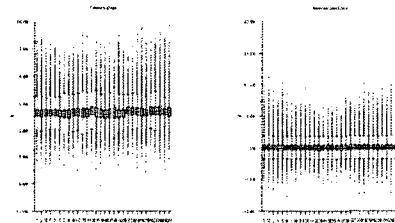
- Discovering useful genetic markers on the following
 - Stage discrimination: among four stages
 - New stage discrimination
 - Recurrence: recur vs. non-recur
 - Survival time in discretized form

Materials & Methods

- DNA chip data
 - Yonsei CMRC: 100 arrays (all tumor samples)
- Data Filtering
 - Background subtraction: removing the negative-valued spots
 - Signal flags filtering
- Normalization
 - Blockwise Lowess normalization: data fraction 0.5
 - Multi-slide scaling
- Analysis Software
 - GenPlex™ v1.2 was used to analyze the data in all processes
 - GenPlex homepage: www.genplex.co.kr

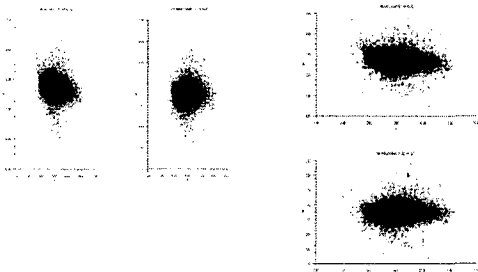
Box Plot

- Before / After Lowess Normalization



MA Plot

- Before / After Lowess Normalization

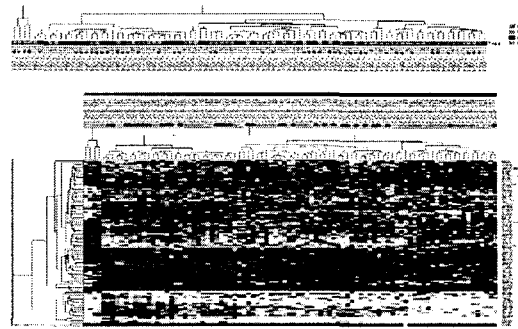


Analysis on Clinical Stage

Preprocessing

- Divided into 4 classes
 - I: 8 samples
 - II: 22 samples
 - III: 29 samples
 - IV: 33 samples
- Removing the spots: > 20% missing at least in a class
- Imputation: filled with group average
- Total 13,828 genes remained

Differentially Expressed Genes



Method: One-way ANOVA, Asymptotic distribution (Pvalue < 0.01), 288 genes selected
Hierarchical Clustering, complete linkage, Euclidean distance

