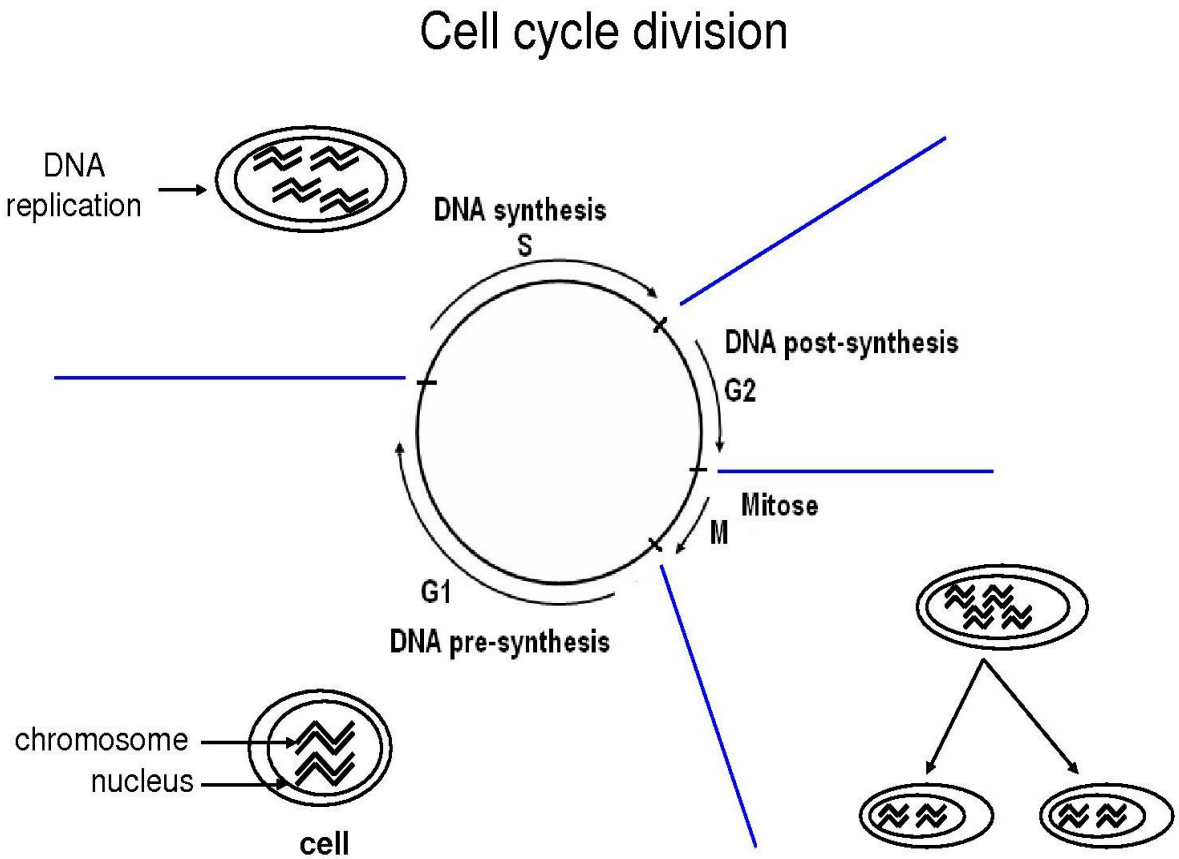
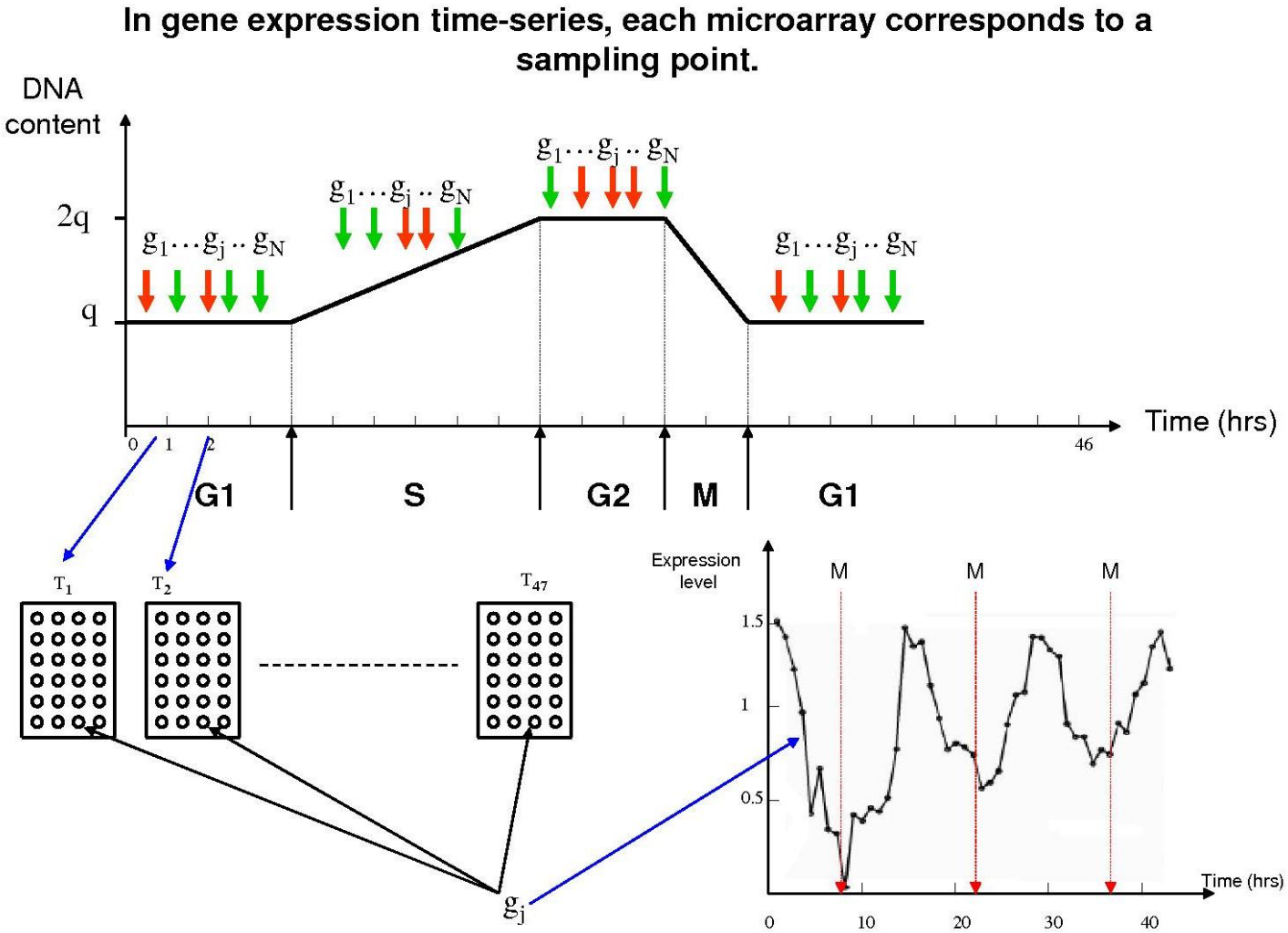
Time Series Clustering: application on cell cycle genes expression profiles

Ahlame Douzal (Ahlame.Douzal@imag.fr)
AMA, LIG, Université Joseph Fourier
Master 2R - MOSIG (2011)

