

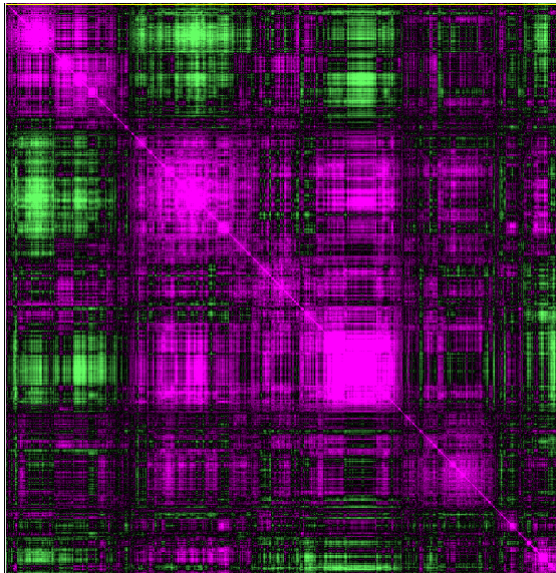
Practical Bioinformatics

Mark Voorhies

5/15/2015

- Indentation matters

Clustering exercises – Visualizing the distance matrix



Loading and re-loading your functions

```
# Use import the first time you load a module  
# (And keep using import until it loads  
# successfully)
```

```
import my_module
```

```
my_module.my_function(42)
```

```
# Once a module has been loaded, use reload to  
# force python to read your new code
```

```
reload(my_module)
```

Setting Canopy's working/import directory

OS X

- Open a terminal
- `cd path/to/working/directory`
- `env PYTHONPATH="$PYTHONPATH:$PWD" canopy`

Windows (or OS X)

- Start canopy
- `%cd path/to/working/directory`
- `import sys, os`
- `sys.path.append(os.getcwd())`

Pearson similarity

$$s(x, y) = \frac{1}{N} \sum_i^N \left(\frac{x_i - x_{offset}}{\phi_x} \right) \left(\frac{y_i - y_{offset}}{\phi_y} \right)$$

$$\phi_G = \sqrt{\sum_i^N \frac{(G_i - G_{offset})^2}{N}}$$

Pearson similarity

$$s(x, y) = \sum_i^N \left(\frac{x_i - x_{\text{offset}}}{\phi_x} \right) \left(\frac{y_i - y_{\text{offset}}}{\phi_y} \right)$$

$$\phi_G = \sqrt{\sum_i^N (G_i - G_{\text{offset}})^2}$$

Pearson similarity

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

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

$$d_{uncentered}(x, y) = 1 - s(x, y)$$

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

$$d_{uncentered}(x, y) = 1 - s(x, y)$$

Euclidean distance

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

Clustering exercises – Negative controls

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

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

```
def shuffleGenes(self, seed = None):
    """ Shuffle expression matrix by row. """
    import random
    if (seed != None):
        random.seed(seed)
    indices = range(len(self.genes))
    random.shuffle(indices)
    genes = [self.geneName[i] for i in indices]
    self.geneName = genes
    annotations = [self.geneAnn[i] for i in indices]
    self.geneAnn = genes
    num = [self.num[i] for i in indices]
    self.num = num
```

Clustering exercises – Negative controls

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

Clustering exercises – Negative controls

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

```
def shuffleRows(self, seed = None):
    """Permute ratio values within rows."""
    import random
    if (seed != None):
        random.seed(seed)
    for i in self.num:
        random.shuffle(i)
```

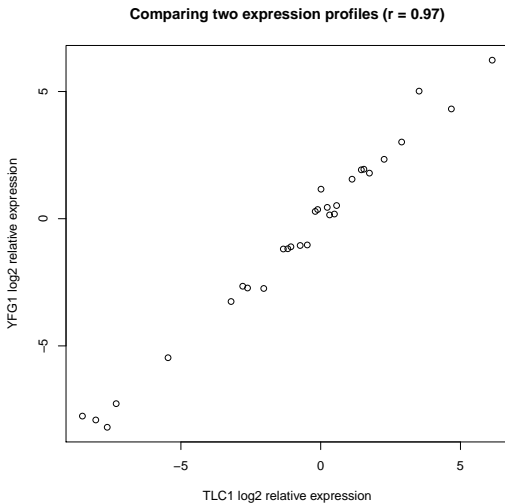
Clustering exercises – Negative controls

