Predictive Medicine by Cytomics

G.Valet

Max-Planck-Institut für Biochemie, Martinsried, Germany

Goal

wanted: *therapy related individualized disease course predictions*

typical: *group oriented predictions of disease prognosis like:*

- cytogenetic abnormalities and CD antigens stratify AML patients into high and low risk groups
- *problem:* no individualized predictions are typically obtained from group oriented data evaluation (e.g. Kaplan-Meier statistics)

VAL

Data Evaluation Strategies

predictions: individualized future

- data pattern classification (data sieving)
  following exhaustive information extraction
  from measured primary data

prognosis: patient group oriented future

- statistical (Kaplan-Meier, Bayes)
- cluster
- multivariate
- principal component
- fuzzy logic
- neural network

Future of Patient Groups

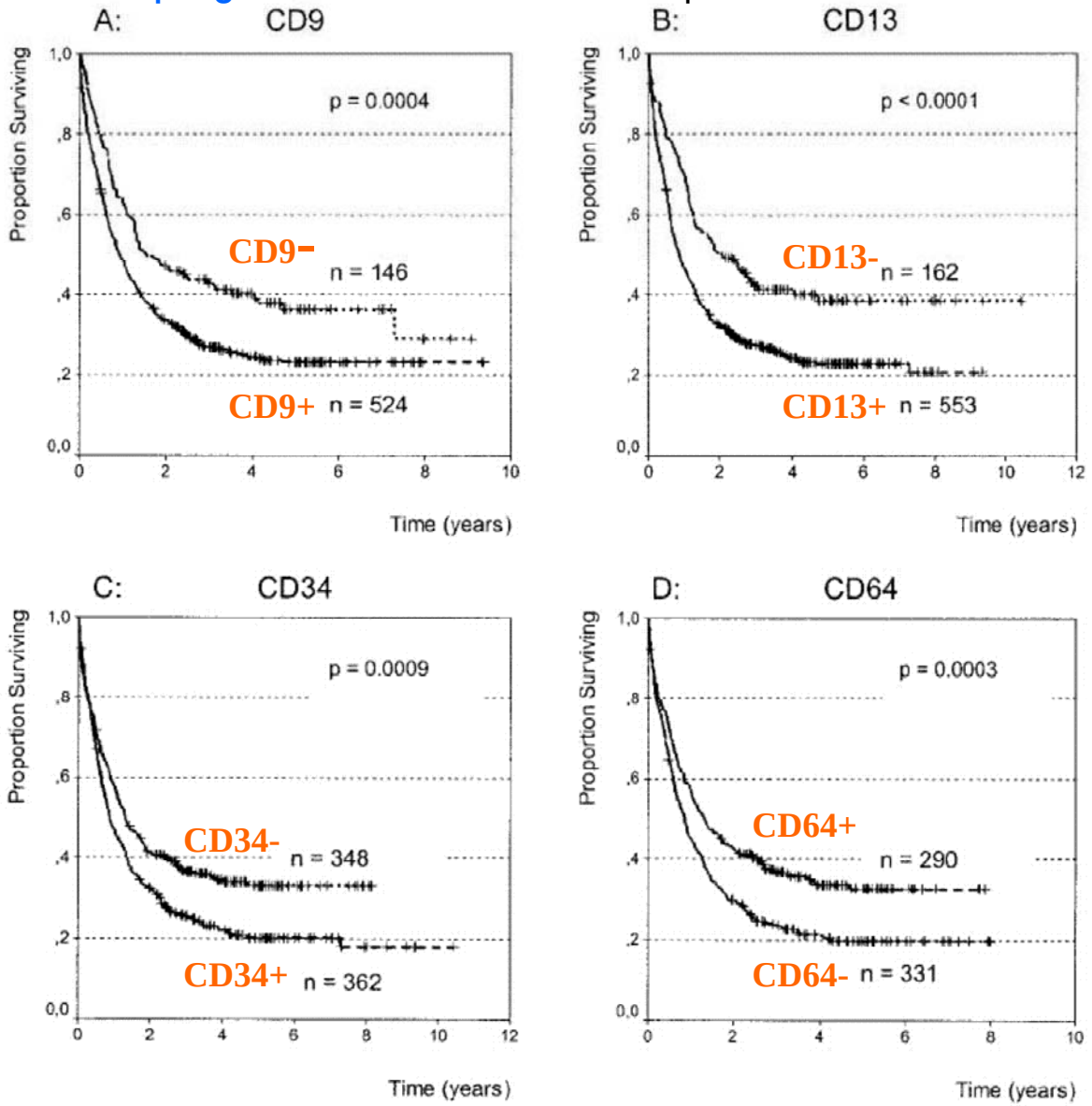
Kaplan-Meier Statistics

(patient stratification)

