

Finding Robust Biomarkers

KITP, March 15, 2005

1



1 IAS
2 IBM
3 UMDNJ

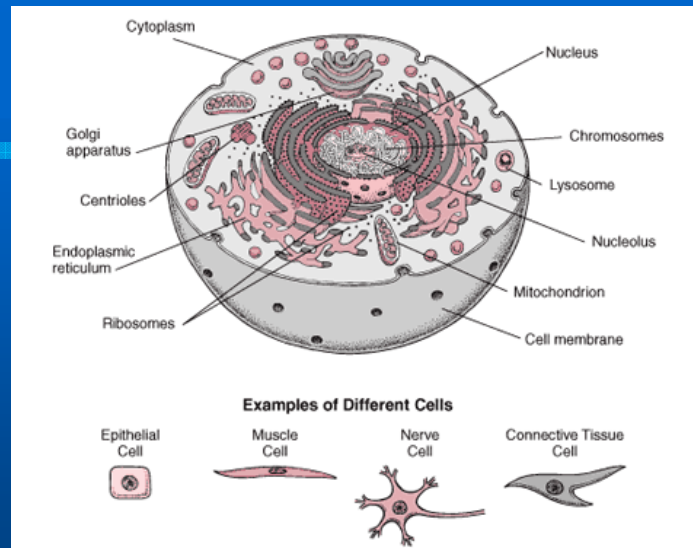
Gabriela Alexe¹
Gyan Bhanot^{1,2}
Babu Venkataraghavan¹
Gustavo Stolovitzky²
Arnold Levine^{1,3}

2

Overview

- Brief Introduction to Mol Bio, Mass Spec and Gene Arrays
- Motivation for Present Analysis
- Pattern-based meta-classifier
- Case studies
 - Prostate CA from mass spec data
 - FL/DLBCL Gene Array data

3



4

Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data

HUMAN CHROMOSOMES

The diagram illustrates the human karyotype, showing 22 pairs of autosomes and one pair of sex chromosomes (X and Y). A detailed view of a chromosome shows its structure, including the centromere, telomere, and chromatid.

The human haploid genome contains 3,000,000,000 DNA nucleotide pairs, divided among twenty two (22) pairs of autosomes and one pair of sex chromosomes.

A chromosome is formed from a single DNA molecule that contains many **genes**. A chromosomal DNA molecule contains three specific nucleotide sequences which are required for replication: a DNA *replication origin*; a *centromere* to attach the DNA to the *mitotic spindle*; a *telomere* located at each end of the linear chromosome.

Bases: Purines and Pyrimidines

The diagram shows the chemical structures of five nitrogenous bases:

- Adenine (A):** A purine base with a double-ring structure. An arrow points to the 1' carbon of either pentose.
- Guanine (G):** A purine base with a double-ring structure. An arrow points to the 1' carbon of either pentose.
- Thymine (T):** A pyrimidine base with a single-ring structure. An arrow points to the 1' carbon of deoxyribose.
- Cytosine (C):** A pyrimidine base with a single-ring structure. An arrow points to the 1' carbon of either pentose.
- Uracil (U):** A pyrimidine base with a single-ring structure. An arrow points to the 1' carbon of ribose.

Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data

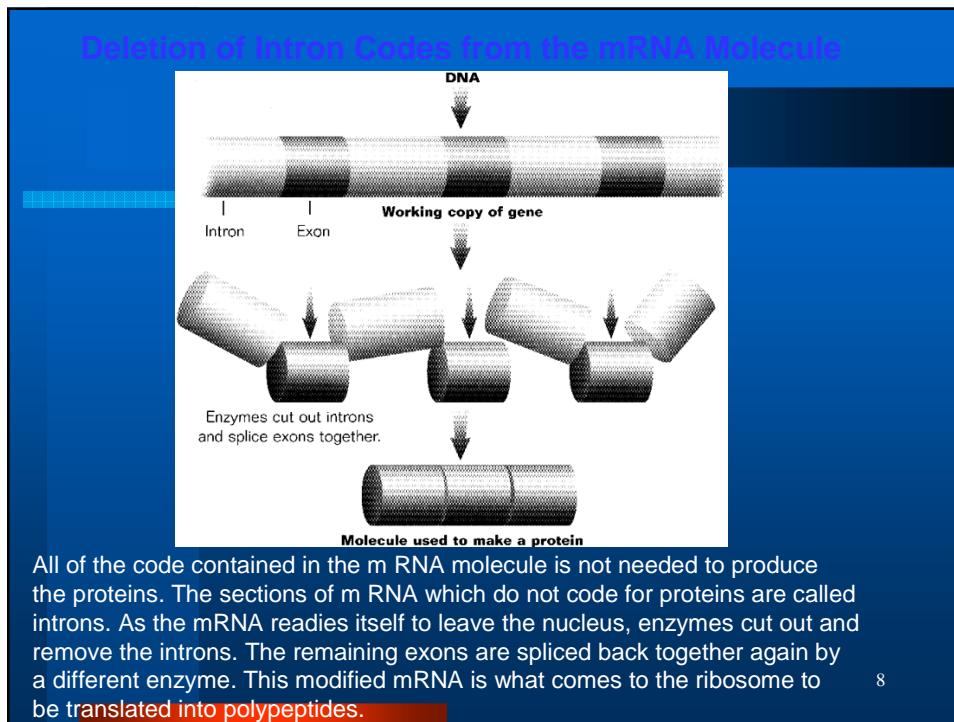
a The genetic code, each amino acid is coded for by three mRNA bases arranged in a specific sequence.

b The second base is at the top of the chart.

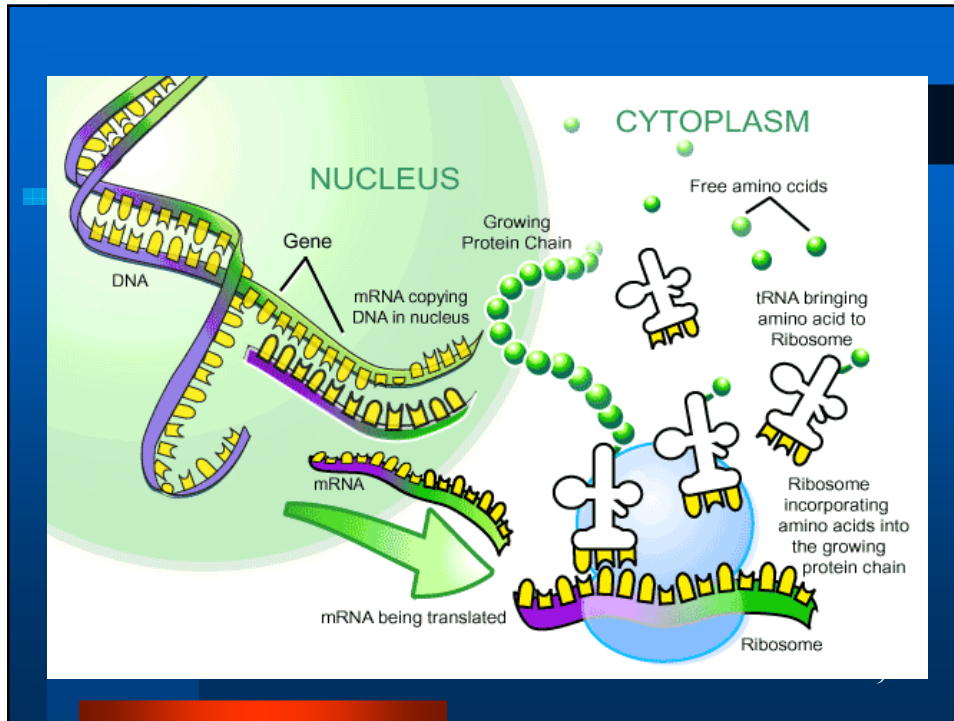
c The third base in the codon is found along the right side of the chart.

The first base in a codon is found along the left side of the chart.

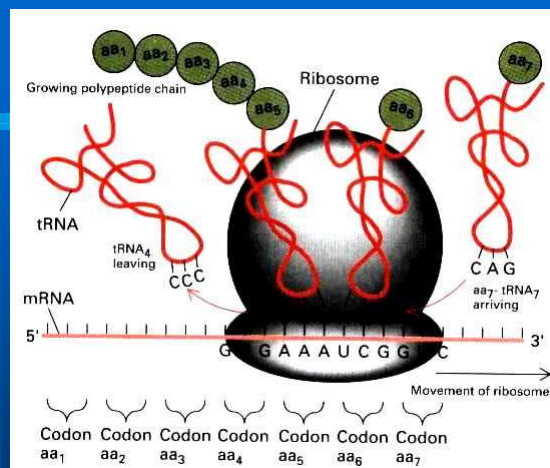
1st ↓	2nd →	U	C	A	G	
U	Phenylalanine Phenylalanine Leucine Leucine	Serine Serine Serine Serine	Tyrosine Tyrosine Stop Stop	Cysteine Cysteine Stop Tryptophan	U C A G	
C	Leucine Leucine Leucine Leucine	Proline Proline Proline Proline	Histidine Histidine Glutamine Glutamine	Arginine Arginine Arginine Arginine	U C A G	
A	Isoleucine Isoleucine Isoleucine Methionine	Threonine Threonine Threonine Threonine	Asparagine Asparagine Lysine Lysine	Serine Serine Arginine Arginine	U C A G	
G	Valine Valine Valine Valine	Alanine Alanine Alanine Alanine	Aspartic acid Aspartic acid Glutamic acid Glutamic acid	Glycine Glycine Glycine Glycine	U C A G	



Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data



Translation: synthesis of protein by ribosomes in cytoplasm



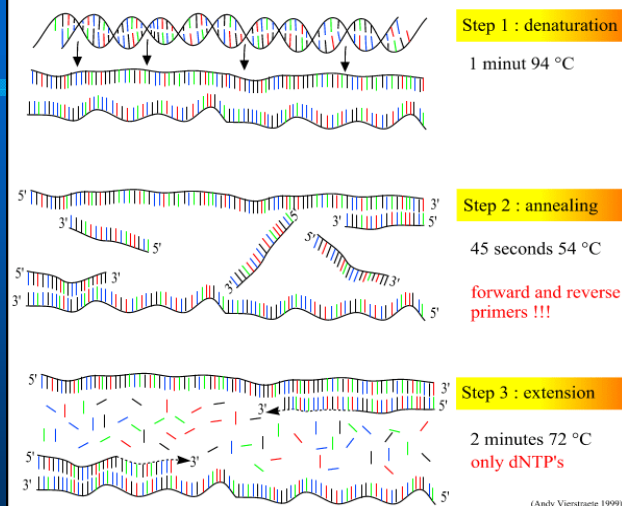
Ribosome: Takes in m-RNA, read it from 5' to 3' end in codon units (1 codon = 3 bps) and attaches the appropriate t-RNA so that the correct amino acid is placed in the growing polypeptide chain. The process terminates at stop codon on mRNA and the newly completed polypeptide is released from ribosome.