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

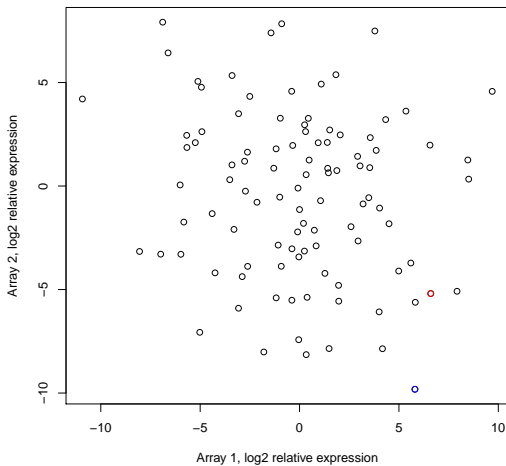
```
def shuffleRows(self, seed = None):
    """Permute ratio values within rows."""
    import random
    if (seed != None):
        random.seed(seed)
    for i in self.num:
        random.shuffle(i)

def shuffleCols(self, seed = None):
    """Permute ratio values within columns."""
    import random
    if (seed != None):
        random.seed(seed)
    # Transpose the expression matrix
    cols = []
    for col in xrange(len(self.num[0])):
        cols.append([row[col] for row in self.num])
    # Shuffle
    for i in cols:
        random.shuffle(i)
    # Transpose back to original orientation
    self.num = []
    for row in xrange(len(cols)):
        self.num.append([col[row] for col in row])
```

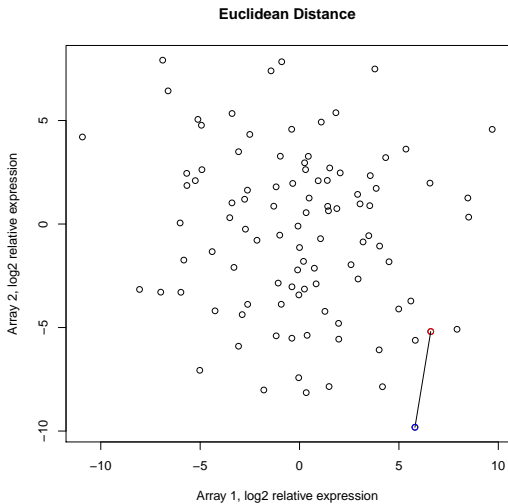

Comparing all measurements for two genes



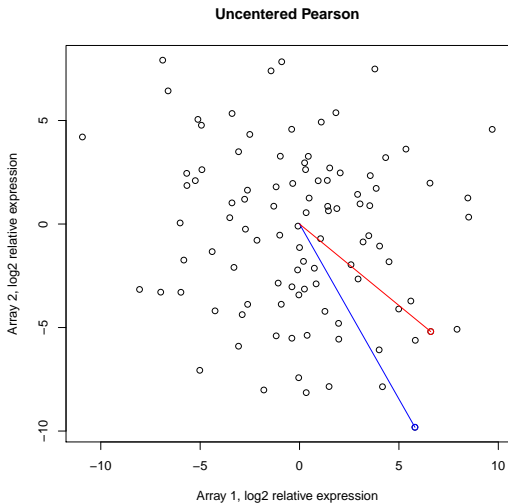
Comparing all genes for two measurements



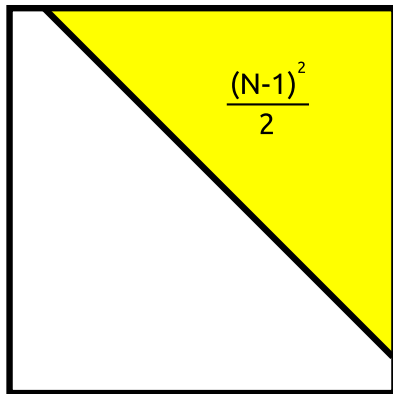
Comparing all genes for two measurements



