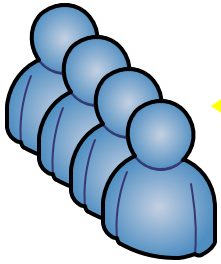
Simple Decision Rules for Classifying Human Cancers from Gene Expression Profiles

Aik Choon TAN
Post-Doc Research Fellow
actan@jhu.edu

Prof. Raimond L. Winslow rwinslow@jhu.edu, Director, ICM & CCBM,
Prof. Donald Geman geman@jhu.edu,
Prof. Daniel Naiman daniel.naiman@jhu.edu,
Lei Xu leixu@jhu.edu,
Troy Anderson troy_anderson@jhu.edu

The Institute for Computational Medicine (ICM) and
Center for Cardiovascular Bioinformatics and Modeling (CCBM),
Johns Hopkins University

Biomarkers Discovery Workflow

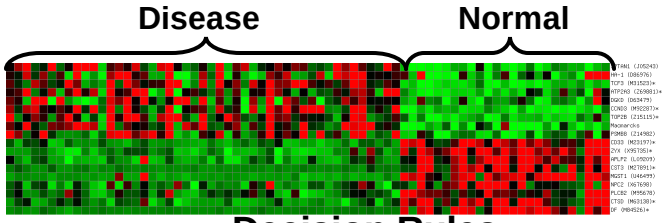


Patients

Clinical Applications

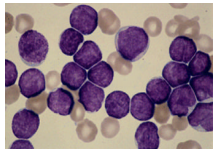
Follow-up Study

Candidate Biomarkers

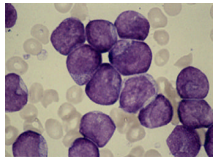


Decision Rules

Sample Collection



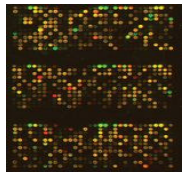
Disease



Normal

Transcriptomics Pipeline

Experiments



Gene Expression Profiling

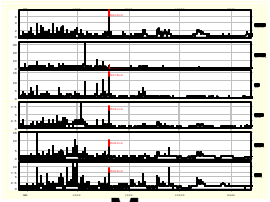
Store

MAGE-DB2

Machine Learning

Query

***Relative Expression Reversal* Classifiers**

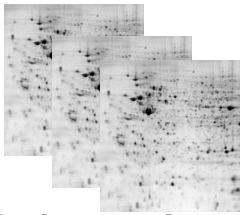


Mass Spectrometry

Store

PROTEIN-DB2

Query



Difference Gel Electrophoresis

Proteomics Pipeline

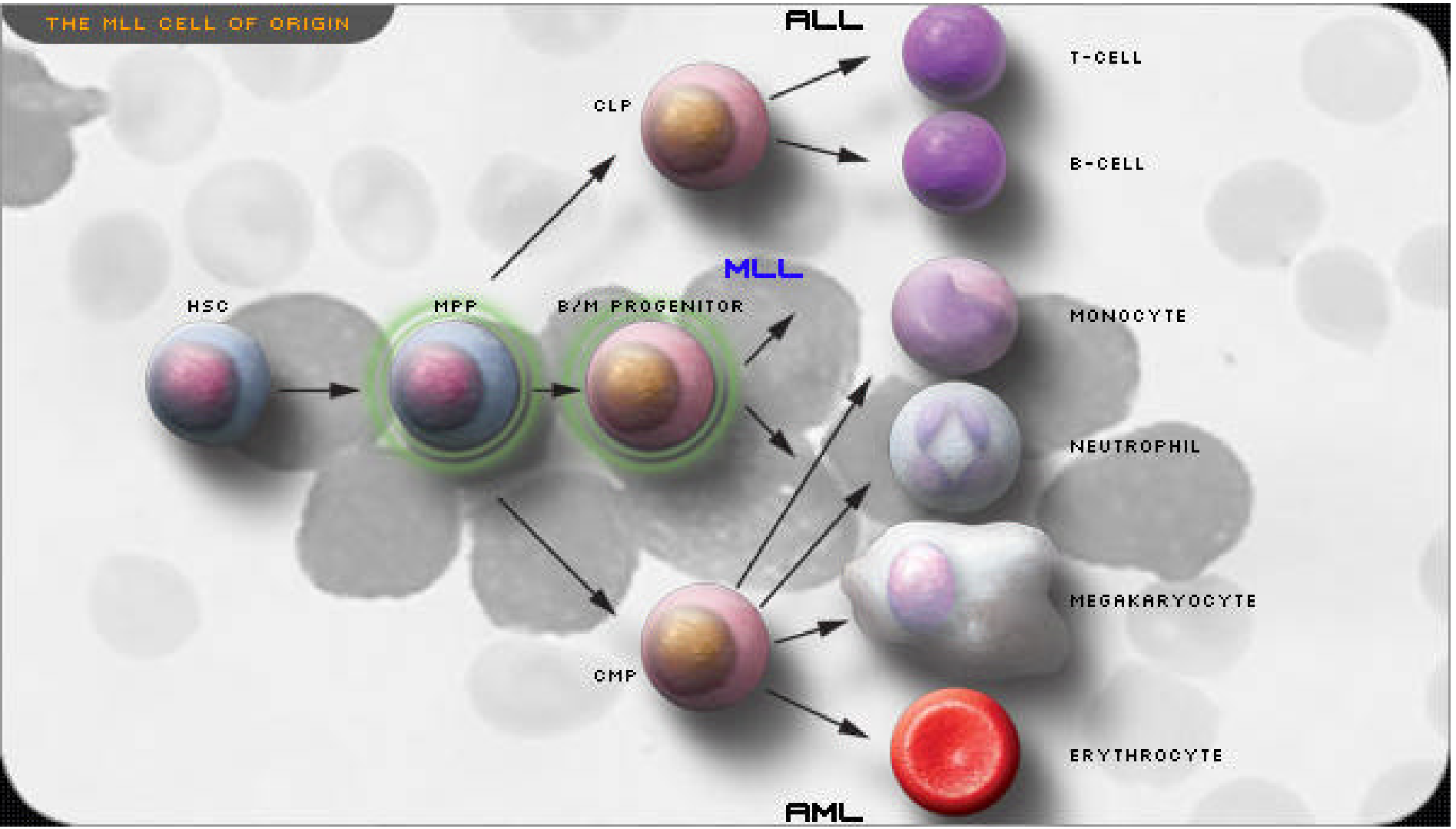
Store

Available at ICM/CCBM

Outline

1) Relative Expression Reversal Classifiers
   - *TSP* classifier
   - *k-TSP* classifier
2) Results on binary & multi-class disease gene expression classification problems
3) Data Integration and Cross-platform analysis
4) Applications to other “–omics” data
5) Conclusions

