

Practical Bioinformatics

Mark Voorhies

6/18/2010

- `import` vs. `reload`
- indexing multi-dimensional arrays
- defining class methods
- handling exceptions

Adding an output function to our class

```
def write(self , filename):  
    fp = open(filename , "w")  
    fp.write(self.header)  
    for i in range(len(self.genes)):  
        fp.write(self.genes[i])  
        fp.write("\t"+self.annotations[i])  
        for j in self.ratios[i]:  
            if (j == None):  
                fp.write("\t"+"")  
            else :  
                fp.write("\t%f" % j)  
    fp.write("\n")
```

- 1 Write a function to calculate the uncentered Pearson distance between two gene profiles

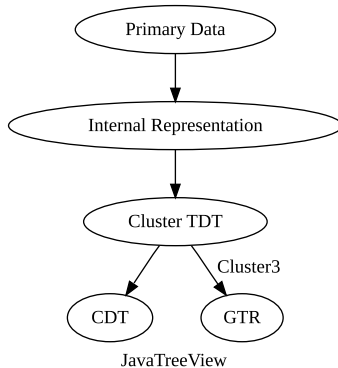
$$d(x, y) = 1 - \frac{\sum_i^N (x_i - x_{offset})(y_i - y_{offset})}{\sqrt{\sum_i^N (x_i - x_{offset})^2} \sqrt{\sum_i^N (y_i - y_{offset})^2}} \quad (1)$$

- 2 Amend the function to calculate the centered Pearson (or another distance metric from the Cluster3 manual)
- 3 Write a function to calculate all pairwise distances for the yeast expression profiles for a particular distance function.

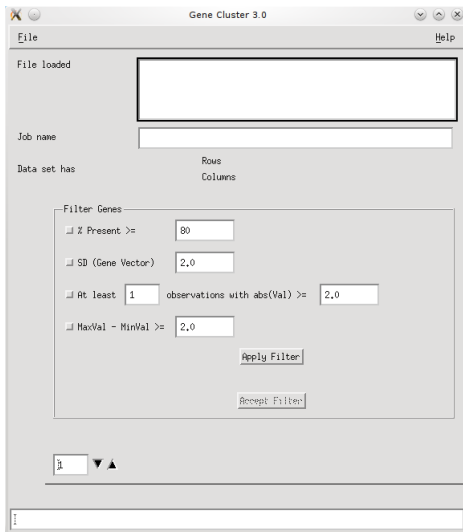
Special cases for Pearson in Cluster3

```
int flag = 0;
/* flag will remain zero if no nonzero combinations of mask1 and mask2 are
 * found.
 */
int i;
for (i = 0; i < n; i++)
{ if (mask1[index1][i] && mask2[index2][i])
  { double term1 = data1[index1][i];
    double term2 = data2[index2][i];
    double w = weight[i];
    result += w*term1*term2;
    denom1 += w*term1*term1;
    denom2 += w*term2*term2;
    flag = 1;
  }
}
if (!flag) return 0.;
if (denom1==0.) return 1.;
if (denom2==0.) return 1.;
result = result / sqrt(denom1*denom2);
result = 1. - result;
return result;
```

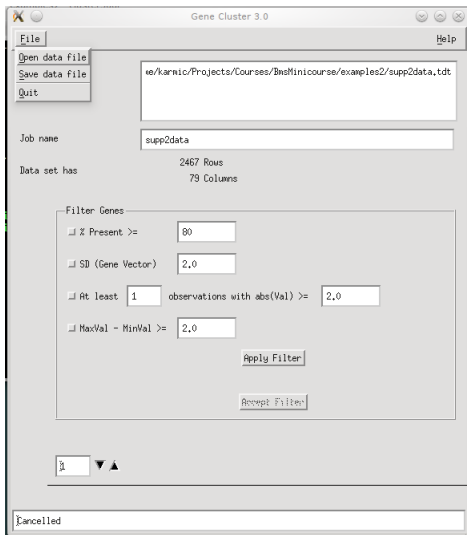
Clustering protocol



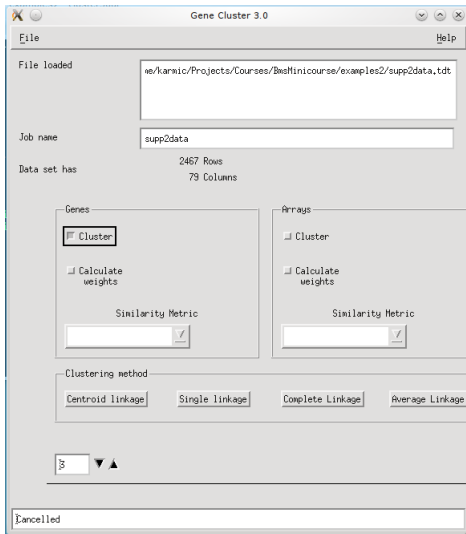
Using the Cluster3 GUI



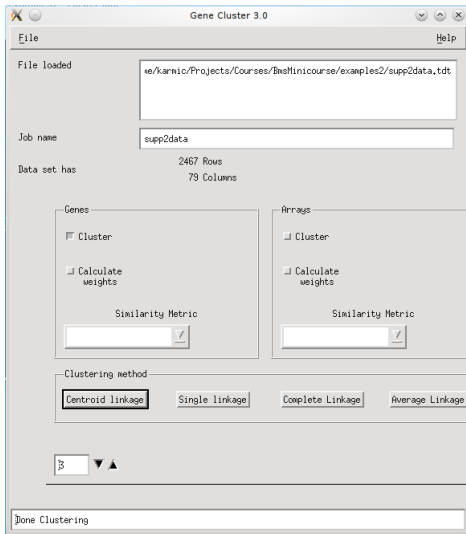
Load your data



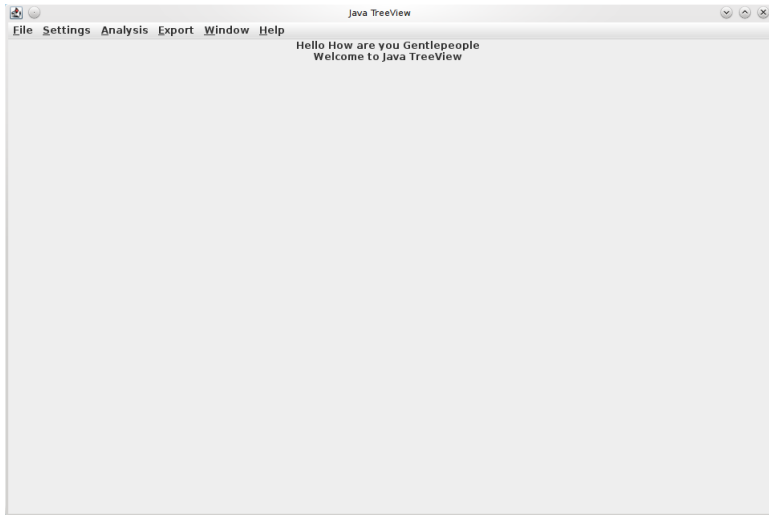
Choose distance function



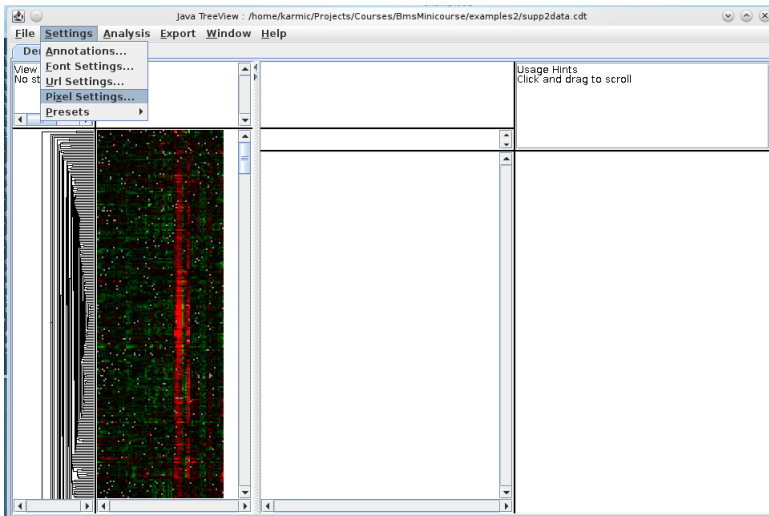
Choose linking method



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 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 section shows sample IDs (e.g., `alpha 0`, `alpha 14`, etc.) and a "Usage Hints" box with the text "Click and drag to scroll". Below this is a list of gene names and their associated biological processes, such as `GDH3 GLUTAMATE BIOSYNTHESIS NADP-GLUTAMAT`, `GDH1 GLUTAMATE BIOSYNTHESIS GLUTAMATE DEH`, and `SEC18 SECRETION NSF; VESICLE FUSION`.

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 be included in the annotation table. 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 visible portion of 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	UFEL	SECRETION	ER MEMBRANE
YDR192C	NUP42	NUCLEAR PROTEIN TARGETIN	NUCI
YDL224C	WHI4	CELL SIZE	PUTATIVE RNA
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. 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 "Settings" menu is open, showing options like "Annotations...", "Font Settings...", "Url Settings...", "Pixel Settings...", and "Presets". The "Presets" submenu is also open, listing "Gene Url Presets..." (highlighted), "Array Url Presets...", "Dendrogram Color Presets...", "KnnDendrogram Color Presets...", "Karyoscope Color...", "Karyoscope Coordinates...", and "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). A red vertical bar highlights a specific column in the heatmap. The "Usage Hints" panel on the right side of the window contains the following text: "Click and drag to scroll".

Gene ID	Gene Name	Gene Description	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	TFIID 14;
YNL106C	INP52	ENDOCYTOSIS (PUTATIVE)	INO1
YML069W	POB3	DNA REPLICATION (PUTATIV	BINI
YDR481C	PHO8	PHOSPHATE METABOLISM	VACI
YFL021W	GAT1	NITROGEN CATABOLISM	TRANS
YDR284C	DPP1	PHOSPHOLIPID METABOLISM	DIAI
YOR405W	MFP20	PROTEIN SYNTHESIS	RIBOSOI
YAL028C	DPS2	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	HISTOF
YKL201C	MNN4	PROTEIN GLYCOSYLATION	PHO;
YNL039W	TFC5	TRANSCRIPTION	TFIIIB 9;
YOR290C	SNF2	TRANSCRIPTION	COMPONENT
YNL272C	SEC2	SECRETION	GDP/GTP EXCI
YOR075W	LEF1	SECRETION	ER MEMBRANE
YDR192C	NUP42	NUCLEAR PROTEIN TARGETIN	NUCI
YDL224C	WHI4	CELL SIZE	PUTATIVE RN;
YER112W	USS1	MRNA SPLICING	U6 SNRNP
YOR105W	REF2	MRNA 3' END PROCESSING	UNKI
YER107C	GLE2	NUCLEAR PROTEIN TARGETIN	NUCI
YHR208W	BAT1	BRANCHED CHAIN AMINO ACI	TRAI
YER069W	MOT2	MATING	TRANSCRIPTION;
YDR149C	KG02	TCA CYCLE	2-OXOGLUTAR;
YDR204W	CO04	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 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 main window is titled "Dendrogram" and contains a "View Status" section with "Select Node to view annotator". Below this is a dendrogram and a heatmap. A "Presets" dialog box is open in the foreground, titled "Modify Url Presets". It has a table with columns "Enabled", "Header", "Name", "Template", and "Default?".

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/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: Save, Cancel

Usage Hints: Click to select node - use arrow keys to navigate tree

Gene Array

YAL062W GDH3 GLUTAMATE BIOSYNTHESIS NADH
YOR375C GDH1 GLUTAMATE BIOSYNTHESIS GLU

YER190W MET2 PWW3 5' ENR. PROCESSING UNKI
YER107C GLE2 NUCLEAR PROTEIN TARGETIN NUCI
YHR208W BAT1 BRANCHED CHAIN AMINO ACI TRAI
YER066W MOT2 MATING TRANSCRIPTION
YDR148C KGD2 TCA CYCLE 2-OXOGLUTAR
YDR204W COQ4 UBIQUINONE BIOSYNTHESIS UNKI
YKR866C CFP1 OXIDATIVE STRESS RESPON CYTI
YGR190C PDX1 GLYCOLYSIS PYRUVATE DEI
YTL146C ECM37 CELL WALL BIOGENESIS UNKI
YJR106W ECM27 CELL WALL BIOGENESIS UNKI

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", "Flip Gene Tree Node", "Align to Tree...", "Compare to...", "Remove comparison", "Summary Window...", "Dendrogram", "Alignment", "KnnDendrogram", "Karyoscope", "Scatterplot", "ArrayTreeAnno", "GeneTreeAnno", "Remove Current", and "Detach Current". The "GeneTreeAnno" option is highlighted. The main window displays a dendrogram on the left and a heatmap on the right. The heatmap has a grid of colored cells (green, red, black) representing data points. To the right of the heatmap is a list of gene names and their associated biological processes, such as "YAL063W GLUTAMATE BIOSYNTHESIS NADF", "YOR375C GDH1 GLUTAMATE BIOSYNTHESIS GLU", "YBR080C SEC18 SECRETION NSF; VESICLI", "YMR072W ABF2 MITOCHONDRIAL GENOME MAI (PU", "YDR311W RHO3 CYTOSKELETON GTP-BIND", "YOR274C TFBI TRANSCRIPTION TFIIH 75", "YML106C INP52 ENDOCYTOSIS (PUTATIVE) INO3", "YML069W POB3 DNA REPLICATION (PUTATIV BINI", "YDR481C PHO8 PHOSPHATE METABOLISM VACI", "YFL021W GAT1 NITROGEN CATABOLISM TRANSC", "YDR284C DPP1 PHOSPHOLIPID METABOLISM DIAI", "YDR495W MRP20 PROTEIN SYNTHESIS RIBOSOM", "YAL029C DRS2 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 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 WHI4 CELL SIZE PUTATIVE RW", "YER112W USS1 MRNA SPLICING U6 SNRNP", "YOR185W REF2 MRNA 3'-END PROCESSING UNKI", "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 UNKI", "YKR066C COP1 OXIDATIVE STRESS RESPON CYT", "YGR083C POK1 GLYCOLYSIS PYRUVATE DE", "YJL146C ECM37 CELL WALL BIOGENESIS UNKI", "YJR106W ECM27 CELL WALL BIOGENESIS UNKI". A "Usage Hints" box in the top right corner says "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". The "Detach Current" option is highlighted. The main window area contains a table with columns: "NODEID", "LEFT", "RIGHT", "CORRELAT...", "NAME", and "ANNOTATI...". The table contains 25 rows of data. Above the table, there are input fields for "Name" and "Annotation".

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 (YFR028C)
Column: 49 (sp0_10)
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
YML069W 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
YNL225C CNM67 CYTOSKELETON SPINDLE
YOR092C CDC10 CYTOKINESIS GTP BINDING
YLR210W CLB4 CELL CYCLE G2/M CYCLIN
YLR314C CDC3 CYTOKINESIS SEPTIN
YBR045C GIP1 GLUCOSE REPRESSION (PUTAT
YDL159W CLB3 CELL CYCLE G2/M CYCLIN
YDR118W APC4 CELL CYCLE ANAPHASE-PF
YDR253C MET32 METHIONINE METABOLISM TRF
YMR190W 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

