On utility of **gene set signatures** in **gene expression-based class prediction**

Minca Mramor, Marko Toplak, Gregor Leban,
Tomaž Curk, Janez Demšar and Blaž Zupan

Class Prediction & Background knowledge

Central to machine learning research

Inclusion of background knowledge:

- increase model stabilty
- increase predictve accuracy
- increase interpretability

Domain knowledge in systems biology

Sources:

- gene structure & function
- biological pathways
- protein interactions
- literature references

analysis of high-throughput data

(DNA microarrays, proteomics data, SNP analysis)

Gene expression microarrays

54 data instances (samples)

9698 attributes (genes)

RNF14	72.700	34.700	25.600	158.400	18.000	59.800	54.500	23.600	80.600	52.800	32.600	60.900	19.700	81.000	42.600	262.500
RNF10	43.600	86.000	118.800	130.600	208.900	172.600	69.800	127.700	42.400	50.500	48.700	92.300	145.000	?	142.000	179.600
RNF11	65.000	91.100	206.200	402.200	188.200	195.300	108.100	240.300	184.600	124.500	135.900	110.300	255.000	?	292.900	473.800
RNF13	107.700	98.200	55.400	71.400	54.700	88.700	116.700	32.200	80.000	52.800	42.300	51.400	31.500	49.700	17.700	75.100
NDP	11.300	16.600	25.100	30.900	5.600	2.700	12.500	11.100	18.500	18.500	12.600	40.000	8.800	?	18.900	9.400
PMM2	193.700	177.600	183.700	94.000	152.800	229.300	167.500	156.400	168.900	112.300	111.300	101.400	141.100	33.600	161.800	105.400
ASS1	68.200	249.000	281.900	330.900	271.200	333.300	279.600	74.100	124.700	370.300	282.900	336.700	292.400	357.600	133.800	200.000
NCBP1	36.700	22.900	12.800	9.500	21.700	8.900	28.000	9.700	47.300	23.600	28.300	44.400	17.500	22.300	10.900	7.100
NID2	?	5.200	14.000	7.400	9.500	12.900	8.900	24.100	7.900	3.700	12.500	16.400	18.700	?	9.200	7.400
CAMK1	70.800	90.100	?	?	104.600	1.900	?	?	?	?	?	?	?	?	?	?
SPR	19.100	?	0.100	22.900	24.700	?	27.500	?	3.800	50.300	6.100	0.600	?	0.100	16.100	43.600
CAMK4	6.800	17.100	12.033	39.600	9.050	23.800	18.200	18.450	18.000	7.800	10.900	12.100	15.450	24.850	27.500	10.133
ZC3H13	19.300	11.900	17.800	22.900	1.900	23.800	22.200	7.200	21.500	25.300	22.000	17.500	9.800	6.800	16.000	32.000
RNF115	190.900	131.300	307.900	145.200	209.200	205.200	321.000	212.400	179.000	225.900	178.500	354.100	271.800	202.100	163.800	203.700
ZC3H15	183.500	293.300	406.100	217.100	218.600	194.600	190.600	344.200	269.700	253.800	272.300	225.300	381.100	261.900	401.200	171.400
ZC3H14	28.900	30.800	21.950	13.000	19.200	10.000	26.800	36.300	38.300	15.900	43.900	27.900	3.900	30.900	15.350	20.900
SPN	8.800	?	32.400	26.100	?	28.000	?	?	4.200	16.000	29.300	?	?	63.700	7.400	21.400
AB028973	?	?	8.300	?	2.200	?	2.300	?	3.800	17.400	?	?	?	6.600	?	6.700
GRIN1	7.100	30.900	16.600	19.500	79.200	30.400	69.600	?	10.100	69.600	?	0.300	123.750	73.900	0.900	23.350
D26155	153.000	116.900	75.700	42.500	120.100	60.300	124.400	24.200	111.400	118.000	105.800	107.200	30.100	106.700	103.000	87.600
M95929	264.600	108.100	95.900	106.800	89.000	284.000	208.300	93.500	274.900	287.700	98.400	59.600	374.300	138.200	132.500	54.000
DHX8	97.000	45.300	86.200	61.900	94.300	60.600	89.900	70.100	61.600	38.600	40.100	57.200	82.800	?	105.600	63.700
DHX9	169.200	371.650	368.950	237.250	280.400	194.900	231.950	312.950	337.850	248.250	325.500	263.000	316.000	319.600	348.450	193.300
TCOF1	274.000	285.000	290.000	329.050	234.450	277.100	277.950	281.400	277.050	270.450	220.100	319.400	238.800	488.750	283.900	278.500
W22110	200.200	157.200	191.600	212.900	130.800	180.400	147.600	140.400	125.700	226.900	122.200	269.100	179.400	220.400	236.700	179.000
OR2H4P	41.400	47.800	?	35.200	21.300	23.300	11.200	18.000	49.700	27.200	29.700	10.800	43.900	43.800	33.100	40.700
XPC	168.800	28.300	55.650	54.250	75.200	75.150	103.100	70.450	56.950	79.050	94.700	82.600	56.000	126.900	59.850	70.950
SP1	25.600	13.700	26.100	15.000	24.500	15.200	8.100	17.600	14.300	16.700	19.900	12.000	17.700	17.200	23.000	7.300
XPA	31.100	22.167	47.800	55.950	46.200	67.150	70.100	55.200	35.867	34.467	13.400	96.450	47.867	48.500	55.833	33.433
PNMT	501.900	502.500	467.300	588.100	427.800	467.800	314.300	364.000	388.200	551.400	409.300	327.400	629.300	951.800	638.100	427.100
outcome	remission	remission	remission	remission	remission	remission	remission	remission	relapse	relapse	relapse	relapse	relapse	relapse	relapse	relapse