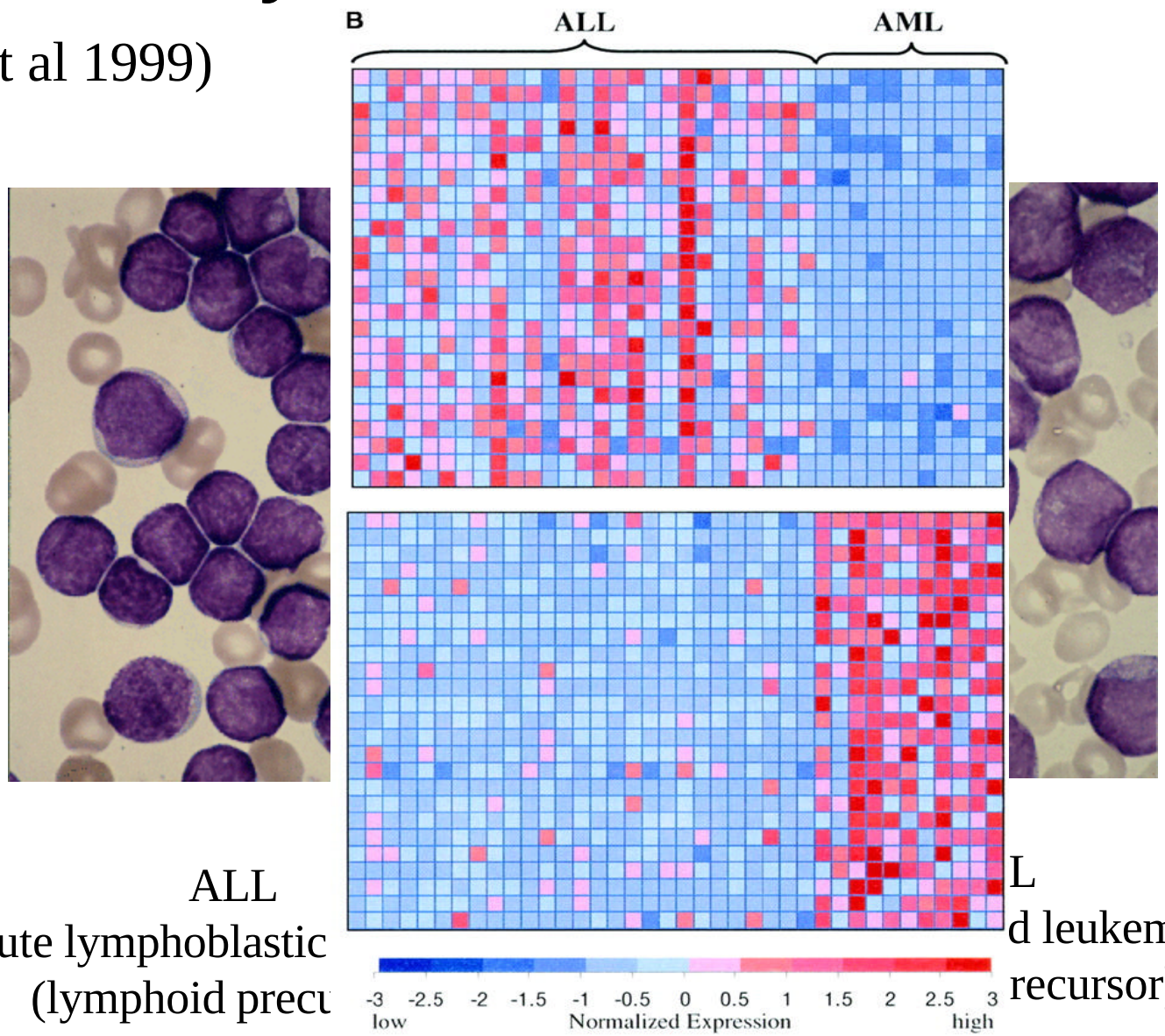
Disease Classification



From http://research.dfci.harvard.edu/korsmeyer/Home.html

Microarray Gene Expression Profiles

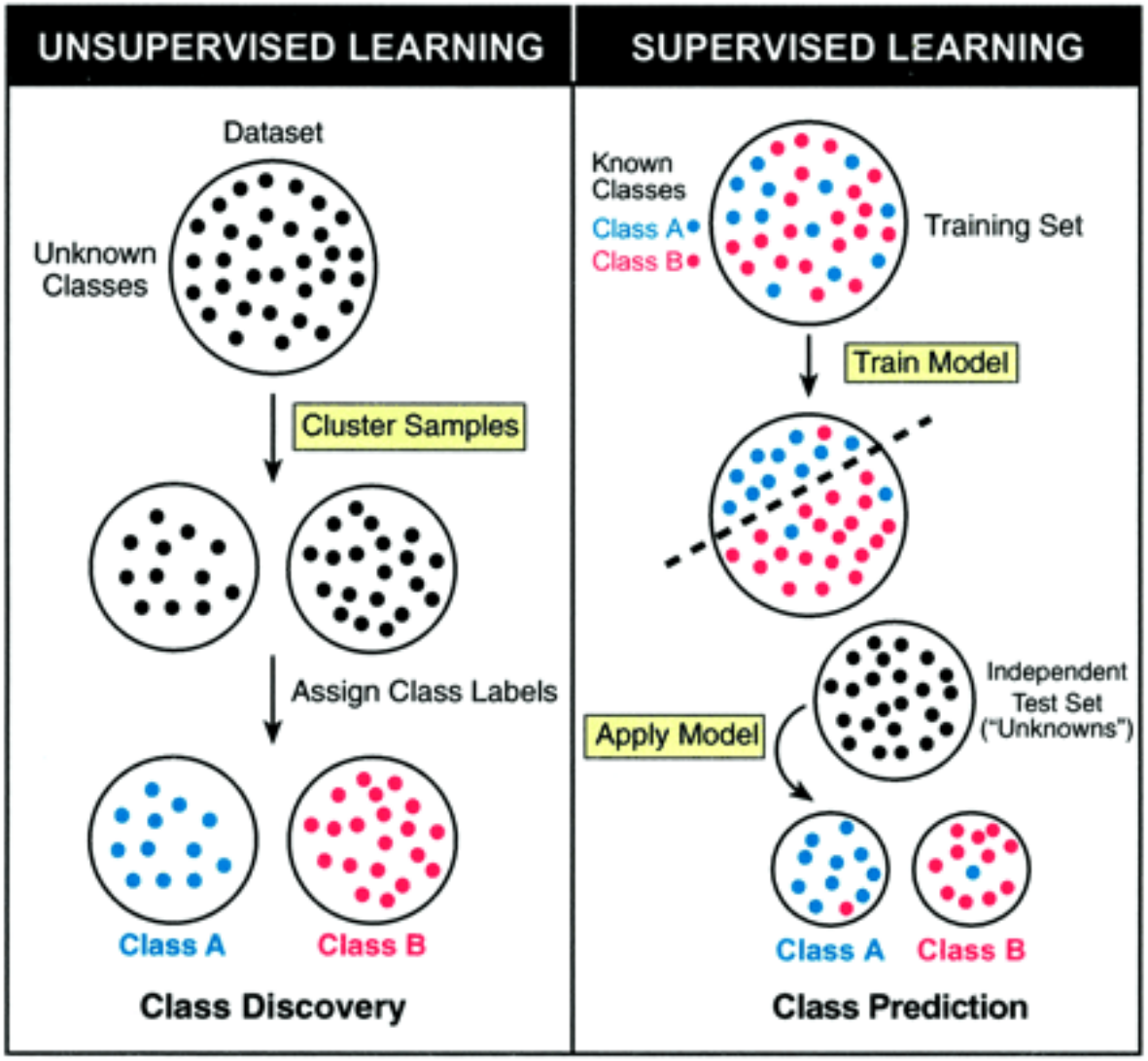
(Golub et al 1999)



ALL
acute lymphoblastic
(lymphoid precu

L
d leukemia
recursor)

Learning Approaches



(Ramaswamy and Golub 2002)

Gene Expression Profiles

$P \times N$ matrix

N arrays (N = {1,2,...N})

Y = {Cancer, Normal}

Label	Cancer	Normal	…	Cancer
Geneid	Array 1	Array 2	…	Array N
$\mathbf{g}_1$	103.02	58.79	…	101.54
$\mathbf{g}_2$	40.55	1246.87	…	1432.12
…	…	…	…	…
$\mathbf{g}_P$	78.13	66.25	…	823.09

P genes

P = {1,...,P}

Microarray Data Analysis

- A $P \times N$ matrix where
  - P is the number of genes
  - N is the number of experiments
  - The columns are “gene expression profiles”

Sample Size Dilemma

- Small *N* (typically tens to hundreds)
- Large *P* (typically thousands)
- **Consequence:** Standard methods in machine learning often lead to over-fitting and inflated estimates of performance.

Interpretability Dilemma (Biological Perspective)

- The “decision boundary” generated by standard machine learning methods is often highly complex.
- **Examples:** support vector machines, neural networks, random forests, nearest neighbors.
- **Consequence:** Decision-making is a mystery and does not readily generate hypotheses or suggest follow-up studies.

Relative Expression Reversal Classifiers

- *Pairwise rank*-based *comparisons* (relative expression values *within each array*)
- Generates *accurate* and *simple* decision rules
  - ***TSP*** classifier: Top Scoring Pair
  - ***k-TSP*** classifier: *k*-disjoint Top Scoring Pairs
- *Data driven*, *parameter-free* learning algorithm
- *Performance comparable to or exceeds* that of other machine learning methods
- *Easy to interpret*, *facilitating follow-up study (small number of genes)*

(Tan *et al.*, 2005, Bioinformatics, 21:3896-3904)

Rank-based Classification

- **Novelty:** Replace the measured expression values by their *ranks within profiles*, hence obtaining *invariance to normalization*.

- **Example:** Differentiate between classes by finding pairs of genes whose ordering typically changes from Normal to Disease.

- **Simple Interpretation:** *Inversion of mRNA (protein) abundance*.

Statistical Formulation

- The expression profile is a random vector

$$\mathbf{X} = (X_1,\ldots,X_P)$$

- The true class is also a r.v. Y, say $Y=1$ (Disease) or $Y=2$ (Normal)
- A classifier is a mapping $\boldsymbol{f}$ from $\mathbf{X}$ to $\{1,2\}$.
- Training data: A $P \times N$ matrix $\boldsymbol{S}$ whose columns represent $N = N_1+N_2$ samples of $(\mathbf{X}, Y)$, with N_1 (resp. N_2) samples for which $Y=1$ ($Y=2$).

Statistical Formulation (cont)

- **Learning algorithm:** A mapping from the training set $\boldsymbol{S}$ to a classifier $\boldsymbol{f}$ based on $\boldsymbol{S}$.
- **Generalization error:** $e(\boldsymbol{f}) = P(\boldsymbol{f}(\mathbf{X})?Y)$. This depends on $\boldsymbol{S}$ and the distribution of $(\mathbf{X},Y)$ and is extremely hard to estimate.
- **Dilemmas:**
  - $N << P$
  - $\boldsymbol{f}(\mathbf{X})$ is too complex and hard to interpret

Gene Expression Comparisons

- **Features:** $Z_{ij} = 1_{\{X_i < X_j\}}, \quad 1 \le i < j \le P.$
- **Feature Score:**

$$\Delta_{ij} = \left| P(X_i < X_j \mid Y = 1) - P(X_i < X_j \mid Y = 2) \right|$$

$$\approx \left| \frac{N_{ij}^{(1)}}{N_1} - \frac{N_{ij}^{(2)}}{N_2} \right|.$$

where

$$N_{ij}^{(k)} = \left| \{1 \le m \le N : Y_m = k, X_{im} < X_{jm}\} \right|, k = 1,2.$$

TSP Algorithm

1

	n1	n2	n3	n4
	Cancer	Cancer	Normal	Normal
g1	1000	789	356	45
g2	289	150	500	1000
g3	634	450	220	150
g4	367	455	150	50
g5	2500	1800	1900	2100

2

	n1	n2	n3	n4
	Cancer	Cancer	Normal	Normal
g1	2	2	3	5
g2	5	5	2	2
g3	3	4	4	3
g4	4	3	5	4
g5	1	1	1	1

3

P(g1 > g2 | Cancer) = 0/2 = 0 P(g1 > g2 | Normal) = 2/2 = 1
$?_{12}$ = |P(g1>g2|Cancer) – P(g1 > g2|Normal)| = |0 – 1| = 1

P(g1 > g3 | Cancer) = 0/2 = 0 P(g1 > g2 | Normal) = 1/2 = 0.5
$?_{13}$ = |P(g1>g3|Cancer) – P(g1 > g3|Normal)| = |0 – 0.5| = 0.5

P(g1 > g4 | Cancer) = 0/2 = 0 P(g1 > g4 | Normal) = 1/2 = 0.5
$?_{14}$ = |P(g1>g4|Cancer) – P(g1 > g4|Normal)| = |0 – 0.5| = 0.5

P(g1 > g5 | Cancer) = 2/2 = 1 P(g1 > g5 | Normal) = 2/2 = 1
$?_{15}$ = |P(g1>g5|Cancer) – P(g1 > g5|Normal)| = |1 – 1| = 0
…
P(g4 > g5 | Cancer) = 2/2 = 1 P(g4 > g5 | Normal) = 2/2 = 1
$?_{45}$ = |P(g4>g5|Cancer) – P(g4 > g5|Normal)| = |1 – 1| = 0

4

High ?

$?_{12}$ = 1
.
.
.
$?_{13}$ = 0.5
$?_{14}$ = 0.5
.
.
.
$?_{15}$ = 0
$?_{45}$ = 0

Low ?

TSP Classifier

- Select only the top scoring pairs :
  - $\{ (i^*, j^*): ?_{i^*j^*} = ?_{max} \}$

- *TSP* classifier (h_{TSP}) is based on these pairs:
  - *Example*: Let *all* the top scoring pairs "vote" (Geman et al, 2004)
  - *Example*: Select one unique top scoring pair, based on maximizing difference in ranks (i, j) (Tan et al, 2005)

- Prediction: Suppose $P_{ij}(\text{Normal}) > P_{ij}(\text{Disease})$, $\mathbf{X}_{\text{new}}$ = new profile:

$$y_{new} = h_{TSP}(\mathbf{X}_{new}) = \begin{cases} \text{Normal, if } R_{i,new} > R_{j,new} \\ \text{Disease, otherwise.} \end{cases} \quad (1)$$

  - If, on the other hand, if $P_{ij}(\text{Disease}) > P_{ij}(\text{Normal})$, then the decision rule is reversed.

