Supplementary results for "Integrating functional knowledge during sample clustering for microarray data using unsupervised decision trees"

Henning Redestig[1,2], Dirk Repsilber[2], Florian Sohler[3] and Joachim Selbig[2,4]

1: Corresponding author
2: Max Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany
3: Institute for Informatics, Ludwig Maximilian's University, Amalienstraße 17, 80333 Munich, Germany
4: University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany

Additional Real Data

The dataset presented in **Mootha et. al.** *Nat. Genet.* (2003) contains gene expression data from diabetes type 2 patients, glucose intolerant patients and normal patients. The authors could not find any reliably differentially expressed genes in the dataset but managed, with the use of a gene class based hypothesis test (that was introduced in the same publication) to extract some information.
The dataset was clustered using GO-UDTs but no significant splits could be found in any gene class and thus no tree obtained. This result show an advantage of incorporating a bootstrap test in the clustering. Where no structure can be found one also do not obtain any result.

The well-known dataset by **Golub et. al.** *Science* (1999) was also clustered. In the original publication the authors analyzed gene expression from patients with acute lymphoblastic leukemia of T and B-cell lineage (T-ALL and B-ALL) along with patients with acute myeloid leukemia (AML).
The training data used in that study was floored and filtered as in the the original publication. Samples were clustered using 2000 bootstrap replicates, false discovery rate was set to 0.05 and the minimum similarity index *w* was set to 0.6. The result showed a clear reconstruction of the available patient subgroup and many of the reported gene class are directly related to the studied disease. In total four samples are heterogeneously clustered based on the biological background which is the same 'error' that was achieved in the original publication using a two-cluster self-organizing map. *S* in the tree is the goodness score measuring the bimodality of sample distribution in the given gene class, and *Size* is the amount of genes in it.

Tables of five top gene classes found to be significant by the global test

Table 5: A table of the top five gene classes based on the p-value from globaltest at each split in the hierarchical clustering tree obtained by clustering the **Chiaretti data**.

	Node	Rank	ID	Term	log10(*p*)	Size	Present in most similar CAPI-UDT node
	A	1	GO:0046649	lymphocyte activation	-55	79	Yes; node B
		2	GO:0030097	hemopoiesis	-55	74	Yes; node B
		3	GO:0001775	cell activation	-55	93	Yes; node B
		4	GO:0045321	immune cell activation	-55	92	Yes; node B

		5	GO:0006959	humoral immune response	-54	171	
	B	1	GO:0016337	cell-cell adhesion	-13	143	
		2	GO:0006959	humoral immune response	-13	171	
		3	GO:0007283	spermatogenesis	-11	88	
		4	GO:0007276	gametogenesis	-9	116	
		5	GO:0019953	sexual reproduction	-9	143	
	C	1	GO:0007599	hemostasis	-4	98	
		2	GO:0006904	vesicle docking during exocytosis	-4	17	
		3	GO:0050878	regulation of body fluids	-4	112	
		4	GO:0007229	integrin-mediated signaling pathw..	-4	62	
		5	GO:0001501	skeletal development	-4	150	
	D	1	GO:0007417	central nervous system developmen..	-11	117	
		2	GO:0006694	steroid biosynthesis	-11	64	
		3	GO:0045595	regulation of cell differentiatio..	-10	49	
		4	GO:0007420	brain development	-10	39	
		5	GO:0007548	sex differentiation	-10	39	Yes; node A
	E	1	GO:0008203	cholesterol metabolism	-10	59	
		2	GO:0016125	sterol metabolism	-10	64	
		3	GO:0006664	glycolipid metabolism	-9	22	
		4	GO:0008285	negative regulation of cell proli..	-8	177	
		5	GO:0008202	steroid metabolism	-8	127	

Table 6: A table of the top five gene classes based on the p-value from globaltest at each split in the PAM clustering obtained by clustering the **Chiaretti data**.