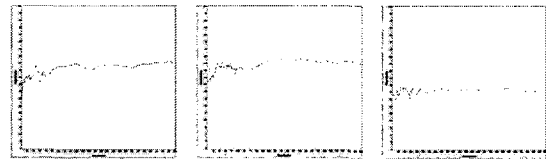
DEG from 4 classes: KEGG

Gene ID	Description
100000001	Gene description
100000002	Gene description
100000003	Gene description
100000004	Gene description
100000005	Gene description
100000006	Gene description
100000007	Gene description
100000008	Gene description
100000009	Gene description
100000010	Gene description
100000011	Gene description
100000012	Gene description
100000013	Gene description
100000014	Gene description
100000015	Gene description
100000016	Gene description
100000017	Gene description
100000018	Gene description
100000019	Gene description
100000020	Gene description
100000021	Gene description
100000022	Gene description
100000023	Gene description
100000024	Gene description
100000025	Gene description
100000026	Gene description
100000027	Gene description
100000028	Gene description
100000029	Gene description
100000030	Gene description
100000031	Gene description
100000032	Gene description
100000033	Gene description
100000034	Gene description
100000035	Gene description
100000036	Gene description
100000037	Gene description
100000038	Gene description
100000039	Gene description
100000040	Gene description
100000041	Gene description
100000042	Gene description
100000043	Gene description
100000044	Gene description
100000045	Gene description
100000046	Gene description
100000047	Gene description
100000048	Gene description
100000049	Gene description
100000050	Gene description
100000051	Gene description
100000052	Gene description
100000053	Gene description
100000054	Gene description
100000055	Gene description
100000056	Gene description
100000057	Gene description
100000058	Gene description
100000059	Gene description
100000060	Gene description
100000061	Gene description
100000062	Gene description
100000063	Gene description
100000064	Gene description
100000065	Gene description
100000066	Gene description
100000067	Gene description
100000068	Gene description
100000069	Gene description
100000070	Gene description
100000071	Gene description
100000072	Gene description
100000073	Gene description
100000074	Gene description
100000075	Gene description
100000076	Gene description
100000077	Gene description
100000078	Gene description
100000079	Gene description
100000080	Gene description
100000081	Gene description
100000082	Gene description
100000083	Gene description
100000084	Gene description
100000085	Gene description
100000086	Gene description
100000087	Gene description
100000088	Gene description
100000089	Gene description
100000090	Gene description
100000091	Gene description
100000092	Gene description
100000093	Gene description
100000094	Gene description
100000095	Gene description
100000096	Gene description
100000097	Gene description
100000098	Gene description
100000099	Gene description
100000100	Gene description

Method: One-way ANOVA, Asymptotic distribution (Pvalue < 0.01), 288 genes selected

Classification: whole computation

- Predicting Clinical Stage from the transcriptional profile
- Input data: 13,828 genes
- Error Estimation: incomplete LOOCV
- Maximum accuracy: around 63%



BSS/WSS: 59.8% with 180 genes
PM with Euclidean distance

KW H-test 63% with 220 genes
PM with Euclidean distance

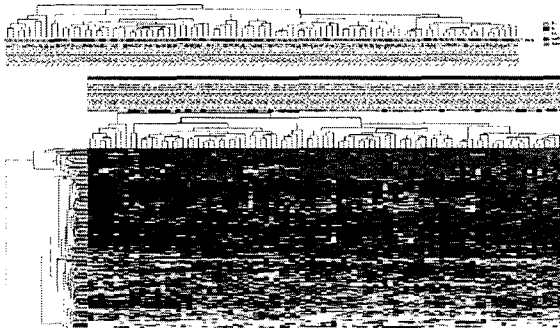
Re T-test 43% with 26 genes
PM with Euclidean distance

Analysis on New Stage

Preprocessing

- Divided into 5 classes
 - IA and IB: 7 samples
 - II: 24 samples
 - IIIA: 22 samples
 - IIIB: 13 samples
 - IV: 26 samples
- Removing the spots: > 20% missing at least in a class
- Imputation: filled with group average
- Total 13,817 genes remained

DEGs from 5 classes



Method: One-way ANOVA, Asymptotic distribution (Pvalue < 0.01), 255 genes selected
 Hierarchical Clustering, complete linkage, Euclidean distance