(Tan *et al*., 2005, Bioinformatics, 21:3896-3904)

Initial Conclusions

- There may be many pairs of genes with an informative ordering
  - Motivation for *k-TSP*
- The *TSP* classifier is sensitive to ***S*** for small samples but invariant to normalization
  - Motivation for “data integration”

k-TSP Classifier

- Uses exactly *k* top disjoint pairs in prediction.
- *k* is determined by internal cross-validation
- Ensemble learning – to combine the discriminating power of many "weaker" rules to make more reliable predictions.
- Prediction:
  - Suppose $\mathbf{X}_{\text{new}}$ = new profile, each gene pair (i_u, j_u), $u = 1,\ldots, k$, votes according (1).
  - The *k-TSP* classifier $h_{k\text{-}TSP}$ employs an unweighted majority voting procedure to obtain the final prediction of y_{new}.

(Tan *et al*., 2005, Bioinformatics, 21:3896-3904)

Microarray Data Sets

(Binary class Problems)

Data set	Platform	# genes	# samples C_1	# samples C_2	Reference
Colon	cDNA	2,000	40 (T)	22 (N)	(Alon et al. 1998)
Leukemia	Affy	7,129	25 (AML)	47 (ALL)	(Golub et al. 1999)
CNS	Affy	7,129	25 (C)	9 (D)	(Pomeroy et al. 2002)
DLBCL	Affy	7,129	58 (D)	19 (F)	(Shipp et al. 2002)
Lung	Affy	12,533	150 (A)	31 (M)	(Gordon et al. 2002)
Prostate1	Affy	12,600	52 (T)	50 (N)	(Singh et al. 2002)
Prostate2	Affy	12,625	38 (T)	50 (N)	(Stuart et al. 2004)
Prostate3	Affy	12,626	24 (T)	9 (N)	(Welsh et al. 2001)
GCM	Affy	16,063	190 (C)	90 (N)	(Ramaswamy et al. 2001)

(Multi-class Problems)

Data set	Platform	# classes	# genes	# samples Training	# samples Testing	Reference
Leukemia1	Affy	3	7,129	38	34	(Golub et al. 1999)
Lung1	Affy	3	7,129	64	32	(Beer et al. 2002)
Leukemia2	Affy	3	12,582	57	15	(Armstrong et al. 2002)
SRBCT	cDNA	4	2,308	63	20	(Khan et al. 2001)
Breast	Affy	5	9,216	54	30	(Perou et al. 2000)
Lung2	Affy	5	12,600	136	67	(Bhattacharjee et al. 2001)
DLBCL	cDNA	6	4,026	58	30	(Alizadeh et al. 2000)
Leukemia3	Affy	7	12,558	215	112	(Yeoh et al. 2002)
Cancers	Affy	11	12,533	100	74	(Su et al. 2001)
GCM	Affy	14	16,063	144	46	(Ramaswamy et al. 2001)

Results

(LOOCV Binary Class Problems)

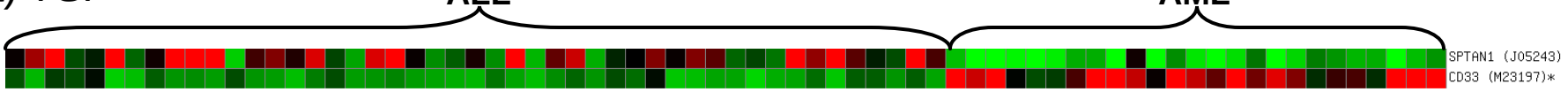
Method	Leukemia	CNS	DLBCL	Colon	Prostate1	Prostate2	Prostate3	Lung	GCM	Average
TSP	93.80	77.90	**98.10**	**91.10**	**95.10**	67.60	97.00	98.30	75.40	88.26
k-TSP	95.83	**97.10**	97.40	90.30	91.18	75.00	97.00	98.90	85.40	**92.01**
DT	73.61	67.65	80.52	80.65	87.25	64.77	84.85	96.13	77.86	79.25
NB	**100.00**	82.35	80.52	58.06	62.75	73.86	90.91	97.79	84.29	81.17
k-NN	84.72	76.47	84.42	74.19	76.47	69.32	87.88	98.34	82.86	81.63
SVM	98.61	82.35	97.40	82.26	91.18	76.14	**100.00**	**99.45**	**93.21**	91.18
PAM	97.22	82.35	85.71	85.48	91.18	**79.55**	**100.00**	**99.45**	79.29	88.91

Number of Informative Genes

Method	Leukemia	CNS	DLBCL	Colon	Prostate1	Prostate2	Prostate3	Lung	GCM
TSP	2	2	2	2	2	2	2	2	2
k-TSP	18	10	2	2	2	18	2	10	10
DT	2	2	3	3	4	4	1	3	14
PAM	2296	4	17	15	47	13	701	9	47

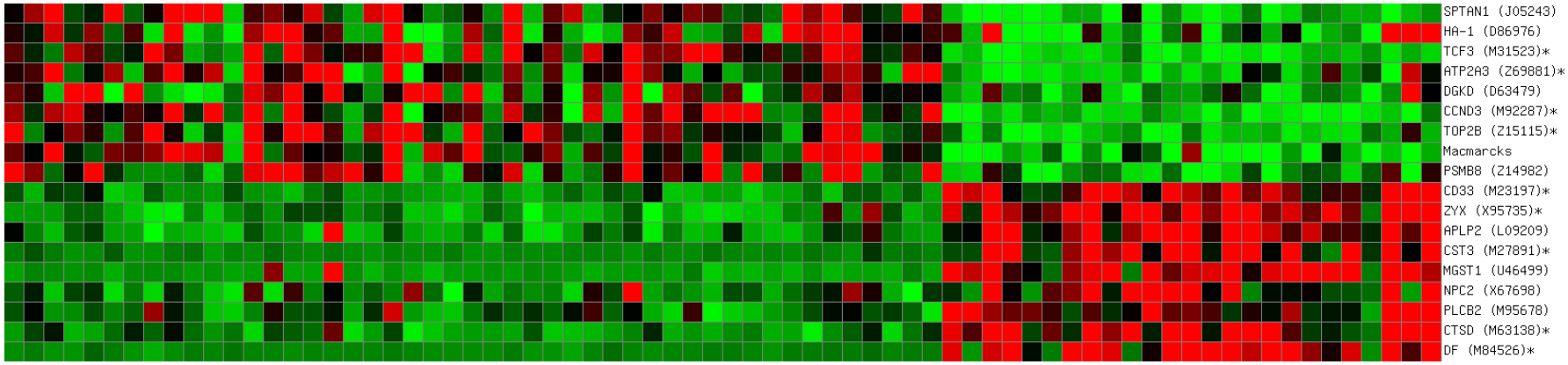
(Tan *et al*., 2005, Bioinformatics, 21:3896-3904)

(a) *TSP*



IF SPTAN1 ≥ CD33* **THEN** ALL; **ELSE** AML Δ = 0.9787

(b) *k-TSP*



IF SPTAN1 ≥ CD33* **THEN** ALL; **ELSE** AML Δ = 0.9787
IF HA-1 ≥ ZYX* **THEN** ALL; **ELSE** AML Δ = 0.9787
IF TCF3* > APLP2 **THEN** ALL; **ELSE** AML Δ = 0.9574
IF ATP2A3* ≥ CST3* **THEN** ALL; **ELSE** AML Δ = 0.9387
IF DGKD > MGST1 **THEN** ALL; **ELSE** AML Δ = 0.9387
IF CCND3* ≥ NPC2 **THEN** ALL; **ELSE** AML Δ = 0.9387
IF TOP2B* > PLCB2 **THEN** ALL; **ELSE** AML Δ = 0.9387
IF Macmarcks ≥ CTSD* **THEN** ALL; **ELSE** AML Δ = 0.9362
IF PSMB8 ≥ DF* **THEN** ALL; **ELSE** AML Δ = 0.9200

* Genes previously identified by Golub *et al* (1999)

(Tan *et al*., 2005, Bioinformatics, 21:3896-3904)

Results

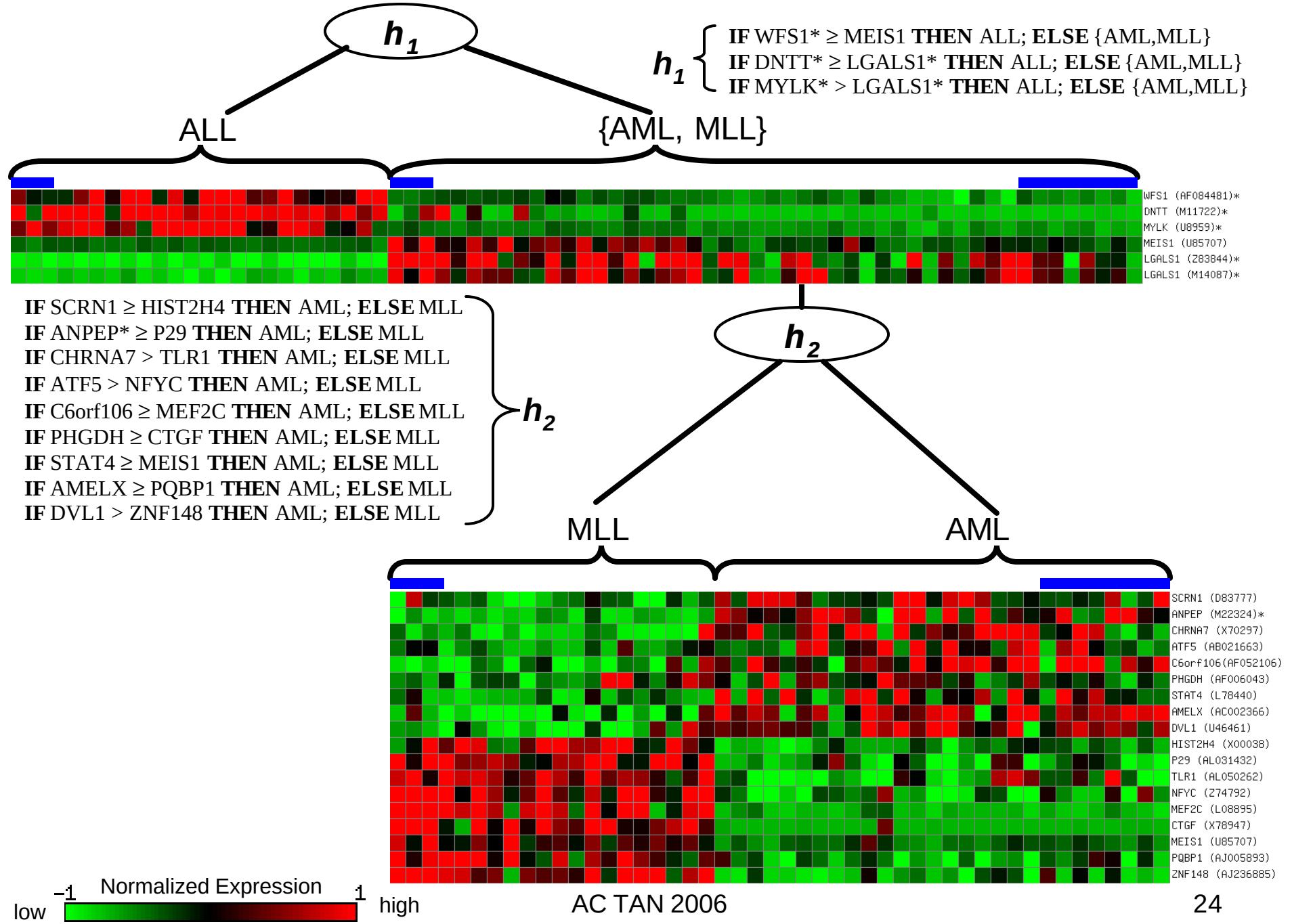
(Test Accuracy for Multi-Class Problems)