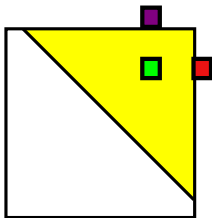
Comparing all genes for two measurements



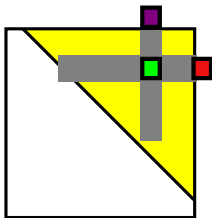
Measure all pairwise distances under distance metric



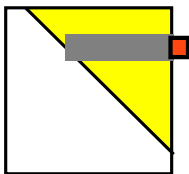
Hierarchical Clustering



Hierarchical Clustering



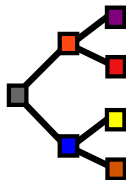
Hierarchical Clustering



Hierarchical Clustering



Hierarchical Clustering



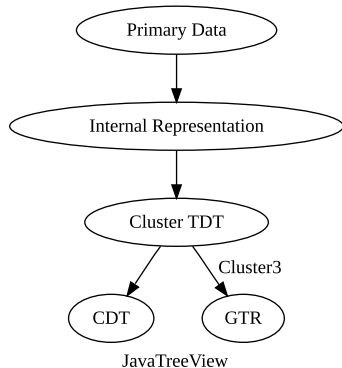
- Running Cluster3 from the command line
 - `/Applications/Cluster.app/Contents/MacOS/Cluster`
 - `/Program Files/Stanford University/Cluster3/Cluster.com`
- Command-line programs are like functions
- “man program” is like “help(function)”
- Use the `subprocess` module to run command-line programs from within Python.

USAGE: cluster [options]

| | |
|-------------|--|
| -f filename | File loading |
| -u jobname | Allows you to specify a different name for the output files (default is derived from the input file name) |
| -g [0..8] | Specifies the distance measure for gene clustering 0: No gene clustering 1: Uncentered correlation 2: Pearson correlation 3: Uncentered correlation, absolute value 4: Pearson correlation, absolute value 5: Spearman's rank correlation 6: Kendall's tau 7: Euclidean distance 8: City-block distance (default: 0) |
| -m [msca] | Specifies which hierarchical clustering method to use m: Pairwise complete-linkage s: Pairwise single-linkage c: Pairwise centroid-linkage a: Pairwise average-linkage (default: m) |

Scripting the Protocol

```
from subprocess import check_call
check_call(
    # Which program to run
    ("cluster",
    # Input file
    "-f", "supp2data.tdt",
    # Output prefix
    "-u", "supp2data.Uncentered.Complete",
    # Clustering method: complete linkage
    "-m", "m",
    # Distance function: uncentered Pearson
    "-g", "1"))
```



Using the Cluster3 GUI

Gene Cluster 3.0

File Help

File loaded

Job name

Data set has Rows Columns

Filter Genes

- % Present >= 80
- SD (Gene Vector) 2,0
- At least 1 observations with abs(Val) >= 2,0
- MaxVal - MinVal >= 2,0