GDS1059: Analysis of mononuclear cells from 54 chemotherapy treated patients less than 15 years of age with acute myeloid leukemia (AML). Results identify expression patterns associated with complete remission and relapse with resistant disease.

Gene sets as background knowledge

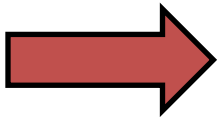
GENE SETS – groups of related genes

(gene structure, molecular function, biological pathways)

1_AND_2_METHYLNAPHTHALENE_DEGRADATION	ADH1A	ADH1A ///	ADH1B	ADH1C	ADH4	ADH6	ADH7	ADHFE1	
41BBPATHWAY	ATF2	CHUK	IFNG	IKBKB	IL2	IL4	JUN	MAP3K1	MAP3K5
ACE2PATHWAY	ACE2	AGT	AGTR1	AGTR2	CMA1	COL4A1	COL4A2	COL4A3	COL4A4
ACE_INHIBITOR_PATHWAY_PHARMGKB	ACE	AGT	AGTR1	AGTR2	BDKRB2	KNG1	NOS3	REN	
ACETAMINOPHENPATHWAY	CYP1A2	CYP2E1	CYP3A	NR1I3	PTGS1	PTGS2			
ACETYLCHOLINE_SYNTHESIS	ACHE	CHAT	CHKA	PCYT1A	PDHA1	PDHA2	PEMT	SLC18A3	
ACHPATHWAY	AKT1	BAD	CHRNB1	CHRNG	FOXO3A	MUSK	PIK3CA	PIK3R1	PTK2

Explorative analysis:

- functional annotations (gene ontology)
- enrichment analysis



Gains in:

- stability & robustness
- insight into the investigated problem

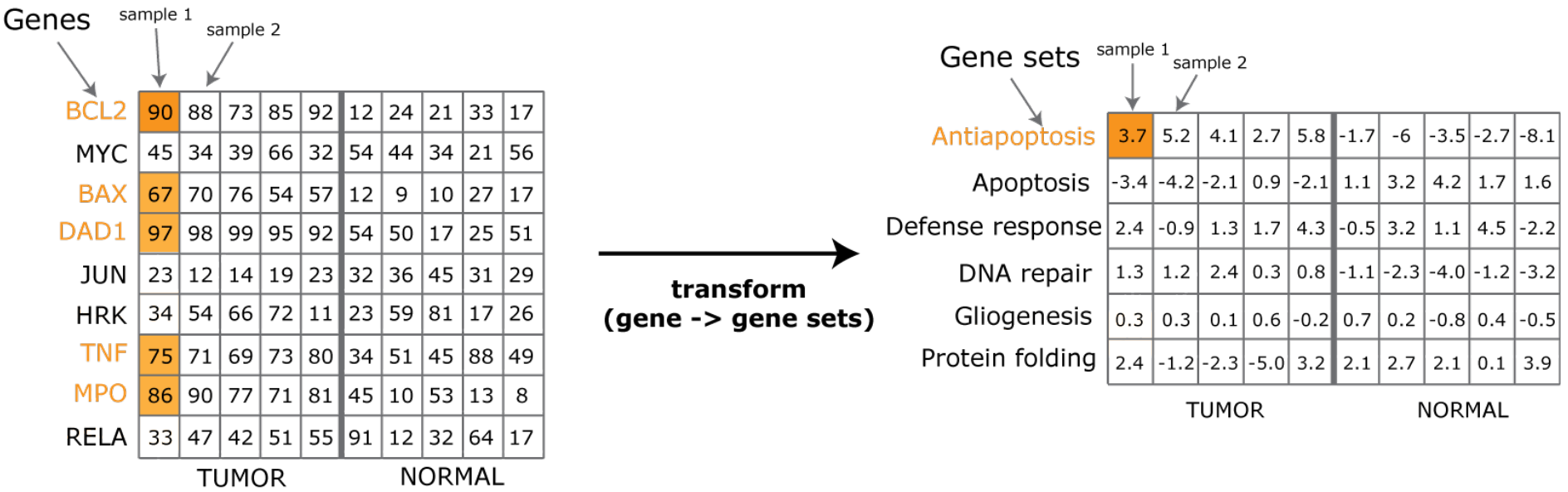
Goal

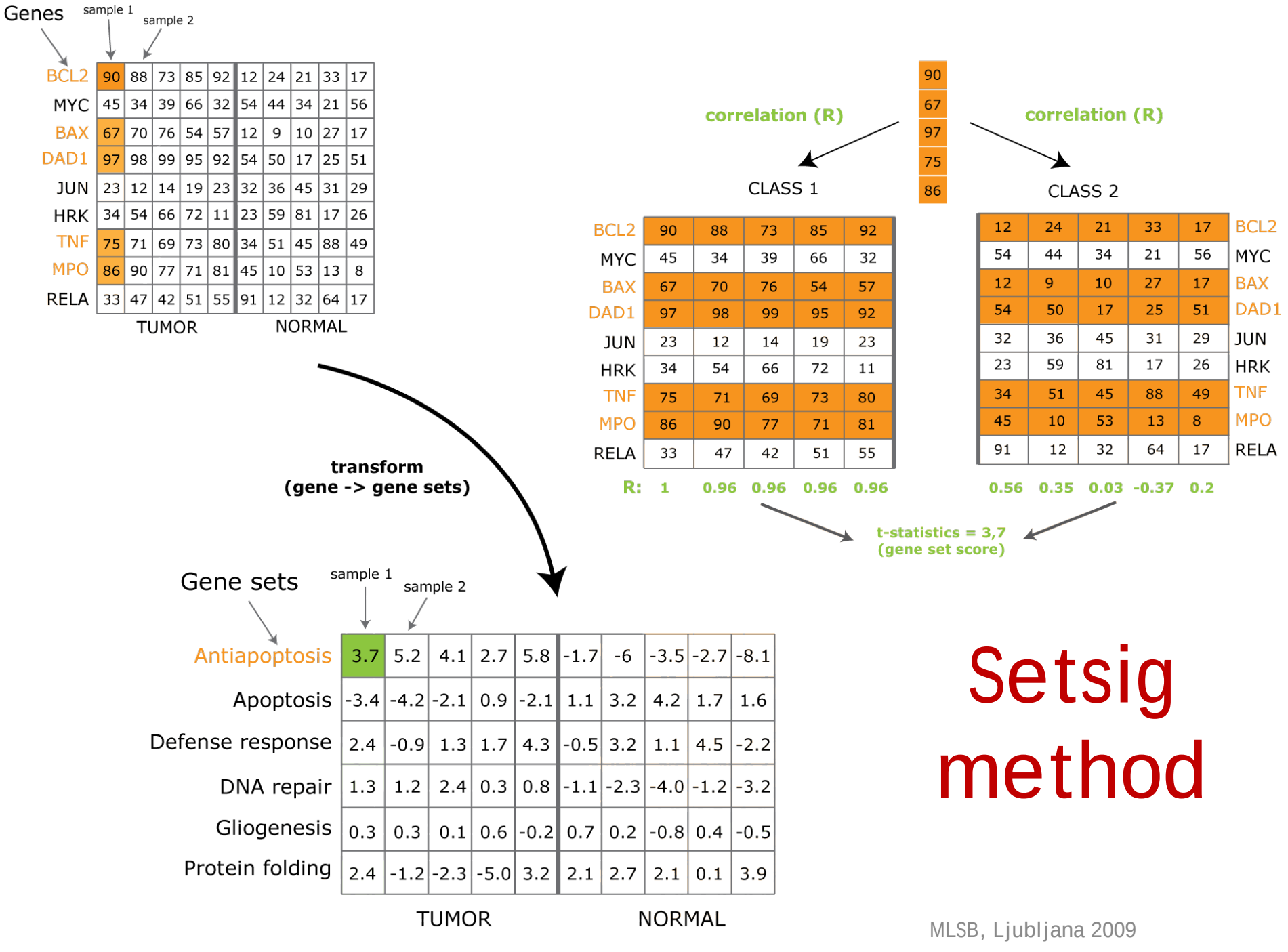
Use gene sets in inference of class prediction models – Setsig method