	Node	Rank	ID	Term	log10(*p*)	Size	Present in most similar CAPI-UDT node
	A	1	GO:0006916	anti-apoptosis	-12	122	
		2	GO:0001501	skeletal development	-12	150	
		3	GO:0043066	negative regulation of apoptosis	-12	133	
		4	GO:0043069	negative regulation of programmed..	-12	136	
		5	GO:0008643	carbohydrate transport	-12	18	

	B	1	GO:0045934	negative regulation of nucleobase..	-9	149	
		2	GO:0031324	negative regulation of cellular m..	-9	172	
		3	GO:0046942	carboxylic acid transport	-9	57	
		4	GO:0016481	negative regulation of transcript..	-9	142	
		5	GO:0000122	negative regulation of transcript..	-9	79	
	C	1	GO:0006898	receptor mediated endocytosis	-16	33	
		2	GO:0000902	cellular morphogenesis	-16	197	
		3	GO:0050793	regulation of development	-15	164	
		4	GO:0006897	endocytosis	-14	124	
		5	GO:0030036	actin cytoskeleton organization a..	-14	123	
	D	1	GO:0007601	visual perception	-19	173	
		2	GO:0007611	learning and/or memory	-15	25	
		3	GO:0006664	glycolipid metabolism	-11	22	
		4	GO:0009967	positive regulation of signal tra..	-11	86	
		5	GO:0006869	lipid transport	-11	48	
	E	1	GO:0007548	sex differentiation	-11	39	Yes; node A
		2	GO:0042445	hormone metabolism	-10	45	Yes; node A
		3	GO:0042446	hormone biosynthesis	-10	31	Yes; node A
		4	GO:0006700	C21-steroid hormone biosynthesis	-10	17	Yes; node A
		5	GO:0007420	brain development	-9	39	
	F	1	GO:0046649	lymphocyte activation	-55	79	Yes; node B
		2	GO:0030097	hemopoiesis	-55	74	Yes; node B
		3	GO:0001775	cell activation	-55	93	Yes; node B
		4	GO:0045321	immune cell activation	-55	92	Yes; node B
		5	GO:0006959	humoral immune response	-54	171	

Table 7: A table of the top five gene classes based on the p-value from globaltest at each split in the hierarchical clustering tree obtained by clustering the **Spira data**.

	Node	Rank	ID	Term	log10(*p*)	Size	Present in most similar CAPI-UDT node
	A	1	GO:0007254	JNK cascade	-32	44	

		2	GO:0000165	MAPKKK cascade	-32	97	
		3	GO:0008284	positive regulation of cell proli..	-31	149	
		4	GO:0001501	skeletal development	-31	150	
		5	GO:0015698	inorganic anion transport	-31	109	
	B	1	GO:0016311	dephosphorylation	-4	137	
		2	GO:0016485	protein processing	-4	30	
		3	GO:0042990	regulation of transcription facto..	-4	14	
		4	GO:0001558	regulation of cell growth	-4	108	
		5	GO:0016125	sterol metabolism	-4	64	

Table 8: A table of the top five gene classes based on the p-value from globaltest at each split in the PAM clustering obtained by clustering the **Spira data**.

	Node	Rank	ID	Term	log10(*p*)	Size	Present in most similar CAPI-UDT node
	A	1	GO:0008202	steroid metabolism	-16	193	Yes; node A
		2	GO:0046942	carboxylic acid transport	-16	90	
		3	GO:0007586	digestion	-14	66	Yes; node A
		4	GO:0006809	nitric oxide biosynthesis	-14	25	
		5	GO:0006739	NADP metabolism	-14	17	
	B	1	GO:0016042	lipid catabolism	-25	101	
		2	GO:0009582	detection of abiotic stimulus	-25	186	
		3	GO:0001501	skeletal development	-25	195	
		4	GO:0007605	perception of sound	-24	153	
		5	GO:0050982	detection of mechanical stimulus	-24	155	
	C	1	GO:0048511	rhythmic process	-4	42	
		2	GO:0006004	fucose metabolism	-4	27	
		3	GO:0007178	transmembrane receptor protein se..	-4	58	
		4	GO:0007179	transforming growth factor beta r..	-4	39	
		5	GO:0007159	leukocyte adhesion	-4	14	

Tables were generated with LaTeX2HTML