Method	Leuk1	Lung1	Leuk2	SRBCT	Breast	Lung2	DLBCL	Leuk3	Cancers	GCM	Average
HC-*TSP*	**97.06**	71.88	80.00	95.00	66.67	83.58	83.33	77.68	74.32	52.17	78.17
HC-*k*-*TSP*	**97.06**	78.13	**100**	**100**	66.67	94.03	83.33	82.14	82.43	**67.39**	85.12
DT	85.29	78.13	80.00	75.00	73.33	88.06	86.67	75.89	68.92	52.17	76.35
NB	85.29	81.25	**100**	60.00	66.67	88.06	86.67	32.14	79.73	52.17	73.20
k-NN	67.65	75.00	86.67	30.00	63.33	88.06	93.33	75.89	64.86	34.78	67.96
1-vs-1-*SVM*	79.41	**87.50**	**100**	**100**	83.33	97.01	**100**	84.82	83.78	65.22	**88.11**
PAM	**97.06**	78.13	93.33	95.00	**93.33**	**100**	90.00	**93.75**	**87.84**	56.52	**88.50**

Number of Informative Genes

Method	Leuk1	Lung1	Leuk2	SRBCT	Breast	Lung2	DLBCL	Leuk3	Cancers	GCM
HC-*TSP*	4	4	4	6	8	8	10	12	20	26
HC-*k*-*TSP*	36	20	24	30	24	28	46	64	128	134
DT	2	4	2	3	4	5	5	16	10	18
PAM	44	13	62	285	4822	614	3949	3338	2008	1253

(Tan *et al*., 2005, Bioinformatics, 21:3896-3904)


h1
ALL
{AML, MLL}
h1
IF WFS1* ≥ MEIS1 THEN ALL; ELSE {AML,MLL}
IF DNTT* ≥ LGALS1* THEN ALL; ELSE {AML,MLL}
IF MYLK* > LGALS1* THEN ALL; ELSE {AML,MLL}
WFS1 (AF084481)*
DNTT (M11722)*
MYLK (U8959)*
MEIS1 (U85707)
LGALS1 (Z83844)*
LGALS1 (M14087)*
IF SCRN1 ≥ HIST2H4 THEN AML; ELSE MLL
IF ANPEP* ≥ P29 THEN AML; ELSE MLL
IF CHRNA7 > TLR1 THEN AML; ELSE MLL
IF ATF5 > NFYC THEN AML; ELSE MLL
IF C6orf106 ≥ MEF2C THEN AML; ELSE MLL
IF PHGDH ≥ CTGF THEN AML; ELSE MLL
IF STAT4 ≥ MEIS1 THEN AML; ELSE MLL
IF AMELX ≥ PQBP1 THEN AML; ELSE MLL
IF DVL1 > ZNF148 THEN AML; ELSE MLL
h2
h2
MLL
AML
SCRN1 (D83777)
ANPEP (M22324)*
CHRNA7 (X70297)
ATF5 (AB021663)
C6orf106(AF052106)
PHGDH (AF006043)
STAT4 (L78440)
AMELX (AC002366)
DVL1 (U46461)
HIST2H4 (X00038)
P29 (AL031432)
TLR1 (AL050262)
NFYC (Z74792)
MEF2C (L08895)
CTGF (X78947)
MEIS1 (U85707)
PQBP1 (AJ005893)
ZNF148 (AJ236885)
−1
Normalized Expression
1
low
high

“Direct” Data Integration

(Lei Xu *et al*, 2005, Bioinformatics, 21:3905-3911)

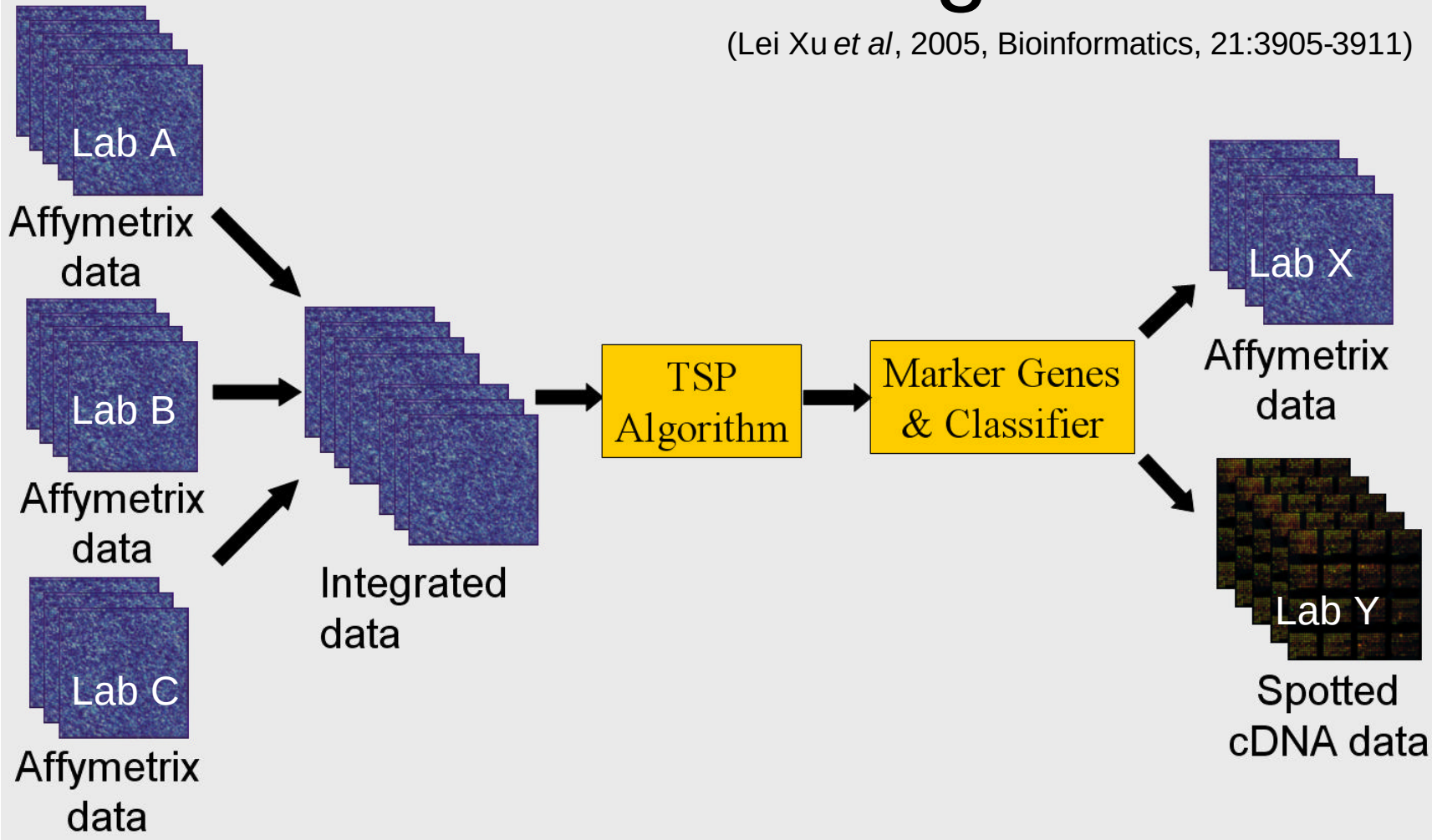


Figure 1: Summary of data integration, marker genes identification and cross-platform validation.

Data Sets

Data Set		Microarray Platform	Number of Probe Sets	No. of Normal Samples	No. of Cancer Samples
Training Set	Singh[6]	Affy. HG_U95Av2	12600	50	52
	Stuart[7]	Affy. HG_U95Av2	12625	50	38
	Welsh[8]	Affy. HG_U95Av2	12626	9	24
Testing Set	LaTulippe[9]	Affy. HG_U95Av2	12626	3	23
	Lapointe[10]	Spotted cDNA	44160/43008*	41	62

* 22 samples (9 normal / 13 cancer) have 44160 probes and 81 samples (32 normal / 49 cancer) have 43008 probes.

(Lei Xu *et al*, 2005, Bioinformatics, 21:3905-3911)

*TSP*s from Data Integration

Training Data Set	Sample Size	Probe Set ID of TSP (HG_U95Av2)	Gene Symbol of TSP	Score of TSP	Classification Accuracy (%)
Welsh	33	39608_at, 32526_at	SIM2, JAM3	1.00	97.0
Stuart	88	41732_at, 456_at	CTNNB1, SMARCD3	0.74	69.3
Singh	102	40282_s_at, 2035_s_at	DF, ENO1	0.90	95.1
Welsh_Stuart*	121	31971_at, 34213_at	TP73L, KIBRA	0.79	77.7
Welsh_Singh	135	37639_at, 32198_at	**HPN**, COMMD4	0.88	83.7
Stuart_Singh	190	37639_at, 41222_at	**HPN, STAT6**	0.75	86.8
Welsh_Stuart_Singh	223	37639_at, 41222_at	**HPN, STAT6**	0.78	88.8

* Welsh_Stuart is the integrated data set of Welsh and Stuart data sets. Other integrated data sets use similar symbols.

(Lei Xu *et al*, 2005, Bioinformatics, 21:3905-3911)

Results on Test Set

Testing Data Set	Microarray Platform	No. of Normal Sample	No. of Cancer Sample	Accuracy (%)	Sensitivity (%)	Specificity (%)
LaTulippe	Affy. HG_U95Av2	3	23	96.2	95.7	100
Lapointe	Spotted cDNA	41	61*	93.1	90.2	97.6
Overall	Cross-platform	44	84	**93.8**	**91.7**	**97.7**

* One of the cancer samples has missing value for HPN and is removed from the testing set.