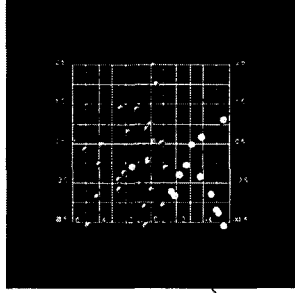
DEGs from 5 classes: KEGG

044170	Apoptosis
044171	Apoptosis - cell cycle
044172	Apoptosis - cell cycle
044173	Apoptosis - cell cycle
044174	Apoptosis - cell cycle
044175	Apoptosis - cell cycle
044176	Apoptosis - cell cycle
044177	Apoptosis - cell cycle
044178	Apoptosis - cell cycle
044179	Apoptosis - cell cycle
044180	Apoptosis - cell cycle
044181	Apoptosis - cell cycle
044182	Apoptosis - cell cycle
044183	Apoptosis - cell cycle
044184	Apoptosis - cell cycle
044185	Apoptosis - cell cycle
044186	Apoptosis - cell cycle
044187	Apoptosis - cell cycle
044188	Apoptosis - cell cycle
044189	Apoptosis - cell cycle
044190	Apoptosis - cell cycle
044191	Apoptosis - cell cycle
044192	Apoptosis - cell cycle
044193	Apoptosis - cell cycle
044194	Apoptosis - cell cycle
044195	Apoptosis - cell cycle
044196	Apoptosis - cell cycle
044197	Apoptosis - cell cycle
044198	Apoptosis - cell cycle
044199	Apoptosis - cell cycle
044200	Apoptosis - cell cycle
044201	Apoptosis - cell cycle
044202	Apoptosis - cell cycle
044203	Apoptosis - cell cycle
044204	Apoptosis - cell cycle
044205	Apoptosis - cell cycle
044206	Apoptosis - cell cycle
044207	Apoptosis - cell cycle
044208	Apoptosis - cell cycle
044209	Apoptosis - cell cycle
044210	Apoptosis - cell cycle
044211	Apoptosis - cell cycle
044212	Apoptosis - cell cycle
044213	Apoptosis - cell cycle
044214	Apoptosis - cell cycle
044215	Apoptosis - cell cycle
044216	Apoptosis - cell cycle
044217	Apoptosis - cell cycle
044218	Apoptosis - cell cycle
044219	Apoptosis - cell cycle
044220	Apoptosis - cell cycle
044221	Apoptosis - cell cycle
044222	Apoptosis - cell cycle
044223	Apoptosis - cell cycle
044224	Apoptosis - cell cycle
044225	Apoptosis - cell cycle
044226	Apoptosis - cell cycle
044227	Apoptosis - cell cycle
044228	Apoptosis - cell cycle
044229	Apoptosis - cell cycle
044230	Apoptosis - cell cycle
044231	Apoptosis - cell cycle
044232	Apoptosis - cell cycle
044233	Apoptosis - cell cycle
044234	Apoptosis - cell cycle
044235	Apoptosis - cell cycle
044236	Apoptosis - cell cycle
044237	Apoptosis - cell cycle
044238	Apoptosis - cell cycle
044239	Apoptosis - cell cycle
044240	Apoptosis - cell cycle
044241	Apoptosis - cell cycle
044242	Apoptosis - cell cycle
044243	Apoptosis - cell cycle
044244	Apoptosis - cell cycle
044245	Apoptosis - cell cycle
044246	Apoptosis - cell cycle
044247	Apoptosis - cell cycle
044248	Apoptosis - cell cycle
044249	Apoptosis - cell cycle
044250	Apoptosis - cell cycle
044251	Apoptosis - cell cycle
044252	Apoptosis - cell cycle
044253	Apoptosis - cell cycle
044254	Apoptosis - cell cycle
044255	Apoptosis - cell cycle
044256	Apoptosis - cell cycle
044257	Apoptosis - cell cycle
044258	Apoptosis - cell cycle
044259	Apoptosis - cell cycle
044260	Apoptosis - cell cycle
044261	Apoptosis - cell cycle
044262	Apoptosis - cell cycle
044263	Apoptosis - cell cycle
044264	Apoptosis - cell cycle
044265	Apoptosis - cell cycle
044266	Apoptosis - cell cycle
044267	Apoptosis - cell cycle
044268	Apoptosis - cell cycle
044269	Apoptosis - cell cycle
044270	Apoptosis - cell cycle
044271	Apoptosis - cell cycle
044272	Apoptosis - cell cycle
044273	Apoptosis - cell cycle
044274	Apoptosis - cell cycle
044275	Apoptosis - cell cycle
044276	Apoptosis - cell cycle
044277	Apoptosis - cell cycle
044278	Apoptosis - cell cycle
044279	Apoptosis - cell cycle
044280	Apoptosis - cell cycle
044281	Apoptosis - cell cycle
044282	Apoptosis - cell cycle
044283	Apoptosis - cell cycle
044284	Apoptosis - cell cycle
044285	Apoptosis - cell cycle
044286	Apoptosis - cell cycle
044287	Apoptosis - cell cycle
044288	Apoptosis - cell cycle
044289	Apoptosis - cell cycle
044290	Apoptosis - cell cycle
044291	Apoptosis - cell cycle
044292	Apoptosis - cell cycle
044293	Apoptosis - cell cycle
044294	Apoptosis - cell cycle
044295	Apoptosis - cell cycle
044296	Apoptosis - cell cycle
044297	Apoptosis - cell cycle
044298	Apoptosis - cell cycle
044299	Apoptosis - cell cycle
044300	Apoptosis - cell cycle

