A probabilistic method for gene expression data

Gunter Ritter

Faculty of Informatics and Mathematics
University of Passau/Germany

ritter@fim.uni-passau.de

Overview

1. The classification model, robustness
2. Variable Selection
3. Gene expression data
4. Analyzing the LEUKEMIA data set

1. The classification model

The classification model of clustering

Observations $x_i \in \mathbb{R}^d$, $i \in 1..n$,
mixing rates π_j, population parameters $\gamma_j \in \Gamma_j$, $j \in 1..g$

$$X_i \sim \sum_{j=1}^{g} \pi_j f_{\gamma_j} \quad \Rightarrow \quad \text{joint likelihood function} = \prod_i \sum_{j=1}^{g} \pi_j f_{\gamma_j}(x_i)$$

The classification model of clustering

Observations $x_i \in \mathbb{R}^d$, $i \in 1..n$,
mixing rates π_j, population parameters $\gamma_j \in \Gamma_j$, $j \in 1..g$

$$X_i \sim \sum_{j=1}^{g} \pi_j f_{\gamma_j} \quad \Rightarrow \quad \text{joint likelihood function} = \prod_i \sum_{j=1}^{g} \pi_j f_{\gamma_j}(x_i)$$

When components are well separated by location: Introduce labels $\ell_i \in 1..g$

$$\prod_i \pi_{\ell_i} f_{\gamma_{\ell_i}}(x_i) = \prod_j \pi_j^{n_j(\boldsymbol{\ell})} \prod_{\ell_i = j} f_{\gamma_j}(x_i)$$

$\Rightarrow$ General MAP criterion to be "maximized" w.r.t. $(\ell_1, \ldots, \ell_n)$

$$\sum_{j=1}^{g} \sum_{\ell_i = j} \log f_{\gamma_j(\boldsymbol{\ell})}(x_i) - n\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell})}{n}, \ldots, \frac{n_g(\boldsymbol{\ell})}{n}\Big)$$

Special case: Ward's sum-of-squares criterion (1963)

Reduction step, normal case

Normal case: Determinant criterion (Symons 1981)

$$-\tfrac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_j(\boldsymbol{\ell}) - n\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell})}{n}, \ldots, \frac{n_g(\boldsymbol{\ell})}{n}\Big)$$

Trimming (Cuesta-Albertos, Gordaliza, Matrán 1997, k-means)

Normal classification criterion with trimming of $n - r$ elements (TDC)

$$-\tfrac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_j(\boldsymbol{\ell}) - r\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell})}{r}, \ldots, \frac{n_g(\boldsymbol{\ell})}{r}\Big)$$

Gallegos & R. (2005, 2009):

"On well separated data sets with sufficiently large clusters (plus outliers), the scale-constrained TDC has a positive asymptotic breakdown point if (i) the correct number of components is assumed and if (ii) sufficiently many and not too many elements are discarded."

Reduction step in the trimmed normal case, TDC

Elementary multi-point reduction step with trimming

// Input: An admissible assignment $\boldsymbol{\ell}$.
// Output: An admissible assignment with improved criterion *or* "steady" *or* "fail."

1. (Estimation) Compute the MLE's $\gamma_j(\boldsymbol{\ell})$ of the parameters for all clusters w.r.t. $\boldsymbol{\ell}$ and compute the posterior probabilities (weights) $u_{i,j} = n_j + f_{\gamma_j(\boldsymbol{\ell})}(x_i)$.

2. (Assignment and trimming) Assign each i to the class j with maximum weight $u_{i,j} \rightsquigarrow \boldsymbol{\ell}_{\text{new}}$ and discard the $n - r$ data points i with smallest weights $u_{i,\ell_{\text{new},i}}$.

3. (Decision) *If* ℓ_{new} is inadmissible then "fail";
   else if $\sum_{i:\ell_{\text{new},i}\neq 0} u_{i,\ell_{\text{new},i}} > \sum_{i:\ell_i\neq 0} u_{i,\ell_i}$ then return the new labeling;
   else "steady".

Iteration $\Rightarrow$ steady assignment $\boldsymbol{\ell}$ $\Rightarrow$ Iterative replication.

2. Variable Selection

Irrelevance

Milligan 1980, Fowlkes, Gnanadesikan, and Kettenring 1988: Noisy variables degrade the analysis.

Model for irrelevance (redundancy and noise) John et al. 1994, Koller and Sahami 1996:

Let $F_1, F_2 \subseteq 1..D$ be disjoint subsets.