Introduction to Gene Arrays and Mass Spectrometry

11

PCR : Polymerase Chain Reaction

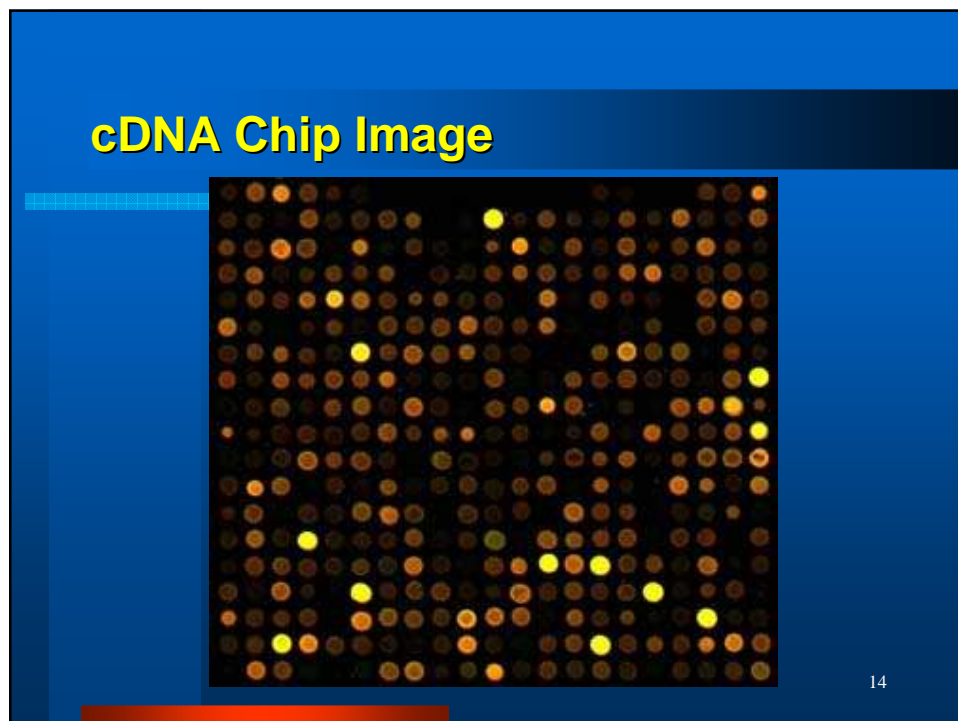
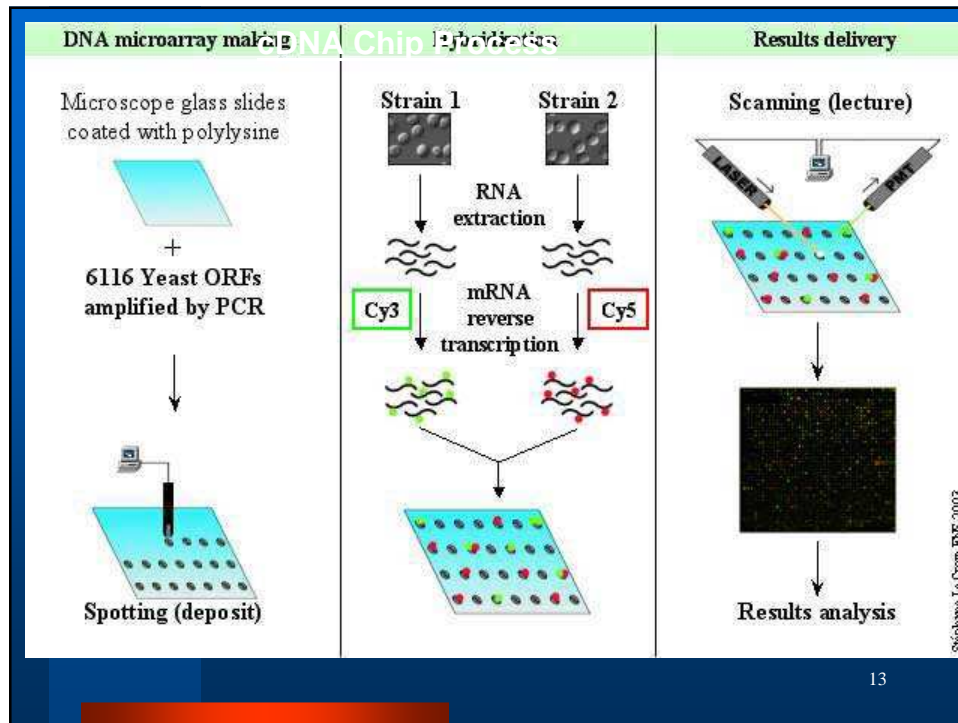
30 - 40 cycles of 3 steps :



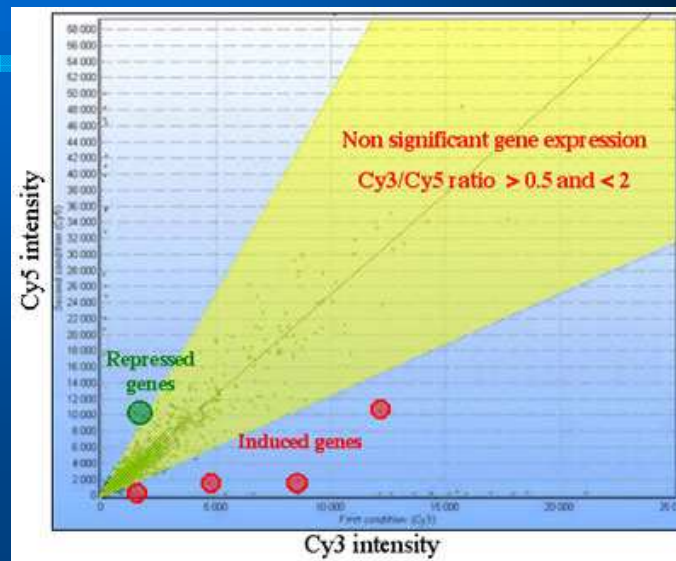
[Animated picture of PCR](#)

12

Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data



Analysis



15

- Detailed Protocols:

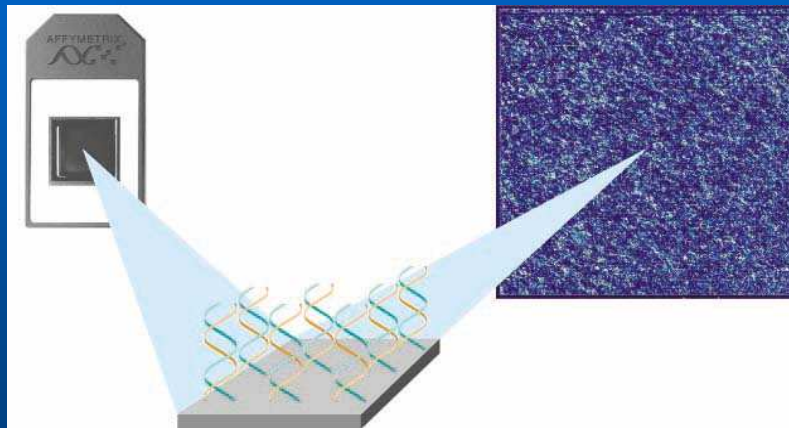
<http://www.microarrays.org/protocols.html>

- Animation of [DNA Microarray Methodology](#)

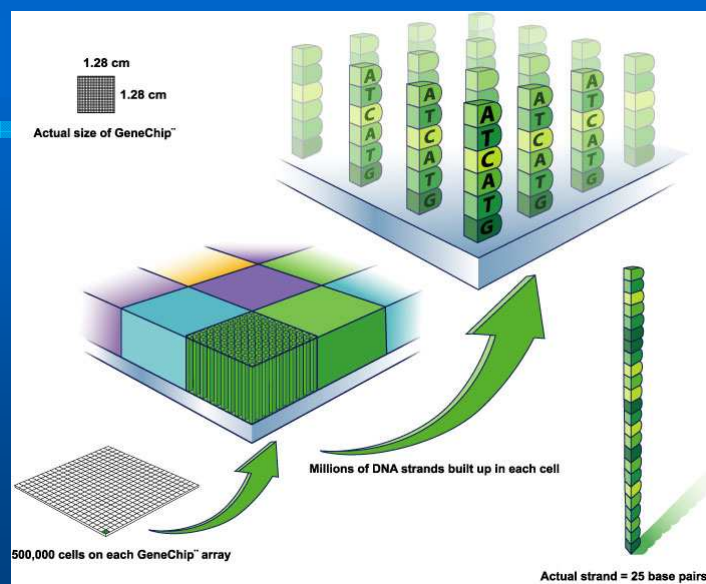
16

Affymetrix Chips

An Alternative is to synthesize the DNA directly onto the matrix

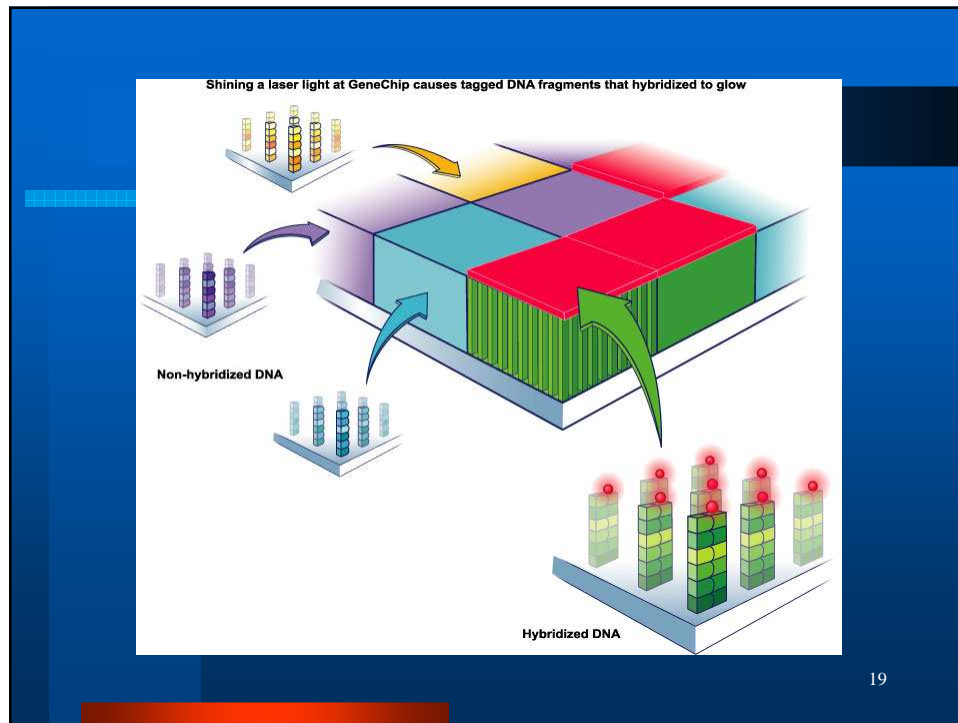


17



18

Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data



Analysis of Chip data

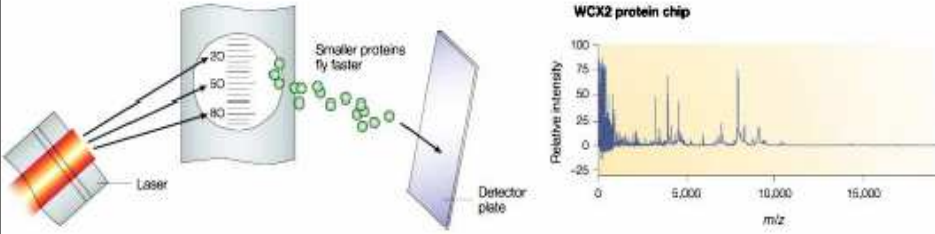
● Free downloads of many analysis tools

- Genes@work:
<http://www.research.ibm.com/FunGen/FGDownloads.htm>
- Broad Institute (MIT) <http://www.broad.mit.edu/resources.html>
- Other Useful Sites to browse:
 - NCBI: <http://workshop.molecularevolution.org/software/ncbi/>
<http://www.ncbi.nlm.nih.gov/genome/guide/human/>
 - Entrez Genome:
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>