Comparisons of Marker *TSP* with Individual *TSP*s

Testing Data Set	TSP	Accuracy (%)	Sensitivity (%)	Specificity (%)
LaTulippe (HG_U95Av2)	Welsh	69.2	69.6	66.7
	Stuart	84.5	82.6	100
	Singh	88.5	87.0	100
	Welsh_Stuart_Singh	96.2	95.7	100
Lapointe (cDNA)	Welsh	70.9	95.2	34.1
	Stuart	43.6	6.7	97.6
	Singh	43.7	6.4	100
	Welsh_Stuart_Singh	93.1	90.2	97.6

(Lei Xu *et al*, 2005, Bioinformatics, 21:3905-3911)

Marker *TSP* for Prostate Cancer

- HPN (Hepsin) [biomarker candidate for prostate cancer]
- STAT6 (Signal transduction and translation protein)

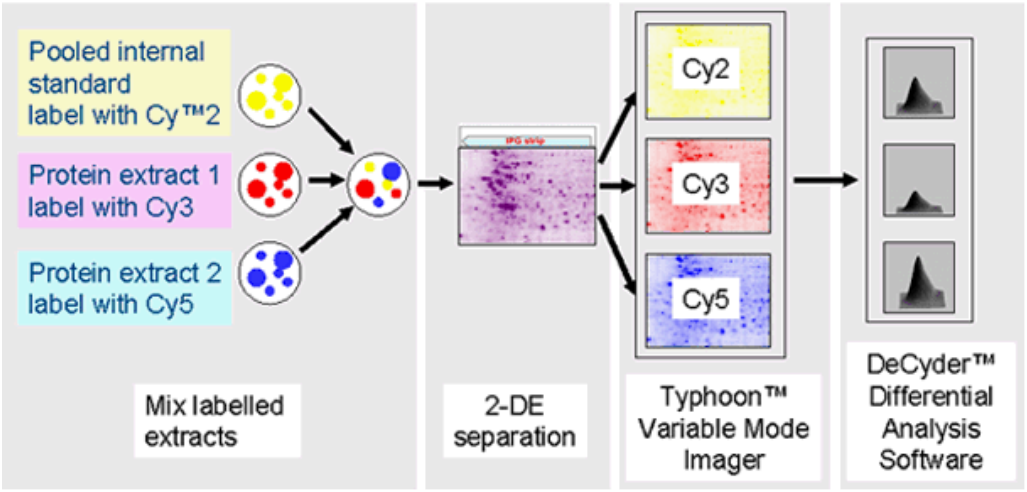
IF HPN > STAT6 **THEN** *Prostate Cancer*
ELSE *Normal*

PSA (Prostate Specific Antigen): Sn = 67.5% – 80% , Sp = 60% - 70%

TSP (HPN, STAT6): Sn = 91.7%, Sp = 97.7% (From this study!)

(Lei Xu *et al*, 2005, Bioinformatics, 21:3905-3911)

DIGE Technology



(From http://www5.amershambiosciences.com)

Proteomics Data

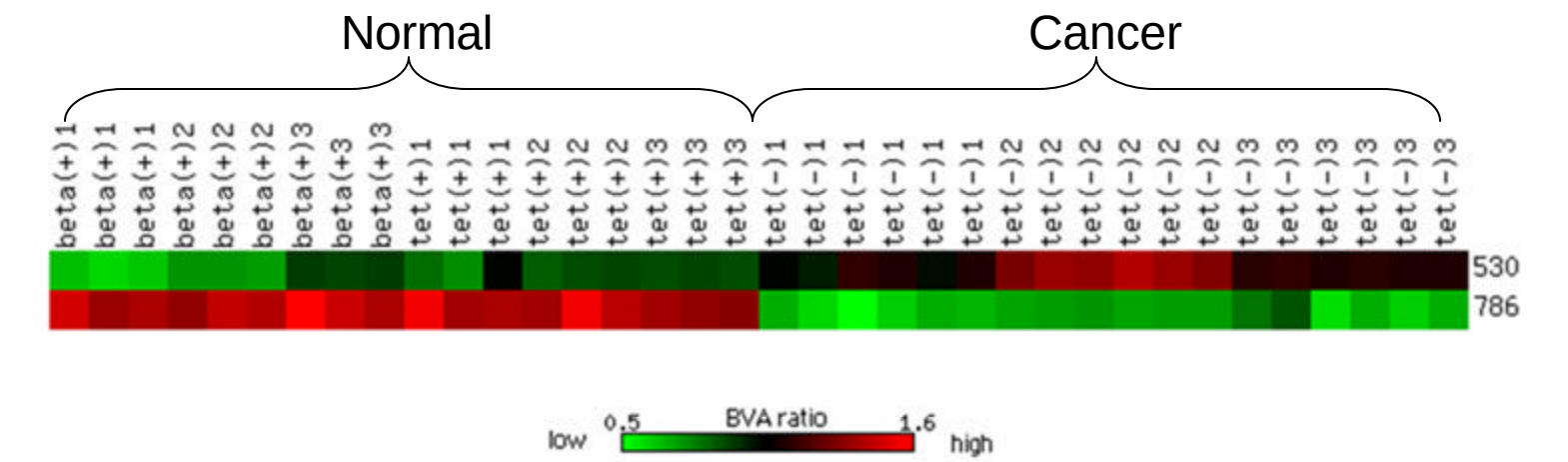
Experimental Settings:

Name:	**Tet (+)**	**Beta(+)**	**Tet (-)**
Model state:	Normal non-proliferating	Normal proliferating	Cancer
Tetracycline present?	yes	yes	no
Beta- estrodiol present?	no	yes	no

Gels: 18 experiments
Cy2 – Internal Standards (18)
Cy3 – Cancer gels (18)
Cy5 – Normal gels (18)
1098 protein spots (BVA ratios from DeCyder software)

(Troy Anderson *et al*)

Decision Rule



Decision Rule:

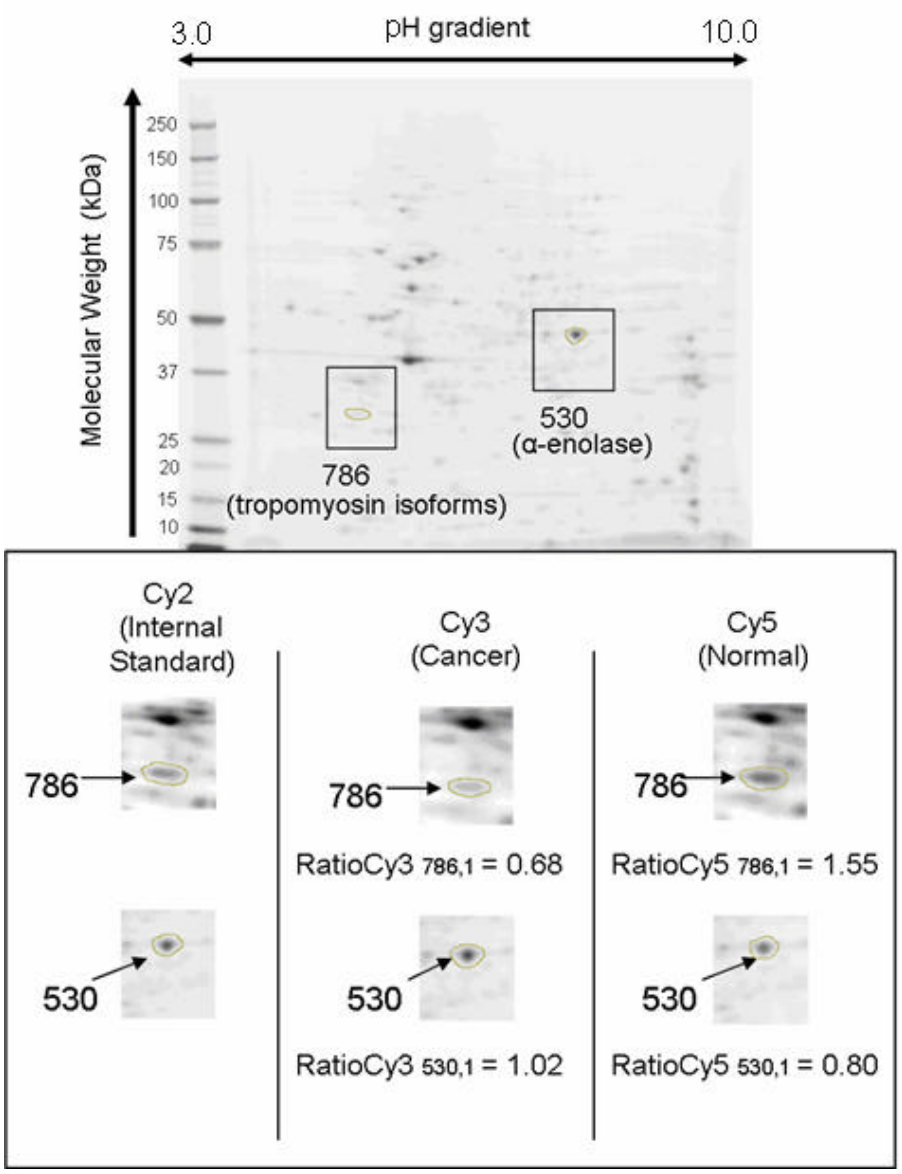
IF $\text{Ratio}_{530} \geq \text{Ratio}_{786}$ **THEN** Cancer,
ELSE Normal.