Method: One-way ANOVA, Asymptotic distribution (Pvalue < 0.01), 255 genes selected

IIIA vs. IIIB (1/5)

- Histopathological classification vs. Genetic profile



Method: Welch t-test
 Asymptotic distribution
 Significance: Pvalue < 0.01
 131 genes selected as DEG
 Sample PCA shown
 Blue ball: IIIA (22 samples)
 Yellow ball: IIIB (13 samples)

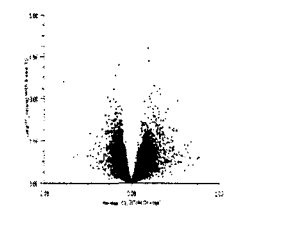
DEGs from IIIA vs. IIIB: KEGG

0441347	Apoptosis
0441348	Apoptosis
0441349	Apoptosis
0441350	Apoptosis
0441351	Apoptosis
0441352	Apoptosis
0441353	Apoptosis
0441354	Apoptosis
0441355	Apoptosis
0441356	Apoptosis
0441357	Apoptosis
0441358	Apoptosis
0441359	Apoptosis
0441360	Apoptosis
0441361	Apoptosis
0441362	Apoptosis
0441363	Apoptosis
0441364	Apoptosis
0441365	Apoptosis
0441366	Apoptosis
0441367	Apoptosis
0441368	Apoptosis
0441369	Apoptosis
0441370	Apoptosis
0441371	Apoptosis
0441372	Apoptosis
0441373	Apoptosis
0441374	Apoptosis
0441375	Apoptosis
0441376	Apoptosis
0441377	Apoptosis
0441378	Apoptosis
0441379	Apoptosis
0441380	Apoptosis
0441381	Apoptosis
0441382	Apoptosis
0441383	Apoptosis
0441384	Apoptosis
0441385	Apoptosis
0441386	Apoptosis
0441387	Apoptosis
0441388	Apoptosis
0441389	Apoptosis
0441390	Apoptosis
0441391	Apoptosis
0441392	Apoptosis
0441393	Apoptosis
0441394	Apoptosis
0441395	Apoptosis
0441396	Apoptosis
0441397	Apoptosis
0441398	Apoptosis
0441399	Apoptosis
0441400	Apoptosis
0441401	Apoptosis
0441402	Apoptosis
0441403	Apoptosis
0441404	Apoptosis
0441405	Apoptosis
0441406	Apoptosis
0441407	Apoptosis
0441408	Apoptosis
0441409	Apoptosis
0441410	Apoptosis
0441411	Apoptosis
0441412	Apoptosis
0441413	Apoptosis
0441414	Apoptosis
0441415	Apoptosis
0441416	Apoptosis
0441417	Apoptosis
0441418	Apoptosis
0441419	Apoptosis
0441420	Apoptosis
0441421	Apoptosis
0441422	Apoptosis
0441423	Apoptosis
0441424	Apoptosis
0441425	Apoptosis
0441426	Apoptosis
0441427	Apoptosis
0441428	Apoptosis
0441429	Apoptosis
0441430	Apoptosis
0441431	Apoptosis
0441432	Apoptosis
0441433	Apoptosis
0441434	Apoptosis
0441435	Apoptosis
0441436	Apoptosis
0441437	Apoptosis
0441438	Apoptosis
0441439	Apoptosis
0441440	Apoptosis
0441441	Apoptosis
0441442	Apoptosis
0441443	Apoptosis
0441444	Apoptosis
0441445	Apoptosis
0441446	Apoptosis
0441447	Apoptosis
0441448	Apoptosis
0441449	Apoptosis
0441450	Apoptosis
0441451	Apoptosis
0441452	Apoptosis
0441453	Apoptosis
0441454	Apoptosis
0441455	Apoptosis
0441456	Apoptosis
0441457	Apoptosis
0441458	Apoptosis
0441459	Apoptosis
0441460	Apoptosis
0441461	Apoptosis
0441462	Apoptosis
0441463	Apoptosis
0441464	Apoptosis
0441465	Apoptosis
0441466	Apoptosis
0441467	Apoptosis
0441468	Apoptosis
0441469	Apoptosis
0441470	Apoptosis
0441471	Apoptosis
0441472	Apoptosis
0441473	Apoptosis
0441474	Apoptosis
0441475	Apoptosis
0441476	Apoptosis
0441477	Apoptosis
0441478	Apoptosis
0441479	Apoptosis
0441480	Apoptosis
0441481	Apoptosis
0441482	Apoptosis
0441483	Apoptosis
0441484	Apoptosis
0441485	Apoptosis
0441486	Apoptosis
0441487	Apoptosis
0441488	Apoptosis
0441489	Apoptosis
0441490	Apoptosis
0441491	Apoptosis
0441492	Apoptosis
0441493	Apoptosis
0441494	Apoptosis
0441495	Apoptosis
0441496	Apoptosis
0441497	Apoptosis
0441498	Apoptosis
0441499	Apoptosis
0441500	Apoptosis