20

Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data

Mass Spec



The diagram illustrates the SELDI-TOF process. On the left, a laser beam is directed at a protein chip. The chip has a scale with markers at 20, 50, and 80. Proteins are shown binding to the chip surface. A label 'Smaller proteins fly faster' points to the ions being ejected. On the right, a detector plate captures the ions. To the right of the detector is a mass spectrum graph titled 'WCX2 protein chip'. The y-axis is 'Relative intensity' ranging from -25 to 100. The x-axis is 'm/z' ranging from 0 to 15,000. The spectrum shows several peaks, with the most prominent ones between 5,000 and 10,000 m/z.

SELDI-TOF (Surface Enhanced Laser Desorption Ionization - Time of Flight) : Robotic sample dispenser applies 1 μ L serum to surface of protein-binding chip. Subset of proteins bind to chip surface. Bound proteins treated with a MALDI (Matrix-Assisted Laser-Desorption Ionization) matrix, washed and dried. Chip inserted into a vacuum chamber and irradiated with a laser. Laser desorbs adherent proteins, which are launched as ions. 'Time of flight' (TOF) is a measure of m/z .

WCX2 (weak cation-exchange surface)

21

Motivation for Present Analysis

- Promise of Gene Chips and Mass Spec not translated to treatment
- No consensus on method, lack of robustness, not accepted by medical community
- NEED FOR ROBUST PREDICTION

22

Prostate cancer

- Prostate cancer accounts for over 25% of all cancers in men (395,000 new cases/year)
- Risk increases with age
- Increased incidence since 80's (probably from increased use of PSA screening)
- Screening role is controversial

23

Diagnosis

Low risk : Minimal disease, stage T1c or limited T2a (tumor covers < 1/2 of one lobe), Gleason score below 7 and PSA < 10 ng/mL

Medium risk: Stage T2b (tumor can be felt on both sides of prostate but no evidence of spread), Gleason score =7, 10 ng/mL < PSA < 20 ng/mL

High risk: Spread to surrounding tissue - Stage T3 or Gleason score > 7 or PSA level > 20 ng/mL

The Stages of Prostate Cancer

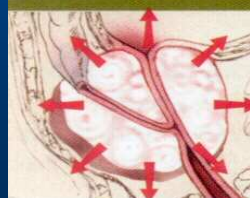


Stage T1c
The tumour can't be felt by DRE



Stage T2a
Tumour involves one lobe

Stage T2b
Tumour involves both lobes



Stage T3-T4
Tumour has spread to surrounding tissue

Cancer is a Proteomic Disease Should be Visible in Serum Protein

- CA cell interacts with tissue microenvironment
- Genetic mutations modify pathways to create survival advantage for cell
- Tumor growth, invasion, angiogenesis pathways are modified
- Pathogenic pathways extend to tumor / host interface

The diagram illustrates the complex signaling pathways between a carcinoma cell and its microenvironment. Key components include:

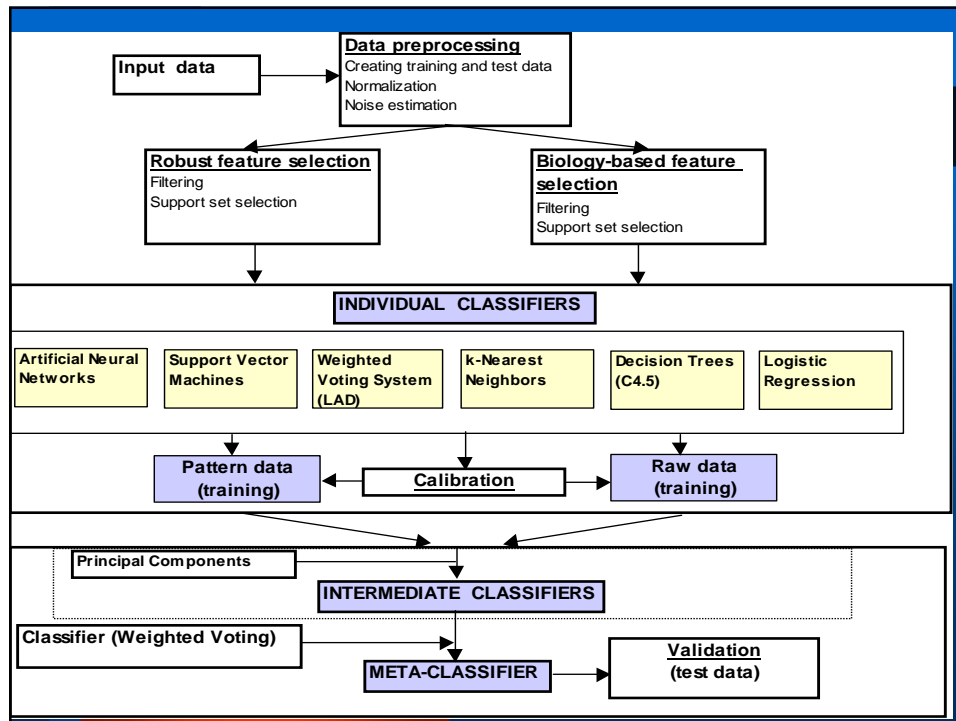
- Extracellular Matrix:** Fibroblasts and Endothelial cells release factors like TGF-β, uPA, MMP, and VEGF.
- Carcinoma Cell Receptors:** c-MET, TGF-βR, Integrin/RGD, and EGFR.
- Intracellular Pathways:** FAK, PI3K, RAS, and others.
- Outcomes:** Motility (Cytoskeletal and extracellular matrix remodeling), Proliferation (Gene transcription, bypass/override cell-cycle checkpoints), and Survival (Suppression of apoptosis, avoidance of anoikis).

What is the promise of proteomics?

- Early detection
- Accurate staging
- Increased survival/cure rates
- Avoid surgery, morbidity
- Protein pathways may suggest better HT, markers other than PSA and perhaps lead to individualized therapy