LOOCV Results:

Accuracy: 97.2%	(35/36)
Sensitivity: 100%	(18/18)
Specificity: 94.4%	(17/18)

(Troy Anderson *et al*)

Protein Marker Spots



(Troy Anderson *et al*)

High-throughput DNA methylation profiling using universal bead arrays

Marina Bibikova,[1] Zhenwu Lin,[2] Lixin Zhou,[1] Eugene Chudin,[1] Eliza Wickham Garcia,[1] Bonnie Wu,[1] Dennis Doucet,[1] Neal J. Thomas,[3] Yunhua Wang,[2] Ekkehard Vollmer,[5] Torsten Goldmann,[5] Carola Seifart,[6] Wei Jiang,[7] David L. Barker,[1] Mark S. Chee,[1] Joanna Floros,[2,3,4] and Jian-Bing Fan[1,8]

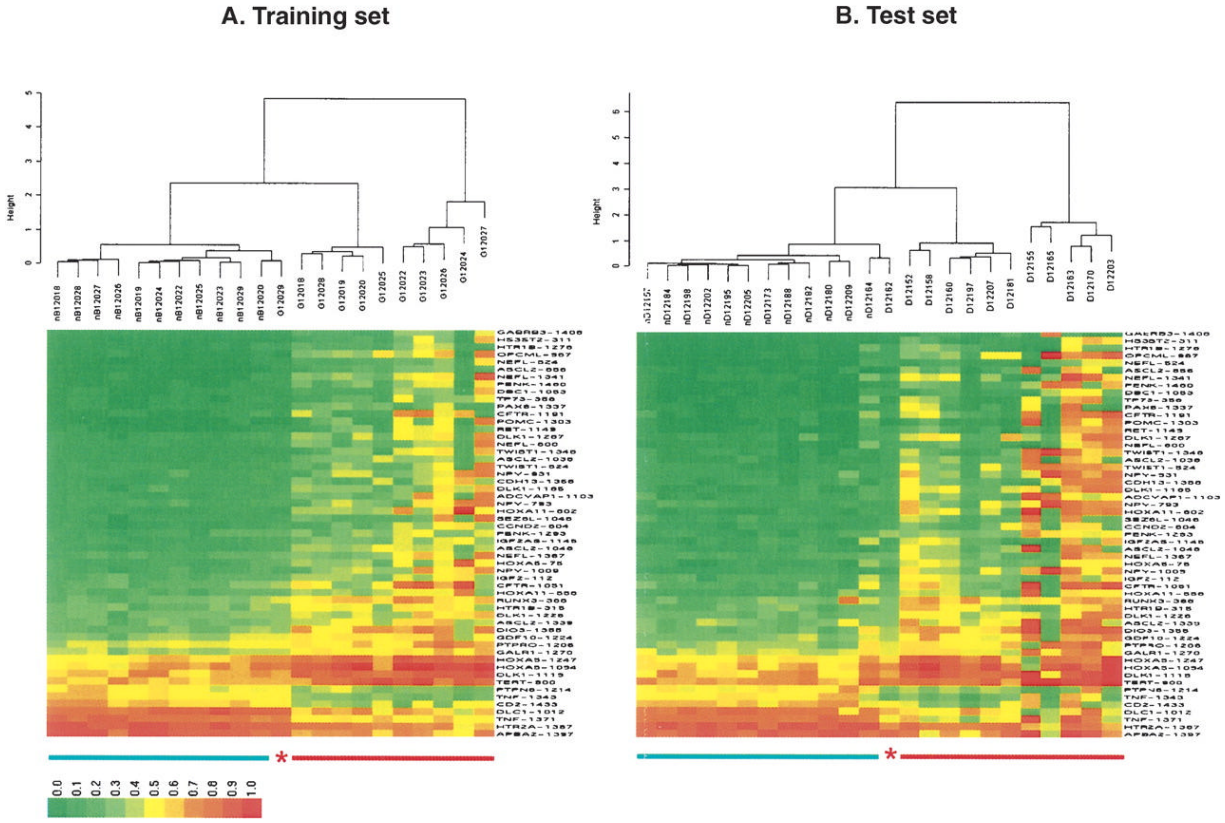
[1]Illumina, Inc., San Diego, California 92121, USA; [2]Department of Cellular and Molecular Physiology, [3]Department of Pediatrics and Health Evaluation Sciences, and [4]Department of Obstetrics and Gynecology, Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033, USA; [5]Clinical and Experimental Pathology, Research Center Borstel, Parkallee, 23845 Borstel, Germany; [6]Department of Internal Medicine, Division of Respiratory Medicine, Philipps-University of Marburg, Baldingerstasse, 35043 Marburg, Germany; [7]The Burnham Institute, La Jolla, California 92037, USA

We have developed a high-throughput method for analyzing the methylation status of hundreds of preselected genes simultaneously and have applied it to the discovery of methylation signatures that distinguish normal from cancer tissue samples. Through an adaptation of the GoldenGate genotyping assay implemented on a BeadArray platform, the methylation state of 1536 specific CpG sites in 371 genes (one to nine CpG sites per gene) was measured in a single reaction by multiplexed genotyping of 200 ng of bisulfite-treated genomic DNA. The assay was used to obtain a quantitative measure of the methylation level at each CpG site. After validating the assay in cell lines and normal tissues, we analyzed a panel of lung cancer biopsy samples ($N = 22$) and identified a panel of methylation markers that distinguished lung adenocarcinomas from normal lung tissues with high specificity. These markers were validated in a second sample set ($N = 24$). These results demonstrate the effectiveness of the method for reliably profiling many CpG sites in parallel for the discovery of informative methylation markers. The technology should prove useful for DNA methylation analyses in large populations, with potential application to the classification and diagnosis of a broad range of cancers and other diseases.

[Supplemental material is available online at www.genome.org.]

Bibikova et al (2006) *Genome Research* 16: 383

Cluster analysis of lung adenocarcinoma samples



•55 CpG sites that are differentially methylated in cancer versus normal tissues with high confidence level (adjusted *P*-value < 0.001) and significant change in absolute methylation level (|β| > 0.15).
•Cancer sample G12029 was mistakenly coclustered with normal samples
•Cancer sample D12162 was coclustered with normal samples
•Normal samples are underlined in green, cancer in red.
•The asterisks indicate misclassified samples.

Bibikova et al (2006) *Genome Research* 16: 383

k-TSP Results

***k-TSP* decision rules:**

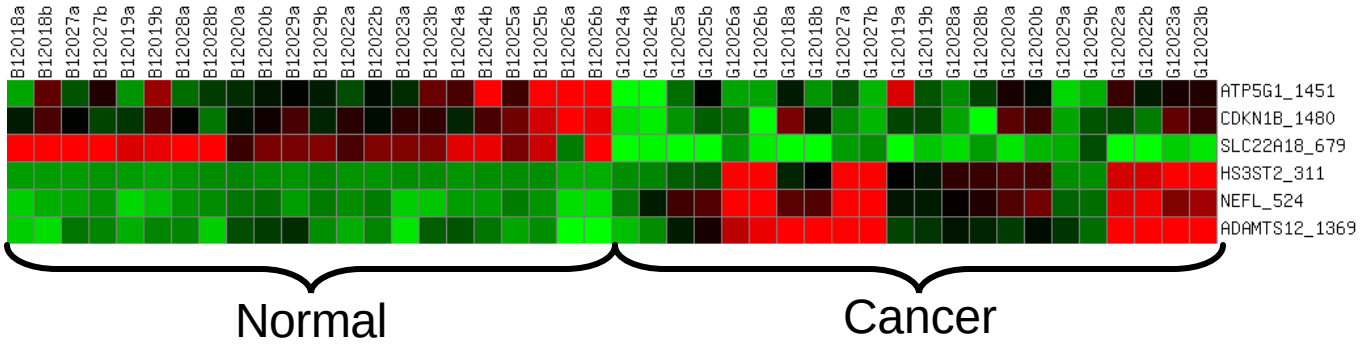
IF ATP5G1_1451 ≥ HS3ST2_311[#] **THEN** Normal, **ELSE** Cancer.

IF CDKN1B_1480 ≥ NEFL_524[#] **THEN** Normal, **ELSE** Cancer.

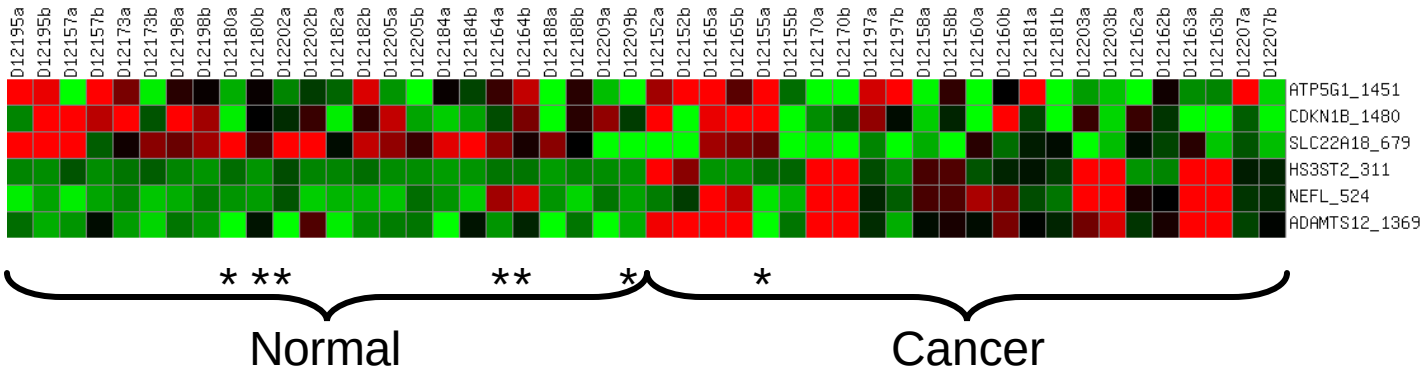
IF SLC22A18_679 ≥ ADAMTS12_1369 **THEN** Normal, **ELSE** Cancer.

[#] CpG sites used in Bibikova et al (adjusted *p*-values = 0.000112, top in their list).

Training Set (LOOCV = 95.5%)



Test Set (Prediction Accuracy = 85.42%)



* The asterisks indicate misclassified samples

Test Results:

Sample	True Class	TSP	KTSP
D12152a	cancer	cancer	cancer
D12152b	cancer	cancer	cancer
D12165a	cancer	normal	cancer
D12165b	cancer	normal	cancer
D12155a	cancer	normal	normal
D12155b	cancer	cancer	cancer
D12170a	cancer	cancer	cancer
D12170b	cancer	cancer	cancer
D12197a	cancer	cancer	cancer
D12197b	cancer	normal	cancer
D12158a	cancer	cancer	cancer
D12158b	cancer	cancer	cancer
D12160a	cancer	cancer	cancer
D12160b	cancer	cancer	cancer
D12181a	cancer	cancer	cancer
D12181b	cancer	cancer	cancer
D12203a	cancer	cancer	cancer
D12203b	cancer	cancer	cancer
D12162a	cancer	normal	cancer
D12162b	cancer	normal	cancer
D12163a	cancer	cancer	cancer
D12163b	cancer	cancer	cancer
D12207a	cancer	cancer	cancer
D12207b	cancer	cancer	cancer
D12195a	normal	normal	normal
D12195b	normal	normal	normal
D12157a	normal	cancer	normal
D12157b	normal	normal	normal
D12173a	normal	normal	normal
D12173b	normal	cancer	normal
D12198a	normal	normal	normal
D12198b	normal	normal	normal
D12180a	normal	cancer	cancer
D12180b	normal	normal	cancer
D12202a	normal	cancer	cancer
D12202b	normal	normal	normal
D12182a	normal	normal	normal
D12182b	normal	normal	normal
D12205a	normal	normal	normal
D12205b	normal	normal	normal
D12184a	normal	normal	normal
D12184b	normal	normal	normal
D12164a	normal	cancer	cancer
D12164b	normal	normal	cancer
D12188a	normal	normal	normal
D12188b	normal	normal	normal
D12209a	normal	normal	normal
D12209b	normal	normal	cancer

