

A Comparison of Clustering Techniques for Gene Expression Data

Michiel de Hoon, Seiya Imoto, Satoru Miyano
Human Genome Center, University of Tokyo



The 10th International Conference on
Intelligent Systems for Molecular Biology
August 3 – 7, 2002, Edmonton, Canada

Genes / microarrays can be clustered based on the similarity between their expression profiles



What do we mean by similarity?

Several distance measures are commonly used.

- **Unnormalized distance functions:**

- Minkowski metrics: Euclidean distance, city-block distance, ...
 - Mahalanobis distance

- Requires some normalization of the expression data

- **Normalized distance functions:**

- Usually based on some correlation coefficient r via $d = 1 - r$

- Pearson correlation, Spearman rank correlation, Kendall's τ

- Can be applied to normalized and unnormalized data

- Information about the magnitude of changes in the gene expression is ignored

- These distances do not satisfy the triangle inequality!

- **Intermediate:**

- Angle between expression data vectors (“uncentered correlation” *)

* M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, “Cluster analysis and display of genome-wide expression patterns”, *Proc. Natl. Acad. Sci. USA*, **Vol. 95**, pp. 14863-14868, 1998

Which distance function should be used?

This is currently an unanswered question.



Pearson:

The Pearson correlation coefficient is used most commonly:

Alon *et al.*, PNAS **96**, 6745-6750 (1999);

Heyer *et al.*, Genome Research **9**, 1106-1115 (1999);

Ewing & Claverie, Proc. PSB **5**, 430-441 (2000);

Zhu & Zhang, Proc. PSB **5**, 479-490 (2000).

The angle between expression data vectors
("uncentered correlation") is also widely used:
Eisen *et al.*, PNAS **95**, 14863-14868 (1998).



Euclid:

However, the Euclidean distance has been reported to perform better:

Slonim *et al.*, RECOMB, 263-272 (2000).

Three types of clustering algorithms are commonly used



- Hierarchical clustering:

- Initially each gene forms one cluster

- On each step, join the two closest clusters

- Continue until only one cluster remains

- Draw a tree diagram

- Several types of hierarchical clustering exist:

- Simple linkage, maximum linkage, centroid linkage, average linkage

- *k*-means clustering:

- Create *k* clusters; assign genes randomly to a cluster

- Iterate: reassign genes to the cluster whose centroid is the closest

- Continue until no further reassignment takes place

- This algorithm should be repeated many times to find the optimal solution.

- The optimal solution has the smallest sum of distances to the cluster centroids.

- Self-Organizing Maps:

- Based on a topology of clusters

Which clustering algorithm should be used?

Various clustering programs are available



- Cluster/TreeView (Michael Eisen, Berkeley/Stanford):

Widely used, especially by biologists

Mainly focused on hierarchical clustering (simple, maximum, centroid linkage)

but also includes *k*-means clustering and 1D Self-Organizing Maps

Good documentation

Recently, the source code (C++) was (partially) released.

We have modified Cluster/TreeView to include an improved *k*-means clustering algorithm.

Available from our website (<http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/index.html>)

- Tamayo's GeneCluster program:

Offers 2D Self-Organizing Maps

Written in Java with nice graphics

The source code is not available: impossible to check the algorithm or to improve the code

- The C Clustering Library:

Numerical routines for hierarchical and *k*-means clustering, as well as 2D Self-Organizing Maps.

Released under the GNU Lesser Public License. Also available for use with Python.

- Statistical programs such as S and R

Different clustering methods produce different clustering solutions



We consider a real-life example to compare different clustering methods

Gene expression levels were measured for 66 patients, each afflicted by one of five disease types:

- A (14 patients)
- B (7 patients)
- C (7 patients)
- D (10 patients)
- E (28 patients)

We only include genes that satisfy the following two conditions:

- The gene expression levels deviate by at least one standard deviation;
- Less than 10% of the data values are missing.

This resulted in 1224 remaining genes.

Pairwise **centroid** linkage clustering with the **Euclidean** distance produces one large cluster ...



Cluster I

E

Cluster II

E

Cluster IV

E

Cluster V

B

Cluster III

A	A	B	C	D	D	E	E	E	E
A	A	B	C	D	D	E	E	E	E
A	A	B	C	D	D	E	E	E	E
A	A	B	C	D	D	E	E	E	E
A	A	B	C	D	D	E	E	E	E
A	A	B	C			E	E	E	E
A	A		C						E

Can we learn anything from this clustering solution?

... while pairwise **complete** linkage clustering with the **Euclidean** distance performs slightly better



Cluster I

E

Cluster II

E

Cluster IV

E

Cluster III

A	A	B	C	D	D	E	E
A	A	B	C	D	D	E	E
A	A	B	C	D	D	E	E
A	A	B	C	D	D	E	E
A	A	B	C	D	D	E	E
A	A	B				E	E
A						E	E

Cluster V

A	E	E
B	E	E
C	E	E
C	E	E
	E	E
	E	

... and **k-means** clustering with the **Euclidean** distance performing better still



Cluster I

A	A	C
A	A	C
A	A	C
A	A	D
A	A	E
A	A	

Cluster II

E
E

Cluster IV

C
E
E

Cluster III

A	C	E	E	E
B	C	D	E	E
B	D	D	E	E
B	D	D	E	E
B	D	D	E	E
B	D	D	E	E
B			E	E

Cluster V

A	E	E
B	E	E
C	E	E
	E	E

k-means with the Euclidean distance; 100,000 trials



Pairwise **complete** linkage clustering with the **uncentered correlation** gives a better result ...



Cluster I

A	A	B
A	A	B
A	A	B
A	A	D
A	A	

Cluster III

C	D	E	E	E
C	D	E	E	E
C	D	E	E	E
C	D	E	E	A
C	D	E	E	A
	D	E	E	B

Cluster II

D	E	A
D	E	A
D	E	C

Cluster IV

B	E
B	E
B	E
C	E

Cluster V

E
E
E
E
E
E



... but again **k-means** gives the clearest result

Cluster I

A	A
A	A
A	A
A	A
A	A
A	A
B	

Cluster III

C	E
C	E
C	E
C	E
C	E
A	E
A	E
A	E
A	B

Cluster IV

E	E	E
E	E	E
E	E	A
E	E	B
E	E	B
E	E	C

Cluster V

D
D
D
D
D
B

Cluster II

D	E	B
D	E	C
D	E	C
D	E	C
D	E	C

k-means with the uncentered correlation

*Note: Finding this solution takes 100,000 trials!
(approximately twelve hours on a PC)*