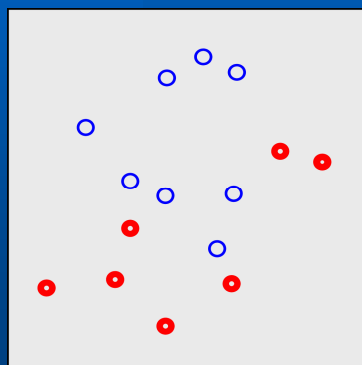
26

Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data

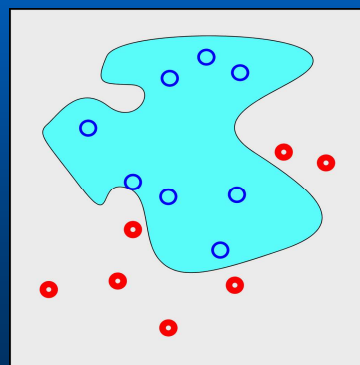


Patterns

Observed dataset



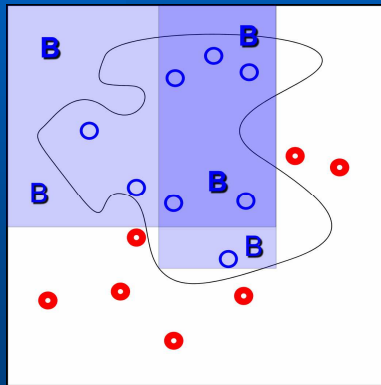
Desired identification



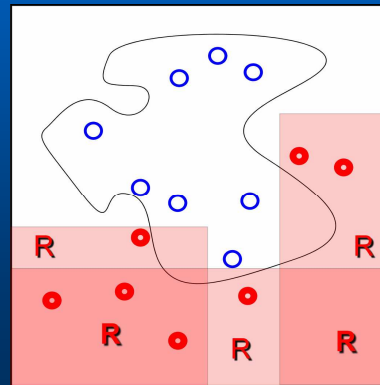
28

Patterns

Positive patterns

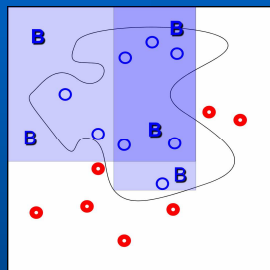


Negative patterns

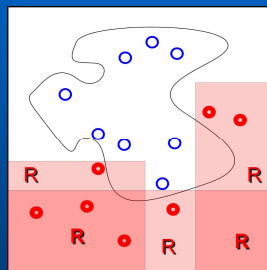


Pattern-based classification

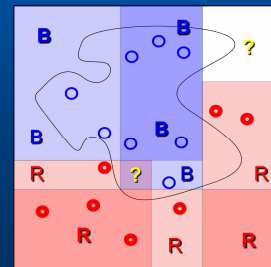
Positive patterns



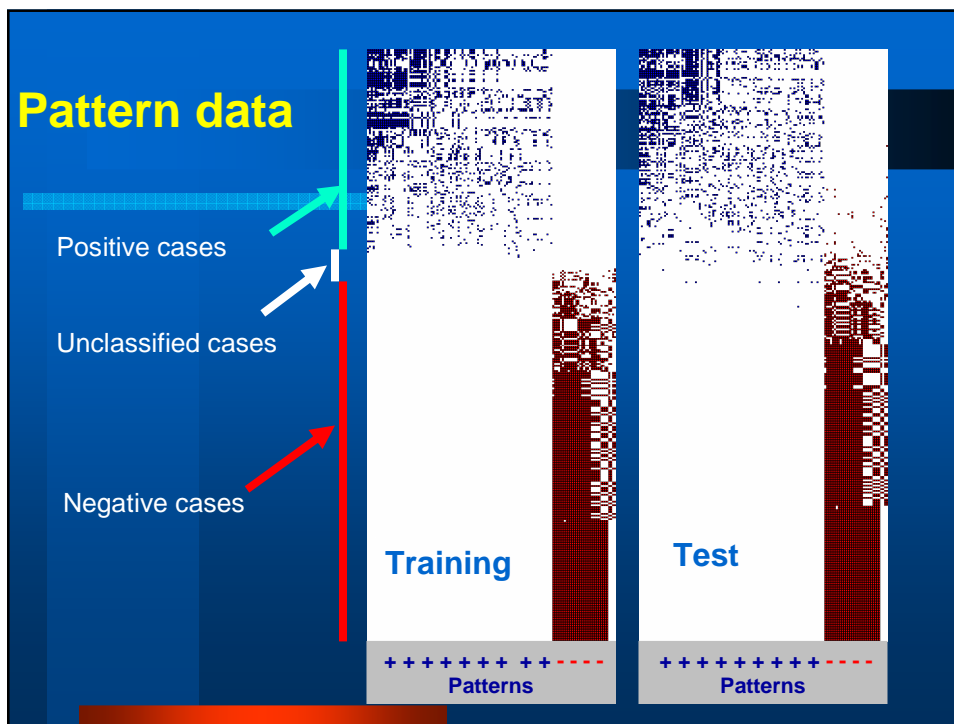
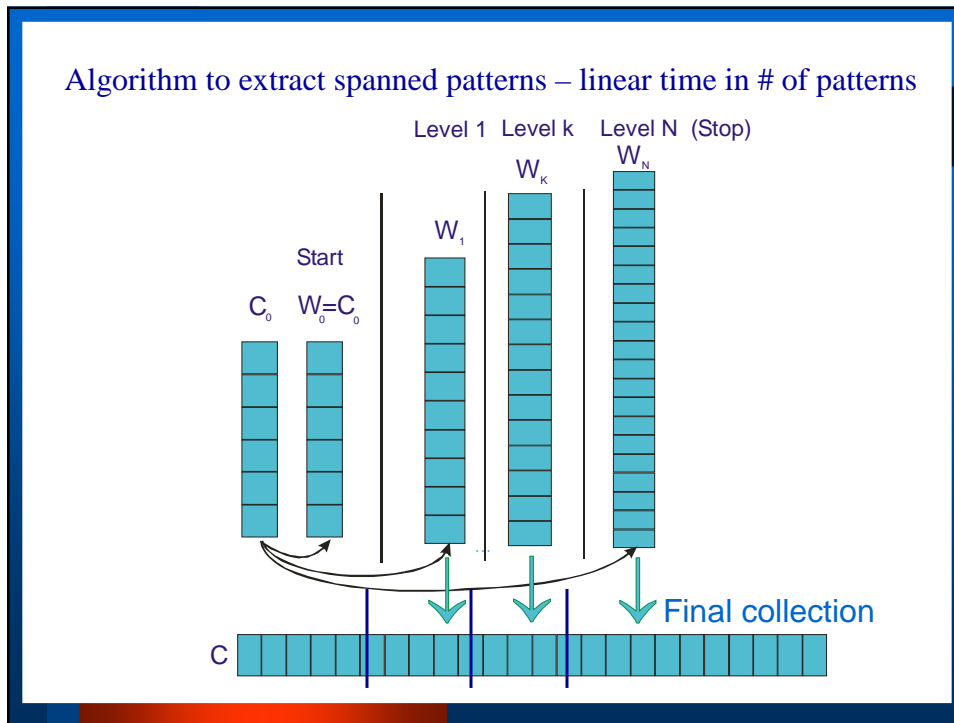
Negative patterns



Model



Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data



Case Study 1: Prostate CA

Data: MALDI-TOF/SELDI-TOF C16 Chipergen Biosystems
JNCI_Data_7_30_02 <http://ncifdaproteomics.com/>

Cases:
253 controls
69 prostate cancer

Features: 15,154 peptides

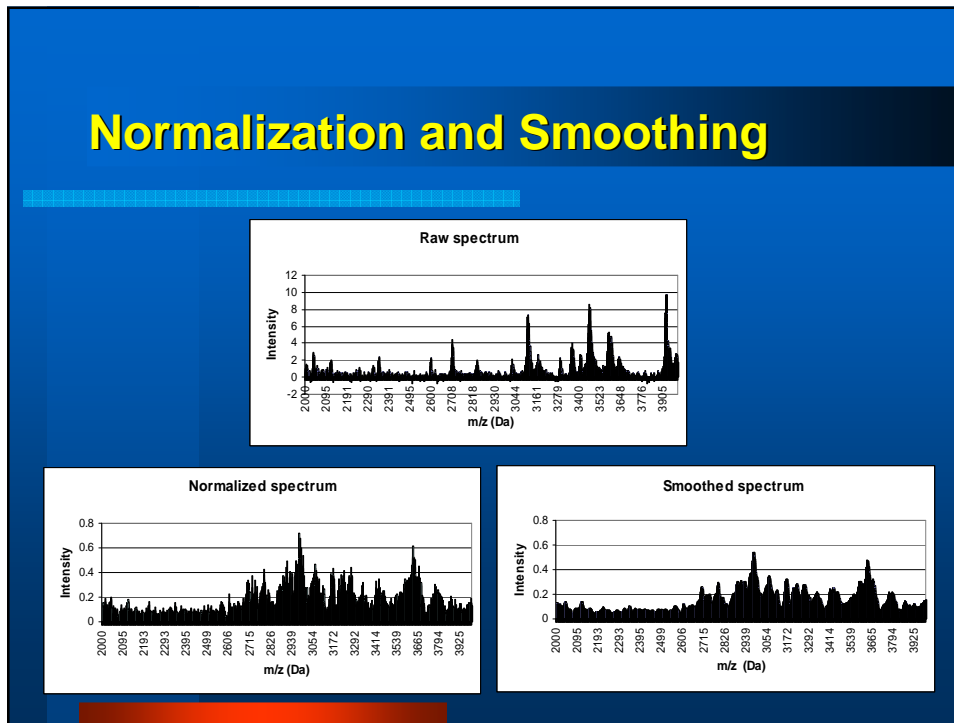
33

Data preprocessing

- Data comes with baseline subtracted, peaks aligned
- Randomly extract 2/1 training/test sets in both phenotypes
- Normalize and smooth:
$$x \rightarrow \frac{x - \text{mean}(x)}{\text{stdev}(x)}$$
- Estimate Sample Noise as median of variances

34

Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data



Data reduction: Step 1

Univariate filtering (original data + 100 perturbed datasets)

- signal-to-noise (top 25%)
- F-test (p-value 0.001)
- t-test (q-value 0.001)
- Pearson correlation with ideal phenotype (top 25%)

Select only features that pass ALL tests

Leaves us with only 512 m/Z values

36

Data reduction: Step 2

- **Peptides with nearby m/Z values highly correlated**
- **Binning**
 - *bin label* = average m/z
 - *bin intensity / sample* = average intensity sample / bin
 - *bin weight* = # peptides in bin
- **39 data bins**
 - (22 m/Z < 2000 kDa & 17 m/Z > 2000 kDa)

37

Data reduction: Step 3

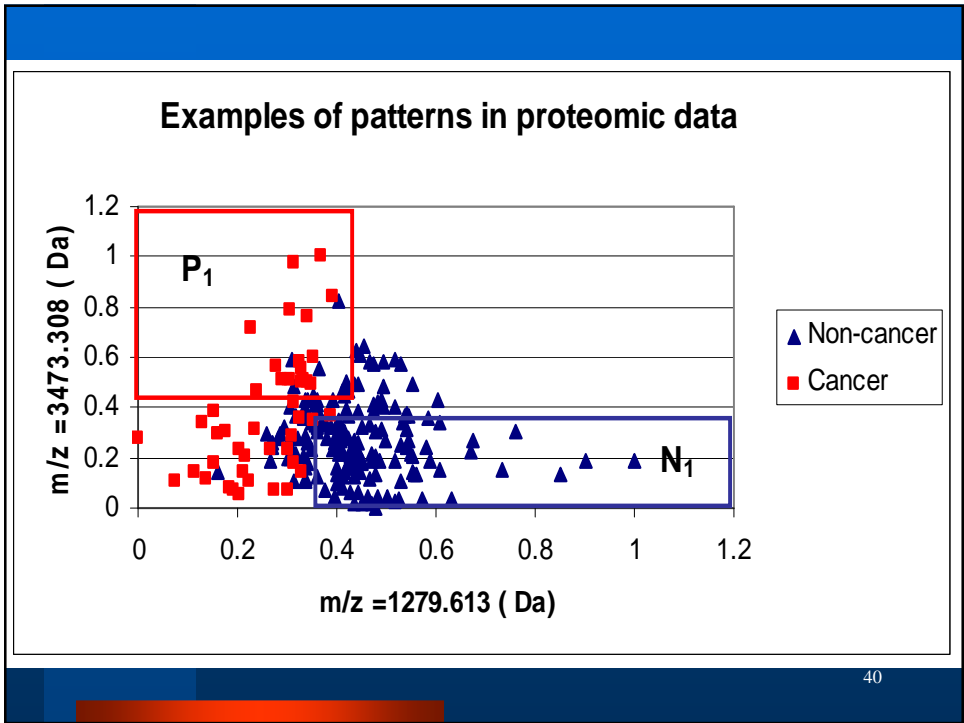
- **Pattern-based selection**
 - Create large collections of patterns
 - Sort m/Z bins by their frequency in patterns
 - Iterate until # bins relatively small
 - record weighted voting sensitivity/specificity at each step
- **Final support set has 11 bins**

38

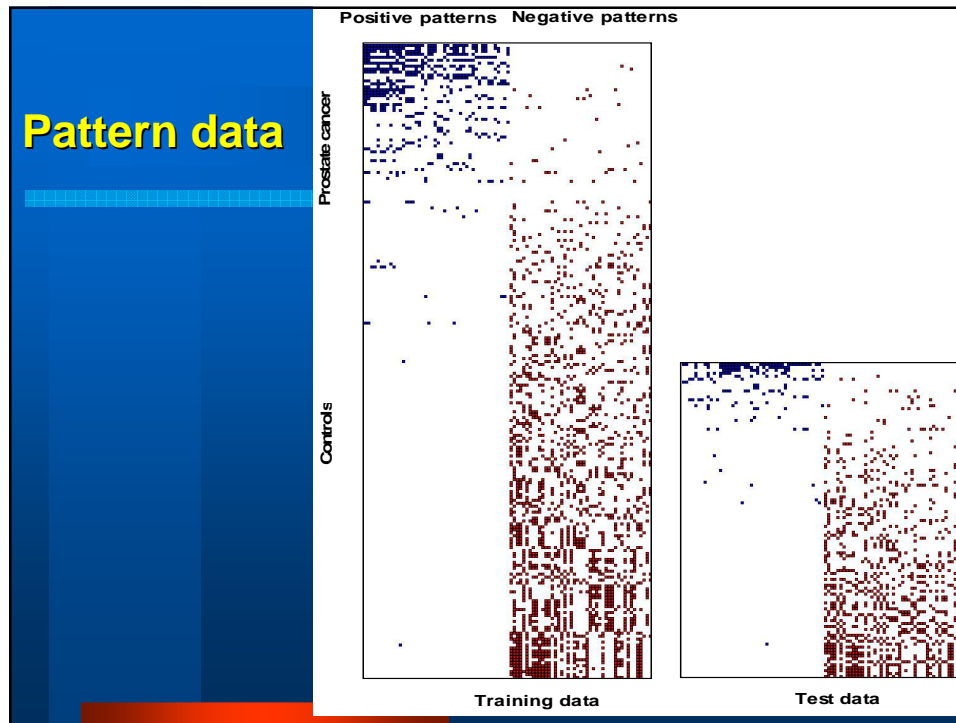
Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data