Test the gene-set based models:

- across a larger set of data sets
- across different transformation methods
- comparisson with gene based models

Gene set transformation




Genes
sample 1
sample 2
BCL2 90 88 73 85 92 12 24 21 33 17
MYC 45 34 39 66 32 54 44 34 21 56
BAX 67 70 76 54 57 12 9 10 27 17
DAD1 97 98 99 95 92 54 50 17 25 51
JUN 23 12 14 19 23 32 36 45 31 29
HRK 34 54 66 72 11 23 59 81 17 26
TNF 75 71 69 73 80 34 51 45 88 49
MPO 86 90 77 71 81 45 10 53 13 8
RELA 33 47 42 51 55 91 12 32 64 17
TUMOR
NORMAL
90
67
97
75
86
correlation (R)
correlation (R)
CLASS 1
CLASS 2
BCL2 90 88 73 85 92
MYC 45 34 39 66 32
BAX 67 70 76 54 57
DAD1 97 98 99 95 92
JUN 23 12 14 19 23
HRK 34 54 66 72 11
TNF 75 71 69 73 80
MPO 86 90 77 71 81
RELA 33 47 42 51 55
R: 1 0.96 0.96 0.96 0.96
12 24 21 33 17 BCL2
54 44 34 21 56 MYC
12 9 10 27 17 BAX
54 50 17 25 51 DAD1
32 36 45 31 29 JUN
23 59 81 17 26 HRK
34 51 45 88 49 TNF
45 10 53 13 8 MPO
91 12 32 64 17 RELA
0.56 0.35 0.03 -0.37 0.2
t-statistics = 3,7
(gene set score)
transform
(gene -> gene sets)
Gene sets
sample 1
sample 2
Antiapoptosis 3.7 5.2 4.1 2.7 5.8 -1.7 -6 -3.5 -2.7 -8.1
Apoptosis -3.4 -4.2 -2.1 0.9 -2.1 1.1 3.2 4.2 1.7 1.6
Defense response 2.4 -0.9 1.3 1.7 4.3 -0.5 3.2 1.1 4.5 -2.2
DNA repair 1.3 1.2 2.4 0.3 0.8 -1.1 -2.3 -4.0 -1.2 -3.2
Gliogenesis 0.3 0.3 0.1 0.6 -0.2 0.7 0.2 -0.8 0.4 -0.5
Protein folding 2.4 -1.2 -2.3 -5.0 3.2 2.1 2.7 2.1 0.1 3.9
TUMOR
NORMAL
Setsig method

Related work

Unsupervised approaches:

- Mean* and Median* (Guo *et al.*, 2005)
- Principal component analysis* (Liu *et al.*, 2007) ,
- Singular value decomposition (Tomfohr *et al.*, 2005 and Bild *et al., 2006)*

Supervised approaches:

- Partial least squares (Liu *et al.*, 2007)
- PCA with relevant gene selection (Chen *et al.*, 2008)
- Activity scores based on condition-responsive genes* (Lee *et al.*, 2009)
- Gene Set Analysis (Efron and Tibshirani, 2007)
- ASSESS* (Edelman *et al.*, 2006)