Method: Welch's T-test Asymptotic distribution (Pvalue < 0.01), 131 genes selected

IIIA vs. IIIB: Volcano Plot

- IIIA (reference) vs. IIIB

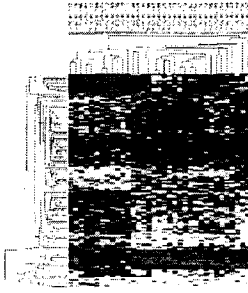


Method: Volcano Plot,
 Welch t-test (Pvalue < 0.01),
 Fold change cutoff: 2-fold

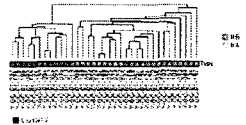
Gene No.	ID	Average log2	Average Pval	P-value	Gene_Symbol	Gene Title	Chromosome	Chromosomal Pos.	GO Biological Process	GO Cellular Component	GO Molecular Function
1	029203	1.66	2.66	0.01	CDK2	Cyclin dependent kinase 2	5p13	125,000,000	cell cycle process	cytoplasm	protein kinase activity
2	029204	1.66	2.66	0.01	CDK2	Cyclin dependent kinase 2	5p13	125,000,000	cell cycle process	cytoplasm	protein kinase activity
3	029205	1.66	2.66	0.01	CDK2	Cyclin dependent kinase 2	5p13	125,000,000	cell cycle process	cytoplasm	protein kinase activity

IIIA vs. IIIB (2/5)

- Histopathological classification vs. Genetic profile

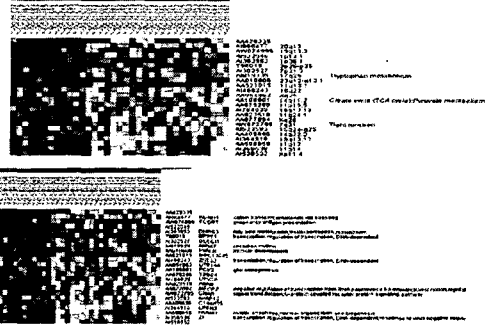


Method: Hierarchical Clustering
 Average Linkage
 Distance measure: Euclidean
 131 genes selected as DEG