Collection of bins selected by the filtering approach

Bin #	95% CI		Representative m/z	# m/z values included in bin	Average intensity	
					Cancer cases	Non-cancer cases
1	254.990	254.990	254	1	0.46	0.58
2	272.946	274.224	273	10	0.30	0.48
3	278.140	278.140	278	1	0.37	0.26
4	277.450	281.740	279	6	0.50	0.36
5	359.761	364.819	362	6	0.40	0.63
6	363.235	367.040	365	4	0.45	0.29
7	364.596	373.480	369	5	0.51	0.35
8	415.070	415.070	415	1	0.51	0.37
9	416.590	416.590	416	1	0.45	0.34
10	428.810	431.635	430	4	0.39	0.28
11	470.605	484.855	477	4	0.61	0.72
12	491.060	504.800	497	2	0.37	0.52
13	500.556	504.031	502	11	0.44	0.26
14	515.725	519.245	517	2	0.32	0.18
15	538.877	542.457	540	6	0.44	0.28
16	586.070	586.070	586	1	0.54	0.64
17*	875.799	875.811	875	2	0.12	0.02
18	895.780	895.780	895	1	0.25	0.14
19*	935.025	935.025	935	10	0.18	0.04
20	952.510	952.510	952	1	0.16	0.04
21	980.950	980.950	980	1	0.26	0.14
22*	1106.310	1106.310	1106	1	0.18	0.06
23*	2009.127	2010.538	2009	24	0.19	0.31
24	2052.057	2052.987	2052	25	0.17	0.30
25	3108.330	3108.330	3108	1	0.30	0.23
26*	3370.392	3371.423	3370	29	0.22	0.09
27*	3471.836	3472.552	3472	41	0.23	0.11
28	3504.961	3505.606	3505	3	0.24	0.13
29*	4096.067	4098.015	4097	34	0.28	0.38
30	4117.572	4118.074	4117	20	0.29	0.38
31	4625.511	4629.172	4627	25	0.36	0.28
32	4853.500	4853.500	4853	1	0.41	0.31
33	5241.970	5241.970	5241	1	0.33	0.23
34*	6713.381	6714.365	6713	63	0.26	0.12
35*	6805.956	6806.306	6806	26	0.22	0.11
36*	6951.030	6951.685	6951	44	0.25	0.13
37*	7085.121	7085.540	7085	38	0.28	0.14
38	7119.308	7120.018	7119	32	0.15	0.08
39	9217.333	9220.654	9218	44	0.39	0.31



Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data

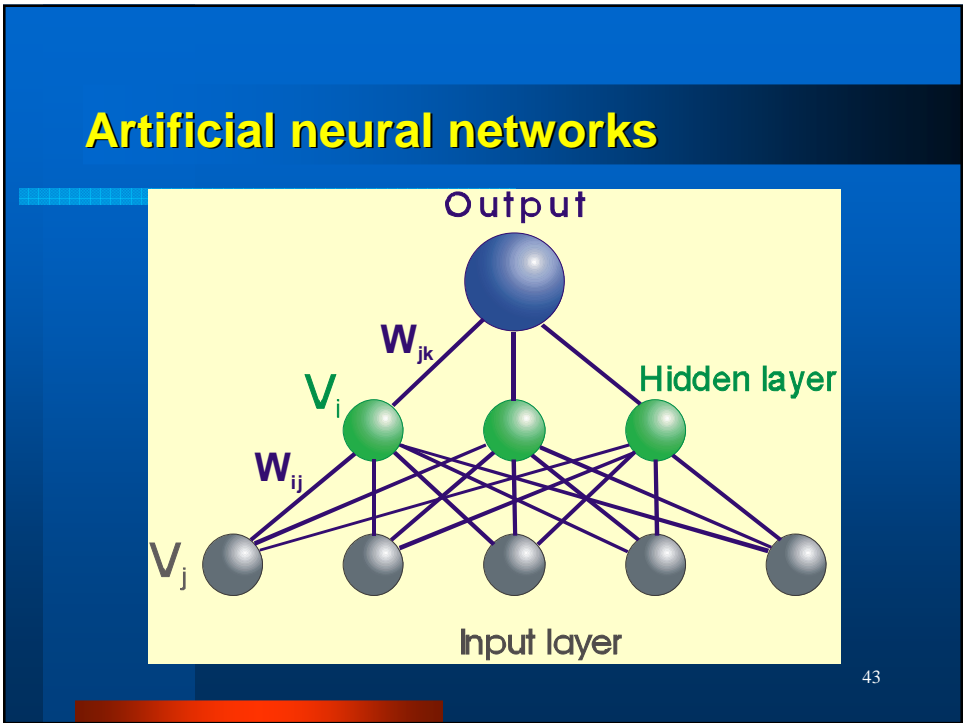


Individual classifiers

ANN, SVM, WV, KNN, CART, LR

Trained / calibrated (leave-one-out):
raw data
pattern data

42



Linear support vector machines

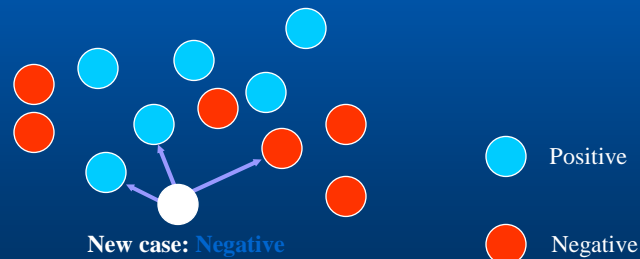
Find a maximum margin hyperplane in pattern space (Vapnik)

(P) $\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m z_i$
 $w\tilde{x} \geq b+1 - z_i, \tilde{x}$ positive
 $w\tilde{x} \leq b-1 + z_i, \tilde{x}$ negative
 $z_i \geq 0, i=1, \dots, m$

(D) $\min \left(\frac{1}{2} \sum_i y_i y_j \alpha_i \alpha_j x_i x_j - \sum_i \alpha_i \right)$
 $s.t. \sum_i y_i \alpha_i = 0, C \geq \alpha_i \geq 0, i=1, 2, \dots, m$

k-Nearest neighbors

- Training data : samples in normalized peptide space
- Prediction for test data: The dominant class of the k-nearest neighbors in Euclidean metric



45

Weighted voting

Pattern data:

- each pattern P is a voter
- weight = fraction of correctly classified cases by the pattern
- each test case: compute sum of weights of triggered positive patterns and negative patterns
- classify by highest weight

46

Logistic regression

- Dataset of two phenotypes (e.g., cancer vs. non-cancer)
- Transform into logit space
- Find phenotype predictor as a linear combination of data values in logit space

Inightful Miner

47

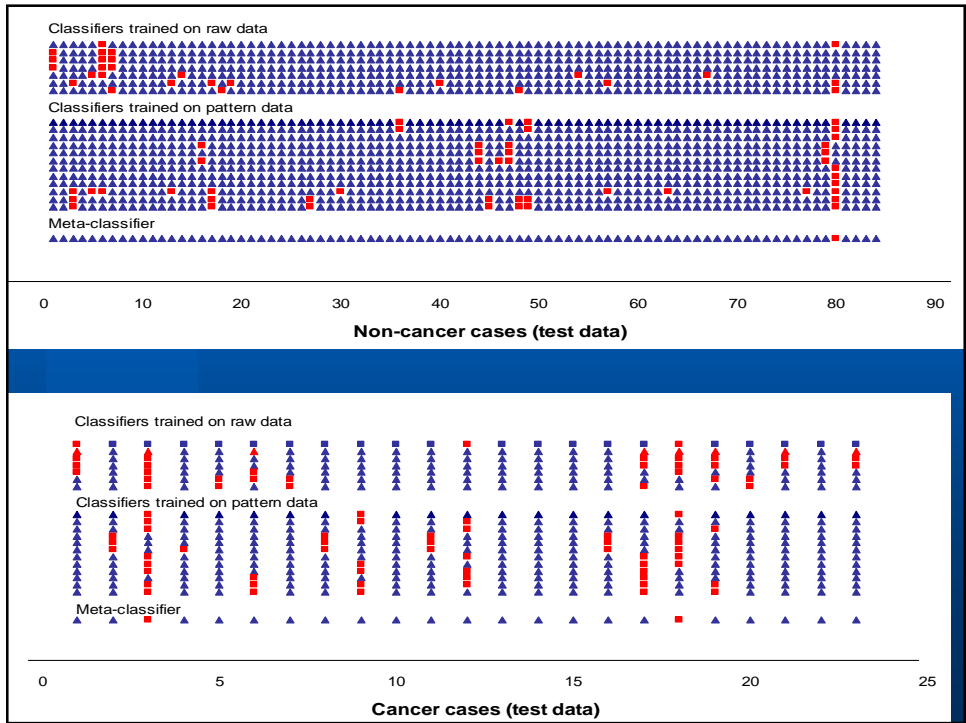
Decision trees / forests

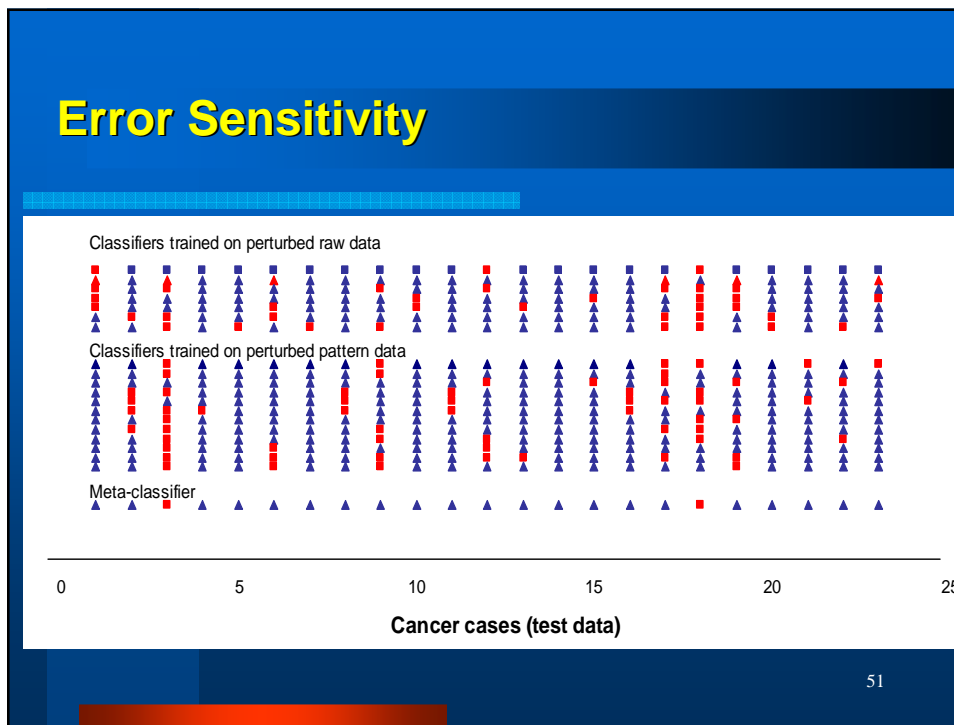
- Find rules in training data:
 - find root peptide which best classifies samples by phenotype
 - iterate on each branch to find two new peptides which best split each branch by phenotype
 - if necessary prune weak support nodes
- CART = Classification and Regression Trees (Entropy / Gini)
- Many trees = forest
- Metaclassifier based on several forests with 50 trees