Experimental design

Data sets

30 data sets from Gene Expression Omnibus (GEO):
- 2 diagnostic classes
- at least 20 samples
- 20 - 187 samples
- 932 – 34700 genes

preprocessing:
$\mu = 0$, $\sigma^2 = 1$

Gene sets

Molecular signature data base (Subramanian *et al.*, 2005)

biological knowledge collections:
C2 - canonical pathways (639)
C5 - gene ontology (1221)

gene set size:
5 < genes < 200

Experimental design
predictive models

original data - GENES

transformed data - GENE SETS:

- Setsig
- Mean
- Median
- PCA
- CORGs
- ASSESS

learners:

- support vector machines
- k-nearest neighbors
- logistic regression

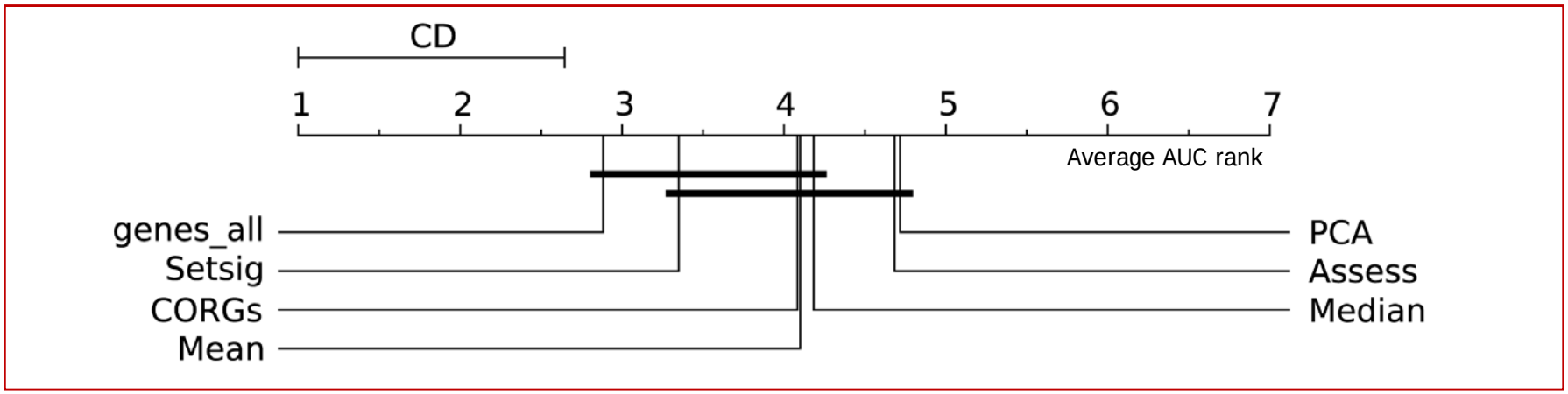
leave-one-out validation

area under ROC (AUC)

Results

Critical distance graph (Demšar, 2006)

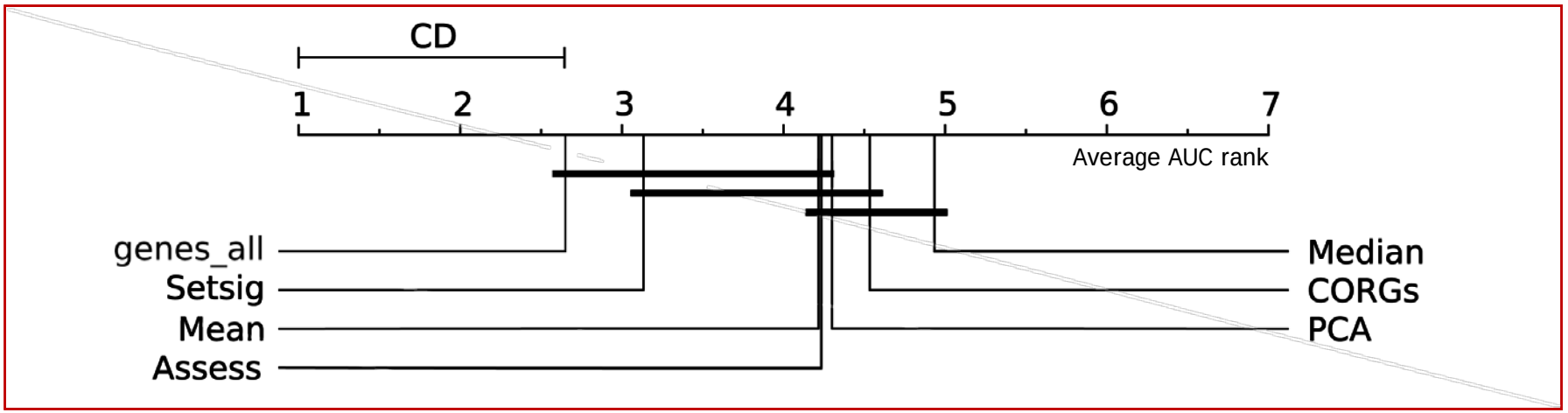
Support vector machines:



Results

Critical distance graph (Demšar, 2006)

Logistic regression:



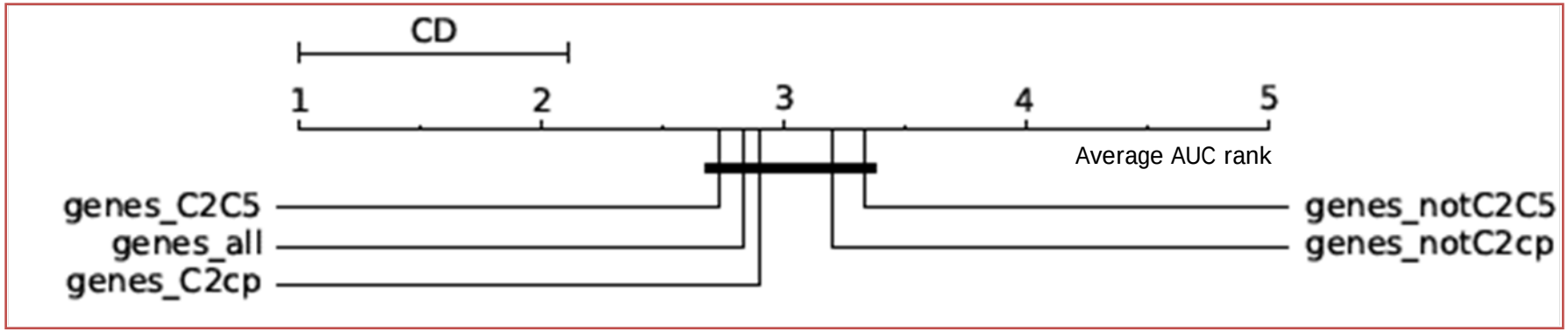
Surprising? Yes.

1. Gene sets in explorative data analysis – increase stability and robustness of results

2. Contradict current reports:
   - Edelman *et al*, 2006 (ASSESS, 6 data sets)
   - Lee *et al*, 2009 (CORGs, 7 data sets)
   - Efron & Tibshirani, 2007 (GSA, 1 data set)

Why worse performance?

1. Do gene sets include class-informative genes?



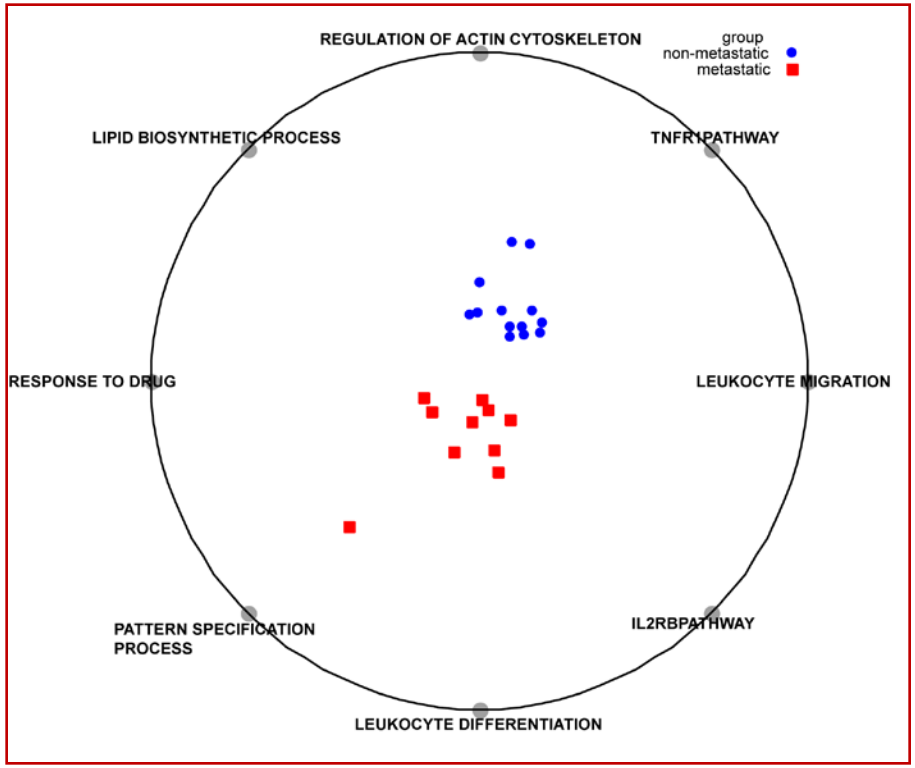
Why worse performance?

2. Gene set signature transformation loses information.
3. Number of samples is too low to estimate gene set scores.
4. Gene sets and pathways are not specific enough to distinguish between different cancer types.

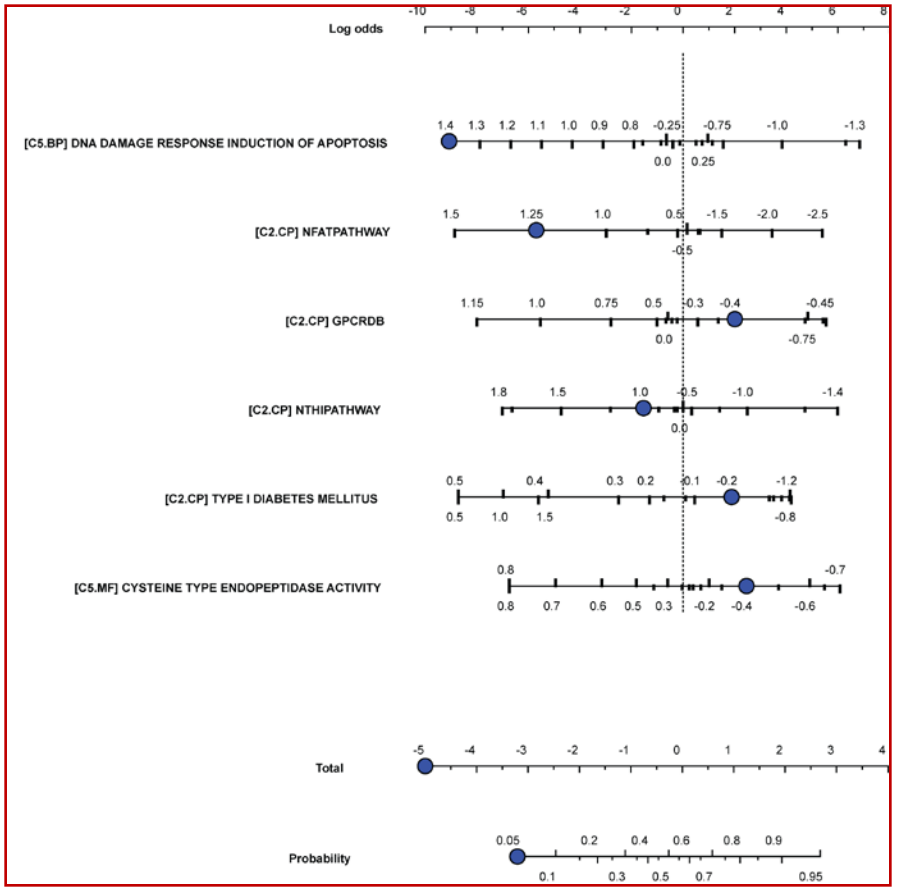
Gene set based class prediction models

- worse/similar performance (Setsig)
- additional insight

VizRank (Mramor *et al.*, 2007



Naive Bayes normogram (Možina *et al.*, 2004)



Thanks to…

- Marko Toplak
- Janez Demšar
- Tomaž Curk
- Gregor Leban
- Blaž Zupan

- Gregor Rot
- Lan Umek
- Aleš Erjavec
- Miha Štajdohar
- Lan Žagar
- Črt Gorup
- Ivan Bratko