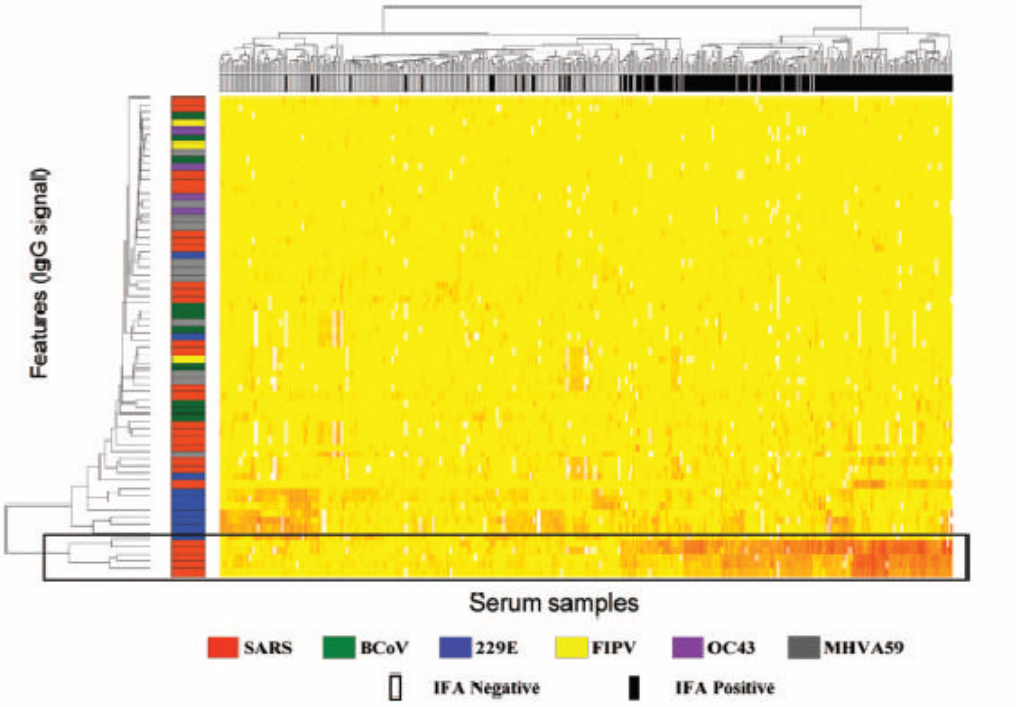
Severe acute respiratory syndrome diagnostics using a coronavirus protein microarray

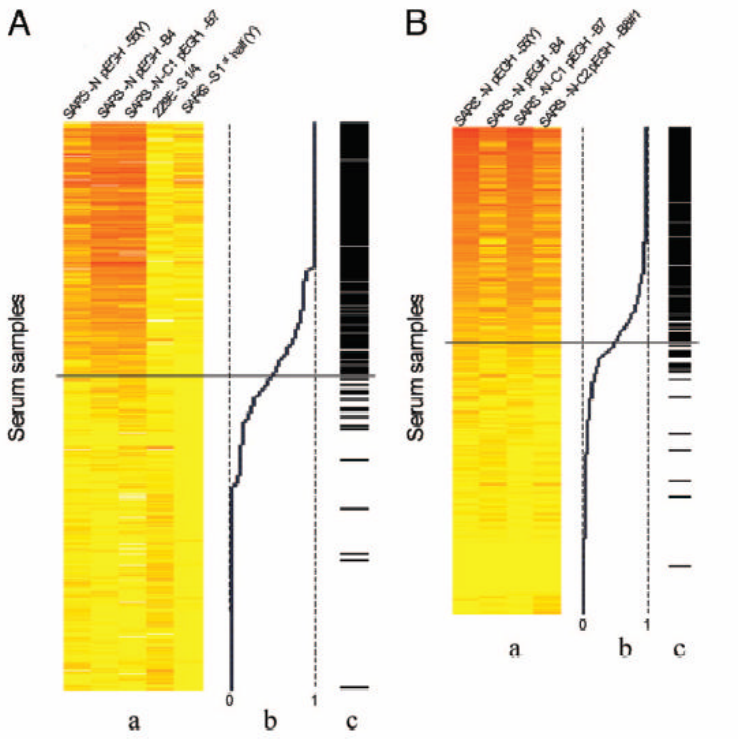
Heng Zhu*†‡§, Shaohui Hu†‡, Ghil Jona*‡, Xiaowei Zhu‡¶, Nate Kreiswirth‖, Barbara M. Willey‖, Tony Mazzulli‖, Guozhen Liu†**, Qifeng Song†, Peng Chen†, Mark Cameron‖, Andrea Tyler‖, Jian Wang†, Jie Wen†, Weijun Chen†, Susan Compton††, and Michael Snyder*¶‡‡

PNAS (2006) 103(11): 4011-4016

Departments of *Molecular, Cellular, and Developmental Biology and ††Comparative Medicine, and ¶Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520; †Biochip Platform Division, Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China; ‖Department of Microbiology, Mount Sinai Hospital, Toronto, ON, Canada M5G 1X5; and **College of Life Sciences, Agricultural University of Hebei, Hebei, Baoding 071001, China

Communicated by Dieter Söll, Yale University, New Haven, CT, January 16, 2006 (received for review October 10, 2005)





AC TAN 2006

Unsupervised Learning

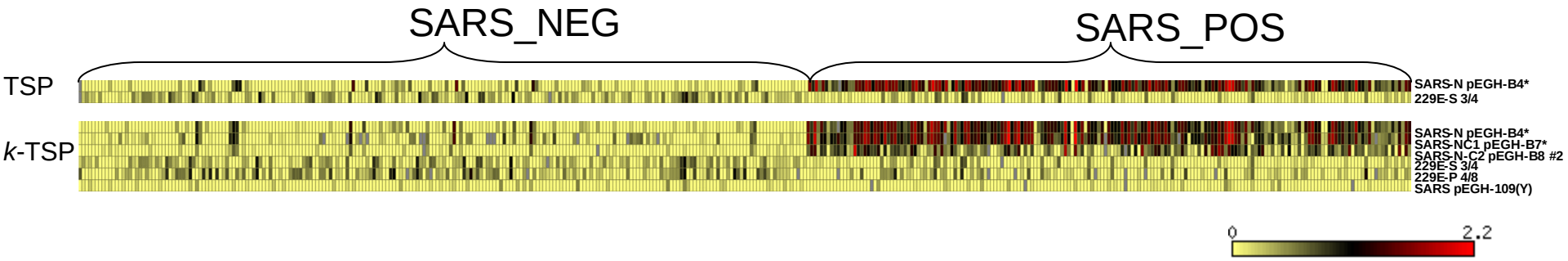
k-NN

LR

Results on SARS Data

(Zhu et al 2006 PNAS)

(A) Canadian Sera Data (Training Set) [Data obtained from Supporting Table 4]



***TSP* Decision Rule:**

IF SARS-N pEGH-B4* > 229E-S 3/4 THEN SARS_POS, ELSE SARS_NEG

***k-TSP* Decision Rules:**

IF SARS-N pEGH-B4* > 229E-S 3/4 THEN SARS_POS, ELSE SARS_NEG
IF SARS-NC1 pEGH-B7* > 229E-P 4/8 THEN SARS_POS, ELSE SARS_NEG
IF SARS-N-C2 pEGH-B8 #2 > SARS pEGH-109(Y) THEN SARS_POS, ELSE SARS_NEG

* Proteins identified by k-NN & LR methods in Zhu et al

Leave-One-Out Cross-Validation:

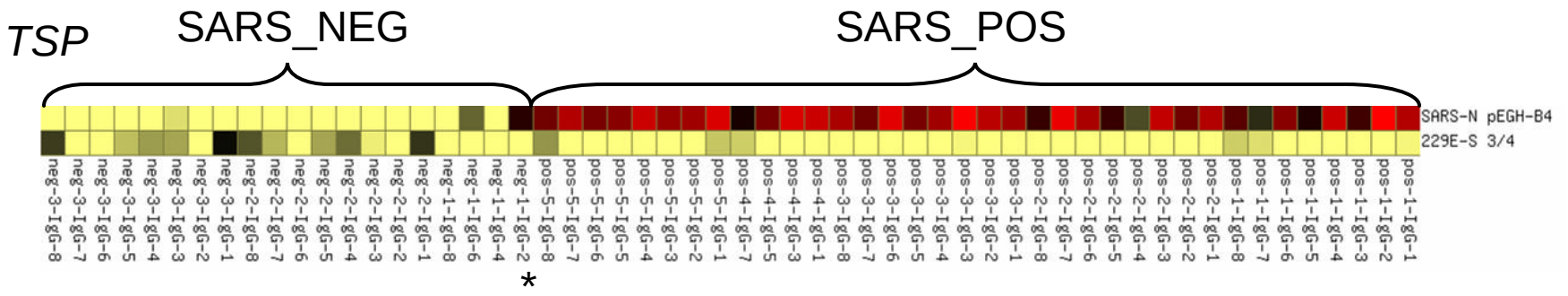
TSP accuracy: 87.7%

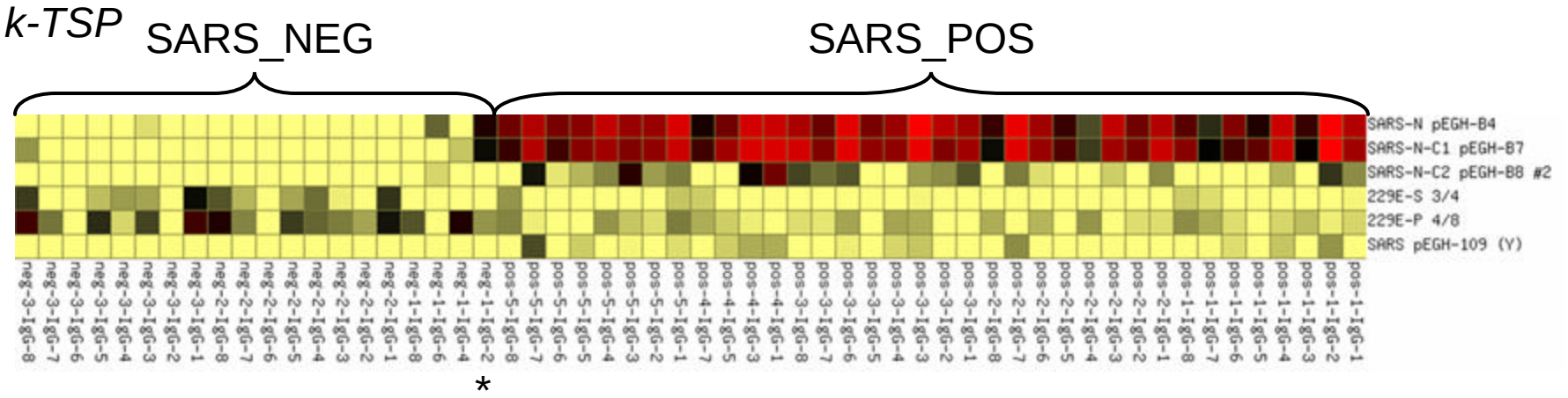
k-TSP accuracy: 90%

Results on the SARS Data

(Zhu et al 2006 PNAS)