48

Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data

Meta-classifier weight = $N \cdot (Sp-50) \cdot (Se-50)$						
Classifier		Weight	Training		Test	
			Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Trained on raw data	ANN	0.061	93.48	98.82	86.96	97.62
	SVM (linear kernel)	0.000	50.00	98.82	65.22	96.43
	SVM (quadratic kernel)	0.012	58.70	98.82	69.57	96.43
	KNN (k=5)	0.04	84.78	89.35	69.57	96.43
	LOGISTIC_REGRESSION	0.072	100.00	100.00	82.61	94.05
	CART (Entropy splitting index)	0.058	91.30	98.82	73.91	91.67
	CART (Gini splitting index)	0.053	86.96	100.00	78.26	94.05
Trained on pattern data	Weighted voting (comprehensive pattern collection size)	0.071	100.00	99.41	86.96	95.24
	Weighted voting (medium pattern collection size)	0.071	100.00	99.41	86.96	96.43
	Weighted voting (comprehensive pattern collection size)	0.069	100.00	97.63	86.96	98.81
	ANN (comprehensive pattern collection size)	0.042	80.43	98.22	78.26	95.24
	ANN (medium pattern collection size)	0.038	78.26	96.45	78.26	96.43
	ANN (comprehensive pattern collection size)	0.042	82.61	94.08	73.91	94.05
	SVM (linear kernel)	0.071	100.00	99.41	82.61	98.81
	SVM (quadratic kernel)	0.071	100.00	99.41	82.61	98.81
	KNN (k=5)	0.072	100.00	100.00	82.61	98.81
	LOGISTIC_REGRESSION	0.069	100.00	97.63	86.96	88.10
	CART (Entropy splitting index)	0.043	82.61	95.27	73.91	91.67
	CART (Gini splitting index)	0.043	82.61	95.27	78.26	91.67
	META-CLASSIFIER			100.00	99.41	91.30





Data Sets

- Shipp MA, et al, Nature Med.; 8(1), 68-74.
- Deisboeck T.S. et al., in press (preprint: <http://www.wkap.nl/prod/a/Stolovitzky.pdf>).
- Alexe G, Alexe S, Axelrod DE, Boros E, Weissmann D, Hammer PL, Artificial Intelligence in Medicine (in press)

53

Non-Hodgkin lymphomas

FL low grade non-Hodgkin lymphoma
t(14;18) translocation: over-expression of anti-apoptotic **bcl2**
(also in DLBCL, MALT)
25-60% FL cases evolve to DLBCL

DLBCL high grade non-Hodgkin lymphoma
t(14;18) translocation, over-expression of **bcl6**, small-cleaved cells
less than 2 years survival if untreated

Known biomarkers for FL transformation to DLBCL

- p53/MDM2 (Moller et al., 1999)
- p16 (Pyniol, 1998)
- p38MAPK (Elenitoba-Johnson et al., 2003)
- c-myc (Lossos et al., 2002)

54

Lymphoma datasets

Data: WI (Shipp et al., 2002) Affy HuGeneFL
 CU (DallaFavera Lab, Stolovitzky, 2005) Affy Hu95Av2

Samples:
 WI: 58 DLBCL & 19 FL
 CU: 14 DLBCL & 7 FL

Genes:
 WI: 6817
 CU: 12581

55

Robust support set

	Genesymbol	Shippetal	Genes@Vik	t-test	p53reg/leuc	Biological function
SEPP1	*	*	*	*		oxidative stress
TXNIP	*	*	*	*		metastases suppressor
DNASE1L3	*	*	*	*		apoptosis
CDH11	*	*	*	*		cell adhesion
LUCA15	*	*	*	*		apoptosis
GPR18	*	*	*	*		signaling pathway
CLU	*	*	*	*		apoptosis
LY9	*	*	*	*		cell adhesion
RHOH	*	*	*	*		T-cell differentiation
ELF2	*	*	*	*		transcription
CCNG2	*	*	*	*	*	cell cycle
CR2	*	*	*	*	*	complement activation
CDKN2D	*	*	*	*	*	cell cycle
PPP2R5C	*	*	*	*	*	signal transduction
G18	*	*	*	*	*	cell growth
LY86	*	*	*	*	*	apoptosis
ARPC1B	*	*	*	*	*	cell motility
MCM7	*	*	*	*	*	cell cycle
BCL2A1	*	*	*	*	*	apoptosis
IMPDH2	*	*	*	*	*	GMP biosynthesis
RRP45	*	*	*	*	*	immune response
STAT1	*	*	*	*	*	NF-kappaB cascade
DLG7	*	*	*	*	*	cell-cell signaling
SLC1A5	*	*	*	*	*	transport
TUBB2	*	*	*	*	*	microtubule movement
PSMA6	*	*	*	*	*	protein catabolism
PSMC1	*	*	*	*	*	spinocerebellar ataxia
LGALS3	*	*	*	*	*	sugar binding
CLTA	*	*	*	*	*	transport
PAGA	*	*	*	*	*	cell proliferation

Examples of FL and DLBCL patterns

Pattern	Gene Symbol				Prevalence (%)			
	GPR18	CLU	DLG7	MCM7	Training set		Test set	
					Pos	Neg	Pos	Neg
P1			>-1.13	>-0.62	97	0	91	23
P2	≤ 0.91			>-0.77	95	0	79	31
N1		>-0.26		≤ -0.55	0	100	3	54

WI training data:

Each DLBCL case satisfies at least one of the patterns P1 and P2

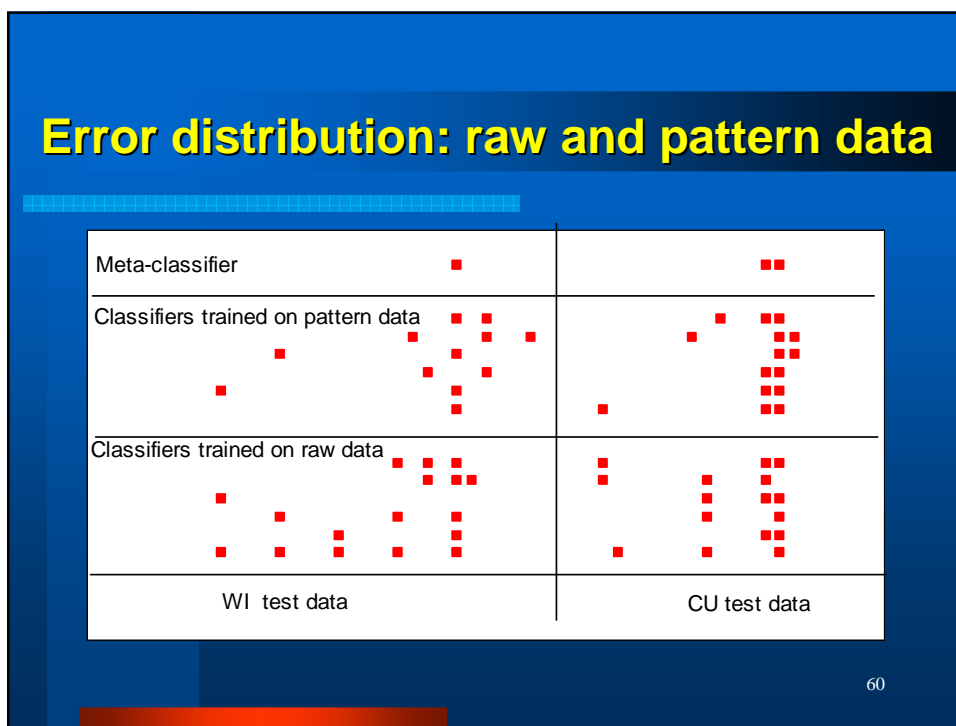
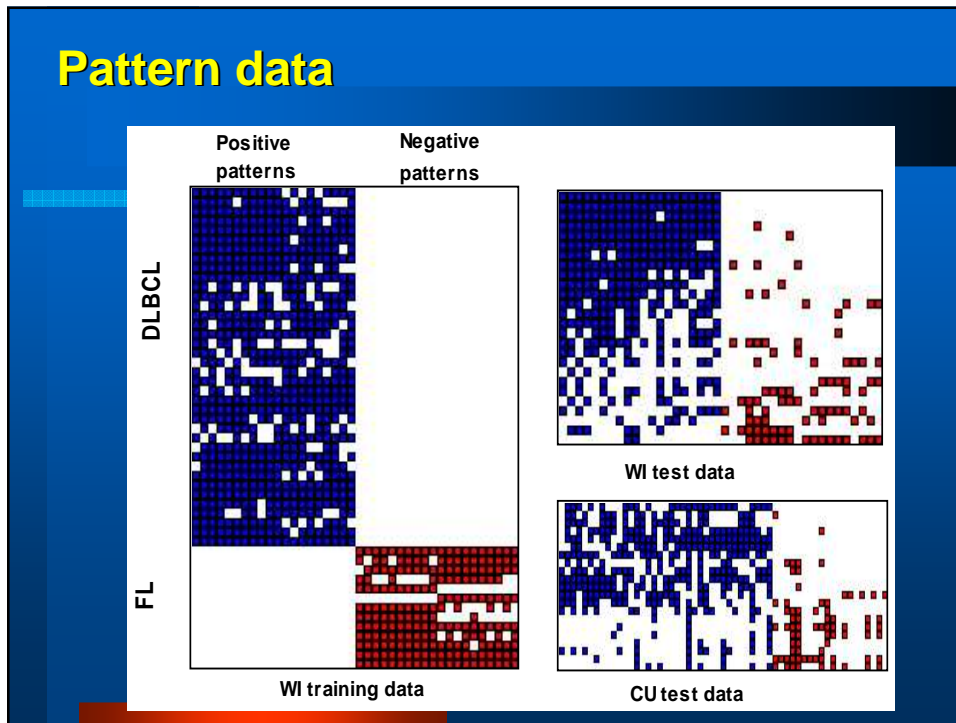
Each FL case satisfies the pattern N1 (and none of the patterns P1 and P2)