Kaplan-Meier Analysis AML

Repp et al. Cytometry 53B:11-19 (2003)

prognosis = statistical future -> patient stratification



Individualized Future

Algorithmic triple matrix data pattern analysis (data sieving)

Individualized Prediction by Data Pattern Classification

classification goals:

- accuracy for correct predictions and diagnosis
- multiplicity to account for the many combinatorial possibilities of genotype and exposure influences on the molecular cell phenotype

a.) disease classification masks
(schematic 10 param.)

	disease course prediction
0 0 0 0 0 0 0 0 0 0	stationary
+ + + + + + + + + +	improvement
− − − − − − − − − −	deterioration

b.) ***accuracy*** at low risk of random coincidence between patient and disease classification masks:

0.0017% $(1/3^{10})$ probability for random coincidence with 10 parameter masks and

0.046% $(1/3^{7})$ for random coincidence with 7 parameter masks

c.) patient classification masks
(some examples)

0 − 0 0 0 + 0 0 − 0	stationary
0 0 + 0 0 + 0 − 0 0	stationary
− 0 0 0 + 0 − 0 0 0	stationary
+ + + 0 0 0 0 0 0 0	stationary
− 0 − 0 0 0 0 0 − 0	stationary
+ 0 + − 0 + + + 0 0	improvement
0 − − 0 + + − − − 0	deterioration
. . .	...

d.) ***multiplicity*** of patient classification masks at correct prediction in case of partial coincidences like 7 out of 10 parameters:

$\frac{10! * 2^3}{7! * 3!}$ = 960 possible patient classification masks for each disease state as potential result of genotype and exposure influences on molecular cell phenotype or other parameters

patient classification:
highest positional coincidence with anyone of the disease classification masks

VAL

CLASSIF1 Classification

Triple Matrix Transformation of Parameter Values

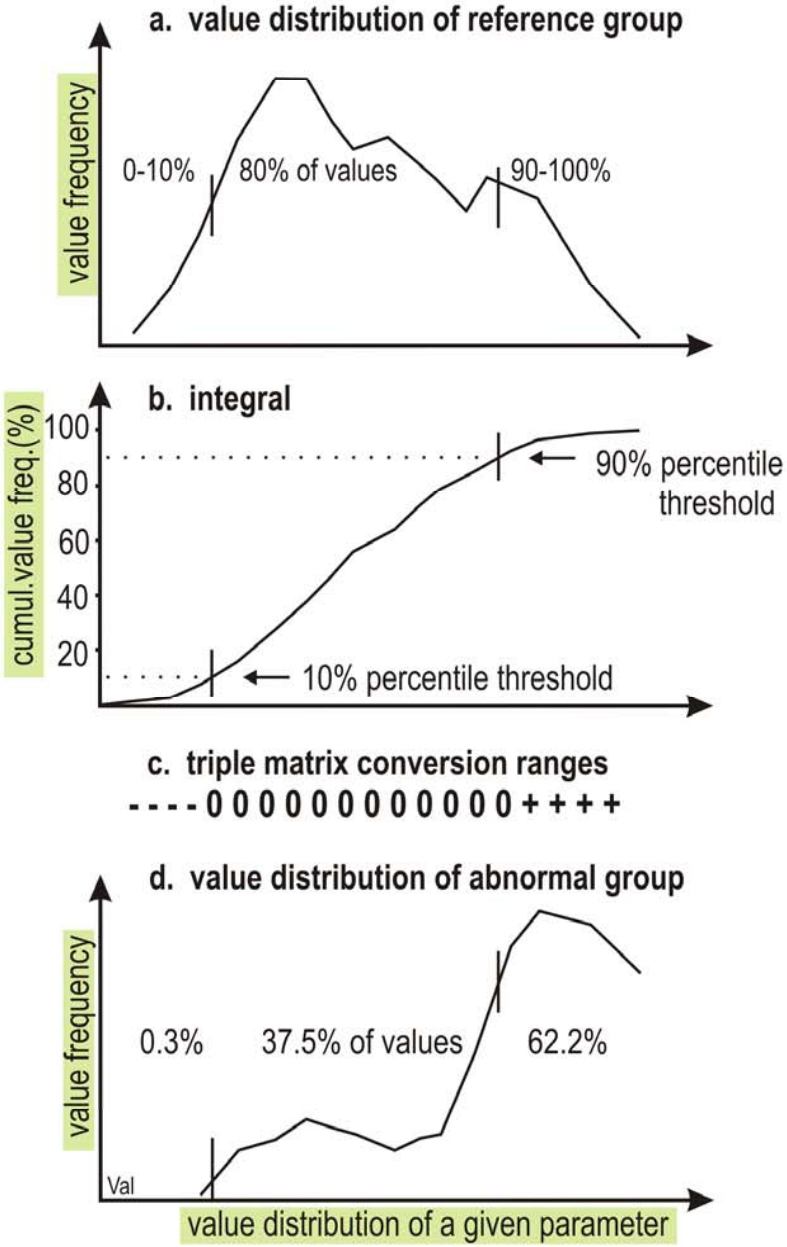
using percentile thresholds of references patients

patient	numeric database columns 1	2	3	triple matrix database columns 1	2	3
1	40.1	4.02	18.38	0	+	0
2	39.5	4.20	5.25	-	+	-
3	41.2	3.46	35.53	0	-	+
4	53.2	3.78	30.72	+	0	+
5	41.3	3.95	27.46	0	0	0
6	79.3	3.80	16.29	+	0	0
7	48.0	3.98	28.33	0	0	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.

http://www.biochem.mpg.de/valet/classif1.html

VAL

Principle of Triple Matrix Data Pattern Classification



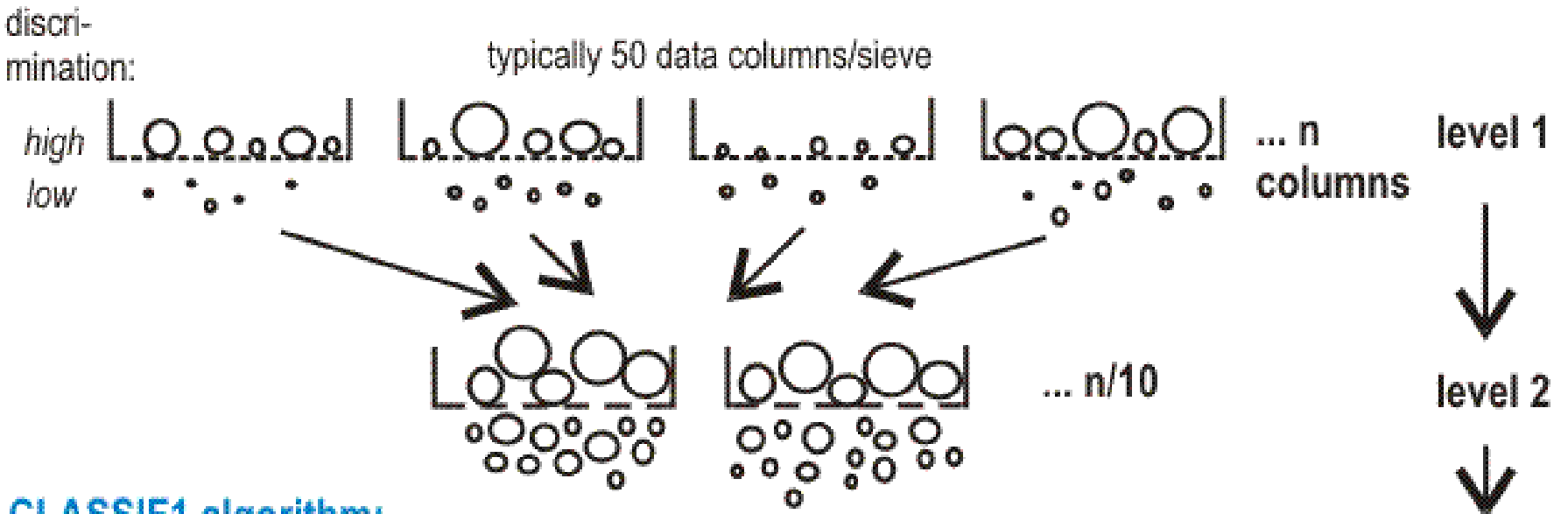
- *numeric* parameter values of for example 10 normal reference individuals will transform on average into eight '0', one ' -' and one '+' *triple matrix characters* using 10% and 90 % percentile thresholds