III A vs. III B (3/5)

- Histopathological classification vs. Genetic profile



III A vs. III B (4/5)

- Histopathological classification vs. Genetic profile

[Classification]

Gene Selection: Welch's T-test (131 genes)
 Classifier: Nearest Centroid (Euclidean distance)
 Error Estimation: LOOCV
 Significance: Permutation test (Pvalue < 0.001)
 Misclassification: Y-GC-01-122

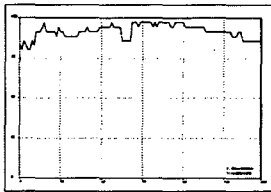
Predicted	Real	Class 1	Class 2	Total
Class 1		12	0	12
Class 2		1	22	23
Indeterminate		0	0	0
Sum		13	22	35
Accuracy		92.31%	100%	92.31%

III A vs. III B (5/5)

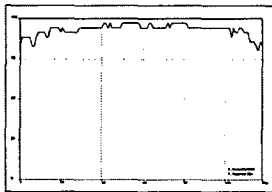
- Histopathological classification vs. Genetic profile

[Classification: whole computation]

Input data: 13,817 genes
 Gene Selection: BSS/WSS
 Error Estimation: incomplete LOOCV



weighted 5-NN with Euclidean distance



Prototype Matching with Euclidean distance

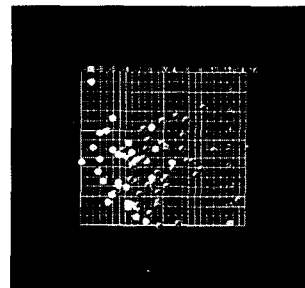
Analysis on Recurrence

Preprocessing

- Divided into 2 classes
 - Recurrence: 34 samples (NED < 60)
 - Non-recurrence: 45 samples (NED ≥ 60)
- Removing the spots: > 20% missing at least in a class
- Imputation: filled with group average
- Total 13,886 genes remained

Differentially Expressed Genes

- Recur vs. Non-recur



Method: Welch t-test
 Asymptotic distribution
 Significance: Pvalue < 0.01
 220 genes selected as DEG
 Sample PCA shown

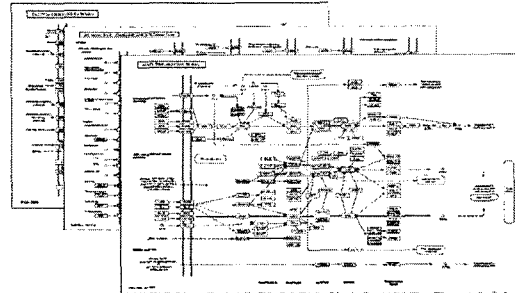
Blue ball: non-recur (45 samples)
 Yellow ball: recur (34 samples)

DEGs: KEGG Pathway

hs01110	Aspartate degradation
hs01120	Aspartic acid metabolism
hs01130	Aspartic acid transport
hs01140	Aspartic acid synthesis
hs01150	Aspartic acid transport
hs01160	Aspartic acid synthesis
hs01170	Aspartic acid synthesis
hs01180	Aspartic acid synthesis
hs01190	Aspartic acid synthesis
hs01200	Aspartic acid synthesis
hs01210	Aspartic acid synthesis
hs01220	Aspartic acid synthesis
hs01230	Aspartic acid synthesis
hs01240	Aspartic acid synthesis
hs01250	Aspartic acid synthesis
hs01260	Aspartic acid synthesis
hs01270	Aspartic acid synthesis
hs01280	Aspartic acid synthesis
hs01290	Aspartic acid synthesis
hs01300	Aspartic acid synthesis
hs01310	Aspartic acid synthesis
hs01320	Aspartic acid synthesis
hs01330	Aspartic acid synthesis
hs01340	Aspartic acid synthesis
hs01350	Aspartic acid synthesis
hs01360	Aspartic acid synthesis
hs01370	Aspartic acid synthesis
hs01380	Aspartic acid synthesis
hs01390	Aspartic acid synthesis
hs01400	Aspartic acid synthesis
hs01410	Aspartic acid synthesis
hs01420	Aspartic acid synthesis
hs01430	Aspartic acid synthesis
hs01440	Aspartic acid synthesis
hs01450	Aspartic acid synthesis
hs01460	Aspartic acid synthesis
hs01470	Aspartic acid synthesis
hs01480	Aspartic acid synthesis
hs01490	Aspartic acid synthesis
hs01500	Aspartic acid synthesis