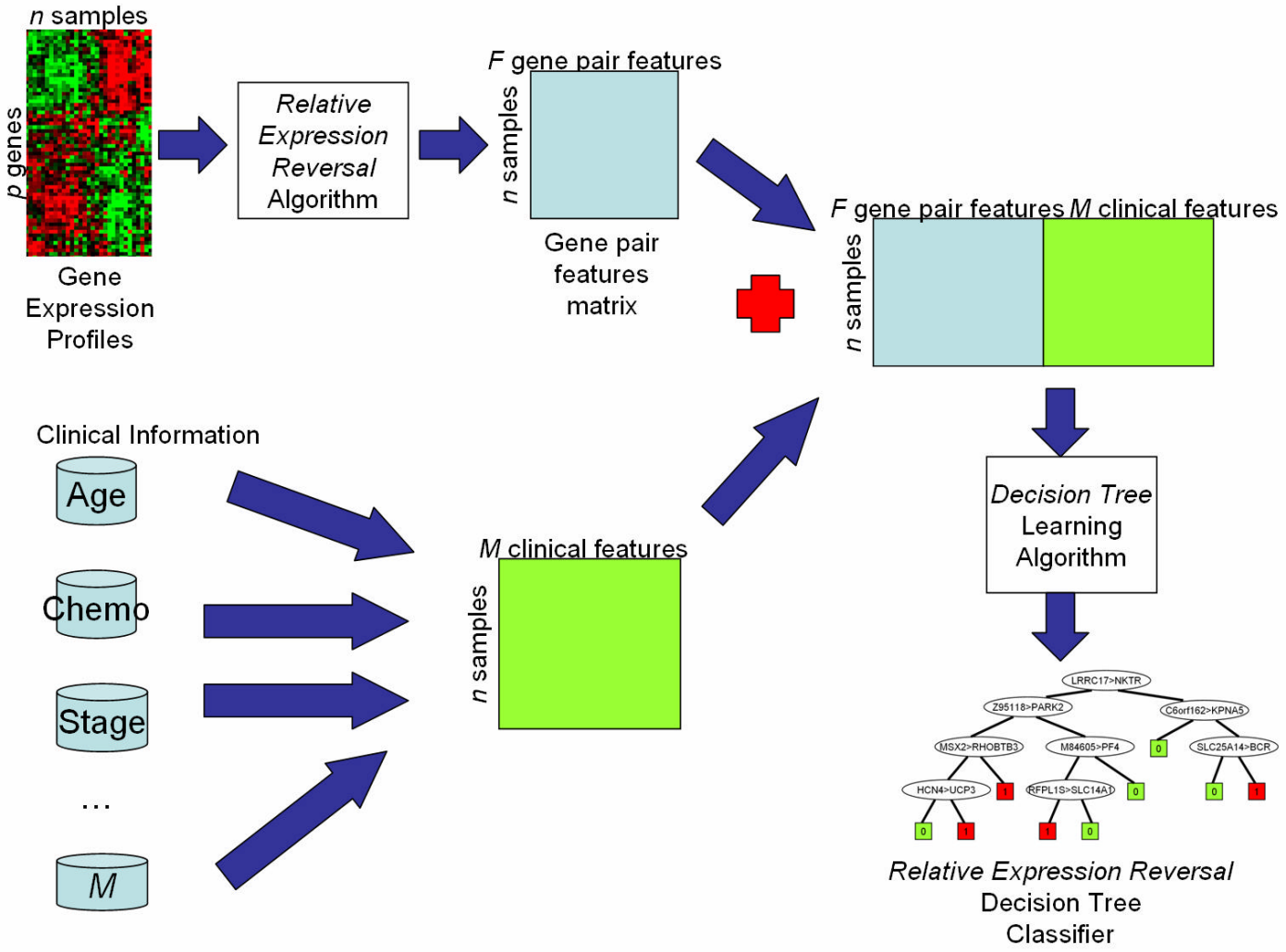
(B) Chinese Sera Data (Testing Set) [Data obtained from Supporting Table 6]





Both classifiers misclassified 1 sample (*) on the Chinese SARS data (56 samples)

Integrating Gene Expression Data and Clinical Information for Cancer Outcome Prediction



Preliminary Results

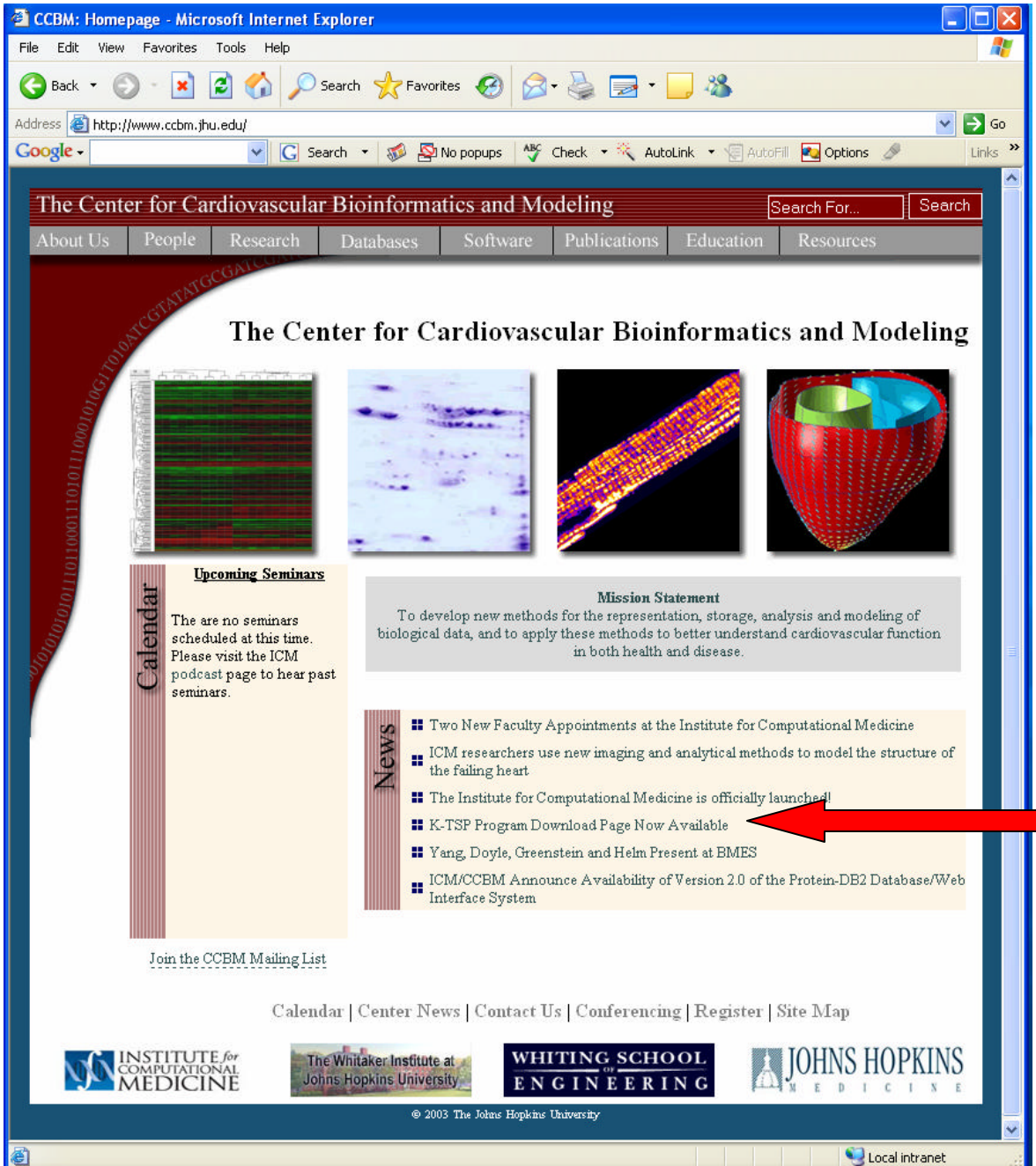
Data Sets:

	Beer et al	West et al	Huang et al	Bullinger et al	Rosenwald et al	Ovarian	Pittman et al	Lung	Miller et al	Ma et al
#genes	7129	7129	12625	6283	7399	22215	12625	54613	44928	22575
#samples	86	49	89	100	240	133	171	85	236	60
Dead	24	15	36	66	138	61	43	43	181	28
Alive	62	34	53	34	102	72	128	42	55	32
# Clinical features	7	5	8	12	2	1	1	2	7	8
Cancer	Lung	Breast	Breast	Leukemia	Lymphoma	Ovarian	Breast	Lung	Breast	Breast

LOOCV Accuracy:

	Beer et al	West et al	Huang et al	Bullinger et al	Rosenwald et al	Ovarian	Pittman et al	Lung	Miller et al	Ma et al
TSP	44.19	48.98	46.07	35.00	63.33	64.70	51.50	20.00	32.20	41.70
***k*-TSP**	66.28	55.10	42.70	60.00	61.25	77.40	56.70	47.10	51.30	46.70
DT(50-TSP + Clinical)	60.47	83.67	56.18	80.00	57.50	72.18	70.18	60.00	74.15	90.00

http://www.ccbm.jhu.edu



Software Availability

Conclusions

- Bioinformatics tools to facilitate biomarkers discovery
- *k-TSP* is comparable with the state-of-the-art classifiers (*PAM*, *SVM*) in classifying gene expression profiles
- *k-TSP* generates simple and accurate decision rules
  - Biological significance
  - Easy to interpret
  - Potential clinical applications
- Allow “direct” data integration without performing normalization
- Allow cross-platform analysis
- Applicable to a wide-range of high-throughput data

Acknowledgements

- Prof. Raimond Winslow
- Prof. Donald Geman
- Prof. Daniel Naiman
- Lei Xu
- Troy Anderson
- DIMACS Travel Fellowships

- THANK YOU !

EMAIL: actan@jhu.edu

References:

- D. Geman, C. d'Avignon, D.Q. Naiman and R.L. Winslow (2004). Classifying gene expression profiles from pairwise mRNA comparison. Statistical Applications in Genetics and Molecular Biology, 3: Article 19.
- A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow and D. Geman (2005). Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics, 21(20): 3896-3904.
- L. Xu, A.C. Tan, D.Q. Naiman, D. Geman and R.L. Winslow (2005). Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics, 21(20): 3905-3911.