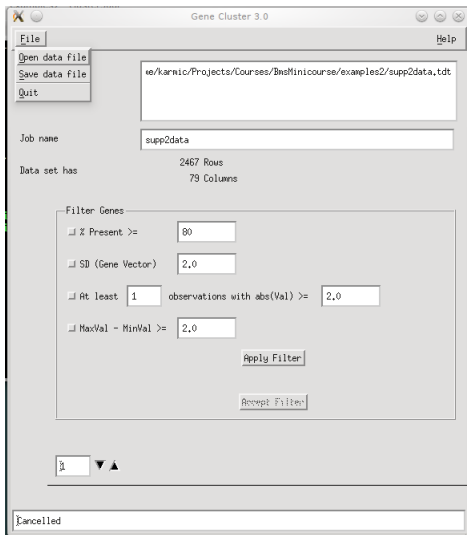
Apply Filter

Accept Filter

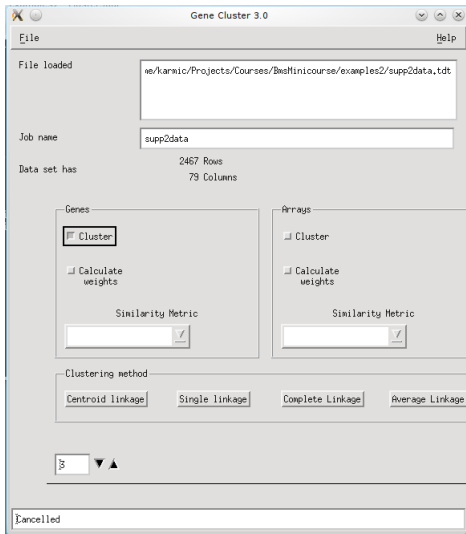
1

1

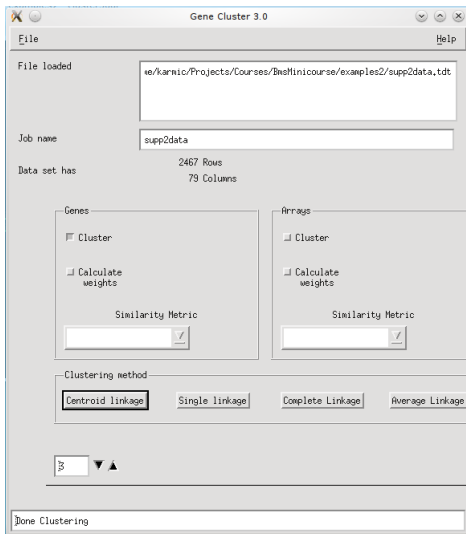
Load your data



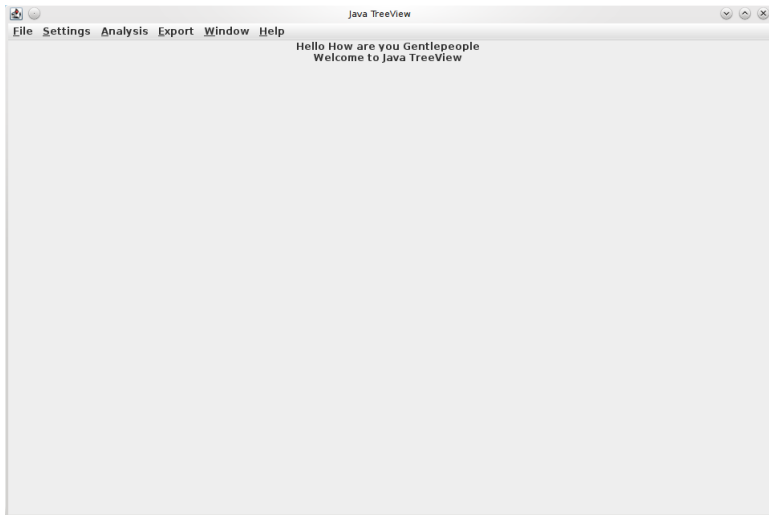
Choose distance function



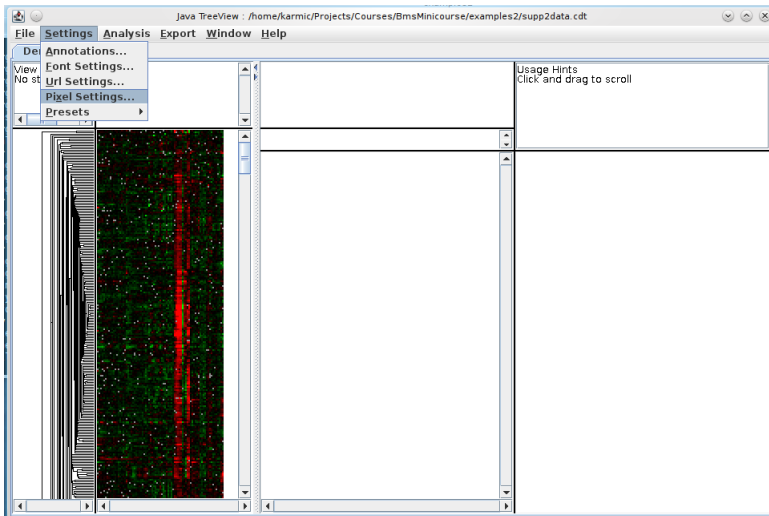
Choose linking method



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 indicates the file path: `/home/karmic/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 like **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 each cell's color (red, green, black) represents the expression level of a gene in a specific sample.
- Right Pane:** A list of gene annotations. The top of this pane includes the text "Usage Hints" and "Click and drag to scroll". The list contains gene IDs and their corresponding biological functions, such as `GDH3 GLUTAMATE BIOSYNTHESIS NADP-GLUTAMAT`.