Method: Welch's T-test, Asymptotic distribution (Pvalue < 0.01), 57220 genes known

Most Common Pathways



Calcium Signaling Pathway: 6 genes involved

MAPK Signaling Pathway: 6 genes involved

Neurexine Ligand Receptor Interaction: 6 genes involved

Classification: whole computation

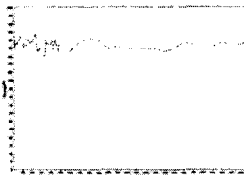
- Predicting recurrence from the transcriptional profile

Input data: 13,886 genes

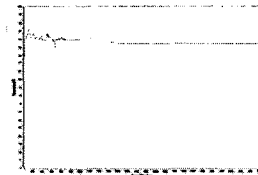
Gene Selection: two-sample t-test

Error Estimation: incomplete LOOCV

Maximum accuracy: around 85%



83.5% with 47 genes
weighted 5-NN with Euclidean distance



86% with 14 genes
Prototype Matching with Euclidean distance

Analysis on Survival

Preprocessing (1)

- Divided into 2 classes
 - Class 1: 41 samples (Survival Time \leq 60 months)
 - Class 2: 49 samples (ST > 61months)
- No Imputation: 10,430 genes remained

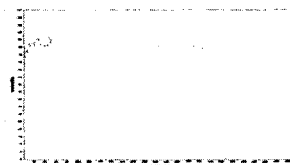
Classification: whole computation

- Predicting survival from the transcriptional profile

Input data: 10,430 genes

Error Estimation: incomplete LOOCV

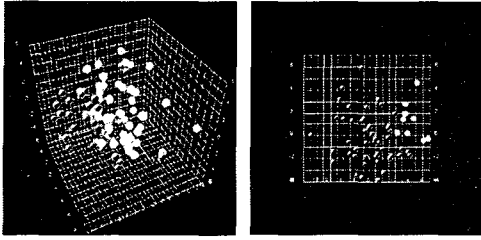
Maximum accuracy: around 82%



Two-sample t-test: 82.2% with 46 genes
PM with Euclidean distance

Differentially Expressed Genes

- Survival Time: discretized case



Method: Welch's t-test, Asymptotic distribution (P -value < 0.01), 201 genes selected
Sample PCA & clustering shown

Preprocessing (2)

- Divided into 3 classes
 - Class 1: 23 samples (Survival Time ≤ 24 months)
 - Class 2: 18 samples (25 months $<$ Survival Time ≤ 60 months)
 - Class 3: 49 samples (ST > 61 months)
- No Imputation: 10,430 genes remained

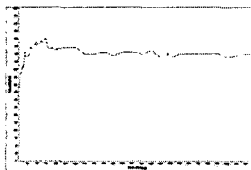
Classification: whole computation

- Predicting survival from the transcriptional profile

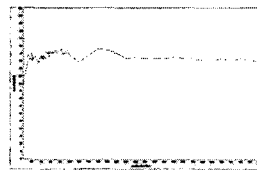
Input data: 10,430 genes

Error Estimation: incomplete LOOCV

Maximum accuracy: around 80%



KW H-test 80% with 55 genes
PM with Euclidean distance



BSS-WSS 73.3% with 69 genes
PM with Euclidean distance