(a) The subset F_2 is irrelevant w.r.t. F_1 if L is conditionally independent of X_{F_2} given X_{F_1}, that is, P-a.s for all j,

$$P[L = j \mid X_{F_1}, X_{F_2}] = P[L = j \mid X_{F_1}].$$

(b) The subset F_2 is irrelevant if it is irrelevant w.r.t. its complement.

Relevance decomposition

(Gallegos & R. 2014)

(c) A subset $F \subseteq 1..D$ is structural if no subset $\emptyset \subset C \subseteq F$ is irrelevant w.r.t. $F \setminus C$.

"Let the real random variables X_i, $i \in 1..D$, have a strictly positive and continuous joint Lebesgue density $f_{(X_1,...,X_D)}$. There exists exactly one structural subset $F \subseteq 1..D$ with irrelevant complement."

Normal case

Assume $\emptyset \subset F \subset 1..D,\ E = \complement F$, covariance matrix $\mathrm{V}X_F$ invertible.

$X^{(j)} = (X_F^{(j)}, X_E^{(j)})$ normal; parametrize the normal family by a regression model:

$$X_E^{(j)} = m_{j,E|F} + G_{j,E|F} X_F^{(j)} + U_{E|F}^{(j)}$$

(a) If X is a normal mixture then the following statements are equivalent.

  (i) The subset E is irrelevant;

  (ii) the parameters $G_{j,E|F}$, $m_{j,E|F}$, and $V_{j,E|F}$ do not depend on j.

(b) In this case the common regression matrix $G_{E|F} = G_{j,E|F}$ can be computed from X, namely $G_{E|F} = \mathrm{Cov}(X_E, X_F)(\mathrm{V}X_F)^{-1}$.

Robust classification–and–selection criterion

Classification–and–selection likelihood:

$$\sum_{j=1}^{g} \sum_{i:\ell_i=j} \log N_{m_j,V_j}(x_{i,F}) - r\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{r}, \ldots, \tfrac{n_g(\boldsymbol{\ell})}{r}\big) + \sum_{i \in R} \log N_{m_{E|F},V_{E|F}}(x_{i,E} - G_{E|F} x_{i,F}).$$

Partial maximization w.r.t. parameters $\rightsquigarrow$ criterion

$$-\tfrac{1}{2} \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_{j,F}(\boldsymbol{\ell}) - r\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{r}, \ldots, \tfrac{n_g(\boldsymbol{\ell})}{r}\big) - \tfrac{r}{2} \log\det S_{R,E|F}.$$

Remains to be maximized w.r.t. F and $\boldsymbol{\ell}$.

Procedure (wrapper)

// Input: Subset $F \subseteq 1\,..\,D$, $|F| = d$, admissible $\boldsymbol{\ell}$ with $n - r$ discards, and value of the criterion.

// Output: New quantities F_{new} and $\boldsymbol{\ell}_{\text{new}}$, with improved criterion *or* "stop".

1. *(Estimation)* Compute the sample mean vectors $\overline{x}_j(\boldsymbol{\ell})$ and scatter matrices $S_j(\boldsymbol{\ell})$, $1 \le j \le g$, and the total scatter matrix S_R, $R =$ regular observations.

2. *(Selection)* Minimize

$$h(F') = \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_{j,F'}(\boldsymbol{\ell}) - r \log\det S_{R,F'} \le h(F)$$

w.r.t. F', $|F'| = d$. Denote the minimizer by F_{new}.

Procedure

3. Use the quantities from step 1 to compute the MLE's of the regression parameters (G, m, V) w.r.t. $\boldsymbol{\ell}$ and the new subsets F_{new} and $E_{\text{new}} = \complement F_{\text{new}}$. Let

$$\begin{aligned} u_{i,j} = & \log n_j - \tfrac{1}{2} \log\det S_{j,F_{\text{new}}}(\boldsymbol{\ell}) \\ & - \tfrac{1}{2}(x_{i,F_{\text{new}}} - \overline{x}_{j,F_{\text{new}}}(\boldsymbol{\ell}))^\top S_{j,F_{\text{new}}}(\boldsymbol{\ell})^{-1}(x_{i,F_{\text{new}}} - \overline{x}_{j,F_{\text{new}}}(\boldsymbol{\ell})) \\ & - \tfrac{1}{2}(x_{i,E_{\text{new}}} - m - Gx_{i,F})^\top V^{-1}(x_{i,E_{\text{new}}} - m - Gx_{i,F}). \end{aligned}$$

4. *(Assignment and trimming)* Compute an admissible, trimmed assignment $\boldsymbol{\ell}_{\text{new}}$ using a reduction step with trimming based on the statistics $u_{i,j}$.