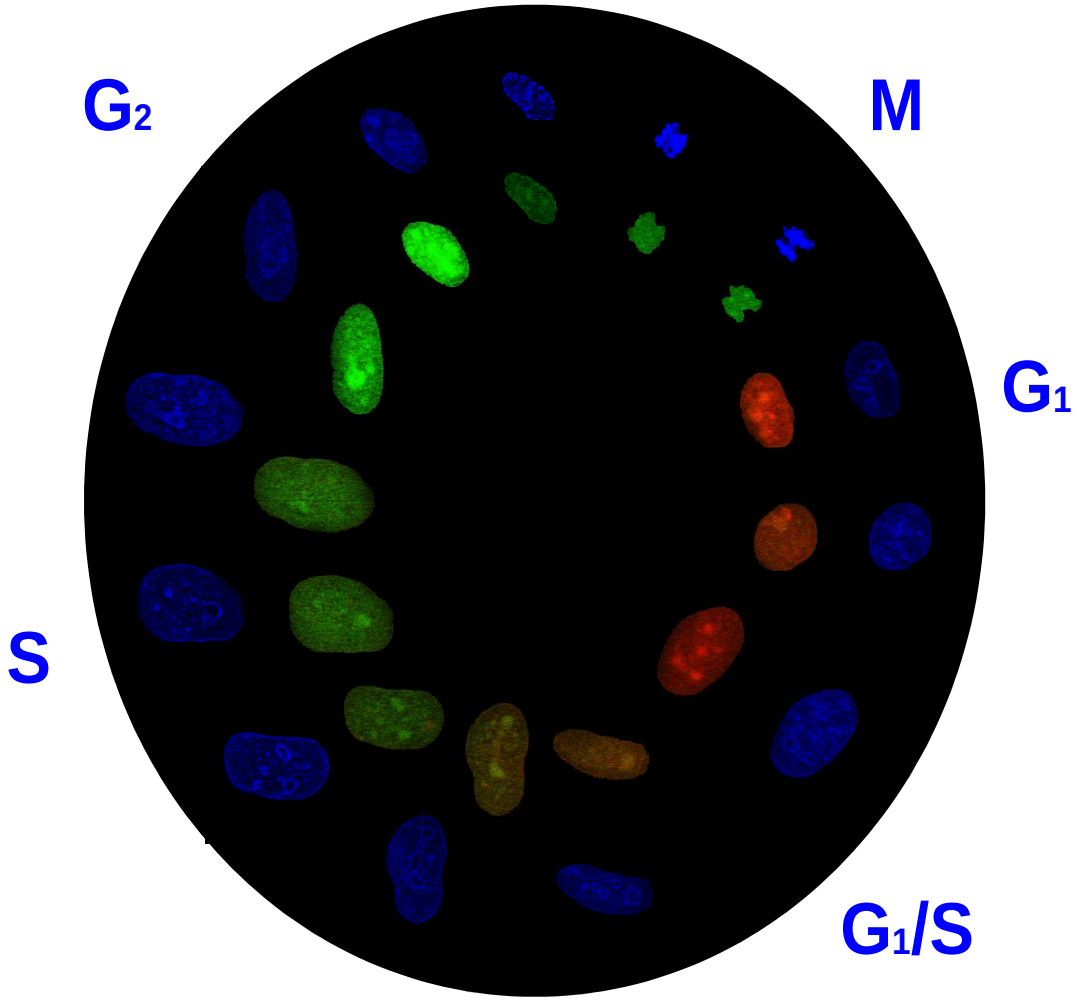
Problem statement



Gene expression data and cell cycle division



Problem statement



Objectives

- Identification of the cell cycle expressed genes
- Determine of differentially expressed genes:
  - which genes are involved in different type of cells (cancer versus healthy cells) ?

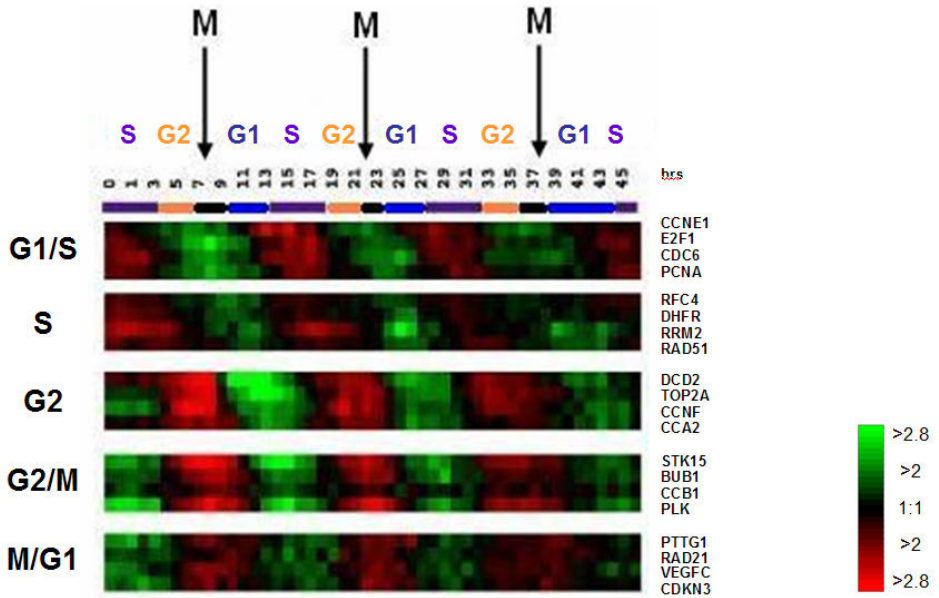
Clustering and classifying have proven helpful to:

- Extract main groups of behaviour expressions
- Identify new genes (unlabeled, unknown)
- Identify new genes relationships: co-expressed genes, co-regulated genes, understand genes functions,

Conventional Approach

Hela data (Whitfield et al. 2002)