At the bottom of the window, there are navigation controls including arrows and a search icon.

Select annotation columns

The screenshot shows the Java TreeView application interface. The main window displays a dendrogram on the left and a heatmap in the center. A dialog box titled 'Annotation Settings' is open, showing a list of columns to include in the annotation. The columns listed are: **GID**, **ORF**, **NAME**, and **GWEIGHT**. The dialog also has tabs for 'Array Tree' and 'Gene Tree', and a 'Close' button.

The annotation table on the right side of the window lists gene IDs and their corresponding annotations. The table is as follows:

| Gene ID | Annotation 1 | Annotation 2 | Annotation 3 |
|---------|--------------|-------------------------------|----------------|
| YAL062W | GDH3 | GLUTAMATE BIOSYNTHESIS | NADP |
| YOR375C | GDH1 | GLUTAMATE BIOSYNTHESIS | GLUF |
| YBR080C | SEC18 | SECRETION | NSF; VESICLE |
| YMR072W | ABF2 | MITOCHONDRIAL GENOME MAI (PUF | |
| YIL119W | RH03 | CYTOSKELETON | GTP-BIND; |
| YDR311W | TFB1 | TRANSCRIPTION | TFIIH 75 |
| YGR274C | TAF145 | TRANSCRIPTION | TFIID 145 |
| YNL106C | INP52 | ENDOCYTOSIS (PUTATIVE) | INOR |
| YML069W | POB3 | DNA REPLICATION (PUTATIVE) | BINE |
| YDR481C | PH08 | PHOSPHATE METABOLISM | VACI |
| YFL021W | GAT1 | NITROGEN CATABOLISM | TRANS |
| YDR284C | DDP1 | PHOSPHOLIPID METABOLISM | DIAI |
| YDR405W | MFP20 | PROTEIN SYNTHESIS | RIBOSOM |
| YAL028C | DRS2 | TRANSPORT | CA(2+) |
| YBL043W | ECM13 | CELL WALL BIOGENESIS | UNKI |
| YMR055C | BUB2 | CELL CYCLE, CHECKPOINT | UNKI |
| YJL006C | CTK2 | CELL CYCLE | CYCLIN-LIKE |
| YGR252W | GCN5 | CHROMATIN STRUCTURE | HISTO |
| YKL201C | MNN4 | PROTEIN GLYCOSYLATION | PHO |
| YNL035W | TFP5 | TRANSCRIPTION | TFIIIB 9K |
| YOF280C | SMF2 | TRANSCRIPTION | COMPONENT |
| YNL272C | SEC2 | SECRETION | GDP/GTP EXC |
| YOR075W | LEF1 | SECRETION | ER MEMBRANE |
| YDR192C | NUP42 | NUCLEAR PROTEIN TARGETIN | NUCI |
| YDL224C | WHI4 | CELL SIZE | PUTATIVE RN |
| YER112W | USS1 | MRNA SPLICING | UG SMRNP |
| YDR195W | REF2 | MRNA 3'-END PROCESSING | UNKI |
| YER107C | GLE2 | NUCLEAR PROTEIN TARGETIN | NUCI |
| YHF208W | BAT1 | BRANCHED CHAIN AMINO ACI | TRAI |
| YER068W | MOT2 | MATING | TRANSCRIPTION; |
| YDR149C | KGD2 | TCA CYCLE | 2-OXOGLUTAR |
| YDR204W | COO4 | UBIQUINONE BIOSYNTHESIS | UNKI |
| YKR068C | OCPI | OXIDATIVE STRESS RESPON | CYTI |
| YGR193C | FOX1 | GLYCOLYSIS | PYRUVATE DEI |
| YIL146C | ECM37 | CELL WALL BIOGENESIS | UNKI |
| YJL106W | ECM27 | CELL WALL BIOGENESIS | UNKI |