5. *(Decision)*
   If F_{new} and $\boldsymbol{\ell}_{\text{new}}$ improve the criterion then return F_{new}, $\boldsymbol{\ell}_{\text{new}}$,
   else "stop".

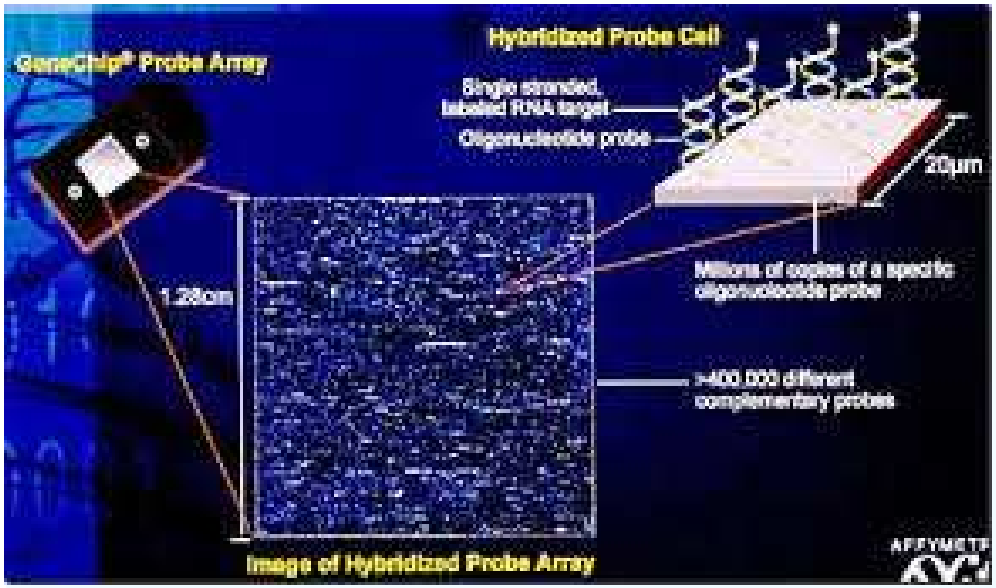
Diagonal model: minimum in step 2 easily attained by sorting.

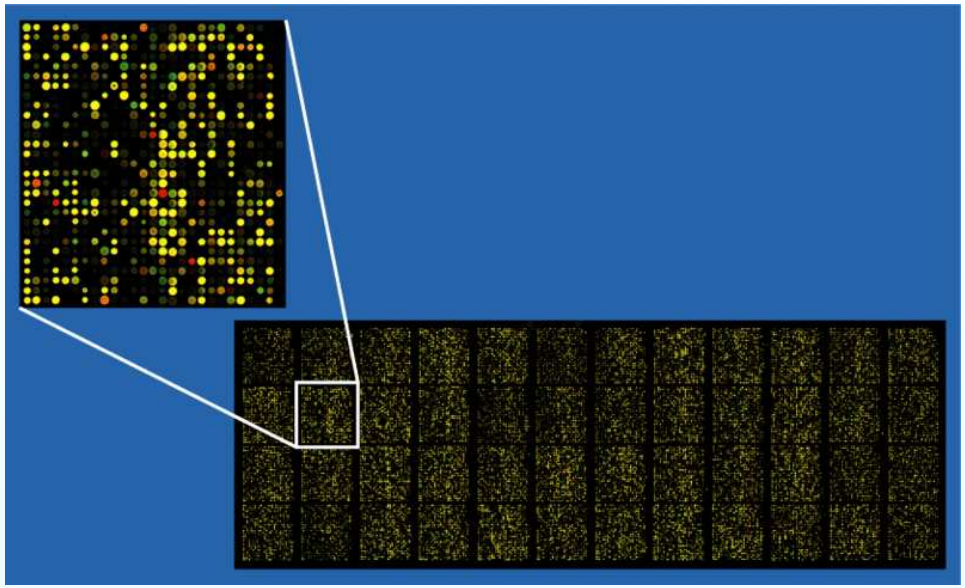
3. Gene expression data

Genes

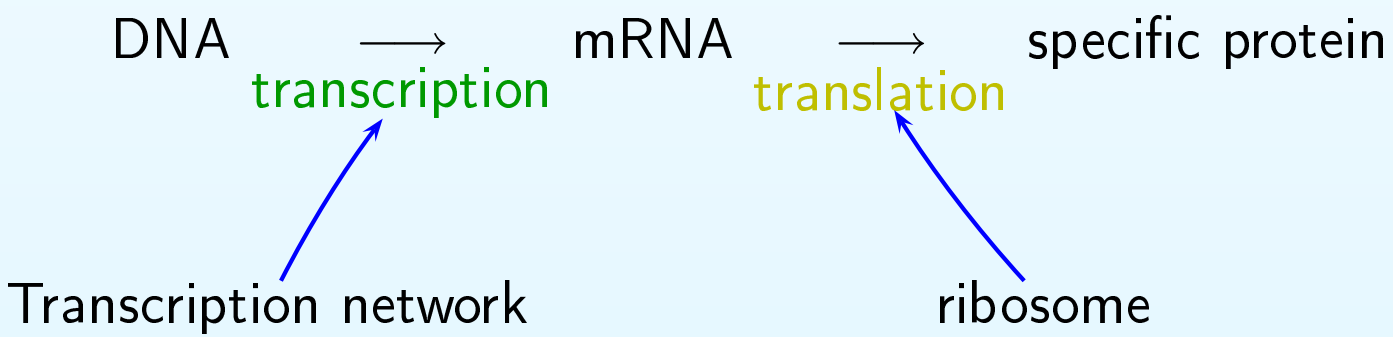
Measuring presence of thousands of genes in a specific tissue (cells of one kind)

Microarrays or gene chips with thousands of spots, each one specific for one gene





Presence = activity (ribosome)

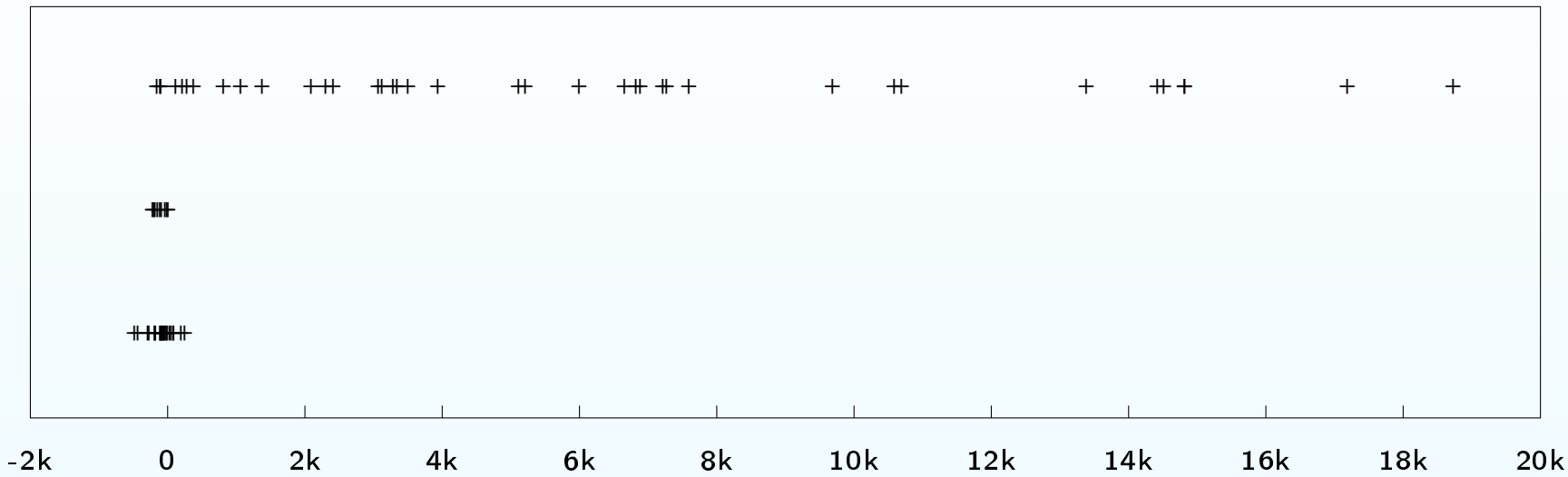


LEUKEMIA, Golub et al., SCIENCE 286 (1999)

Affymetrix Hgu6800 chip, 7129 genes

J02876	-1426	-1150	-1784	-1418	-1180	-979	-389	-1128	-786	-750	-1127	-711	-1046	-890	-541	-815
J02883	234	167	421	206	139	74	102	244	133	173	157	58	21	253	96	182
J02888	493	513	740	509	483	515	630	554	790	320	631	360	396	439	278	512
J02902	2447	2527	2785	2390	2833	2066	2627	2634	3610	1996	2744	1151	2234	2607	2115	1835
J02906	956	923	1658	772	770	485	1401	1572	792	952	653	538	762	669	301	932
J02923	1673	2550	3322	2310	2934	1479	1618	355	2927	1468	991	2075	1652	3515	2653	1589
J02943	294	337	341	226	78	218	324	442	102	173	202	117	186	268	146	308
J02963	334	208	215	240	88	188	193	145	53	190	386	153	122	184	70	259
J02973	-363	-63	-517	-165	-62	-177	-270	-307	-332	-312	-222	-206	-257	-306	-34	-153
J03040	65	-212	-106	74	-53	-161	-68	-286	-160	50	-189	116	21	1	-25	145
J03068	330	189	189	521	131	227	373	278	69	50	56	97	83	119	128	190
J03069	1879	1533	2154	1302	674	2587	1740	3550	1420	1408	1930	1175	1588	1787	1428	2381
J03133	367	211	410	366	137	259	386	433	328	106	231	176	167	289	169	236
J03161	836	609	792	496	490	451	1078	442	831	411	450	198	1141	567	525	759
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X56932	21121	21146	23203	22197	22414	24183	25128	25438	21732	24313	23553	22286	21630	23816	19980	24890
X56997	11777	13550	13412	13358	13239	12392	12064	10009	13536	15069	13017	8309	11219	14125	15412	14593
X57025	89	-2	199	24	24	12	-73	97	43	22	14	-35	33	34	18	69
X57129	58	191	-22	-43	-8	24	-31	-5	-78	115	4	396	255	-140	-29	-84
X57206	292	-184	-38	-13	99	-231	-92	158	266	-109	-103	-199	193	730	-32	747
X57303	244	265	288	307	96	158	174	196	303	184	204	234	200	278	155	184
X57346	1873	2057	2551	2231	3229	3010	1918	1728	4119	1113	1170	965	1328	2634	3714	1393
X57398	565	1087	636	598	1596	545	704	1042	1528	375	1059	459	1036	846	1093	1026
X57522	1342	718	53	613	376	417	446	284	441	71	314	441	609	743	2125	348
X57766	738	508	938	726	626	583	777	1157	931	664	447	441	250	520	567	555

Differential expression



X82240

B-cell ALL (acute lymphatic leukemia),
T-cell ALL,
AML (acute myeloid leukemia)

Tasks

Task	given	aim	method
(a)	Marker genes	conditions	discriminant analysis
(b)	Conditions	marker genes	variable selection
(c)	—	conditions	cluster analysis
(d)	—	conditions and marker genes	cluster analysis with variable selection

Tasks (a) and (b) are supervised, (c) and (d) unsupervised

(a) Bayesian discriminant rule

(b) t-test (Golub et al. 1999, McLachlan et al. 2006), Wilcoxon-Mann-Whitney test for two unpaired samples (Dettling and Bühlmann, 2002), massively multiple testing problem (Efron and Tibshirani, 2002); review Kadota and Shimizu 2011

(c) k-means (Tavazoie et al. 1999, Handl et al. 2005), mixture model (Yeung et al. 2001)

(d) "gene clustering": hierarchical (Eisen et al. 1998), mixture model Ghosh and Chinnaiyan 2002; "dimension reduction": factor models (McLachlan et al. 2002, Wang et al. 2009)

Some overexpressed genes

Type	GenBank accession	M-W count	location	name	function
B-ALL	U05259	26	19q13.2-3	HMB1	immunoglobulin
	M89957	48	17q23	B29	B-cell receptor
	L33930	52		CD24	signal transducer
	M84371	53		CD19	B-cell specific surface protein
	Z49194	55	11q23	OBF1	B-cell coactivator of octamer–binding transcription factors
	X82240	67	14q32	TCL1	oncogene
T-ALL	X04145	0		T3G	T-cell receptor
	X03934	2		T3D	T-cell antigen receptor
	M23323	2		CD3E	membrane protein
AML	M23197	13	19q13.3	CD33	differentiation antigen of myeloid progenitor cells
	X95735	25	7q34-35	zyxin	zyxin-related protein
	M27891	26	20p11.22-21	CST3	cystein proteinase inhibitor

4. Analyzing the LEUKEMIA data set

Preprocessing

(a) Expression levels are intensities $\longrightarrow$ log-transformation

$$\log(|x| + 1))$$

(b) Since number of genes excessive $\longrightarrow$ univariate filter first

Select skewed genes

$$s(\mathbf{x}) = \frac{\sum_i (x_i - \operatorname{med} \mathbf{x})}{\operatorname{med}_i |x_i - \operatorname{med} \mathbf{x}|} \geq \beta,$$

for eleven values $0.18 \leq \beta \leq 0.85$.

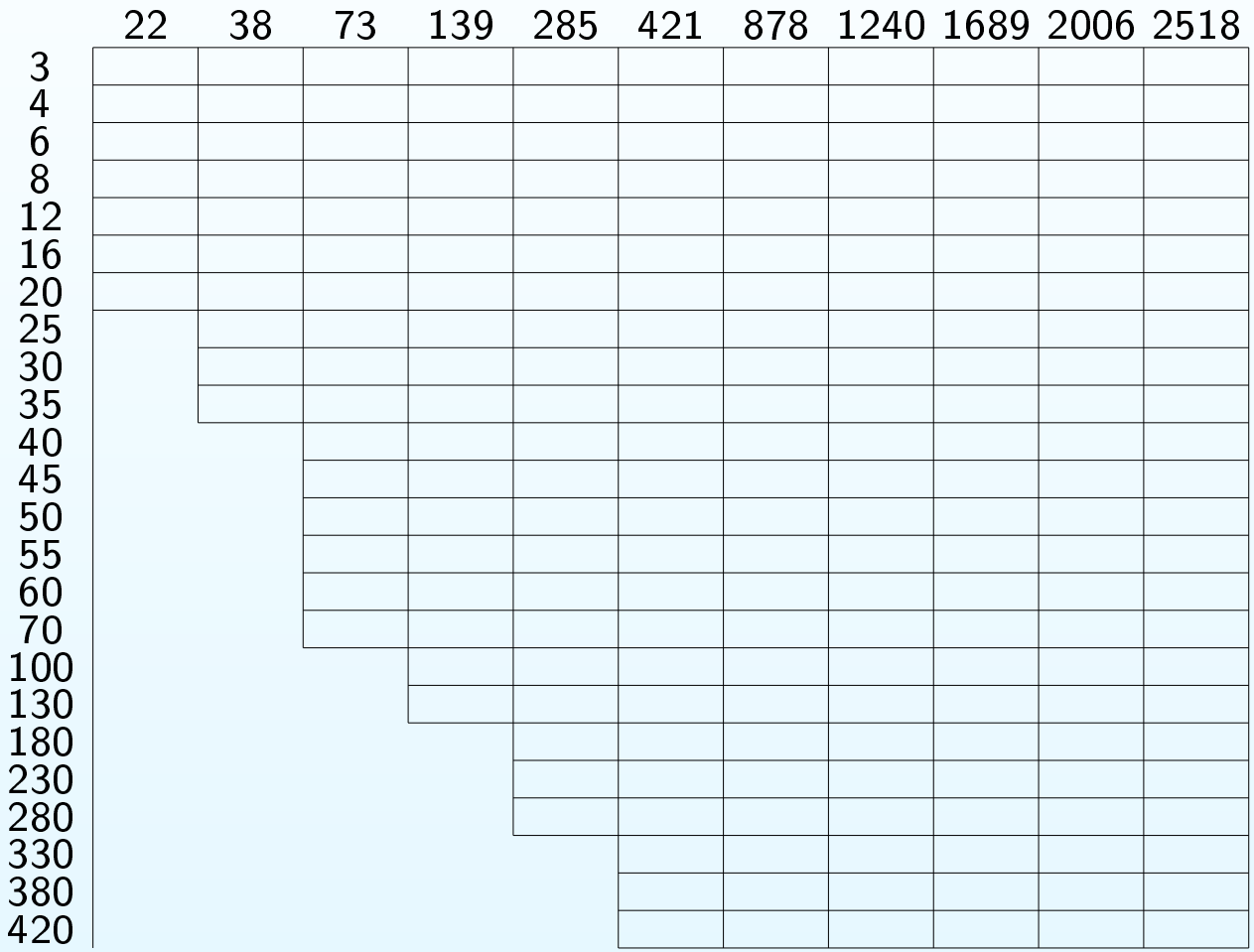
$\Rightarrow$ Nested data sets between 22 and 2518 genes.

Consensus sets and partitions

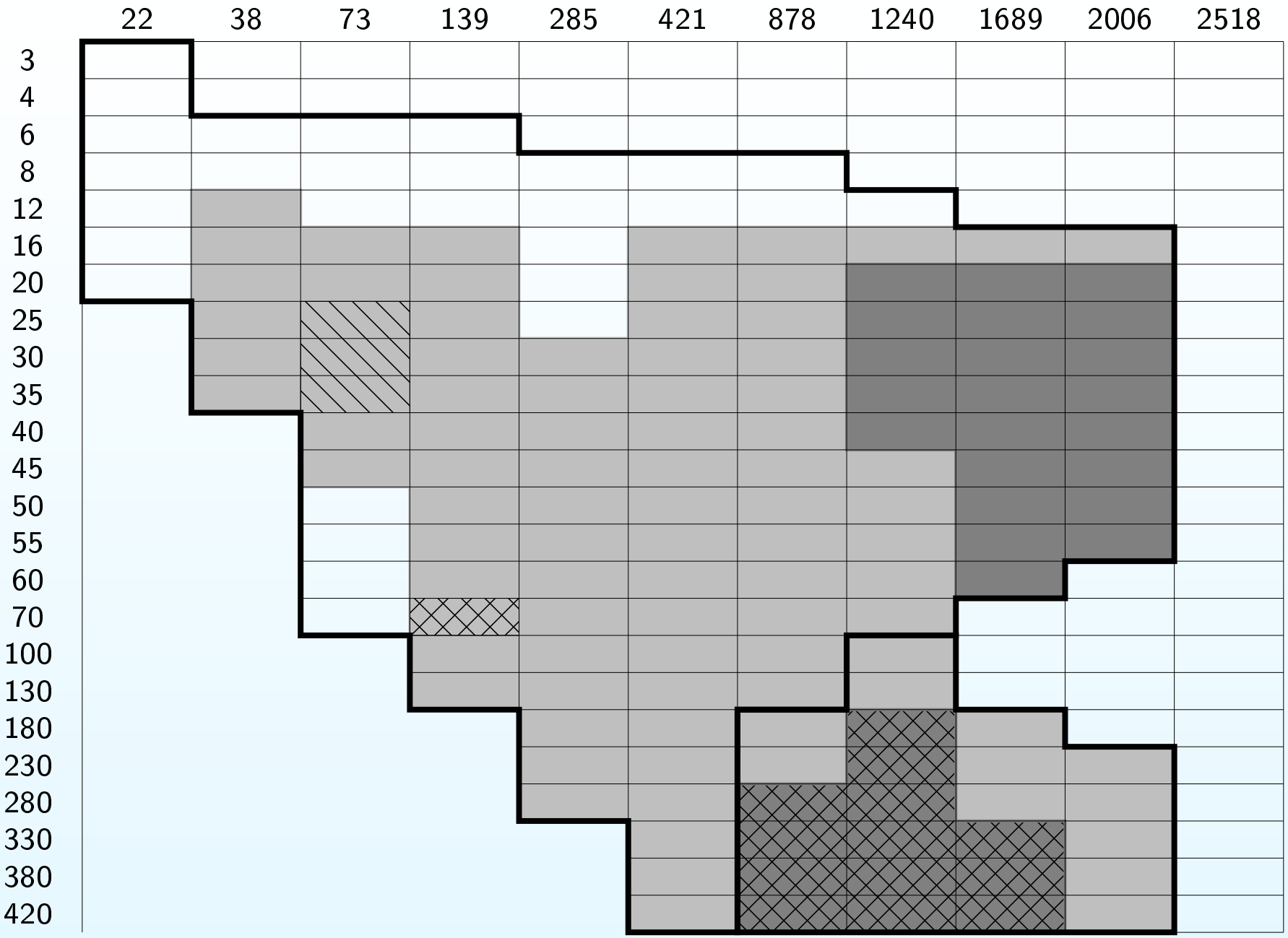
Stability: The α-consensus set of a partition $\mathcal{C} \in \mathcal{P}$ is

$$\{\mathcal{C}' \in \mathcal{P} \mid \mathrm{ARI}(\mathcal{C}', \mathcal{C}) \geq \alpha\}.$$