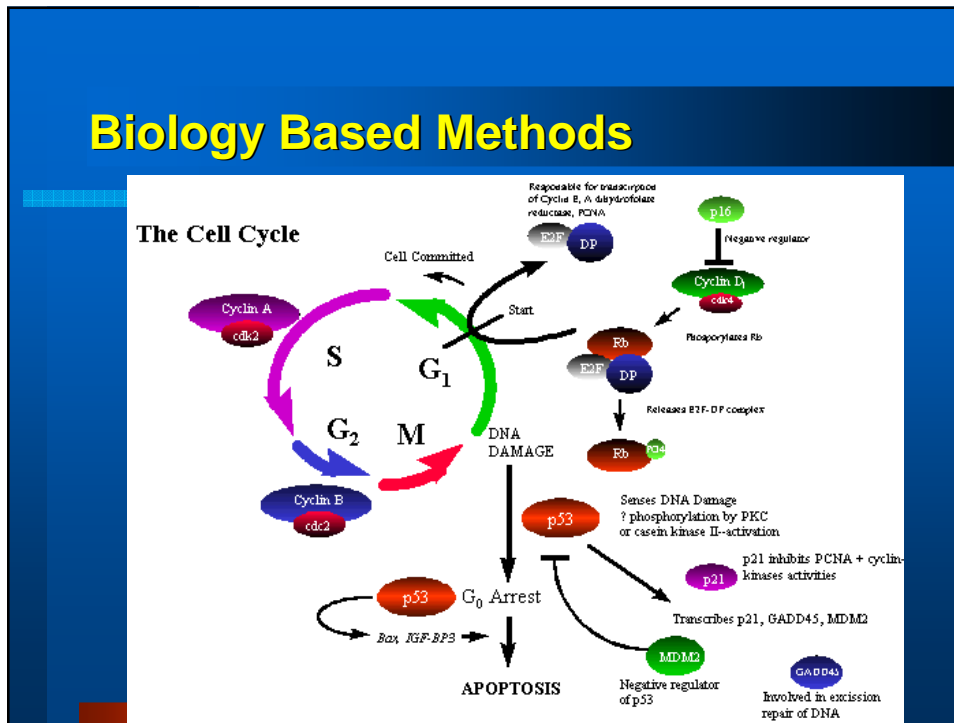
57

What are These Genes?

- GPR18 G protein-coupled receptor: signaling molecule, expressed in lymphoid/blood, spleen, testis/ targeted drugs Claritin
- CLU Clusterin (cancer gene) (Trouwagos et al, Cancer Res 2003): glycoprotein implicated in tumor formation, in vivo cancer progression, aging, Alzheimer. The only very low IR stress responsive gene, under negative control of p53 (down-regulated in cancer tissue - so far known in colon, esophageal etc)
- DLG7 (cancer gene): cell cycle regulated, over-expressed in human hepatocellular cells, may play a role in carcinogenesis (Tsou et al., Oncogene 2003)
- MCM7 (cancer gene) minichromosome maintenance deficient 7 / DNA replication factor, cell cycle control

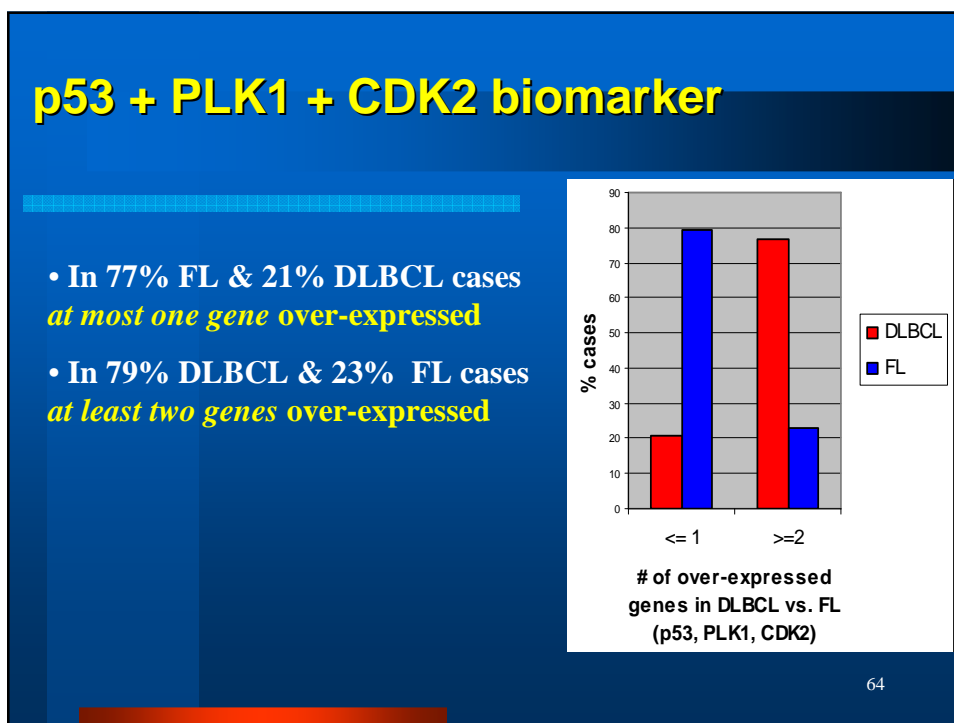
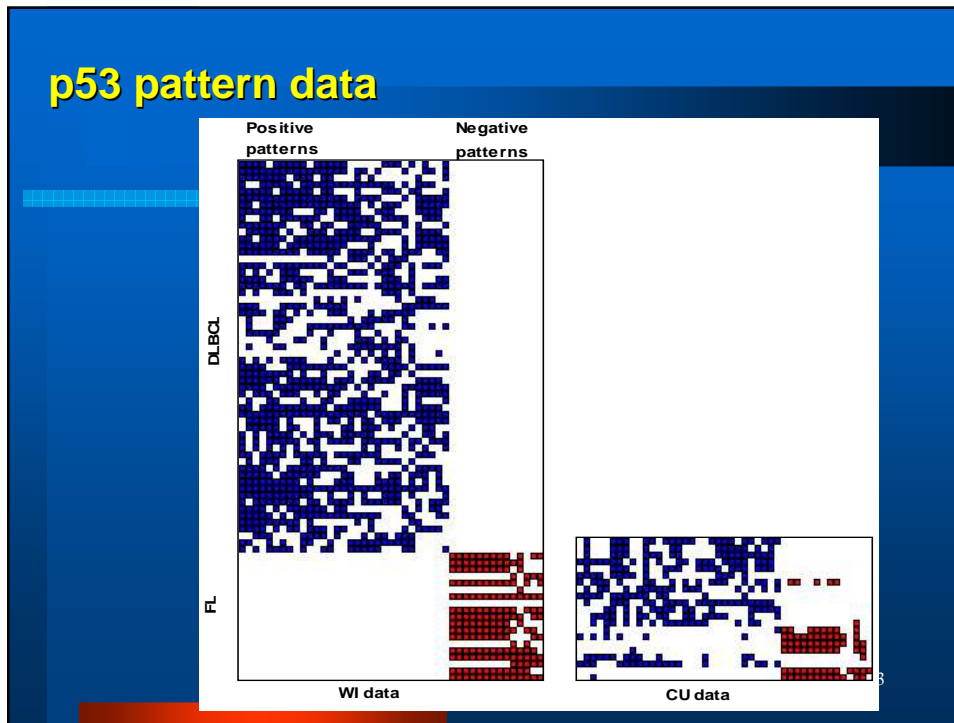
58





Support set of 90 p53 responsive genes

Gene symbol			
CCNB1	EPRS	PMAIP1	E2F3
MCM7	GSK3B	ACAA2	MDM4
BRCA1	COL6A1	E2F5*	AMPD2
BCL2A1	HRAS	POLA	RBBP4
PPP2R4	SERPING1	HMGB2	CCNG2*
EIF2S2	CCNA2	PSMB5	HARS
COMT	CCT6A	ACTA2	CASP6
IARS	PRKDC	INSR	RPS6KA1
MPI	CAD	SNRPA	GRP58
ALAS1	TNFRSF1B	G1P2	TP53
MRPL3	ZNF184*	IMPDH1	SMAD2
NCF2	ALDOA	MAP2K2	ATP5C1
AARS	KARS	TOP2A	TIMP3
KIF11	MAD2L1	CXCL1	THBS2
CDK4	GOT1	BAG1	MYCBP
ATP1B1	CDC25B	TOP1	DTR
CDC20	PSMA1	MAP4	TIMP3
PRIM1	KIAA0101	FDFT1	CBS
CDC2	PCNA	MTA1	CDKN2D*
TOP2A	TCF3	CDKN1A	RELA
CDK2	CYC1	HLAE*	
MYC	UPP1	PLK1	
CCNE1	TOPBP1	CDK7	



What are these Genes?