- similary, parameter values of 10 *abnormal* individuals will transform at the same thresholds on average into four '0' and six discriminatory '+' characters according to the value distribution graph of the lower panel

- definition of a *disease classification mask* containing the characters '0' and '-' for normals and '+' for abnormals provides the correct classification of 90% *normals* as normals and of 60% *abnormals* as abnormals

- disease classification masks with typically between 5 and 25 classification parameters increase the discrimination potential and allow accurate diagnosis and predictions on therapy dependent future disease course in individual patients from patient classification masks with accuracies >95% or >99%

Information Enrichment by Data Sieving



CLASSIF1 algorithm:

- sequential data partitioning into 50 parameter (data columns) wide data sieves
- select 5 most discriminatory parameters in each sieve by percentile analysis
- combine selected parameters, repartition, sieve & repeat procedure until 50 columns or less remain
- classify remaining data columns for most discriminatory triple matrix data pattern
- algorithm characteristics: surface data mining (no models), unsupervised exhaustive knowledge extraction to access unknown knowledge spaces

GVAL

Classication Matrix

goal: optimize diagonal sum

clinical outcome	CLASSIF1 prediction (%)		
	stationary	improve	deteriorate
stationary	100.0	0.0	0.0
improve	0.0	100.0	0.0
deteriorate	0.0	0.0	100.0
neg/pos predval	100.0	100.0	100.0

procedure:

- classify all parameters
- cross validate: sequential temporary removal of single parameters
- reclassification
- retain increase or decrease of diagonal sum
- reinsert parameter
- finally: remove parameters decreasing diagonal sum

val

Individualized Pretherapeutic Identification of High Risk Acute Myeloid Leukemia (AML) Patients

Valet G.1), Repp R.2), Link H.3), Ehninger G.4), Gramatzki M.2) and SHG-AML'96 Group

1) Max-Planck-Institut für Biochemie, Martinsried, 2) Med.Klinik III Univ. Erlangen, 3) Klinikum Kaiserslautern, 4) Med.Klinik & Poliklinik I Univ.Dresden

Cytometry (2003) 53B:4-10

SHG-96 Multicenter AML Trial: Parameters

clinical parameters	cytogenetics		CD-antigens		
pat_age	cytogen	t1_7	CD1	CD34	GlyA
pat_sex	abnorm	t3_3	CD2	CD36	MPO7
zz_mio	del5	t6_9	CD4	CD38	TH126
leuk	del7	t9_11	CD7	CD41	TC12
LDH	dely	t9_22	CD9	CD42	TC25
flt3	mono	t15_17	CD10	CD45	TDT
clinic_nr	inv3	abn12	CD11b	CD56	cyCD3
patient_nr	tri8	abn11	CD14	CD58	HLA-DR
OS	tri11	aberr_m	CD15	CD61	
	tri13	inv16	CD19	CD64	(n=36)
(n=6)	tri21	other_ab	CD20	CD75	
	tri22		CD24	CD95	
	t8_21	(n=25)	CD32	CD96	
	t8_16		CD33	CD117	

= survival indicator

classification parameters: n=67

= not used for > 5year survival (less than 10% available values)

val

Pretherapeutic Identification of High Risk AML Patients

learning set

clinical outcome	pat. (n)	CLASSIF1 prediction (%)	
		>5y surv	non surv
>5y surv	24	100.0	0.0
non surv	269	50.2	49.8
neg/pos predval		15.1	100.0

unknown embedded test set

non surv	165	8.5	91.5

20-75% percentiles, DJ/REPP32.BI4

VAL

AML: Predictive Parameters

5 year survival

#	parameter	S	NS
1	zz_mio	-0	+
2	pat.age	-0	+
3	% CD2	-0	+
4	% CD4	-0	+
5	% CD13	-0	+
6	% CD36	-0	+
7	% CD45	-0	+

selected parameters: 7/67 → 10.5%

VAL

Conclusions

Predictive Medicine by Cytomics

- data pattern classification identifies a significant proportion of high risk patients individually prior to therapy as shown for AML patients as an example
- this is an important prerequiste for individualized/personalized therapy in stratified (Kaplan-Meier) patient groups
- detailed knowledge on the mechanisms of disease induction is not required for this approach

VAL

Data Pattern Classification

for further details see:

https://www.classimed.de/classif1.html
https://www.classimed.de/cellclas.html