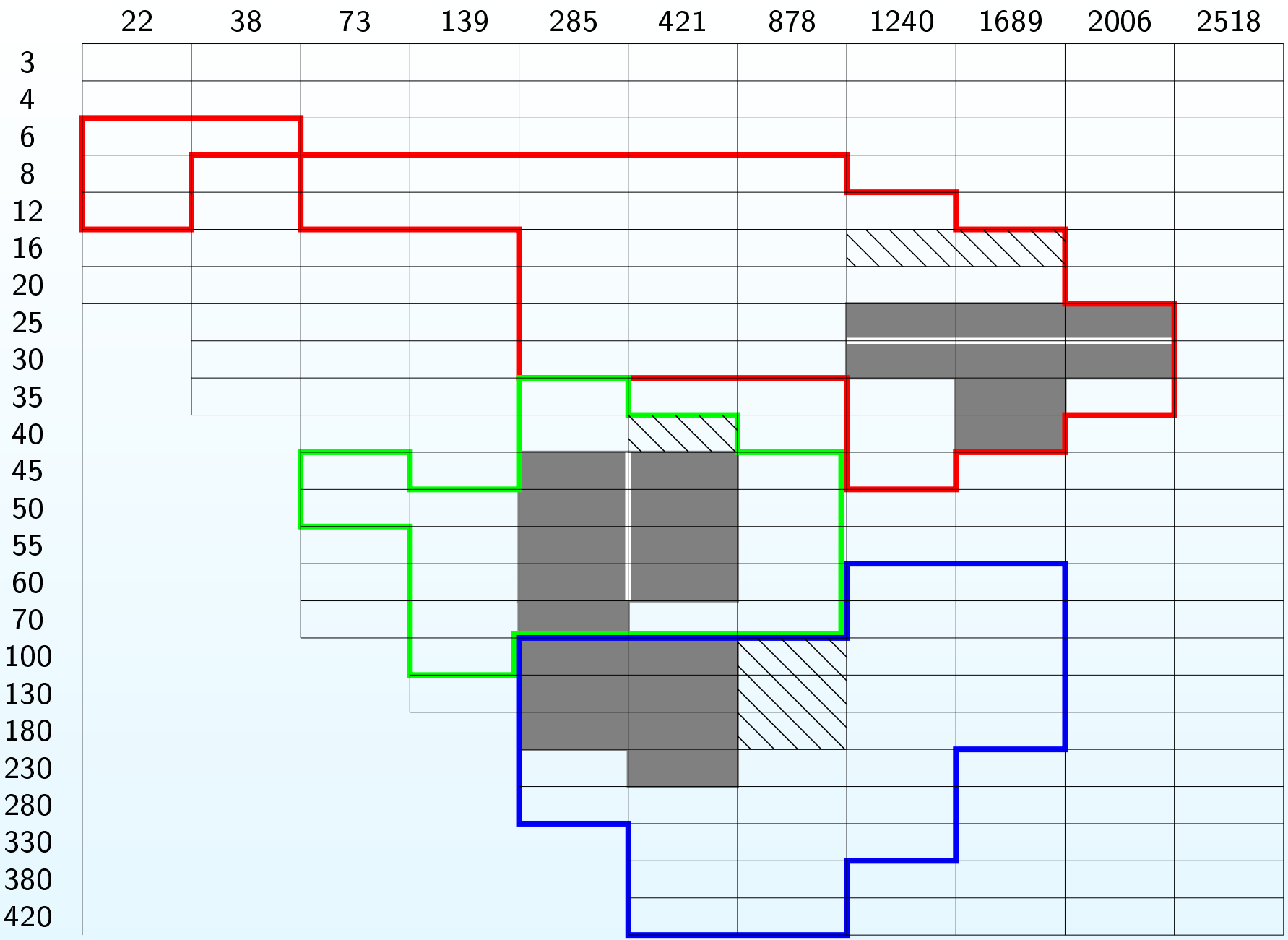
A partition $\mathcal{C}$ is an α-consensus partition if its α-consensus set is large.



Consensus sets, $g = 3$, two discards, $\alpha = 1,\ 0.86,\ 0.76$



Consensus sets, $g = 4$, two discards, $\alpha = 1,\ 0.73$



Finer classification of acute leukemias by immunophenotype

Web site of the American Cancer Society:

- pre-B ALL,
- mature B-cell ALL,
- pre-T ALL,
- mature T-cell ALL,
- AML with certain translocations between chromosomes,
- Acute Promyelocytic Leukemia, APL.

Separation measures and p-values

g	pair	Fisher	linear separ.	Sym. diverg.	–log Helling.	p-value MV S.-W.	p-value lin. separ.
3	1,2	0.999	0.999	161.7	5.84	0.000	0.000
	1,3	1.000	1.000	134.3	10.28	0.000	0.000
	2,3	0.985	0.985	44.0	2.91	0.000	0.000
4	1,2	1.000	1.000	457.6	37.98	0.000	0.000
	1,3	1.000	1.000	128.3	13.17	0.000	0.000
	1,4	0.966	0.966	28.9	2.39	0.001	0.003
	2,3	1.000	1.000	212.5	16.58	0.001	0.000
	2,4	1.000	1.000	399.2	37.04	0.010	0.000
	3,4	0.999	0.999	65.0	5.20	0.002	0.000

Scatter plots