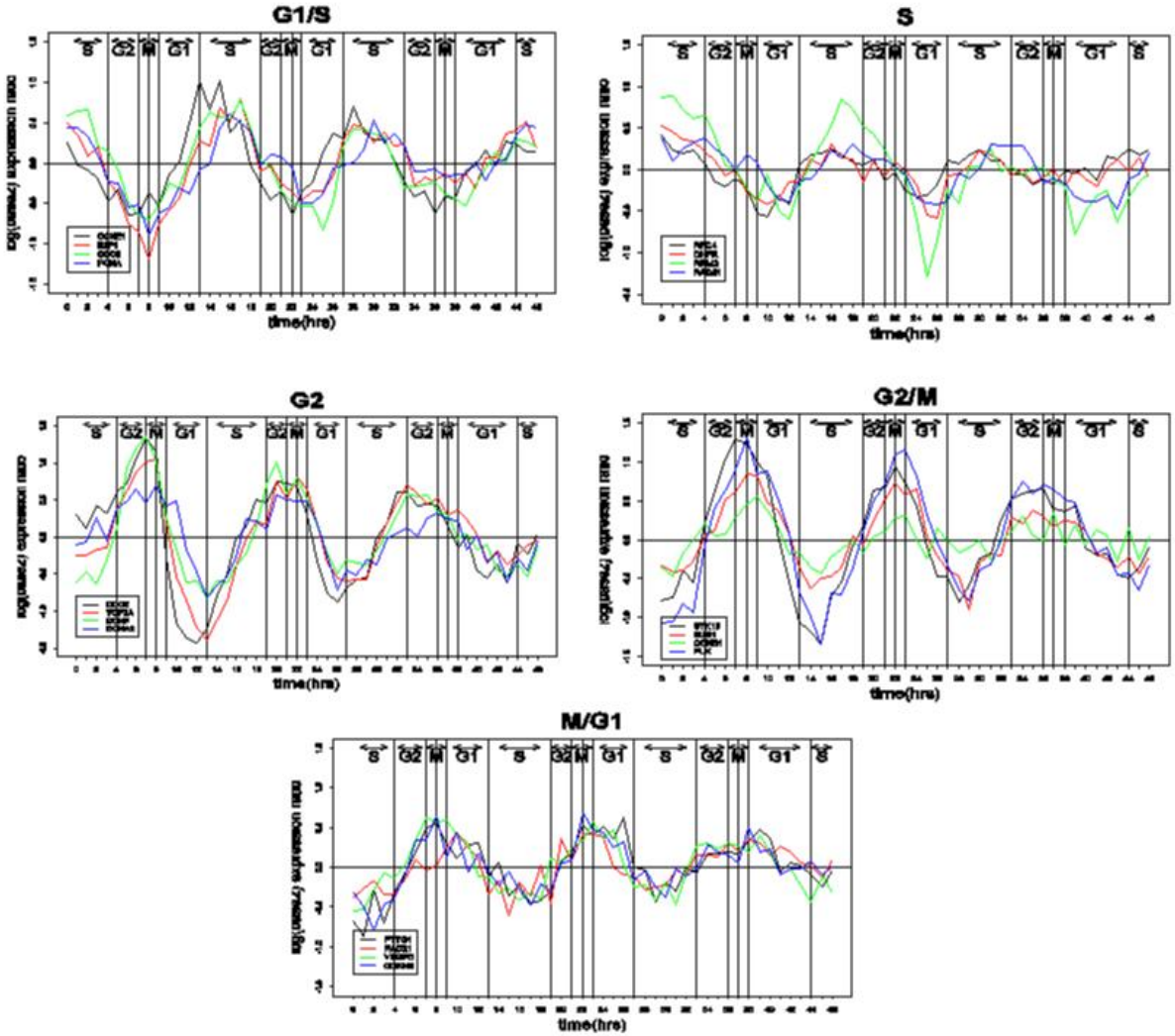
- Studied genes are preprocessed: non-cyclic expression, unexpressed genes and noisy expression
- Experimentally, a set of reference genes are selected



- Each measured gene is assigned to one phase by it's peak similarity to the reference genes: Pearson correlation.

The expression profiles of the 20 reference genes

- The expression profiles of the 20 reference genes, illustrating their peak expression at one phase during three cell division cycles. The double arrowed lines delimit the time duration for four cell cycle phases: G_1, S, G_2 and M.



Aims

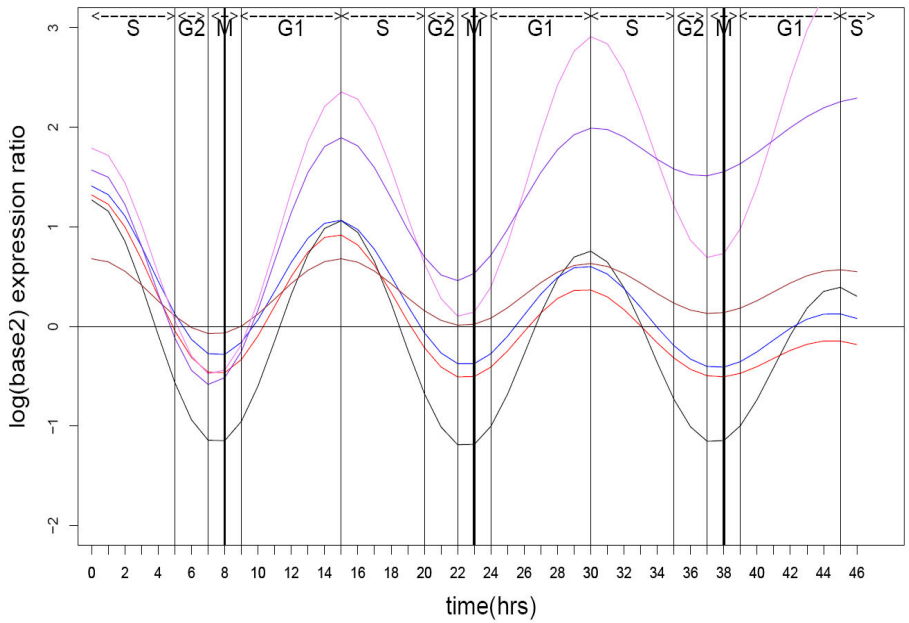
Limitations

- No consensus on the well-characterized genes, experimentation bias
- Similarity ignores the temporal structure, fixed a priori and unjustified

Aims

- Learn the best dissimilarity measure to classify or cluster genes expression profiles.
- Propose a well-founded set of reference genes.

Gene expression progression during the cell division process



- periodic profiles
- variation on: cell-cycle duration, initial amplitude
- amplitude attenuation
- tendency and drifts effects

Which metric for clustering and classifying genes expression profiles ?

- **Values-based metrics ?**
  - DTW, Euclidean distance, Manhattan distance, Fréchet distance, LCSS,...
- **Behavior-based metrics ?**
  - Pearson correlation coefficient, Qualitative distance (slope, derivative comparison), Kendall ratio, temporal correlation coefficient, etc...
- **Values and Behavior-based metrics ?**

Proximity measure specifications

- Genes expression are cyclic profiles
- Time of peak expression determines the cell-cycle phase assignment,
- Genes expression data may include tendency effects, amplitude attenuation,...

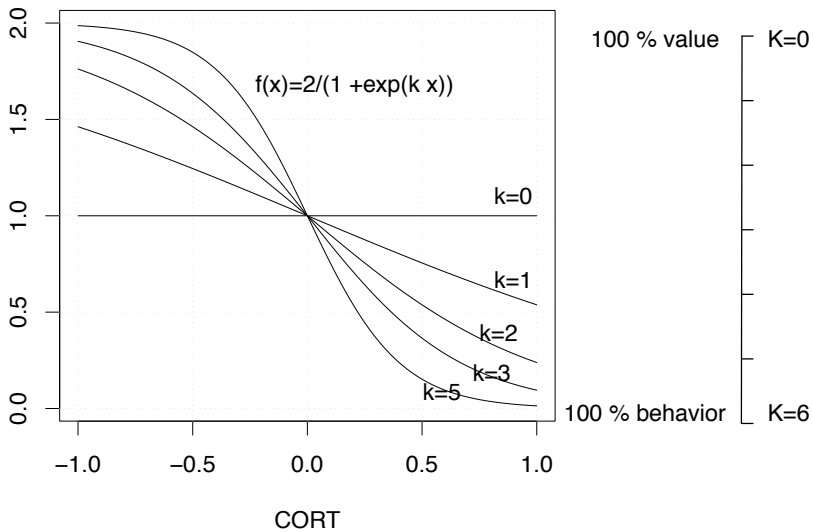
Induced constraints ...