This is similar to the result found with the
centered correlation using *k*-means



Cluster I

A	A
A	A
A	A
A	A
A	A
B	

Cluster III

C	E	E
C	E	E
A	E	E
B	E	E
B	E	E
E	E	E

Cluster IV

E	E	A
E	E	A
E	E	C
E	B	C

Cluster V

D
D
D
D
B
E

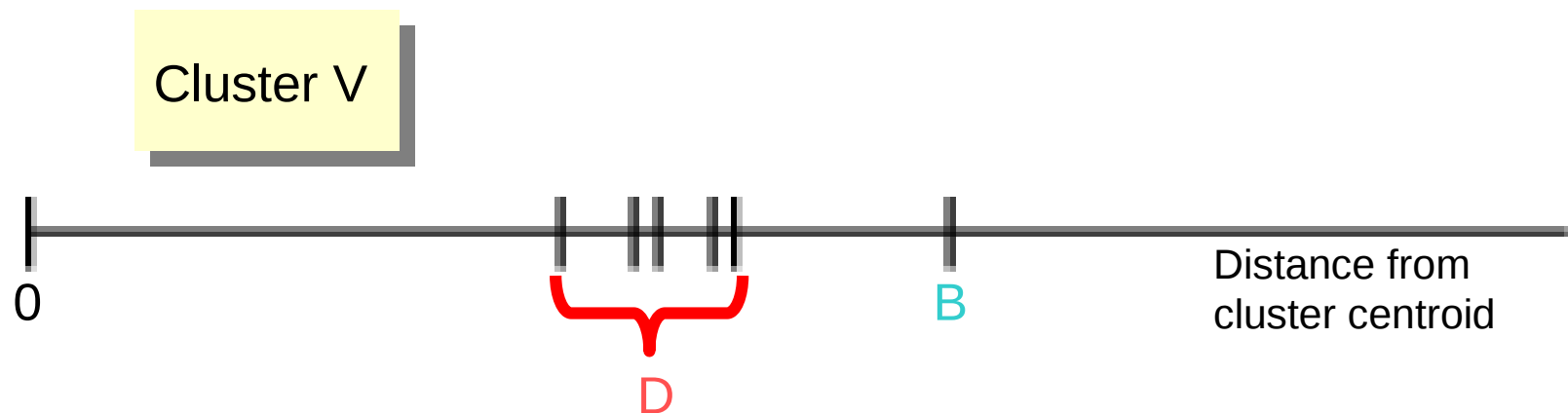
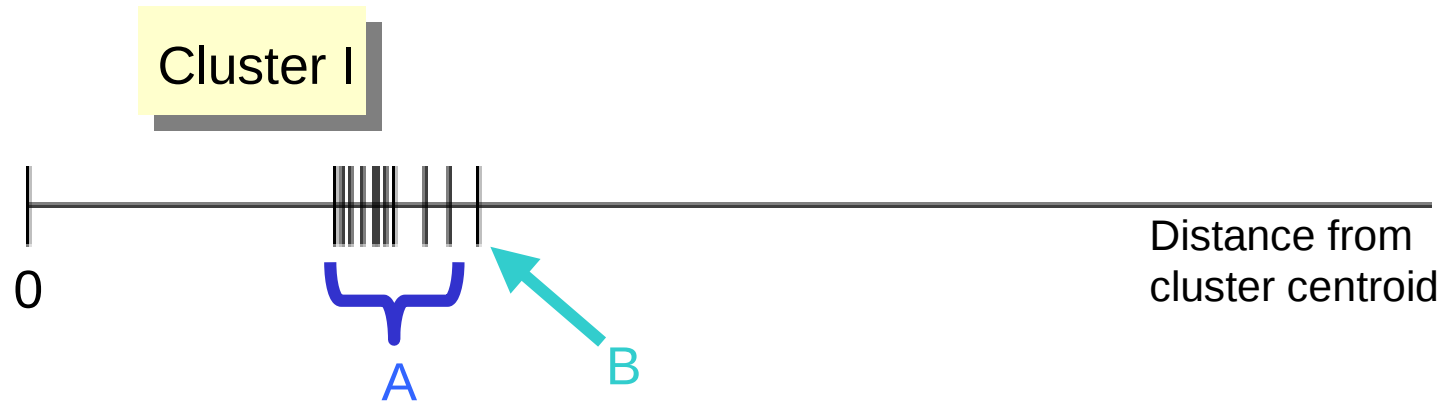
Cluster II

D	E	A
D	E	B
D	E	C
D	E	C
D	E	C
D	E	
	E	

k-means with the centered correlation

*Note: Finding this solution takes 100,000 trials!
(approximately twelve hours on a PC)*

We can understand the *k*-means result better by looking at the distances to the cluster centroid



Our second example considers gene clustering



Cyanobacterium *Synechocystis* sp PCC 6803

(Cyanobase website; Kazusa DNA Research Institute)

Analysis of Cyanobacterial Gene Expression during Acclimation to High Light

Yukako Hihara, Ayako Kamei, Minoru Kanehisa, Aaron Kaplan, Masahiko Ikeuchi

The Plant Cell, Vol. 13, 793-806, April 2001

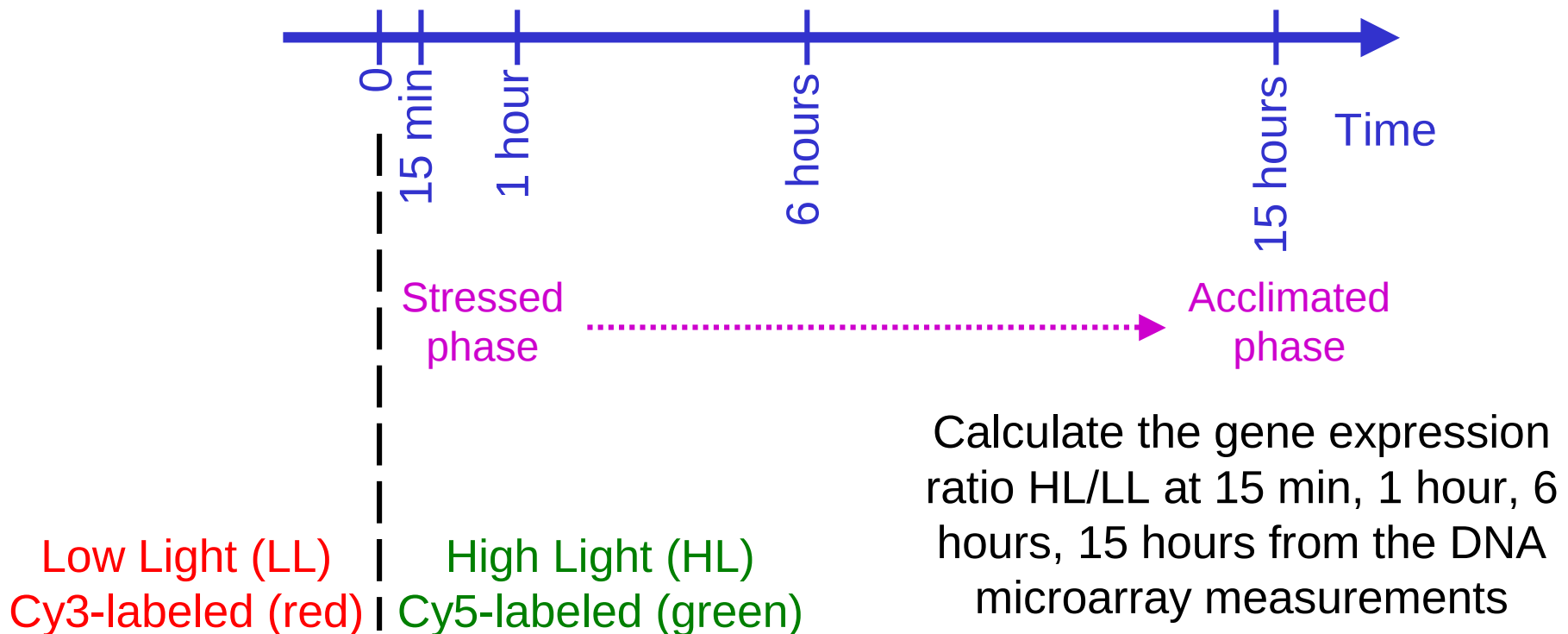
Acclimation from low light to high light:

- Short-term processes take place within several minutes
(State transitions, protective energy dissipation, changes in the efficiency of energy transfer from the harvesting complex to photosystem II, formation of nonfunctional PII reaction centers)
- Long-term processes take place in hours or days
(Changes in the composition, function, and structure of the photosynthetic apparatus and other photosynthesis-related components)

The time-dependent response of Cyanobacterium to high light was measured



Change the light intensity from low to high and examine the response as a function of time



Preprocessing gene expression data



Remove noisy genes

Expression levels close to the background level cannot be distinguished from noise and should therefore not be included in the analysis.

Apply global normalization

Scale the cy5 and cy3 levels by dividing them by their sum over all genes. This avoids errors due to different rates of hybridization.

Remove genes that fail the *t*-test

Perform a *t*-test for each gene to see if its expression ratio is significantly different from unity. If not, that gene was not significantly affected by the experimental manipulation and should be removed from the analysis.

Fit linear spline functions to the gene expression levels

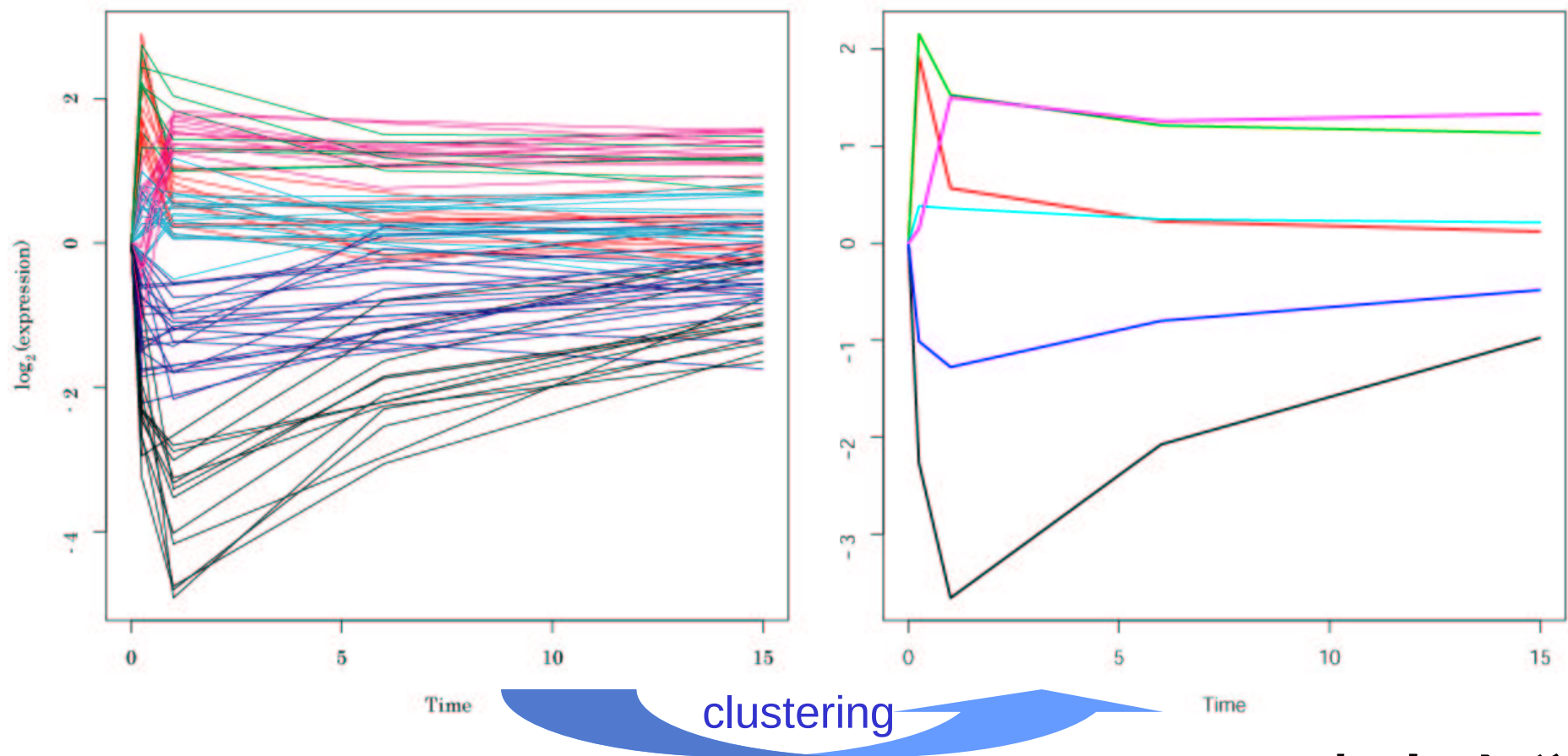
Using a linear spline function instead of the expression data directly reduces the effect of noise present in the data.

These steps have been implemented in Python/C routines.

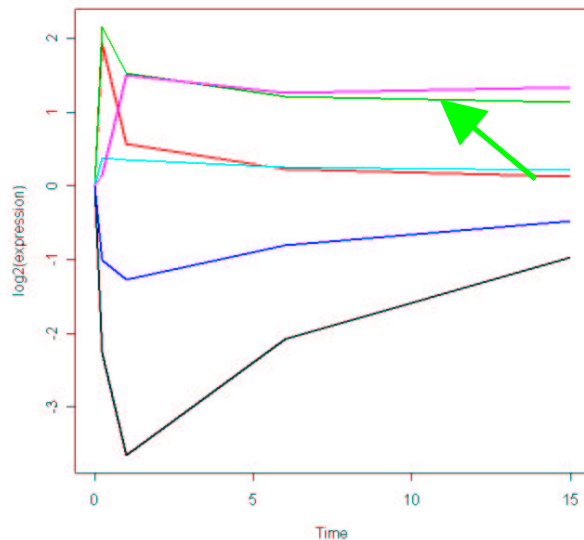
Clustering reveals patterns of gene expression



k-means clustering, using Euclidean distance



k-means clustering using Euclidean distance: Cluster I



sll0322	<i>hypF</i>	transcriptional regulatory protein HypF
sll0927	<i>metX</i>	S-adenosylmethionine synthetase

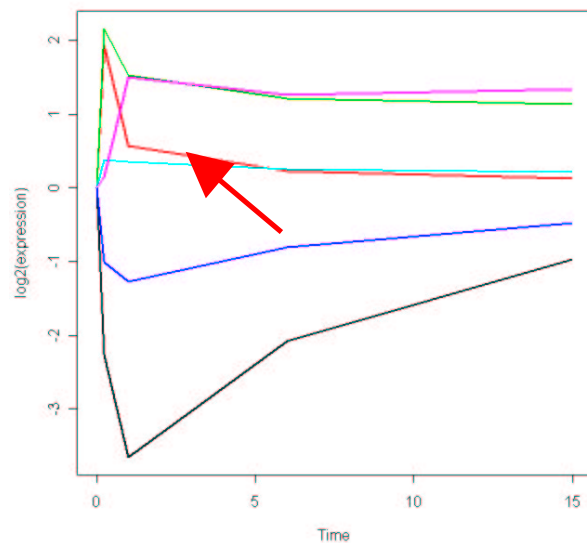
sll1028	<i>ccmK</i>	carbon dioxide concentrating mechanism protein
sll1029	<i>ccmK</i>	carbon dioxide concentrating mechanism protein
sll1031	<i>ccmM</i>	carbon dioxide concentrating mechanism protein

slr1963		hypothetical protein
---------	--	----------------------

slr2075	<i>groES</i>	10kD chaperonin
sll0416	<i>groEL-2</i>	60kD chaperonin 2

10,000 repetitions; solution found 111 times

k-means clustering using Euclidean distance: Cluster II

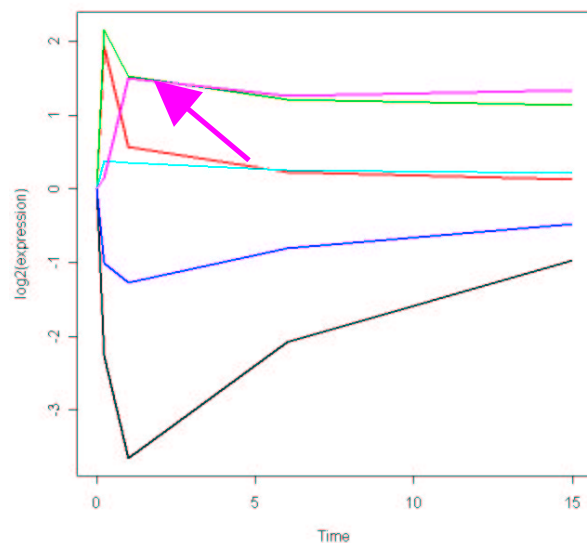


slI0170	<i>dnaK</i>	DnaK protein
slI0430	<i>htpG</i>	heat shock protein

slI0521	<i>ndhG</i>	NADH dehydrogenase subunit 6
slr1280	<i>ndhK</i>	NADH dehydrogenase subunit NdhK

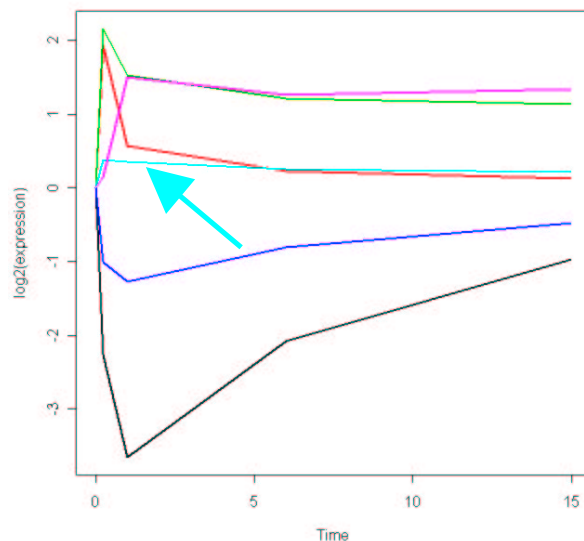
slr0011	<i>rbcX</i>	unknown function
slr1350	<i>desA</i>	fatty acid desaturase
slr1604	<i>ftsH</i>	cell division protein
slr1641	<i>clpB</i>	ClpB protein
slI0814		hypothetical protein
slr0476		hypothetical protein
slr1687		hypothetical protein

k-means clustering using Euclidean distance: Cluster III



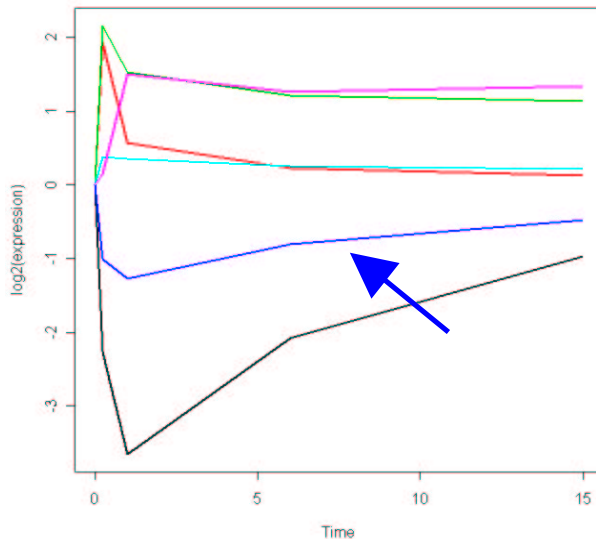
sll1096	<i>rps12</i>	30S ribosomal protein S12
sll1097	<i>rps7</i>	30S ribosomal protein S7
sll1260	<i>rps2</i>	30S ribosomal protein S2
sll1743	<i>rpl11</i>	50S ribosomal protein L11
sll1745	<i>rpl10</i>	50S ribosomal protein L10
sll1799	<i>rpl3</i>	50S ribosomal protein L3
sll1801	<i>rpl23</i>	50S ribosomal protein L23
sll1802	<i>rpl2</i>	50S ribosomal protein L2
sll1804	<i>rps3</i>	30S ribosomal protein S3
sll1807	<i>rpl24</i>	50S ribosomal protein L24
sll1809	<i>rps8</i>	30S ribosomal protein S8
sll1810	<i>rpl6</i>	50S ribosomal protein L6
sll1816	<i>rps13</i>	30S ribosomal protein S13

k-means clustering using Euclidean distance: Cluster IV



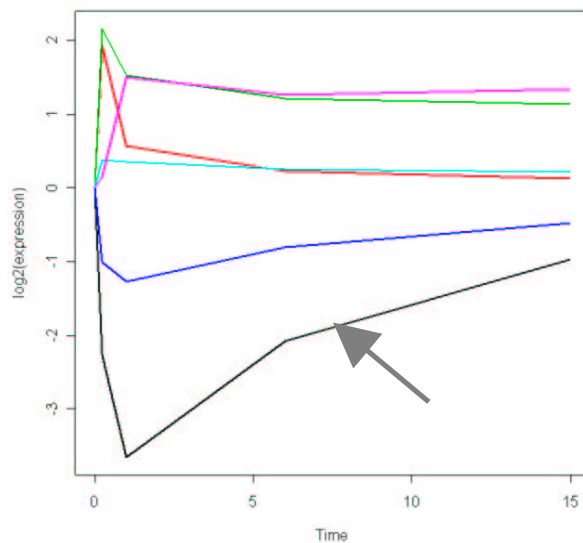
sll0020	<i>clpC</i>	ATP-dependent Clp protease regulatory subunit
sll0144	<i>pyrH</i>	uridine monophosphate kinase
sll0185		unknown
sll0262	<i>desD(des6)</i>	delta-6 desaturase
sll0414		hypothetical protein
sll0519	<i>ndhA</i>	NADH dehydrogenase subunit 1
sll0587	<i>pykF</i>	pyruvate kinase
sll0680	<i>pstS</i> or <i>phoS</i>	phosphate-binding periplasmic protein precursor (PBP)
sll0854		unknown
sll1002	<i>ycf22</i>	hypothetical protein
sll1327	<i>atpC</i>	ATP synthase g subunit
sll1614	<i>pma1</i>	cation-transporting ATPase (E1-E2 ATPase)
slr0208		unknown
slr0452	<i>ilvD</i>	dihydroxyacid dehydratase
slr0642		integral membrane protein
slr0839	<i>hemH</i>	ferrochelatase
slr1237	<i>codA</i>	cytosine deaminase
slr1469	<i>rnpA</i>	protein subunit of ribonuclease P (RNase P)
slr1513		unknown
slr1557		unknown
slr1718		unknown
ssr2554		unknown

k-means clustering using Euclidean distance: Cluster V



sll0258	<i>psbV</i>	cytochrome c550
sll0617	<i>im30</i>	chloroplast membrane-associated 30 kD protein
sll0851	<i>psbC</i>	photosystem II CP43 protein
sll0901	<i>purE</i> or <i>ade6</i>	phosphoribosylaminoimidazole carboxylase
sll1091		391 aa (43 kD) bacteriochlorophyll synthase subunit
sll1214	<i>PNIL34</i> or <i>AT103</i>	phytochrome-regulated gene
sll1306		hypothetical protein
sll1471	<i>cpcG</i>	phycobilisome rod-core linker polypeptide CpcG
sll1626	<i>lexA</i>	SOS function regulatory protein
sll1694	<i>hofG</i> or <i>hopG</i>	general secretion pathway protein G
sll1712		DNA binding protein HU
ssl3093	<i>cpcD</i>	phycocyanin associated linker protein
slr0151		hypothetical protein
slr0329	<i>xylR</i>	xylose repressor
slr0374		cell division cycle protein
slr1128		erthyrocyte band 7 integral embrane protein, protein 7.2B, stomatin
slr1348	<i>cysE</i>	serine acetyltransferase
slr1459	<i>apcF</i>	phycobilisome core component
slr1545	<i>rpoE</i>	RNA polymerase sigma-E factor
slr1793	<i>talB</i>	transaldolase
slr1853		hypothetical protein
slr1856		anti-sigma B factor antagonist
ssl1533		hypothetical protein

k-means clustering using Euclidean distance: Cluster VI



slr1577	<i>cpcB</i>	phycocyanin b subunit
slr1578	<i>cpcA</i>	phycocyanin a subunit
slr1579	<i>cpcC</i>	phycocyanin associated linker protein
slr1580	<i>cpcC</i>	phycocyanin associated linker protein
slr2051	<i>cpcG</i>	phycobilisome rod-core linker polypeptide
slr0819	<i>psaF</i>	photosystem I subunit III
slr0737	<i>psaD</i>	photosystem I subunit II
slr1655	<i>psaL</i>	photosystem I subunit XI
slr1834	<i>psaA</i>	P700 apoprotein subunit Ia
slr1835	<i>psaB</i>	P700 apoprotein subunit Ib
slr1855		hypothetical protein
slr1986	<i>apcB</i>	allophycocyanin b chain
slr2067	<i>apcA</i>	allophycocyanin a chain

Note: The Euclidean distance distinguishes cluster V and VI.

k-means clustering using Pearson Correlation: Cluster I



sll0262	<i>desD(des6)</i>	delta-6 desaturase (cluster IV)
sll0322	<i>hypF</i>	transcriptional regulatory protein HypF
sll0854		unknown (cluster IV)
sll0927	<i>metX</i>	S-adenosylmethionine synthetase
sll1029	<i>ccmK</i>	carbon dioxide concentrating mechanism protein
sll1031	<i>ccmM</i>	carbon dioxide concentrating mechanism protein
slr0208		unknown (cluster IV)
slr1718		unknown (cluster IV)
slr1963		hypothetical protein
slr2075	<i>groES</i>	10kD chaperonin
sll0416	<i>groEL-2</i>	60kD chaperonin 2

1,000,000 repetitions; solution found 37 times

k-means clustering using Pearson correlation:

Cluster II



sll0020	<i>clpC</i>	ATP-dependent Clp protease regulatory subunit (cluster IV)
sll0170	<i>dnaK</i>	DnaK protein
sll0414		hypothetical protein (cluster IV)
sll0430	<i>htpG</i>	heat shock protein

sll0519	<i>ndhA</i>	NADH dehydrogenase subunit 1 (cluster IV)
sll0521	<i>ndhG</i>	NADH dehydrogenase subunit 6
slr1280	<i>ndhK</i>	NADH dehydrogenase subunit NdhK

sll1002	<i>ycf22</i>	hypothetical protein (cluster IV)
sll1028	<i>ccmK</i>	carbon dioxide concentrating mechanism protein (cluster I)
sll1614	<i>pma1</i>	cation-transporting ATPase (E1-E2 ATPase) (cluster IV)
slr0011	<i>rbcX</i>	unknown function
slr0642		integral membrane protein (cluster IV)
slr1350	<i>desA</i>	fatty acid desaturase
slr1513		unknown (cluster IV)
slr1604	<i>ftsH</i>	cell division protein
slr1641	<i>clpB</i>	ClpB protein
sll0814		hypothetical protein
slr0476		hypothetical protein
slr1687		hypothetical protein

k-means clustering using Pearson correlation:

Cluster III



sll0144	<i>pyrH</i>	uridine monophosphate kinase (cluster IV)
sll0587	<i>pykF</i>	pyruvate kinase (cluster IV)
sll0680	<i>pstS</i> or <i>phoS</i>	phosphate-binding periplasmic protein precursor (PBP) (cluster IV)
sll1327	<i>atpC</i>	ATP synthase g subunit (cluster IV)
sll1096	<i>rps12</i>	30S ribosomal protein S12
sll1097	<i>rps7</i>	30S ribosomal protein S7
sll1260	<i>rps2</i>	30S ribosomal protein S2
sll1743	<i>rpl11</i>	50S ribosomal protein L11
sll1745	<i>rpl10</i>	50S ribosomal protein L10
sll1799	<i>rpl3</i>	50S ribosomal protein L3
sll1801	<i>rpl23</i>	50S ribosomal protein L23
sll1802	<i>rpl2</i>	50S ribosomal protein L2
sll1804	<i>rps3</i>	30S ribosomal protein S3
sll1807	<i>rpl24</i>	50S ribosomal protein L24
sll1809	<i>rps8</i>	30S ribosomal protein S8
sll1810	<i>rpl6</i>	50S ribosomal protein L6
sll1816	<i>rps13</i>	30S ribosomal protein S13
slr1237	<i>codA</i>	cytosine deaminase (cluster IV)
slr1469	<i>rnpA</i>	protein subunit of ribonuclease P (RNase P) (cluster IV)
slr1557		unknown (cluster IV)
ssr2554		unknown (cluster IV)

k-means clustering using Pearson correlation: Cluster IV



sll0185		unknown
slr0839	<i>hemH</i>	ferrochelatase

k-means clustering using Pearson correlation: Cluster V



sll0617	<i>im30</i>	chloroplast membrane-associated 30 kD protein
sll0851	<i>psbC</i>	photosystem II CP43 protein
sll1471	<i>cpcG</i>	phycobilisome rod-core linker polypeptide CpcG
sll1694	<i>hofG</i> or <i>hopG</i>	general secretion pathway protein G
slr0374		cell division cycle protein
slr0737	<i>psaD</i>	photosystem I subunit II (cluster VI)
slr1128		erthyrocyte band 7 integral embrane protein, protein 7.2B, stomatin
slr1348	<i>cysE</i>	serine acetyltransferase
slr1793	<i>talB</i>	transaldolase
ssl1533		hypothetical protein
slr1834	<i>psaA</i>	P700 apoprotein subunit Ia (cluster VI)
slr1835	<i>psaB</i>	P700 apoprotein subunit Ib (cluster VI)

k-means clustering using Pearson correlation:

Cluster VI



sll0258	<i>psbV</i>	cytochrome c550 (cluster IV)
sll0901	<i>purE</i> or <i>ade6</i>	phosphoribosylaminoimidazole carboxylase (cluster IV)
sll1091		391 aa (43 kD) bacteriochlorophyll synthase subunit (cluster IV)
sll1214	<i>PNIL34</i> or <i>AT103</i>	phytochrome-regulated gene (cluster IV)
sll1306		hypothetical protein (cluster IV)
sll1577	<i>cpcB</i>	phycocyanin b subunit
sll1578	<i>cpcA</i>	phycocyanin a subunit
sll1579	<i>cpcC</i>	phycocyanin associated linker protein
sll1580	<i>cpcC</i>	phycocyanin associated linker protein
ssl3093	<i>cpcD</i>	phycocyanin associated linker protein (cluster V)
sll1626	<i>lexA</i>	SOS function regulatory protein (cluster V)
sll1712		DNA binding protein HU (cluster V)
slr2051	<i>cpcG</i>	phycobilisome rod-core linker polypeptide
sll0819	<i>psaF</i>	photosystem I subunit III
slr1655	<i>psaL</i>	photosystem I subunit XI
slr0151		hypothetical protein (cluster V)
slr0329	<i>xylR</i>	xylose repressor (cluster V)
slr0452	<i>ilvD</i>	dihydroxyacid dehydratase (cluster IV)
slr1459	<i>apcF</i>	phycobilisome core component (cluster V)
slr1545	<i>rpoE</i>	RNA polymerase sigma-E factor (cluster V)
slr1853		hypothetical protein (cluster V)
slr1855		hypothetical protein
slr1856		anti-sigma B factor antagonist (cluster V)
slr1986	<i>apcB</i>	allophycocyanin b chain
slr2067	<i>apcA</i>	allophycocyanin a chain

東京大学

k-means clustering using uncentered correlation: Cluster I



slI0262	<i>desD(des6)</i>	delta-6 desaturase (cluster IV)
slI0814		hypothetical protein (cluster II)
slI0322	<i>hypF</i>	transcriptional regulatory protein HypF
slI0854		unknown (cluster IV)
slI0927	<i>metX</i>	S-adenosylmethionine synthetase
slI1028	<i>ccmK</i>	carbon dioxide concentrating mechanism protein
slI1029	<i>ccmK</i>	carbon dioxide concentrating mechanism protein
slI1031	<i>ccmM</i>	carbon dioxide concentrating mechanism protein
slr1963		hypothetical protein
slr2075	<i>groES</i>	10kD chaperonin
slI0416	<i>groEL-2</i>	60kD chaperonin 2
slr0011	<i>rbcX</i>	unknown function (cluster II)
slr0208		unknown (cluster IV)
slr0642		integral membrane protein (cluster IV)
slr1513		unknown (cluster IV)
slr1718		unknown (cluster IV)

100,000 repetitions; solution found 131 times

k-means clustering using uncentered correlation:

Cluster II



sll0519	<i>ndhA</i>	NADH dehydrogenase subunit 1 (cluster IV)
sll0521	<i>ndhG</i>	NADH dehydrogenase subunit 6
slr1280	<i>ndhK</i>	NADH dehydrogenase subunit NdhK
slr1350	<i>desA</i>	fatty acid desaturase
slr1604	<i>ftsH</i>	cell division protein
slr1641	<i>clpB</i>	ClpB protein
slr0476		hypothetical protein
slr1687		hypothetical protein
sll0020	<i>clpC</i>	ATP-dependent Clp protease regulatory subunit (cluster IV)
sll0170	<i>dnaK</i>	DnaK protein
sll0414		hypothetical protein (cluster IV)
sll0430	<i>htpG</i>	heat shock protein
sll1614	<i>pma1</i>	cation-transporting ATPase (E1-E2 ATPase) (cluster IV)

k-means clustering using uncentered correlation: Cluster III



sll0144	<i>pyrH</i>	uridine monophosphate kinase (cluster IV)
sll0587	<i>pykF</i>	pyruvate kinase (cluster IV)
sll0680	<i>pstS</i> or <i>phoS</i>	phosphate-binding periplasmic protein precursor (PBP) (cluster IV)
sll1327	<i>atpC</i>	ATP synthase g subunit (cluster IV)
sll1096	<i>rps12</i>	30S ribosomal protein S12
sll1097	<i>rps7</i>	30S ribosomal protein S7
sll1260	<i>rps2</i>	30S ribosomal protein S2
sll1743	<i>rpl11</i>	50S ribosomal protein L11
sll1745	<i>rpl10</i>	50S ribosomal protein L10
sll1799	<i>rpl3</i>	50S ribosomal protein L3
sll1801	<i>rpl23</i>	50S ribosomal protein L23
sll1802	<i>rpl2</i>	50S ribosomal protein L2
sll1804	<i>rps3</i>	30S ribosomal protein S3
sll1807	<i>rpl24</i>	50S ribosomal protein L24
sll1809	<i>rps8</i>	30S ribosomal protein S8
sll1810	<i>rpl6</i>	50S ribosomal protein L6
sll1816	<i>rps13</i>	30S ribosomal protein S13
slr1237	<i>codA</i>	cytosine deaminase (cluster IV)
slr1469	<i>rnpA</i>	protein subunit of ribonuclease P (RNase P) (cluster IV)
slr1557		unknown (cluster IV)
ssr2554		unknown (cluster IV)

k-means clustering using uncentered correlation: Cluster IV



sll0185		unknown
sll1002	<i>ycf22</i>	hypothetical protein
slr0839	<i>hemH</i>	ferrochelatase

k-means clustering using uncentered correlation: Cluster V



sll0258	<i>psbV</i>	cytochrome c550
sll0901	<i>purE</i> or <i>ade6</i>	phosphoribosylaminoimidazole carboxylase
sll1091		391 aa (43 kD) bacteriochlorophyll synthase subunit
sll1214	<i>PNIL34</i> or <i>AT103</i>	phytochrome-regulated gene
sll1626	<i>lexA</i>	SOS function regulatory protein
sll1712		DNA binding protein HU
slr0452	<i>ilvD</i>	dihydroxyacid dehydratase (cluster IV)
slr1459	<i>apcF</i>	phycobilisome core component
slr1545	<i>rpoE</i>	RNA polymerase sigma-E factor
slr1853		hypothetical protein
slr1855		hypothetical protein (cluster VI)
slr1856		anti-sigma B factor antagonist
slr1986	<i>apcB</i>	allophycocyanin b chain (cluster VI)
slr2067	<i>apcA</i>	allophycocyanin a chain (cluster VI)

k-means clustering using uncentered correlation: Cluster VI



sll0617	<i>im30</i>	chloroplast membrane-associated 30 kD protein (cluster V)
sll0851	<i>psbC</i>	photosystem II CP43 protein (cluster V)
sll1306		hypothetical protein (cluster V)
sll1694	<i>hofG</i> or <i>hopG</i>	general secretion pathway protein G (cluster V)
sll1577	<i>cpcB</i>	phycocyanin b subunit
sll1578	<i>cpcA</i>	phycocyanin a subunit
sll1579	<i>cpcC</i>	phycocyanin associated linker protein
sll1580	<i>cpcC</i>	phycocyanin associated linker protein
ssl3093	<i>cpcD</i>	phycocyanin associated linker protein (cluster V)
sll1471	<i>cpcG</i>	phycobilisome rod-core linker polypeptide CpcG (cluster V)
slr2051	<i>cpcG</i>	phycobilisome rod-core linker polypeptide
sll0819	<i>psaF</i>	photosystem I subunit III
slr0737	<i>psaD</i>	photosystem I subunit II
slr1655	<i>psaL</i>	photosystem I subunit XI
slr1834	<i>psaA</i>	P700 apoprotein subunit Ia
slr1835	<i>psaB</i>	P700 apoprotein subunit Ib
slr0151		hypothetical protein (cluster V)
slr0329	<i>xyIR</i>	xylose repressor (cluster V)
slr0374		cell division cycle protein (cluster V)
slr1128		erthyrocyte band 7 integral embrane protein, protein 7.2B, stomatin (cluster V)
slr1348	<i>cysE</i>	serine acetyltransferase (cluster V)
slr1793	<i>talB</i>	transaldolase (cluster V)
ssl1533		hypothetical protein (cluster V)

Hierarchical clustering using Euclidean distance:

Cluster I



sll1028	<i>ccmK</i>	carbon dioxide concentrating mechanism protein
sll1029	<i>ccmK</i>	carbon dioxide concentrating mechanism protein
sll1031	<i>ccmM</i>	carbon dioxide concentrating mechanism protein
slr1963		hypothetical protein
slr2075	<i>groES</i>	10kD chaperonin
sll0416	<i>groEL-2</i>	60kD chaperonin 2

Hierarchical clustering using Euclidean distance:

Cluster II/IV



sll0170	<i>dnaK</i>	DnaK protein
sll0430	<i>htpG</i>	heat shock protein
sll0521	<i>ndhG</i>	NADH dehydrogenase subunit 6
slr1280	<i>ndhK</i>	NADH dehydrogenase subunit NdhK
slr0011	<i>rbcX</i>	unknown function
slr1350	<i>desA</i>	fatty acid desaturase
slr1604	<i>ftsH</i>	cell division protein
slr1641	<i>clpB</i>	ClpB protein
sll0814		hypothetical protein
slr0476		hypothetical protein
slr1687		hypothetical protein

Hierarchical clustering using Euclidean distance:

Cluster III



sll1096	<i>rps12</i>	30S ribosomal protein S12
sll1097	<i>rps7</i>	30S ribosomal protein S7
sll1260	<i>rps2</i>	30S ribosomal protein S2
sll1743	<i>rpl11</i>	50S ribosomal protein L11
sll1745	<i>rpl10</i>	50S ribosomal protein L10
sll1799	<i>rpl3</i>	50S ribosomal protein L3
sll1801	<i>rpl23</i>	50S ribosomal protein L23
sll1802	<i>rpl2</i>	50S ribosomal protein L2
sll1804	<i>rps3</i>	30S ribosomal protein S3
sll1807	<i>rpl24</i>	50S ribosomal protein L24
sll1809	<i>rps8</i>	30S ribosomal protein S8
sll1810	<i>rpl6</i>	50S ribosomal protein L6
sll1816	<i>rps13</i>	30S ribosomal protein S13
sll0322	<i>hypF</i>	transcriptional regulatory protein HypF (cluster I)
sll0927	<i>metX</i>	S-adenosylmethionine synthetase (cluster I)

Hierarchical clustering using Euclidean distance:

Cluster II/IV



sll0020	<i>clpC</i>	ATP-dependent Clp protease regulatory subunit
sll0144	<i>pyrH</i>	uridine monophosphate kinase
sll0185		unknown
sll0262	<i>desD(des6)</i>	delta-6 desaturase
sll0414		hypothetical protein
sll0519	<i>ndhA</i>	NADH dehydrogenase subunit 1
sll0587	<i>pykF</i>	pyruvate kinase
sll0680	<i>pstS</i> or <i>phoS</i>	phosphate-binding periplasmic protein precursor (PBP)
sll0854		unknown
sll1002	<i>ycf22</i>	hypothetical protein
sll1327	<i>atpC</i>	ATP synthase g subunit
sll1614	<i>pma1</i>	cation-transporting ATPase (E1-E2 ATPase)
slr0208		unknown
slr0642		integral membrane protein
slr0839	<i>hemH</i>	ferrochelatase
slr1237	<i>codA</i>	cytosine deaminase
slr1469	<i>rnpA</i>	protein subunit of ribonuclease P (RNase P)
slr1513		unknown
slr1557		unknown
slr1718		unknown
ssr2554		unknown

Hierarchical clustering using Euclidean distance:

Cluster V



slI0258	<i>psbV</i>	cytochrome c550
slI0617	<i>im30</i>	chloroplast membrane-associated 30 kD protein
slI0851	<i>psbC</i>	photosystem II CP43 protein
slI0901	<i>purE</i> or <i>ade6</i>	phosphoribosylaminoimidazole carboxylase
slI1306		hypothetical protein
slI1471	<i>cpcG</i>	phycobilisome rod-core linker polypeptide CpcG
slI1626	<i>lexA</i>	SOS function regulatory protein
slI1694	<i>hofG</i> or <i>hopG</i>	general secretion pathway protein G
slI1712		DNA binding protein HU
slr0151		hypothetical protein
slr0329	<i>xylR</i>	xylose repressor
slr0374		cell division cycle protein
slr0452	<i>ilvD</i>	dihydroxyacid dehydratase (cluster IV)
slr1128		erthyrocyte band 7 integral embrane protein, protein 7.2B, stomatin
slr1348	<i>cysE</i>	serine acetyltransferase
slr1545	<i>rpoE</i>	RNA polymerase sigma-E factor
slr1793	<i>talB</i>	transaldolase
slr1853		hypothetical protein
slr1856		anti-sigma B factor antagonist
ssl1533		hypothetical protein

Hierarchical clustering using Euclidean distance:

Cluster V/VI



slI1091		391 aa (43 kD) bacteriochlorophyll synthase subunit (cluster V)
slI1214	<i>PNIL34</i> or <i>AT103</i>	phytochrome-regulated gene (cluster V)
ssl3093	<i>cpcD</i>	phycocyanin associated linker protein (cluster V)
slr1459	<i>apcF</i>	phycobilisome core component (cluster V)
slr1855		hypothetical protein (cluster VI)
slr1986	<i>apcB</i>	allophycocyanin b chain (cluster VI)
slr2067	<i>apcA</i>	allophycocyanin a chain (cluster VI)

Hierarchical clustering using Euclidean distance:

Cluster VI



slI1577	<i>cpcB</i>	phycocyanin b subunit
slI1578	<i>cpcA</i>	phycocyanin a subunit
slI1579	<i>cpcC</i>	phycocyanin associated linker protein
slI1580	<i>cpcC</i>	phycocyanin associated linker protein
slr2051	<i>cpcG</i>	phycobilisome rod-core linker polypeptide
slI0819	<i>psaF</i>	photosystem I subunit III
slr0737	<i>psaD</i>	photosystem I subunit II
slr1655	<i>psaL</i>	photosystem I subunit XI
slr1834	<i>psaA</i>	P700 apoprotein subunit Ia
slr1835	<i>psaB</i>	P700 apoprotein subunit Ib



Conclusions

- Different clustering methods yield **different clustering solutions**.
- **k-means** clustering is **more reliable** than hierarchical clustering.
- The *k*-means clustering algorithm may take a huge amount of time.
- We therefore recommend to use hierarchical clustering for preliminary analyses, and to check the results with *k*-means.
- Among the hierarchical clustering routines, **complete linkage clustering** was found to perform best.
- We found that the **uncentered correlation** performs better than the Euclidean distance.
- As we show only two examples, we cannot provide a final answer to the question which clustering method is best. We therefore suggest that in general, clustering results should be verified by using different methods. Differences in the clustering solutions should be included in publications.

Reprint requests



In case you would like to receive a reprint of this poster, please leave us your contact information below.