Appendix

Additional data and (color) figures for manuscript

P. Glenisson, J. Mathys, B. De Moor and Y. Moreau, **Meta-Clustering of Gene Expression data and Literature-based information**, *SIGKDD Explorations*, Spring 2004.

Section 2

Table 1: Motifs considered in the cell-cycle experiment. The consensus sequence for each motif was extracted from the TRANSFAC Database

Motif	sequence
Ace2	GTGTGTGTGTGTGTG
Cbf1p_Tav	TCACGTG
Cbf1p	RTCACRTG
CRE	TKACGTCA
ECB_Tav	TTWCCNNNWNNGGAAA
ECB	TTWCCCNNWNAGGAAA
Fkh1	AAWNRTAAAYARGN
Fkh2	GGNAAWANGTAAACAA
GATA	GATAA
Gcn4	TGACTCA
GRR	CTTCTT
ICRE	WYTTCAYRNTGS
M13	WWWTTTTTCA
M14a	TTTSGCKTY
M14b	TGTTTRKT
M1a	TTCYCWYYG
M26	GTKTNNGTGNG
M27	AWATTTTTC
M3a_Tav	MGATGAG
M3a	CGATGAG
M3b	TGAAAAA
M4	GAARKYTTT
M5	YWKNTCCYNTTT
MATA_A1	TGATGTANNT
MBM	TATTTGCTC
Mbp1	GNRRNANACGCGTNR
MCB_Tav	ACGCGKNR
MCB	ACGCGTNA
Met31_32p_Tav	AAANTGTGG
Met31_32p	AAACTGTGG
Ndd1	GGNAAWWNGTMAACA
Pho4	CACGTG
PRE	TGAAACA
Rap1_short	ACCCAT
Rap1	ACACCCATACATTT
SCB_Tav	MAMGAAA
SCB	CACGAAAA
Ste12	TGAAACA
STRE_Tav	AGGRG
STRE	AGGGG
Swi4	CACGAAA
TCS	CATTCT

Section 3

FATTY ACIDS/LIPIDS/..	NUTRITION	SECRETION	DNA_REPAIR	CELL CYCLE CONTROL	METHIONINE
EPT1	BAT2	EMP24	DUN1	CLB5	MUP1
LPP1	PHO8	SEC28	MSH2	CLB6	MET1
PSD1	AGP1	SLY41	MSH6	CLN1	MET6
SUR1	BAT1	UFE1	OGG1	CLN2	MET10
SUR2	GAP1	ERV25	PMS1	HSL1	MET13
SUR4	DIP5	SSO2	RAD27	PCL1	MET14
AUR1	FET3	GYP6	RAD5	PCL2	MET28
ERG3	FTR1		RAD51	RME1	MET17
LCB3	PFK1	**SPORULATION**	RAD53	SWE1	MET3
ERG2	PHO3	SPO16	RAD54	CLB4	SAM1
ERG5	PHO5	HOP1	RDH54	HSL7	
PMA1	PHO11	HDR1	RHC18	WHI3	**GLYCOSYLATION**
PMA2	PHO12	SPS4	UNG1	ACE2	MNN1
PMP1	PHO84	SSP2	HPR5	CLB1	OCH1
ELO1	RGT2		MEC3	CLB2	PMT1
FAA1	SUC2	**MITOTIC EXIT**	ALK1	CLN3	PMT3
FAA3	SUT1	DBF20		SWI5	PMT5
FAA4	VCX1	APC1		PCL9	PSA1
FAS1	ZRT1	CDC5		SWI4	QRI1
	GLK1	CDC20			SVS1
PSEUDOHYPAE	HXT1	DBF2			SSO1
ELM1	HXT2	MOB1			GDA1
PHD1	HXT4	SPO12			PMI40
TEC1	HXT7	TEM1			ALG7
	TAT1	SIC1			VRG4

Figure 1: Ten cell-cycle groups encompassing 126 genes used to demonstrate the mechanism of clustering genes based on text

Section 6

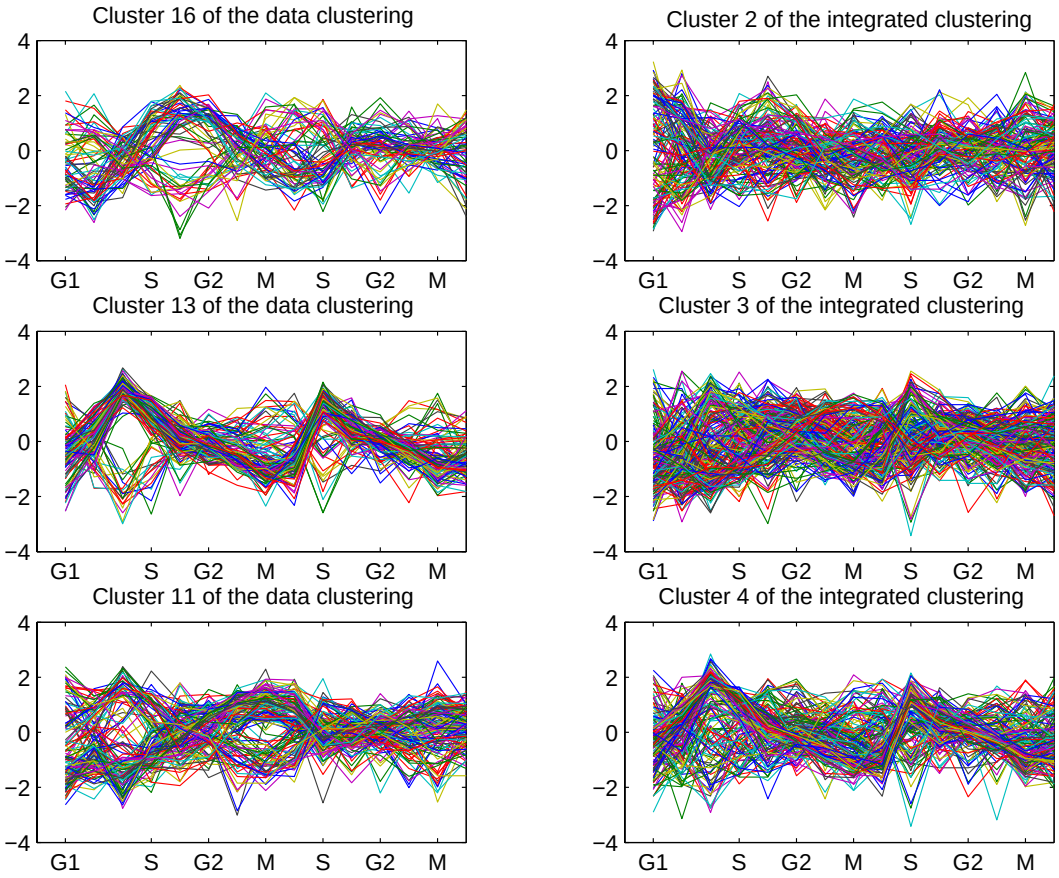


Figure 2: Expression profiles of selected clusters from data and integrated representation

HIGH VAR		HIGH MEAN	
mcm	0.260567	bud	0.158539
spindl	0.257225	chromosom	0.158382
methionin	0.237323	cell_cycl	0.154007
motor	0.175647	methionin	0.139277
kinetochor	0.173833	interact	0.136665
checkpoint	0.152642	spindl	0.130241
dna_replic	0.138802	pathwai	0.128255
methionin_biosynthesi	0.133935	local	0.126685
septin	0.130573	growth	0.117326
sulfoxid	0.118298	transcript	0.11164
kinesin	0.118117	import	0.110614
casein_kinas_i	0.113287	similar	0.107256
cell_cycl	0.112453	mitosi	0.105032
replic	0.11242	control	0.101375
spindl_pole_bodi	0.108965	elem	0.099366
bud	0.107676	homolog	0.098653
tfiid	0.099768	cellular	0.093821
microtubul	0.099754	sensit	0.093494
esteras	0.096126	depend	0.092657
sulfit_reductas	0.095454	regulatori	0.092593

Figure 3: Text profile of data cluster 16

HIGH VAR		HIGH MEAN	
mismatch_repair	0.308635	chromosom	0.188985
cyclin	0.291143	dna_replic	0.159592
dna_replic	0.177855	bud	0.158042
g1	0.170948	cell_cycl	0.157662
repair	0.166447	dna	0.143016
cohesion	0.163375	replic	0.139614
replic	0.162933	interact	0.13199
sister_chromatid_cohesion	0.159212	repair	0.124738
cohesin	0.157841	homolog	0.122083
spindl_pole_bodi	0.151415	pathwai	0.120025
cell_cycl	0.136073	similar	0.117131
telomer	0.135943	phase	0.117092
kinas	0.10929	recombin	0.1144
recombin	0.10821	g1	0.11379
dna_primas	0.104796	progress	0.107689
dna_repair	0.102491	sensit	0.107346
bud	0.100724	cyclin	0.107225
microtubul	0.099773	mitosi	0.105688
checkpoint	0.095579	growth	0.101168
mitosi	0.093322	depend	0.100206

Figure 4: Text profile of data cluster 13

HIGH VAR		HIGH MEAN	
mitosi	0.194852	cell_cycl	0.150875
checkpoint	0.160131	bud	0.148631
exit	0.158122	mitosi	0.13604
hydroxymethylpyrimidin	0.151094	control	0.135622
cell_wall	0.140319	similar	0.131355
cell_cycl	0.138485	pathwai	0.131304
cytokinesi	0.135486	growth	0.12967
mcm	0.134346	chromosom	0.127741
floccul	0.128789	transcript	0.122877
transcript_factor	0.127641	interact	0.118985
dna_replic	0.12763	local	0.112857
bud	0.119962	increas	0.107689
phospholipas_d	0.116157	dna	0.102777
cyclin	0.115533	synthesi	0.100966
mitot	0.103133	mitot	0.099614
zinc	0.101555	kinas	0.097981
anaphas_promot_complex	0.101275	induc	0.09657
protein_kinas	0.099495	import	0.095672
replic	0.099039	depend	0.095235
inositol	0.095915	cell_wall	0.091196

Figure 5: Text profile of data cluster 11

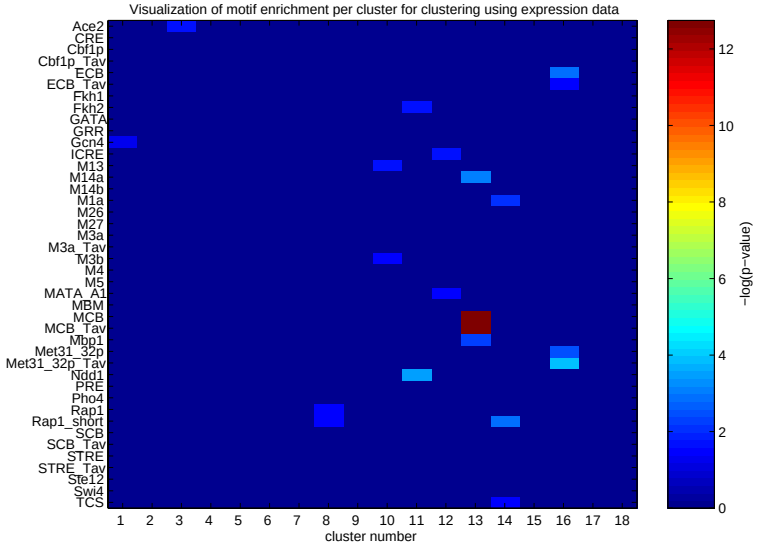


Figure 6: Visualization of motif enrichment for solely expression data (k=18 as obtained from stability analysis). Statistical overrepresentation of a motif in a cluster is computed using the binomial distribution.

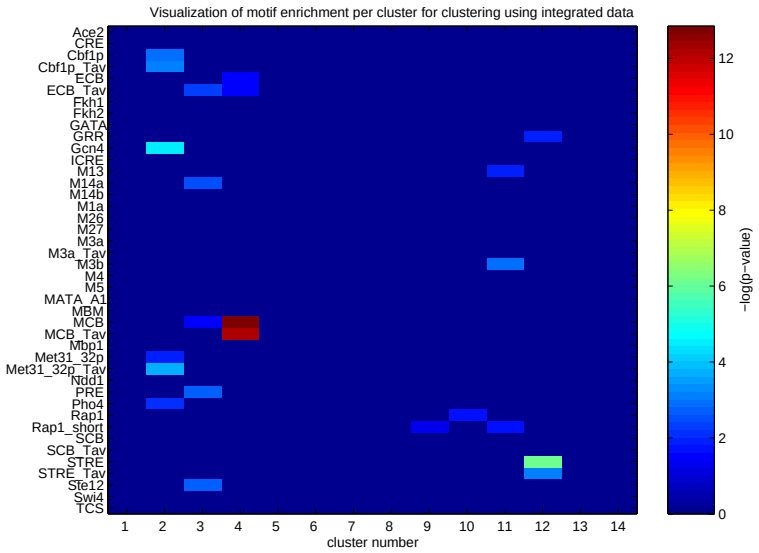


Figure 7: Visualization of motif enrichment for combined data (k=14 as obtained from stability analysis). Statistical overrepresentation of a motif in a cluster is computed using the binomial distribution.