- r should not include time warping
- δ_E is considered for values proximity measure
- *cort* is considered for behavior proximity measure

Values & Behavior-based metrics

$$D_k(S_1, S_2) = f(\text{CORT}(S_1, S_2)) \cdot \delta_E(S_1, S_2) \text{ with } f(x) = \frac{2}{1 + \exp(k\ x)} \quad , \quad k \geq 0$$

k: the contribution of values and of behavior to D



k learned during the classification or the clustering processes

Adaptive clustering: Partitioning around medoids approach

Motivations

- Use the PAM (Partitioning Around Medoids) algorithm to partition the set of genes into 5 clusters (5 cell cycle phases)
- More robust than k-means faced to outliers,
- Provide a more detailed analysis of the obtained partition
- Indicating for each object if it is (width silhouette):
  - well classified (i.e., genes well characterizing a cell cycle phase)
  - or lying on the boundary (genes involved in an inter-phase transition)

Algorithm steps

Learning D_k

- Perform the PAM algorithm based on D_k and for several values of (n, k). Let $P_{n,k}$ be the obtained partition.
- note $P_{n*,k*}$ the optimal partition according to two goodness criteria (asw, wb ratio)

Identification of the well-characterized genes

- Extract a kernel set of the p first genes maximizing the silhouette width,
- Identify the cell cycle phase of the kernel set,
- Assign each cluster to the cell cycle phase of it's kernel set.

Classification of the expressed cell cycle genes

- Assign each gene to the cell cycle phase of the cluster it belongs in.

Application specifications

- Analysis of experimental transcriptomic data from Human cancer cell line (Whitfield et al. 2002)
- Third experimentation
- The expression of 1099 periodically expressed genes through 48 instants covering 3 cell division cycles

http://genome-www.stanford.edu/Human-CellCycle/Hela/

Adaptive Clustering: Identification of the expressed cell cycle genes

Learning the most appropriate D_k

- Perform the PAM algorithm based on D_k for k varying in $[0, 6]$ and n in [4, 10], $P_{n,k}$ the obtained partition for a given (n, k).
- $(n*, k*) = argmax_{n,k}(avgSil(P_{n,k}))$, D_{k*} is the most appropriate dissimilarity.

Identification of the well-characterized genes

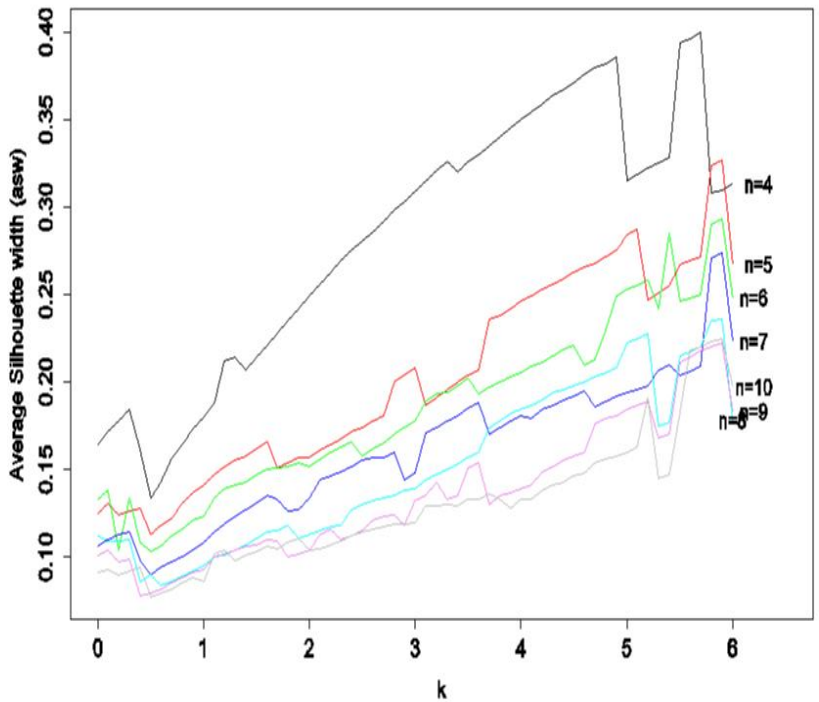
- Extract a kernel set of the N first genes maximizing the silhouette width,
- Identify the cell cycle phase of the kernel set,
- Assign each cluster to the cell cycle phase of it's kernel set.

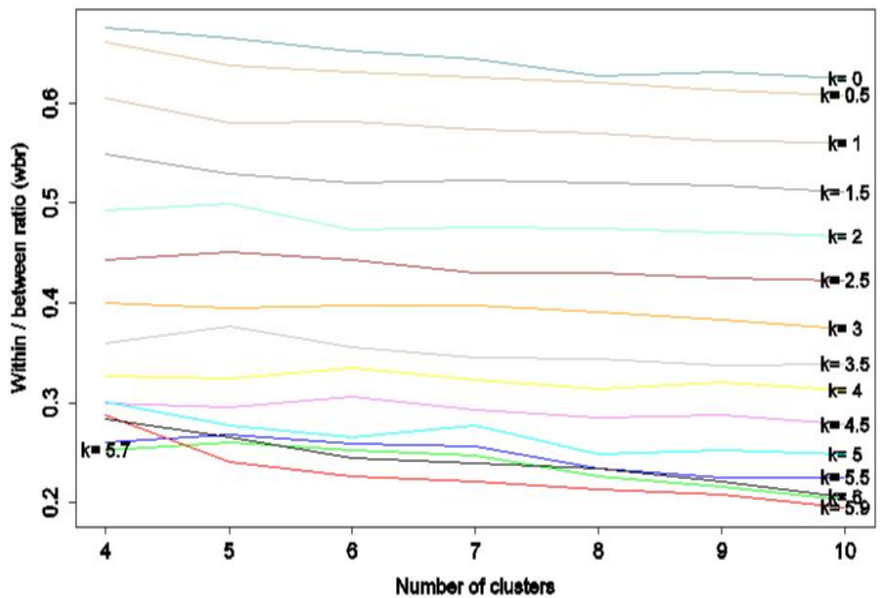
Classification of the expressed cell cycle genes

- Assign each gene to the cell cycle phase of the cluster it belongs in.

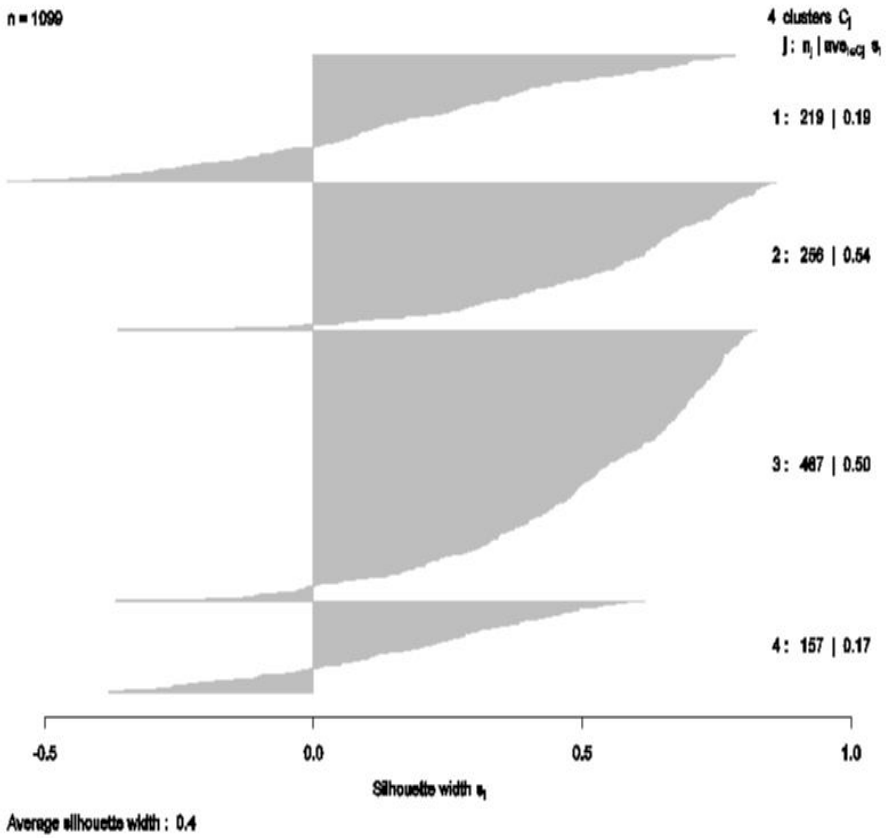
Learning the dissimilarity D_k through the partitioning process