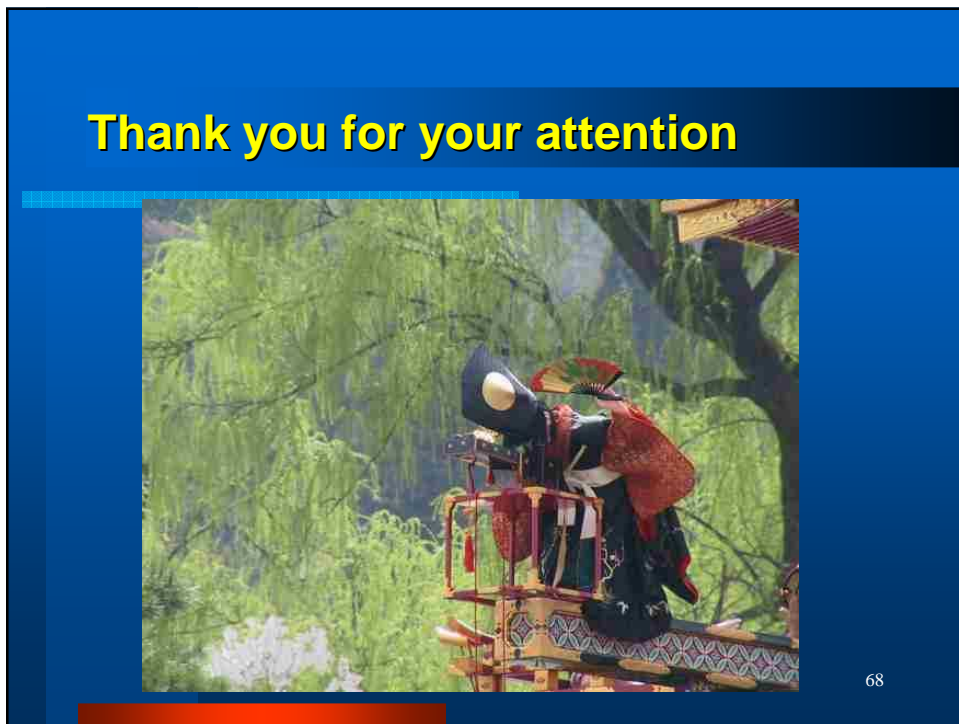
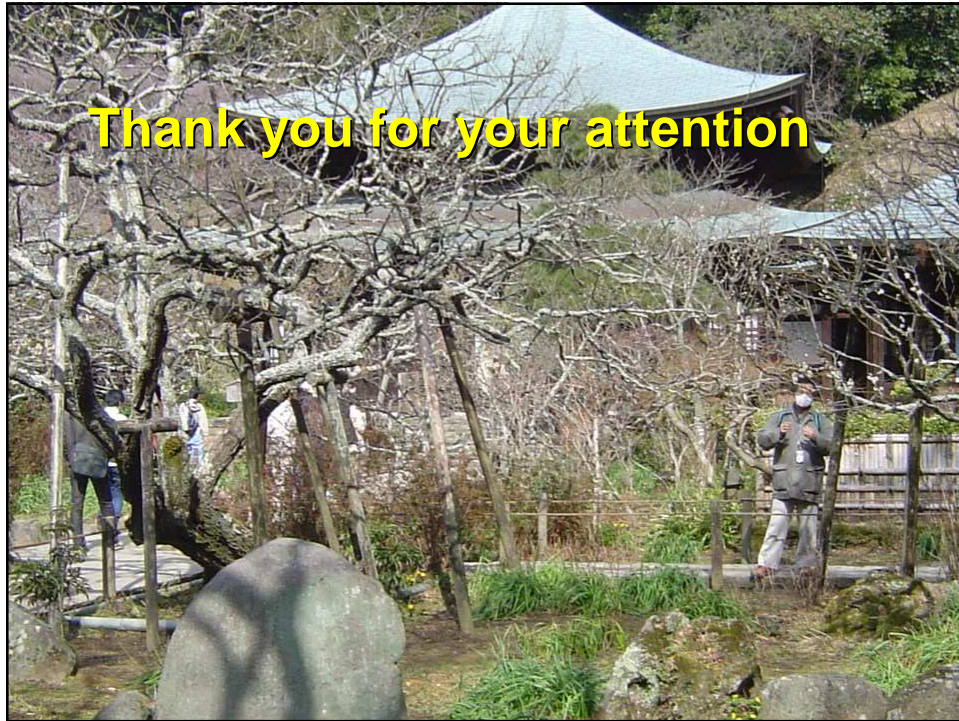
- Plk1 (stp13): polo-like kinase serine threonine protein kinase 13, M-phase specific
- cell transformation, neoplastic, drives quiescent cells into mitosis
- plk1 over-expression in various human tumors
- inhibits p53 (Ando et al., J. Biol. Chemistry, 2004)
- plk1 regulates cdc2/cyclinB through phosphorylation and activation of the cdc25C phosphatase, interacts with mcm7 (Tsvekov et al, J. Biol Chem, 2005)
- Takai et al., Oncogene, 2005: plk1 potential target for cancer therapy, new prognostic marker for cancer
- Mito et al, Leuk Lymph, 2005: plk1 more useful than Ki-67
- cdk2 (p33): cyclin -dependent kinase: G2/M transition of mitotic cell cycle, interacts with cyclins A, B3, D, E

65

Next Steps

- Relate protein biomarkers to pathways
- Extend meta-classifier approach to multi-phenotype classification
- Combine clustering with meta-classifiers
- **CLINICAL APPLICATION !!**

66

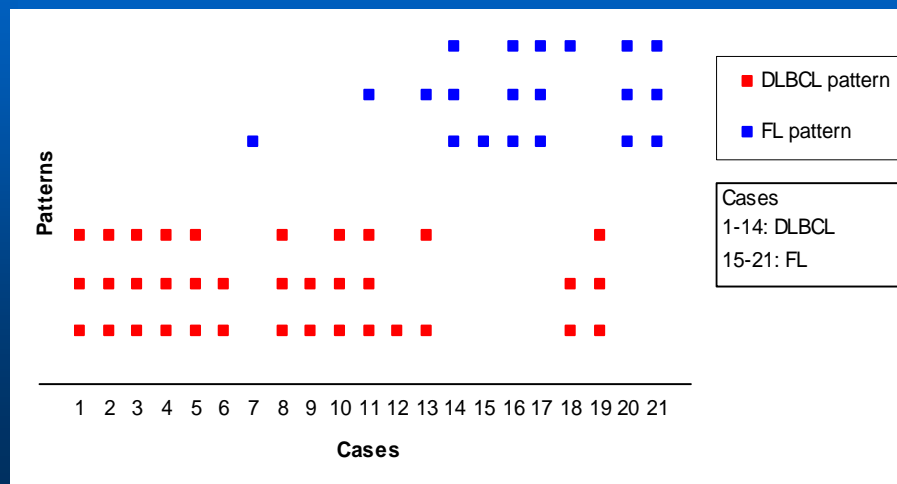


Prostate cancer biomarkers

Study	Chip type	Selected peaks (m/z)
Petricoin et al.	Hydrophobic C16	2092, 2367, 2582, 3080, 4819, 5439, 18220
Adam et al.	IMAC-Cu	4475, 5074, 5382, 7024, 7820, 8141, 9149, 9507, 9656
Qu et al.	IMAC-Cu	3963, 4080, 6542, 6797, 6949, 6991, 7024, 7885, 8067, 8356, 9656, 9720
Banez et al.	WCX2/IMAC-Cu	3972, 8226, 13952, 16087, 25167, 33270
Lehrer et al.	Hydrophobic H4	15200, 15900, 17500
Bhanot et al.	Hydrophobic C16	875, 935, 1106, 2009, 3370, 3472, 4097, 6713, 6806, 6951, 7085

69

Validation of p53 patterns on CU test data



Accurate Diagnosis and Staging: Statistical Techniques to Distinguish Cancer Types from Protein Mass Spec and Gene Array Data

Meta-classifier on multiple support sets

S1: support set of top 100 SNR correlated genes (Shipp et al., 2002)

S2: support set identified with Genes@Work (Stolovitzky, 2005)

S3: support set of top 100 genes w.r. t-test

S4: support set of 30 robust genes (current study)

S5: support set of p53 responsive genes (top 90 SNR correlated)

Missclassified samples	S1	S2	S3	S4	S5	Meta-classifier
DLBCL 15	1		1		1	
DLBCL 21			1		1	
DLBCL 26	1	1	1		1	1
DLBCL 27			1			
DLBCL 29	1	1	1	1	1	1
DLBCL 35	1				1	
DLBCL 36		1				
DLBCL 39	1	1		1	1	1
DLBCL 40					1	
DLBCL 46				1		
DLBCL 52				1		
DLBCL 54					1	
DLBCL 56			1		1	
DLCL 7						
DLCL 13			1		1	
DLCL 14		1	1		1	
FL-DM						
FL-GL	1	1	1	1	1	1
Error rate	6	6	9	5	12	4
Weights	0.26	0.30	0.24	0.30	0.00	

Meta-classifier performance, FL vs DLBCL

Classifier	Weight	Training			Test			
		Sensitivity (%)	Specificity (%)	Error rate (%)	Sensitivity (%)	Specificity (%)	Error rate (%)	
Trained on raw data	ANN	0.08	94.74	92.31	5.88	82.35	84.62	17.02
	SVM	0.08	97.37	92.31	3.92	97.06	76.92	8.51
	kNN	0.09	97.37	100.00	1.96	91.18	84.62	10.64
	WV	0.07	92.11	92.31	7.84	94.12	76.92	10.64
	C4.5	0.06	94.74	84.62	7.84	94.12	69.23	12.77
	LR	0.07	97.37	84.62	5.88	94.12	69.23	12.77
Trained on pattern data	ANN	0.10	100.00	100.00	0.00	97.06	76.92	8.51
	SVM	0.10	100.00	100.00	0.00	97.06	76.92	8.51
	kNN	0.10	100.00	100.00	0.00	100.00	69.23	8.51
	WV	0.10	100.00	100.00	0.00	97.06	76.92	8.51
	C4.5	0.10	100.00	100.00	0.00	91.18	76.92	12.77
	LR	0.05	100.00	76.92	5.88	100.00	61.54	10.64
Meta-classifier		100.00	100.00	0.00	100.00	76.92	6.38	

Slide Preparation

- DNA fragments (probes) chosen from expressed parts of Open Reading Frame (ORF) of genes.
- Fragments amplified by PCR
- Spotted on glass slide coated with polylysine causing DNA fixation by electrostatic interactions.
- DNA denatured to single strand

73

Target Preparation

- mRNA is extracted from two cell cultures which we want to compare
- mRNA transformed to cDNA by reverse transcription.
- cDNA from first culture labelled with green dye
- cDNA from second culture labelled with red dye.

74

Hybridization

- Green and red labelled cDNA mixed together to form “target”
- Placed on matrix of spotted single strand DNA
- Chip incubated overnight at 60 degrees.
- DNA strands in target/probe hybridize to form double stranded DNA.

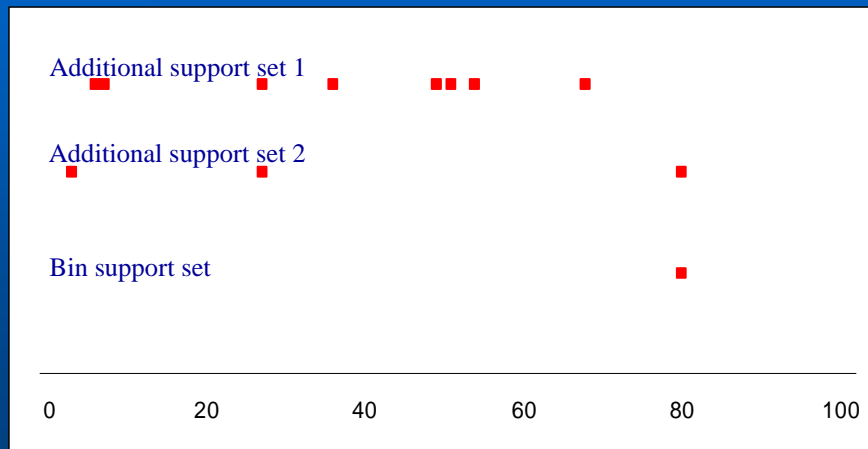
75

Slide Scanning

- Laser used to read each spot and the fluorescent emission is gathered through a photo-multiplier (PMT) coupled to a confocal microscope.
- Two images are obtained in grey scale (fluorescent intensities) and relative (green/red) intensity written to Excel file.
- Image obtained by making two images (green and red and superimposing them to go from green to yellow (equal green and red) to red.

76

Importance of Binning:



Importance of Binning:

