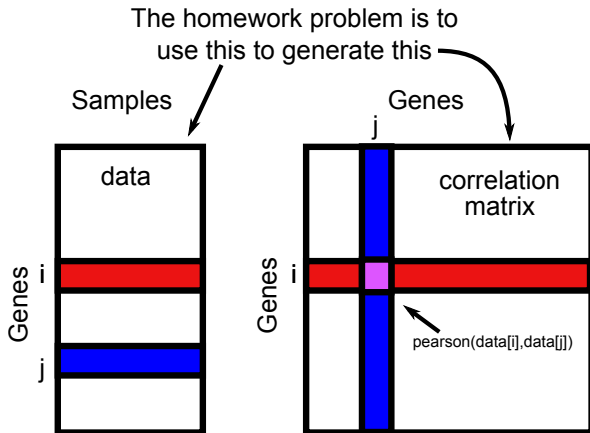


Practical Bioinformatics

Mark Voorhies

5/16/2019

What was Mark even asking for yesterday?



Adding data to a list:

```
mylist = []  
mylist.append(3)  
mylist += [4,5,6]
```

Adding data to a list:

```
mylist = []  
mylist.append(3)  
mylist += [4,5,6]
```

Lists of lists:

```
matrix = [[ 1, 2, 3, 4],  
          [ 5, 6, 7, 8],  
          [ 9,10,11,12]]
```

Pearson similarity

$$s(x, y) = \frac{\sum_i^N (x_i - x_{offset})(y_i - y_{offset})}{\sqrt{\sum_i^N (x_i - x_{offset})^2} \sqrt{\sum_i^N (y_i - y_{offset})^2}}$$

Pearson similarity

$$s(x, y) = \frac{\sum_i^N (x_i - x_{offset})(y_i - y_{offset})}{\sqrt{\sum_i^N (x_i - x_{offset})^2} \sqrt{\sum_i^N (y_i - y_{offset})^2}}$$

Pearson distance

$$d(x, y) = 1 - s(x, y)$$

Pearson similarity

$$s(x, y) = \frac{\sum_i^N (x_i - x_{offset})(y_i - y_{offset})}{\sqrt{\sum_i^N (x_i - x_{offset})^2} \sqrt{\sum_i^N (y_i - y_{offset})^2}}$$

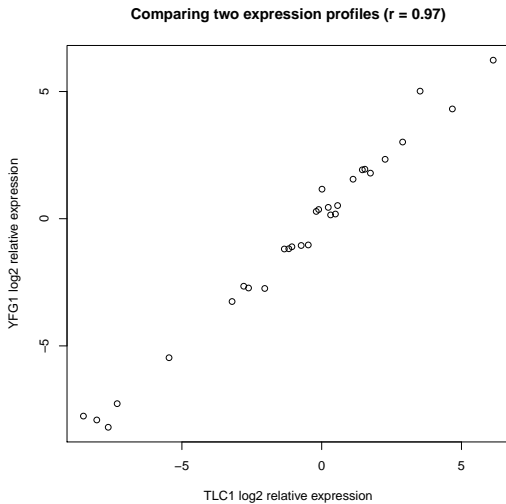
Pearson distance

$$d(x, y) = 1 - s(x, y)$$

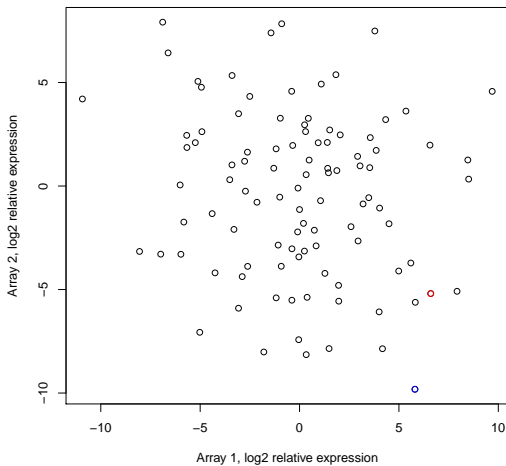
Euclidean distance

$$\frac{\sum_i^N (x_i - y_i)^2}{N}$$

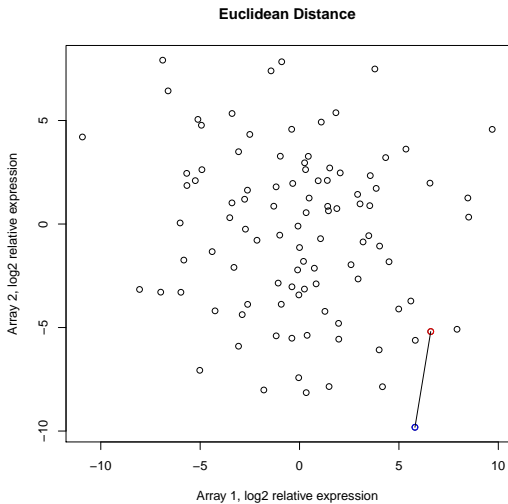
Comparing all measurements for two genes



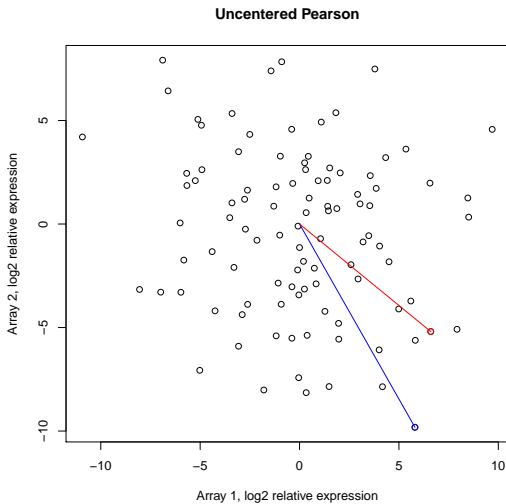
Comparing all genes for two measurements



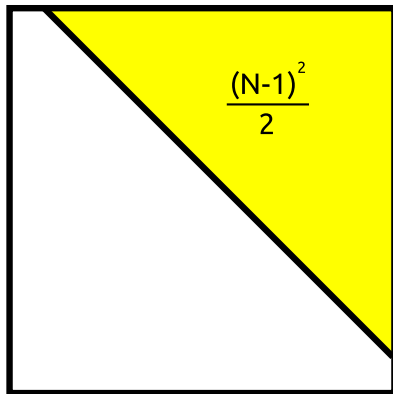
Comparing all genes for two measurements



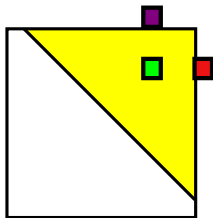
Comparing all genes for two measurements



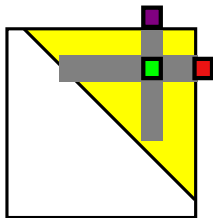
Measure all pairwise distances under distance metric



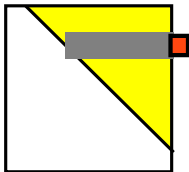
Hierarchical Clustering



Hierarchical Clustering



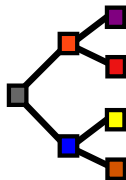
Hierarchical Clustering



Hierarchical Clustering



Hierarchical Clustering

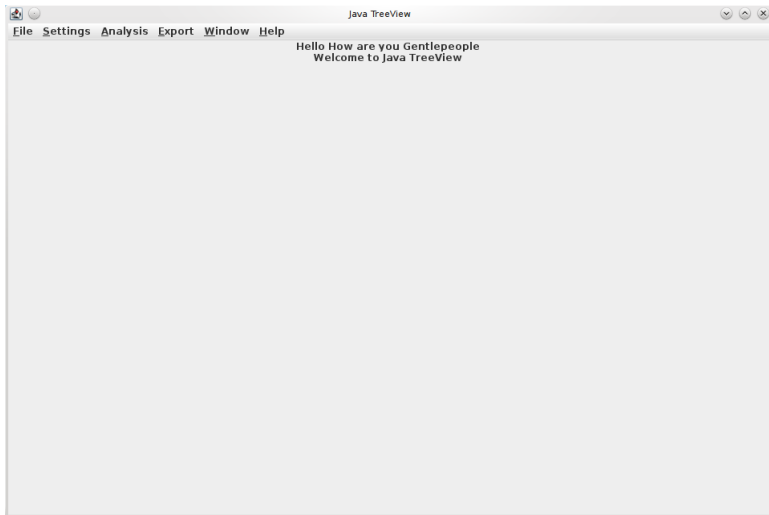


It's hard work at times, but you have to be realistic. If you have a large database with many variables and your goal is to get a good understanding of the interrelationships, then, unless you get lucky, this complex structure is bound to require some hard work to understand.

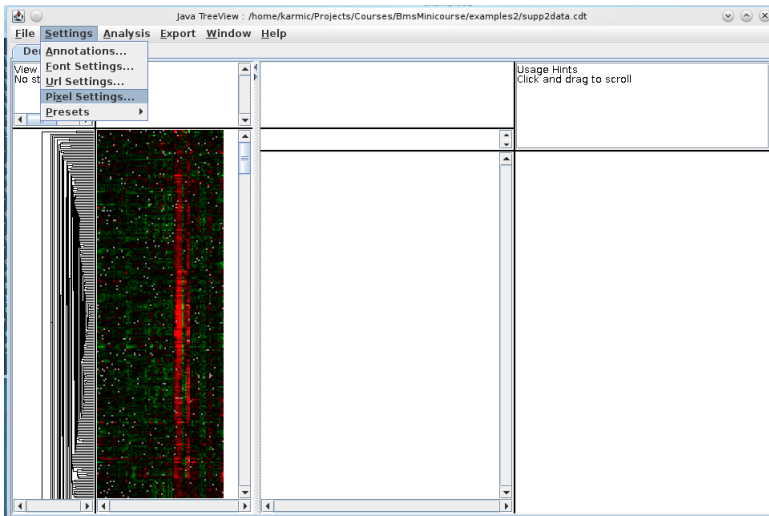
Bill Cleveland and Rick Becker

<http://stat.bell-labs.com/project/trellis/interview.html>

Using JavaTreeView



Adjust pixel settings for global view



Adjust pixel settings for global view

The screenshot shows the Java TreeView application window. The main view displays a heatmap with a dendrogram on the left. A 'Pixel Settings' dialog box is open in the foreground, allowing for adjustments to the heatmap's appearance. The dialog includes the following controls:

- Global:** Radio buttons for 'Fixed Scale' (with input fields for X: 481012658227 and Y: 663964329145) and 'Fill' (selected).
- Zoom:** Radio buttons for 'Fixed Scale' (with input fields for X: 12.0 and Y: 12.0) and 'Fill'.
- Contrast:** A slider with a 'Value' of 3.0.
- LogScale:** A checkbox for 'Log (base 2)' and a 'Center' input field set to 1.0.
- Colors:** Four color selection buttons: 'Positive' (red), 'Zero' (black), 'Negative' (green), and 'Missing' (grey). Below these are 'Load...', 'Save...', and 'Make Preset' buttons, and a dropdown menu currently showing 'RedGreen' and 'YellowBlue' options.
- A 'Close' button at the bottom of the dialog.

Select annotation columns

The screenshot shows the Java TreeView application interface. The title bar reads "Java TreeView : /home/karmac/Projects/Courses/BmsMinicourse/examples2/supp2.data.txt". The menu bar includes "File", "Settings", "Analysis", "Export", "Window", and "Help". The "Settings" menu is open, showing options: "Annotations...", "Font Settings...", "Url Settings...", "Pixel Settings...", and "Presets".

The main window is divided into three panes:

- Left Pane:** A dendrogram showing hierarchical clustering of samples. A red vertical bar highlights a specific cluster of samples.
- Middle Pane:** A heatmap visualization where rows represent genes and columns represent samples. The color scale ranges from black (low expression) to red (high expression).
- Right Pane:** A list of gene annotations. The top of this pane includes "Usage Hints" and "Click and drag to scroll". The gene list includes:

Gene ID	Gene Name	Gene Description
GDH3	GLUTAMATE BIOSYNTHESIS	NADP-GLUTAMAT
GDH1	GLUTAMATE BIOSYNTHESIS	GLUTAMATE DEH
SEC18	SECRETION	NSF; VESICLE FUSION
ABF2	MITOCHONDRIAL GENOME MAI	(PUTATIVE) HM
RH03	CYTOSKELETON	GTP-BINDING PROTEI
TFE1	TRANSCRIPTION	TFIIH 75 KD SUBUNI
TAF145	TRANSCRIPTION	TFIID 145 KD SUBUN
INP52	ENDOCYTOSIS (PUTATIVE)	INOSITOL POLY
POB3	DNA REPLICATION (PUTATIV	BINDS DNA POL
PH08	PHOSPHATE METABOLISM	VACUOLAR ALKA
GAT1	NITROGEN CATABOLISM	TRANSCRIPTION F
DPP1	PHOSPHOLIPID METABOLISM	DIACYLGLYCERO
MFP20	PROTEIN SYNTHESIS	RIBOSOMAL PROTEI
DRS2	TRANSPORT	CA(2+) TRANSPORTING A
ECM13	CELL WALL BIOGENESIS	UNKNOWN
BUB2	CELL CYCLE, CHECKPOINT	UNKNOWN
CTK2	CELL CYCLE	CYCLIN-LIKE
GCN5	CHROMATIN STRUCTURE	HISTONE ACETYLT
MNN4	PROTEIN GLYCOSYLATION	PHOSPHATIDYLI
TFCS	TRANSCRIPTION	TFIIIB 90 KD SUBUN
SNF2	TRANSCRIPTION	COMPONENT OF SWI/S
SEC2	SECRETION	GDP/GTP EXCHANGE FACT
UPE1	SECRETION	ER MEMBRANE T-SNARE
NUP42	NUCLEAR PROTEIN TARGETIN	NUCLEAR PORE
WHI4	CELL SIZE	PUTATIVE RNA BINDING
USS1	MRNA SPLICING	U5 SNRNP PROTEIN
REF2	MRNA 3'-END PROCESSING	UNKNOWN
GLE2	NUCLEAR PROTEIN TARGETIN	NUCLEAR PORE
BAT1	BRANCHED CHAIN AMINO ACI	TRANSAMINASE
MOT2	MATING	TRANSCRIPTIONAL REGULAT
KG02	TCA CYCLE	2-OXOGLUTARATE DEHYDR
COD4	UBIQUITINOME BIOSYNTHESIS	UNKNOWN
CP1	OXIDATIVE STRESS RESPONSI	CYTOCHROME-C
PDX1	GLYCOLYSIS	PYRUVATE DEHYDROGENAS
ECM37	CELL WALL BIOGENESIS	UNKNOWN
ECM27	CELL WALL BIOGENESIS	UNKNOWN

Select annotation columns

The screenshot shows the Java TreeView application window. The main interface is divided into several panes:

- View Status:** Row: 7 (YGR2), Column: 23 (Elu), Value: -0.06
- Dendrogram:** A tree structure on the left side of the heatmap.
- Heatmap:** A grid of colored cells (red, green, black) representing data values for various genes across different conditions.
- Annotation Settings Dialog:** A modal window titled "Annotation Settings" is open, showing a list of "Headers to include" for the annotation. The headers listed are: **GID**, **ORF**, **NAME**, and **GWEIGHT**. The dialog also has tabs for "Array Tree" and "Gene Tree", and a "Close" button.
- Usage Hints:** A text box on the right side of the main window that says "Mouse over to get info".
- Gene List:** A list of gene names and their corresponding annotations is visible on the right side of the heatmap, such as YAL062W, GOH3, GLUTAMATE BIOSYNTHESIS, NADP, etc.

Select URL for gene annotations

The screenshot shows the Java TreeView application window titled "Java TreeView : /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt". The "File" menu is open, and the "Gene Url Presets..." option is selected. A secondary menu is displayed, listing various preset options:

- Gene Url Presets... (Ctrl-P)
- Array Url Presets...
- Dendrogram Color Presets...
- KnnDendrogram Color Presets...
- Karyoscope Color...
- Karyoscope Coordinates...
- Scatterplot Color...

The main window displays a dendrogram on the left and a heatmap on the right. The heatmap has a color scale from 0 (black) to 100 (red). The gene names are listed on the left side of the heatmap, and the corresponding gene annotations are listed on the right side of the heatmap.

Gene	Annotation
YAL062W	GDH3 GLUTAMATE BIOSYNTHESIS NADI
YOR375C	GDH1 GLUTAMATE BIOSYNTHESIS GLJ
YBR080C	SEC18 SECRETION NSF; VESICLI
YMR072W	ABF2 MITOCHONDRIAL GENOME MAI (PU
YTL118W	RHO3 CYTOSKELETON GTP-BIND;
YOR311W	TFB1 TRANSCRIPTION TFIIH 75
YGR274C	TAF145 TRANSCRIPTION TFIIID 14;
YNL106C	INP52 ENDOCYTOSIS (PUTATIVE) INO
YML069W	POB3 DNA REPLICATION (PUTATIV BINI
YDR481C	PHO8 PHOSPHATE METABOLISM VACI
YFL021W	GAT1 NITROGEN CATABOLISM TRANSI
YDR284C	DPP1 PHOSPHOLIPID METABOLISM DIAI
YDR405W	MFP20 PROTEIN SYNTHESIS RIBOSOM
YAL028C	DPS2 TRANSPORT CA (2+) TRAN
YBL043W	ECM13 CELL WALL BIOGENESIS UNKI
YMR055C	BUB2 CELL CYCLE CHECKPOINT UNKI
YJL006C	CTK2 CELL CYCLE CYCLIN-LIKE
YGR252W	GCN5 CHROMATIN STRUCTURE HISTOF
YKL201C	MNN4 PROTEIN GLYCOSYLATION PHO
YNL039W	TF15 TRANSCRIPTION TFIIIB 94
YOR290C	SNF2 TRANSCRIPTION COMPONENT
YNL272C	SEC2 SECRETION GDP/GTP EXCI
YOR075W	UPE1 SECRETION ER MEMBRANE
YDR192C	NUP42 NUCLEAR PROTEIN TARGETIN NU
YDL224C	WHI4 CELL SIZE PUTATIVE RN
YER112W	USS1 MRNA SPLICING U6 SNRNP
YOR109W	REF2 MRNA 3' END PROCESSING UNKI
YER107C	GLE2 NUCLEAR PROTEIN TARGETIN NU
YHR208W	BAT1 BRANCHED CHAIN AMINO ACI TRAI
YER069W	MOT2 MATING TRANSCRIPTION;
YDR149C	KG02 TCA CYCLE 2-OXOGLUTAR;
YDR204W	COO4 UBIQUINONE BIOSYNTHESIS UNKI
YKR069C	CP1 OXIDATIVE STRESS RESPON CYTI
YGR193C	POX1 GLYCOLYSIS PYRUVATE DEI
YTL146C	ECM37 CELL WALL BIOGENESIS UNKI
YJL109W	ECM27 CELL WALL BIOGENESIS UNKI

Select URL for gene annotations

The screenshot shows the Java TreeView application interface. At the top, the title bar reads "java TreeView - /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt". The menu bar includes "File", "Settings", "Analysis", "Export", "Window", and "Help".

The main window is divided into several sections:

- Dendrogram:** Located at the top left, it shows a hierarchical tree structure of nodes.
- View Status:** Below the dendrogram, it says "Select Node to view annotator".
- Heatmaps:** Two heatmaps are visible, showing gene expression data with red and green colors.
- Usage Hints:** A text box on the right says "Click to select node - use arrow keys to navigate tree".
- Presets:** A section below the heatmaps lists various gene annotations, such as "YAL062W GDH3 GLUTAMATE BIOSYNTHESIS" and "YOR375C GDH1 GLUTAMATE BIOSYNTHESIS".

A "Modify Url Presets" dialog box is open in the foreground, showing a table of presets:

Enabled	Header	Name	Template	Default?
<input type="checkbox"/>	*	SGD	http://genome-www4.stanford.edu/cgi-bin/SGD/locus.pl?locus=HEADER	<input checked="" type="radio"/>
<input type="checkbox"/>	*	YPD	http://www.proteome.com/databases/YPD/reports/HEADER.html	<input type="radio"/>
<input type="checkbox"/>	*	WormBase	http://www.wormbase.org/cgi-bin/locate.pl?locus=HEADER&start=0&start=0&ie=utf-8&oe=utf-8	<input type="radio"/>
<input type="checkbox"/>	*	Source CloneID	http://genome-www4.stanford.edu/cgi-bin/SMD/source/sourceResult?option=CloneID	<input type="radio"/>
<input type="checkbox"/>	*	FlyBase	http://flybase.bio.indiana.edu/bin/fbqgenq.html?HEADER	<input type="radio"/>
<input type="checkbox"/>	*	MouseGD	http://www.jax.org/avaw/servlet/SearchTool?query=HEADER&selectedQuery=Genes+and+Markers	<input type="radio"/>
<input type="checkbox"/>	*	GenomeNetEcoli	http://www.genome.ad.jp/dbget-bin/www_bqet?eco:HEADER	<input type="radio"/>
<input type="checkbox"/>		None		<input type="radio"/>

Buttons for "Save" and "Cancel" are at the bottom of the dialog box.

At the bottom of the main window, there are more heatmaps and a list of gene annotations with their corresponding biological processes, such as "YER107C GLE2 NUCLEAR PROTEIN TARGETING NUC", "YHR208B BAT1 BRANCHED CHAIN AMINO ACI TRAI", and "YER066W MOT2 MATING TRANSCRIPTION".

Activate and detach annotation window

The screenshot shows the Java TreeView application window titled "java TreeView : /home/karmac/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt". The interface includes a menu bar (File, Settings, Analysis, Export, Window, Help) and a toolbar. On the left, a vertical toolbar contains various analysis tools, with "GeneTreeAnno" selected. The main workspace is divided into three panes: a dendrogram on the left, a heatmap in the center, and a list of gene annotations on the right. The heatmap shows a grid of colored cells (red, green, black) representing data points for various genes. The gene list on the right includes identifiers like YAL063W, YOR375C, YBR080C, etc., and their corresponding biological functions such as "GLUTAMATE BIOSYNTHESIS", "SECRETION", and "MITOCHONDRIAL GENOME MAI (PU". A "Usage Hints" box in the top right corner states "Click and drag to scroll".

Activate and detach annotation window

The screenshot shows the Java TreeView application window. The title bar reads "Java TreeView : /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt". The menu bar includes "File", "Settings", "Analysis", "Export", "Window", and "Help". The "Analysis" menu is open, showing options like "Find Genes...", "Find Arrays...", "Stats...", "Dendrogram", "Alignment", "KnnDendrogram", "Karyoscope", "Scatterplot", "ArrayTreeAnno", "GeneTreeAnno", "Remove Current", and "Detach Current". The "Detach Current" option is highlighted. The main window displays a table with columns: "NODEID", "LEFT", "RIGHT", "CORRELAT...", "NAME", and "ANNOTATI...". The table contains 24 rows of data, with the 14th row highlighted in blue.

NODEID	LEFT	RIGHT	CORRELAT...	NAME	ANNOTATI...
NODE243...	GENE182...	NODE239...	0.347965		
NODE244...	NODE242...	NODE243...	0.347965		
NODE244...	GENE550X	NODE239...	0.344607		
NODE244...	NODE243...	NODE244...	0.342251		
NODE244...	NODE244...	GENE4X	0.334454		
NODE244...	NODE240...	NODE239...	0.333461		
NODE244...	NODE244...	NODE243...	0.331585		
NODE244...	NODE244...	NODE238...	0.328813		
NODE244...	NODE244...	GENE229...	0.305824		
NODE244...	GENE495X	GENE217...	0.304111		
NODE244...	GENE219...	GENE218...	0.303188		
NODE245...	NODE244...	GENE215X	0.301587		
NODE245...	NODE244...	NODE242...	0.298323		
NODE245...	NODE240...	NODE244...	0.289436		
NODE245...	NODE242...	GENE219...	0.287138		
NODE245...	NODE245...	NODE243...	0.284232		
NODE245...	NODE245...	GENE527X	0.277872		
NODE245...	NODE245...	NODE234...	0.27761		
NODE245...	NODE245...	NODE244...	0.271103		
NODE245...	NODE233...	NODE245...	0.260487		
NODE245...	NODE243...	NODE245...	0.220385		
NODE246...	NODE244...	NODE245...	0.197665		
NODE246...	NODE245...	NODE243...	0.180953		
NODE246...	NODE246...	GENE182...	0.161919		
NODE246...	NODE246...	NODE119...	0.126461		
NODE246...	NODE246...	NODE245...	0.098323		
NODE246...	NODE245...	NODE246...	-0.087409		
NODE246...	NODE246...	NODE246...	-0.354391		

Activate and detach annotation window

Java TreeView : /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt

File Settings Analysis Export Window Help

Dendrogram

View Status
Row: 115 (YOR028C)
Column: 49 (spo11)
Value: 1.34

Usage Hints
Mouse over to get info

cdcl5_170
cdcl5_170
cdcl5_210
cdcl5_250
cdcl5_270
cdcl5_290
spo_9
spo_5
spo_7
spo_9
spo_11
spo5_2
spo5_11
spo_early
spo_mid
heat_0
heat_10
heat_20
heat_40
heat_60
heat_100

YFR028C CDC14 MITOSIS PROTEIN PHOS
YML065W ORC1 DNA REPLICATION ORIGIN F
YIL139C REV7 DNA REPAIR DNA POLYMEF
YNL318C NONE TRANSPORT HEXOSE PERM
YFR023W PES4 DNA REPLICATION UNKNOWN:
YHR015W MIP6 mRNA EXPORT, PUTATIVE RNA
YDR263C DLM7 DNA REPAIR (PUTATIVE) DNA
YLR045C STU2 CYTOSKELETON SPINDLE
YOR033C DHS1 DNA REPAIR EXONUCLEASE
YIL159W BNR1 CYTOSKELETON ACTIN FI
YKL042W SPC42 CYTOSKELETON SPINDLE
YNL225C CNM67 CYTOSKELETON SPINDLE
YCR092C CDC10 CYTOKINESIS GTP BINDING
YLR210W CLB4 CELL CYCLE G2/M CYCLIN
YLR314C CDC3 CYTOKINESIS SEPTIN
YBR045C GIP1 GLUCOSE REPRESSION (PUTA
YDL159W CLB3 CELL CYCLE G2/M CYCLIN
YDR118W APC4 CELL CYCLE ANAPHASE-PF
YDR253C MET32 METHIONINE METABOLISM TRP
YML190W CLK1 CYTOSKELETON SPINDLE
YDR113C PDS1 CELL CYCLE ANAPHASE-T

GeneTreeAnno: /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt

Sporulation

Name Sporulation Annotation Genes upregulated in sporulation

NODEID	LEFT	RIGHT	CORRELAT...	NAME	ANNOTATI...
NODE184...	NODE184...	NODE152...	0.627369	Sporulation	Genes up...
NODE184...	NODE184...	GENE56X	0.627369		
NODE184...	NODE184...	NODE178...	0.627369		
NODE184...	NODE150...	GENE177...	0.627287		

Dock Close

- 1 Write functions to reproduce the shuffling controls in figure 3 of the Eisen paper (removing correlations among genes and/or samples).