- Average silhouette width (left plot) and the within/between ratio (right plot) of $P_{n\,k}$, n from 4 to 10 and k from 0 to 6





Average silhouette width of the genes profiles partition



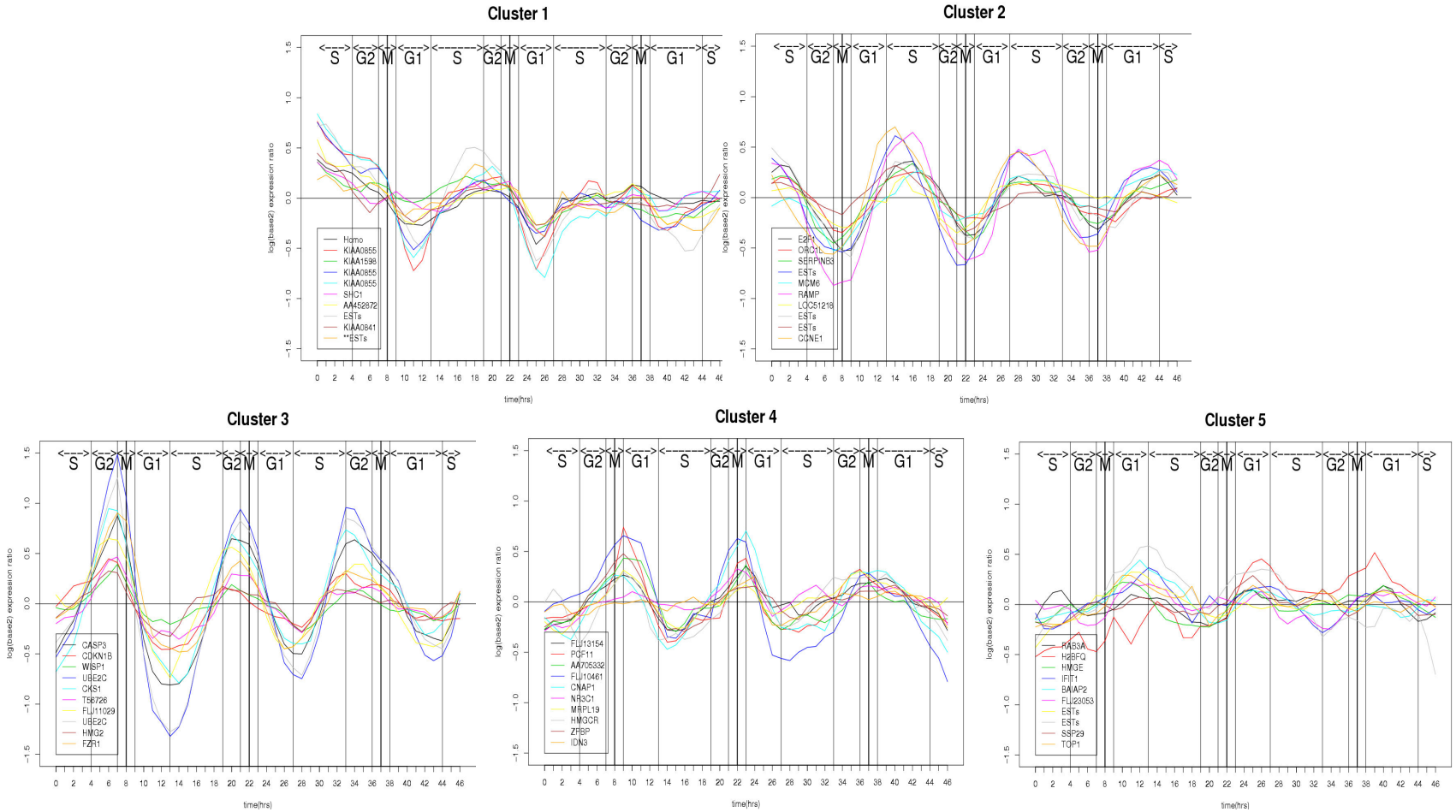
$n* = 4, k* = 5.7$: essentially the behavior separate well genes in 4 clusters
Average Silhouette width=0.4: the clustering structure is reasonable

Identification of the well-characterized genes

-Extract a kernel set of the $n = 10$ first genes maximizing the silhouette width

Cluster Number	Gene Name	Whitfield Assignment	Gene Type	Neighbor Cluster	Silhouette width (sw)	High peaked phase
1	Homo	S	K	2	0.806	S
	KIAA0855	S	K	3	0.697	
	KIAA1598	S	K	2	0.688	
	KIAA0855	S	K	2	0.686	
	KIAA0855	S	K	3	0.681	
	SHC1	S	K	2	0.677	
	AA452872	S	K	3	0.674	
	ESTs	S	K	3	0.665	
	KIAA0841	S	K	2	0.658	
	**ESTs	S	K	3	0.635	
	RRM2	S	R	2	0.586	
	DHFR	S	R	2	0.315	
	RAD51	S	R	3	0.238	
2	E2F1*	G_1/S	K	1	0.832	G_1/S
	ORC1L	G_1/S	K	1	0.829	
	SERPINB3	G_1/S	K	1	0.82	
	ESTs	G_1/S	K	1	0.812	
	MCM6	G_1/S	K	1	0.812	
	RAMP	G_1/S	K	1	0.812	
	LOC51218	G_1/S	K	1	0.802	
	ESTs	G_1/S	K	1	0.794	
	ESTs	G_1/S	K	1	0.794	
	CCNE1	G_1/S	K/R	5	0.786	
	E2F1	G_1/S	R	1	0.775	
	CDC6	G_1/S	R	1	0.682	
	PCNA	G_1/S	R	1	0.625	
	RFC4	S	R	1	0.526	
3	CASP3	G_2	K	4	0.811	G_2/M
	CDKN1B	G_2	K	4	0.807	
	WISP1	G_2	K	4	0.799	
	UBE2C	G_2	K	4	0.788	
	CKS1	G_2	K	4	0.784	
	T56726	G_2	K	4	0.779	
	FLJ11029	G_2	K	1	0.779	
	UBE2C	G_2	K	4	0.779	
	HMG$_2$	G_2	K	4	0.768	
	FZR1	G_2	K	4	0.765	
	CCNF	G_2	R	4	0.757	
	TOP2A	G_2	R	4	0.669	
	CDC2	G_2	R	1	0.618	
	STK15	G_2/M	R	4	0.478	
	CCNA2	G_2	R	4	0.458	
4	FLJ13154	M/G_1	K	3	0.737	M/G_1
	PCF11	M/G_1	K	5	0.717	
	AA705332	G_2/M	K	5	0.695	
	FLJ10461	G_2/M	K	3	0.651	
	CNAP1	G_2/M	K	3	0.599	
	NR3C1	G_2	K	3	0.593	
	MRPL19	M/G_1	K	3	0.585	
	HMGCR	M/G_1	K	3	0.579	
	ZPBP	M/G_1	K	3	0.578	
	IDN3	G_2	K	3	0.576	
	RAD21	M/G_1	R	3	0.433	
	CDKN3	M/G_1	R	3	0.320	
	PTTG$_1$	M/G_1	R	5	0.282	
	BUB1	G_2/M	R	3	0.184	
	VEGFC	M/G_1	R	3	0.148	
	CCNB1	G_2/M	R	3	0.095	
	PLK	G_2/M	R	3	0.003	
5	RAB3A	M/G_1	K	2	0.561	G_1
	H2BFQ	M/G_1	K	2	0.502	
	HMGE	M/G_1	K	4	0.489	
	IFIT1	M/G_1	K	2	0.484	
	BAIAP2	G_1/S	K	2	0.478	
	FLJ23053	G_1/S	K	2	0.475	
	ESTs	M/G_1	K	4	0.429	
	ESTs	G_1/S	K	2	0.407	
	SSP29	G_2/M	K	4	0.398	
	TOP1	M/G_1	K	4	0.394	

Identify the cell cycle phase of each kernel set



k*=5.9, $(S, G_1/S, G_2/M, M/G_1, G_1)$

Assignement of boundary genes

The kernel sets of the partition $P_{n*=4,k=5.7}$

Name	Ada-Assi	Neig	Sw	Final-Assi
MZF1	S	G_1/S	0.049	begin of S
TncRNA	S	G_1/S	0.049	begin of S
CAPS	S	G_1/S	0.041	begin of S
AURKB	S	G_2/M	0.039	G_2
ZFX	S	G_2/M	0.038	G_2
KATNA1	S	G_2/M	0.028	G_2
KBTBD2	S	G_2/M	0.026	G_2
CDKL5	S	G_2/M	0.02	G_2
TTC31	S	G_1/S	0.013	begin of S
LOC134121	S	G_2/M	0.012	G_2
UBL3	S	G_1	0.011	G_1/S
CDKN2C	S	G_2/M	0	G_2
REEP1	S	G_1/S	-0.012	begin of S
TOP2A	S	G_2/M	-0.023	G_2
MICA/HCP5	S	G_2/M	-0.039	G_2
CDH24	S	G_1/S	-0.041	begin of S
ABCC5	S	G_1/S	-0.044	begin of S
RECQL4	G_1/S	S	0.047	begin of S
SLC9A3	G_1/S	S	0.046	begin of S
FLJ13231	G_1/S	S	0.028	begin of S
ESTs	G_1/S	S	0.001	begin of S
EST	G_1/S	G_1	-0.019	end of G_1
HIST1H2AM	G_1/S	S	-0.045	begin of S
BAIAP2	G_1/S	G_1	-0.045	end of G_1
CRLF3	G_1/S	S	-0.05	begin of S

Name	Ada-Assi	Neig	Sw	Final-Assi
NR5A2	G_2/M	G_1	0.05	M
HERPUD2	G_2/M	G_1	0.045	M
AMD1	G_2/M	G_1	0.035	M
NIPBL	G_2/M	G_1	0.012	M
NFIC	G_2/M	G_1	0.008	M
ESTs	G_2/M	G_1	0.006	M
ChGn	G_2/M	G_1	0.003	M
BCLAF1	G_2/M	S	0,001	G_2
WWC1	G_2/M	G_1	-0.003	M
HLA-DOA	G_2/M	G_1	-0.012	M
AGPAT3	G_2/M	G_1	-0.015	M
C20orf199	G_2/M	G_1	-0.017	M
SLC39A10	G_2/M	G_1	-0.02	M
LARP1	G_2/M	G_1	-0.024	M
ANP32B	G_2/M	G_1	-0.026	M
ABHD10	G_2/M	S	-0.029	G_2
FXR1	G_2/M	G_1	-0.032	M
ZNF207	G_1	G_2/M	0.05	M
HSPA2	G_1	G_2/M	0.048	M
PPP2CA	G_1	G_2/M	0.044	M
CEP350	G_1	G_2/M	0.017	M
OC146517	G_1	G_2/M	0.013	M
SAP30BP	G_1	S	0.009	G_1/S
DR1	G_1	G_2/M	0.007	M
TMEM132A	G_1	G_2/M	0.002	M
W85890	G_1	G_2/M	-0.007	M
PCF11	G_1	G_2/M	-0.021	M
DNAJA1	G_1	G_2/M	-0.022	M
TSC22	G_1	G_2/M	-0.023	M
EST	G_1	G_2/M	-0.024	M

Metrics efficiency comparision: Random-Periods model for periodically expressed genes

The sinusoid function characterizing the expected periodic expression of a cell-cycle gene g (Liu et al. (2004)):

$$f(t, \theta_g) = a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} cos(\frac{2\pi t}{Texp(\sigma z)} + \Phi_g) exp(-\frac{z^2}{2}) dz,$$

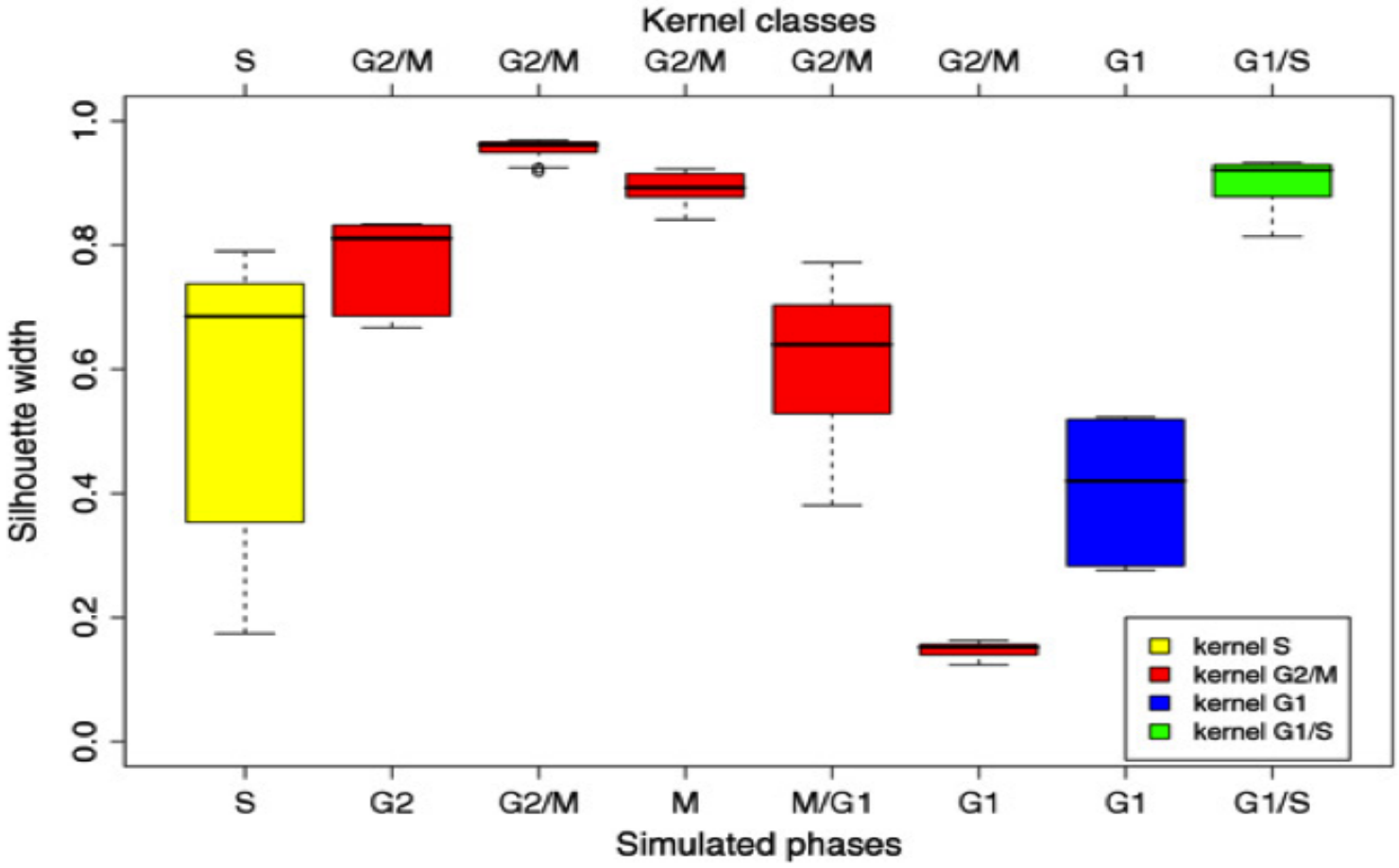
where θ_g is explicitly $(K_g, T, \sigma, \Phi_g, a_g, b_g)$, specific to each gene g

- K_g: initial amplitude of the periodic expression pattern
- T: cell-cycle duration
- σ: governs the rate of attenuation in amplitude
- Φ_g: corresponds to the cell-cycle phase during which the gene undergoes its peak level of transcription
- a_g and b_g: account for any drift (intercepts and slopes, respectively) in a gene's background expression level

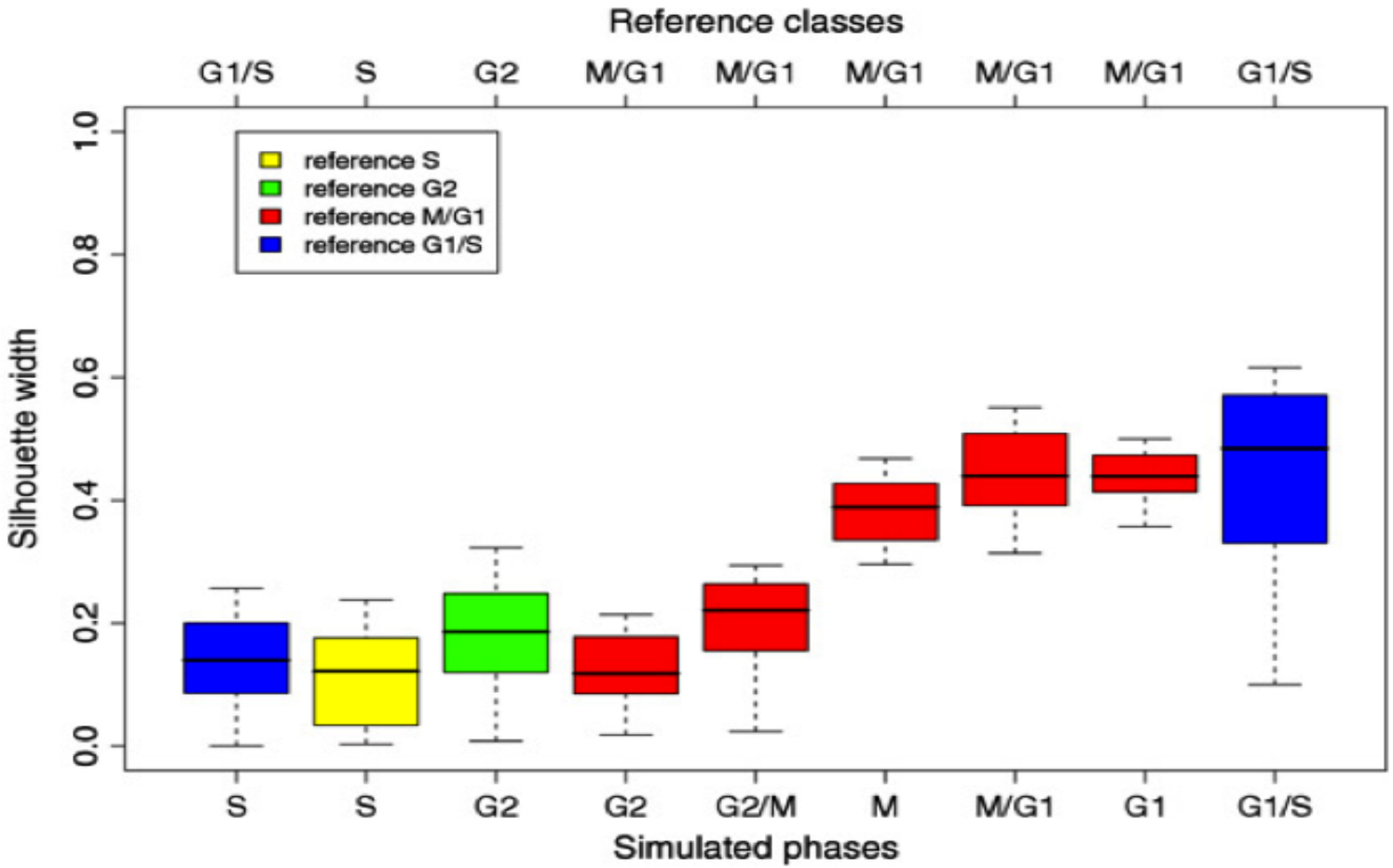
Dissimilarity efficiency for classifying gene expression profiles

Simulated classes	Kernel classes				Reference classes				
	S	G_2/M	G_1	G_1/S	S	G_2	G_2/M	M/G_1	G_1/S
S	100	0	0	0	29	0	0	0	**71**
G_2	0	100	0	0	0	65	0	**35**	0
G_2/M	0	100	0	0	0	0	0	**100**	0
M	0	100	0	0	0	0	0	100	0
M/G_1	0	100	0	0	0	0	0	100	0
G_1	0	**27**	73	0	0	0	0	100	0
G_1/S	0	0	0	100	0	0	0	0	100

Dissimilarity efficiency for classifying gene expression profiles



Dissimilarity efficiency for classifying gene expression profiles

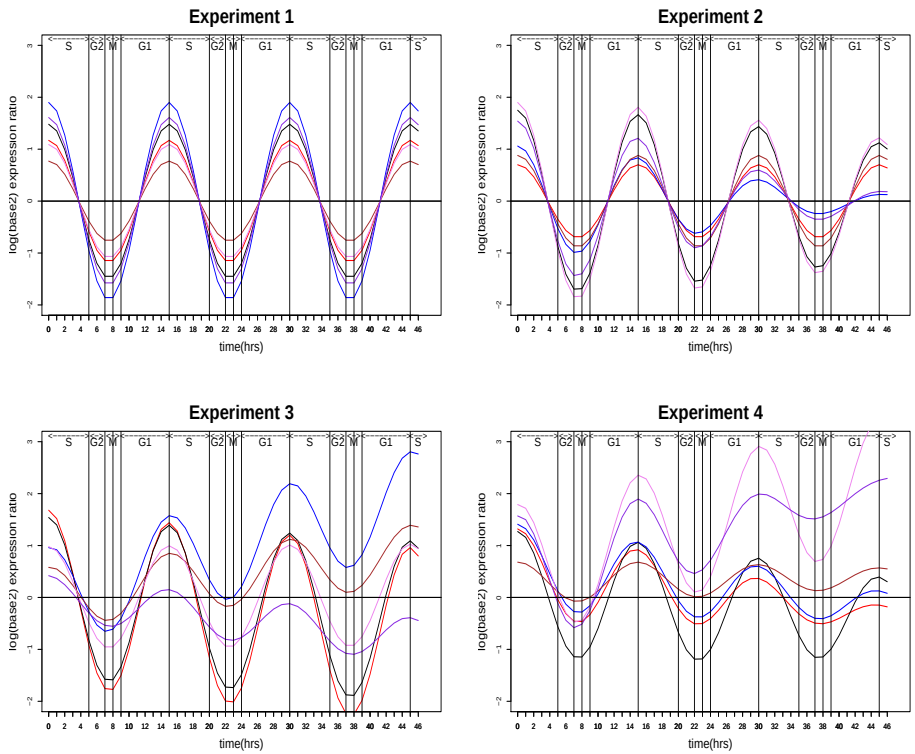


Validation Protocol

- A simulation study based on 20000 genes expression profiles,
- genes are equally generated from 5 classes (five cell-cycle phases)
- each gene expression is observed through 3 cell-cycles on 47 instants,
- Four experiments are simulated (500 genes /experiment)
- 10 samples are generated for each experiment