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 GLU |
| 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 |
| YOR405W | 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 | LEF1 SECRETION ER MEMBRANE |
| YDR192C | NUP42 NUCLEAR PROTEIN TARGETIN NUCL |
| YDL224C | WHI4 CELL SIZE PUTATIVE RN |
| YER112W | USS1 MRNA SPLICING U6 SNRNP |
| YOR109W | REF2 MRNA 3' END PROCESSING UNKI |
| YER107C | GLE2 NUCLEAR PROTEIN TARGETIN NUCL |
| 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 panes:

- 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 dialog box titled "Modify Url Presets" is open in the foreground. It has tabs for "Gene" and "Array". The dialog contains 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 | <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/fbgenq.html?HEADER | <input type="radio"/> |
| <input type="checkbox"/> | * | MouseGD | cs.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_bget?eco:HEADER | <input type="radio"/> |
| <input type="checkbox"/> | | None | | <input type="radio"/> |

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

At the bottom of the main window, there are more heatmaps and a list of gene names and their associated biological processes, such as "YER107C", "YHR208W", "YER066W", "YDR148C", "YDR204W", "YKR866C", "YCR190C", "YTL146C", "YJR106W", "MET2", "GLI2", "BAT1", "MOT2", "KGD2", "COQ4", "COF1", "PDX1", "ECM37", "ECM27", "PWR1", "PWR2", "PWR3", "ERU1", "NUCLEAR PROTEIN TARGETIN", "BRANCHED CHAIN AMINO ACI", "MATING", "TRANSCRIPTION", "TCA CYCLE", "2-OXOGLUTAR", "UBIQUINONE BIOSYNTHESIS", "OXIDATIVE STRESS RESPON", "GLYCOLYSIS", "PYRUVATE DE", "CELL WALL BIOGENESIS", "CELL WALL BIOGENESIS".

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. The "Analysis" menu is open, showing options like "Find Genes...", "Find Arrays...", "Stats...", "Flip Array Tree Node", "Align to Tree...", "Compare to...", "Remove comparison", "Summary Window...", "Dendrogram", "Alignment", "KnnDendrogram", "Karyoscope", "Scatterplot", "ArrayTreeAnno", "GeneTreeAnno", "Remove Current", and "Detach Current".

The main window displays a dendrogram on the left and a heatmap on the right. The heatmap has a grid of colored cells (red, green, black) representing data points. To the right of the heatmap is a list of gene annotations, including:

- YAL063W GDH3 GLUTAMATE BIOSYNTHESIS NADF
- YOR375C GDH1 GLUTAMATE BIOSYNTHESIS GLU
- YBR080C SEC18 SECRETION NSF; VESICLI
- YMR072W ABF2 MITOCHONDRIAL GENOME MAI (PU
- YIL119W RH03 CYTOSKELETON GTP-BIND
- YDR311W TFB1 TRANSCRIPTION TFIIF 75
- YOR274C TAF145 TRANSCRIPTION TFIID 145
- YNL106C INP52 ENDOCYTOSIS (PUTATIVE) INO
- YML069W POB3 DNA REPLICATION (PUTATIV BINI
- YDR481C PH08 PHOSPHATE METABOLISM VACU
- YFL021W GAT1 NITROGEN CATABOLISM TRANS
- YDR284C DPP1 PHOSPHOLIPID METABOLISM DIA
- YDR495W MRP20 PROTEIN SYNTHESIS RIBOSOM
- YAL029C DRS2 TRANSPORT CA(2+) TRAN
- YBL043W ECM13 CELL WALL BIOGENESIS UNK
- YMR055C BUB2 CELL CYCLE, CHECKPOINT UNK
- YJL006C CTK2 CELL CYCLE CYCLIN-LIKE
- YGR252W GCN5 CHROMATIN STRUCTURE HISTO
- YKL201C MNN4 PROTEIN GLYCOSYLATION PHO
- YML039W TFC5 TRANSCRIPTION TFIIB 94
- YOR290C SNF2 TRANSCRIPTION COMPONEN
- YML272C SEC2 SECRETION GDP/GTP EXO
- YOR075W LFE1 SECRETION ER MEMBRANE
- YDR192C NUP42 NUCLEAR PROTEIN TARGETIN NUCL
- YDL224C WH14 CELL SIZE PUTATIVE RW
- YER112W USS1 MRNA SPLICING U6 SNRNP
- YOR185W REF2 MRNA 3'-END PROCESSING UNK
- YER107C GLE2 NUCLEAR PROTEIN TARGETIN NUCL
- YHR208W BAT1 BRANCHED CHAIN AMINO ACI TRAI
- YER069W MOT2 MATING TRANSCRIPTION/O
- YDR148C KGD2 TCA CYCLE 2-OXOGLUTAR
- YDR204W COO4 UBIQUINONE BIOSYNTHESIS UNK
- YKR066C COP1 OXIDATIVE STRESS RESPON CYT
- YGR183C POK1 GLYCOLYSIS PYRUVATE DE
- YIL146C ECM37 CELL WALL BIOGENESIS UNK
- YJR166W ECM27 CELL WALL BIOGENESIS UNK

Usage Hints: 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: "Find Genes..." (Ctrl-G), "Find Arrays..." (Ctrl-A), "Stats..." (Ctrl-S), "Dendrogram", "Alignment", "KnnDendrogram", "Karyoscope", "Scatterplot", "ArrayTreeAnno", "GeneTreeAnno", "Remove Current", and "Detach Current".

Below the menu is a table with the following columns: "Name" and "Annotation". Below that is a larger table with columns: "NODEID", "LEFT", "RIGHT", "CORRELAT...", "NAME", and "ANNOTATI...".

| 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 (YOL)
Column: 49 (SP)
Value: 1.34

Usage Hints
Mouse over to get info

cdcl5_170
cdcl5_170
cdcl5_210
cdcl5_210
cdcl5_250
cdcl5_250
cdcl5_270
cdcl5_290
spo_9
spo_9
spo_5
spo_7
spo_9
spo_11
spo5_2
spo5_2
spo5_11
spo-early
spo- mid
heat_0
heat_10
heat_20
heat_30
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 DLW7 DNA REPAIR (PUTATIVE) DNA
YLR045C STU2 CYTOSKELETON SPINDLE
YOR033C DHS1 DNA REPAIR EXONUCLEASE
YIL159W BNR1 CYTOSKELETON ACTIN FI
YKL042W SPC42 CYTOSKELETON SPINDLE
YML225C CNM67 CYTOSKELETON SPINDLE
YCR092C CDC10 CYTOKINESIS GTP BINDING
YLR210W CLB4 CELL CYCLE G2/M CYCLIN
YLR314C CDC3 CYTOKINESIS SEPTIN
YBR045C GIP1 GLUCOSE REPRESSION (PUT
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-TN

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

Clustering exercises – Scripting Cluster

Modify the clustering protocol script to run Cluster3 multiple times on the same input, varying distance metric and/or clustering method. Be sure to give the output files distinct names.

Clustering exercises – Scripting Cluster

Modify the clustering protocol script to run Cluster3 multiple times on the same input, varying distance metric and/or clustering method. Be sure to give the output files distinct names.

```
metrics = ("None",
          "Uncentered",
          "Pearson",
          "UncenteredAbs",
          "PearsonAbs",
          "Spearman",
          "Kendall",
          "Euclidean",
          "City")
linkage = (("Complete", "m"),
          ("Single", "s"),
          ("Centroid", "c"),
          ("Average", "a"))

# Loop over all 32 possible methods
print "Starting hierarchical clustering runs..."
from subprocess import check_call
for metric in xrange(1, len(metrics)):
    print "    ", metrics[metric], "... "
    for (linkname, link) in linkage:
        print "        ", linkname
        check_call(("cluster", "-f", "shuffled.txt",
                    "-u", ".".join(("shuffled",
                                    metrics[metric],
                                    linkname))),
                    "-m", link, "-g", str(metric)))
```

- 1 If you haven't done so already, read the PNAS paper
- 2 Explore the figure 2 data with Cluster3 and JavaTreeView.
 - Can you find/reproduce the clusters described in the paper?
 - Are the annotations consistent with the current annotations in SGD?
 - Are there other patterns that you can find in the data?
 - What follow-up experiments are prompted by this analysis?