T=15, Φ_g=(0, 5.190, 3.823, 3.278, 2.459)

Experiment number	K_g	σ	b_g	a_g
1	[0.34, 1.33]	0	0	0
2	[0.34, 1.33]	[0, 0.115]	0	0
3	[0.34, 1.33]	0	[-0.05, 0.05]	[0, 0.8]
4	[0.34, 1.33]	[0, 0.115]	[-0.05, 0.05]	[0, 0.8]



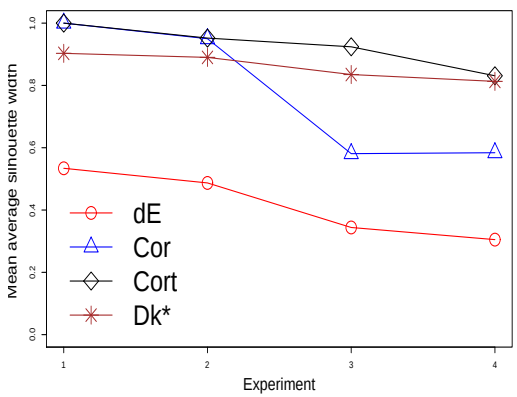
Proximity measures efficiency for clustering genes expression profiles

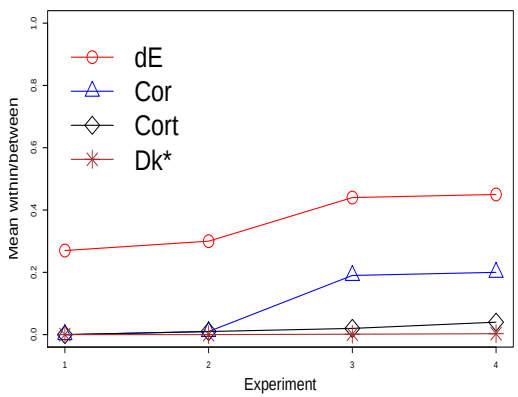
For each experiment $j \in \{1, 2, 3, 4\}$ and for each measure δ_E, COR, and CORT

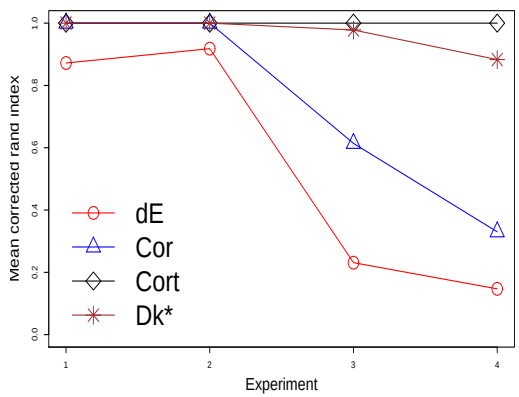
- a *PAM* algorithm is used to partition each sample S_{ij}, $i \in \{1, ..., 10\}$ into 5 clusters (5 cell cycle phases)
- Three goodness criteria: the average silhouette width (asw), the within/between ratio (wbr), and the corrected Rand index (RI)

For the adaptive dissimilarity D_k

- a PAM algorithm is performed for k varying in $[0, 6]$,
- select P^{ij}_{k*}, with $k* = argmax_k(avgSil(P^{ij}_k))$,







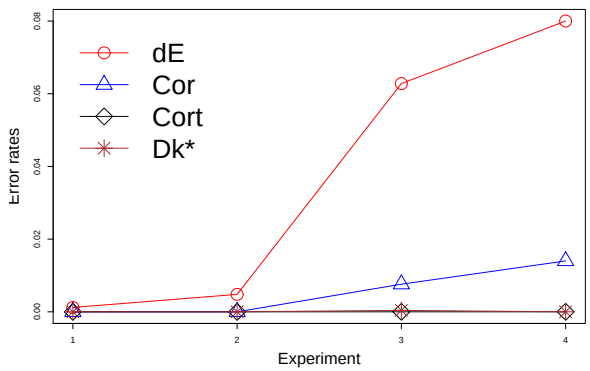
Proximity measure efficiency for classifying gene expression profiles

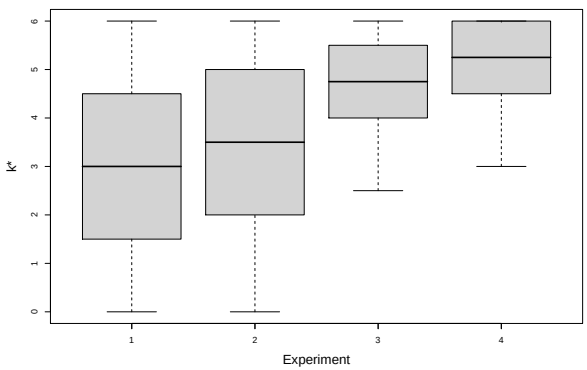
For each experiment $j \in \{1, 2, 3, 4\}$ and for each measure δ_E, COR, and CORT

- A 10-NN algorithm is performed to classify each sample S_{ij}, $i \in \{1, ..., 10\}$
- The misclassification error rate is evaluated

For the adaptive dissimilarity D_k

- the 10-NN algorithm is performed for k varying in $[0, 6]$,
- C^{ij}_{k*} is selected with $k* = argmin_k(ErrorRate(C^{ij}_k))$,





References

- Liu, D., Umbach, D.M., Peddada, S.D., Li, L., Crockett, P.W., Weinberg, C.R., 2004. A random-periods model for expression of cell-cycle genes. Proceedings of the National Academy of Sciences of the USA 101, 72407245.
- Liu, X., Lee, S., Casella, G., Peter, G.F., 2008. Assessing agreement of clustering methods with gene expression microarray data. Computational Statistics and Data Analysis 52 (12), 53565366.
- Maharaj, E.A., 2000. Cluster of time series. Journal of Classification 17, 297314.
- Oates, T., Firoiou, L., Cohen, P.R., 1999. Clustering time series with Hidden Markov models and dynamic time warping. In: Proceedings of the 6th IJCAI-99, Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning, Stockholm, pp. 1721.
- Peddada, S.D., Lobenhofer, L., Li, L., Afshari, C., Weinberg, C., Umbach, D., 2003. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19, 834841.
- Pihur, V., Datta, S., Datta, S., 2007. Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach. Bioinformatics 23 (13), 16071615.

-Whitfield, M.L., Sherlock, G., Murray, J.I., Ball, C.A., Alexander, K.A., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., Botstein, D., 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors molecular. Biology of the Cell 13, 19772000.