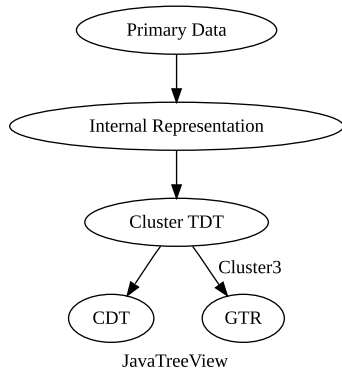
- Running Cluster3 from the command line
 - /Applications/Cluster.app/Contents/MacOS/Cluster
 - /Program Files/Stanford University/Cluster3/Cluster.exe
- Command-line programs are like functions
- “man program” is like “help(function)”
- Use the subprocess module to run command-line programs from within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

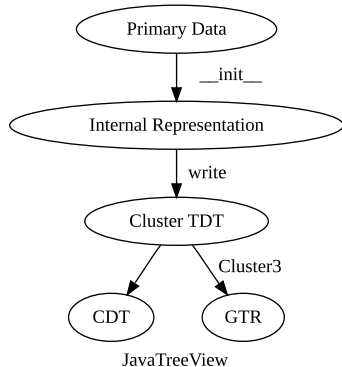
```
from subprocess import check_call
check_call(
    # Which program to run
    ("cluster",
    # Input file
    "-f", "supp2data.tdt",
    # Output prefix
    "-u", "supp2data.Uncentered.Complete",
    # Clustering method: complete linkage
    "-m", "m",
    # Distance function: uncentered Pearson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



Shuffling Genes to Remove Structure

```
from TdtRatios import TdtRatios
data = TdtRatios("supp2data.tdt")
data.shuffleGenes()
data.write("shuffled.tdt")

from subprocess import check_call
check_call(
    ("cluster",
     "-f", "shuffled.tdt",
     "-u", "shuffled.Uncentered.Complete",
     "-m", "m",
     "-g", "1"))
```



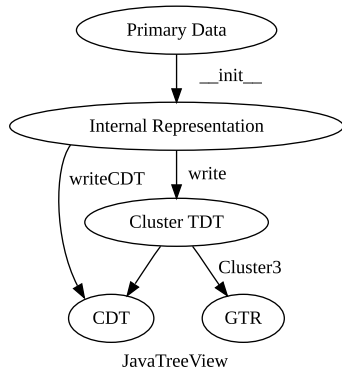
Adding a shuffling function to our class

```
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.genes[i] for i in indices]  
    self.genes = genes  
    annotations = [self.annotations[i] for i in indices]  
    self.genes = genes  
    ratios = [self.ratios[i] for i in indices]  
    self.ratios = ratios
```

Checking that the Shuffling Worked

```
from TdtRatios import TdtRatios
data = TdtRatios("supp2data.tdt")
data.shuffleGenes()
data.write("shuffled.tdt")
data.writeCDT("shuffled.cdt")

from subprocess import check_call
check_call(
    ("cluster",
     "-f", "shuffled.tdt",
     "-u", "shuffled.Uncentered.Complete",
     "-m", "m",
     "-g", "1"))
```



Generating CDT output

```
def write(self, filename):
    fp = open(filename, "w")
    fp.write(self.header)
    for i in range(len(self.genes)):
        fp.write(self.genes[i])
        fp.write("\t"+self.annotations[i])
        for j in self.ratios[i]:
            if(j == None):
                fp.write("\t"+"")
            else:
                fp.write("\t%f" % j)
        fp.write("\n")
```

```
def writeCDT(self, filename):
    """Write CDT file for JavaTreeView."""
    fp = open(filename, "w")
    cols = self.header.rstrip().split("\t")
    fp.write("\t".join(["GID"]+cols[2:]+
                       ["GWEIGHT"]+cols[2:]))
        +"\n")
    fp.write("\t".join(["EWEIGHT"]+[""]*3+
                       ["1.0"]*
                       len(self.ratios)))
        +"\n")
    for i in range(len(self.genes)):
        fp.write("GENE%4dX" % (i+1))
        fp.write("\t"+self.genes[i])
        fp.write("\t"+self.annotations[i])
        fp.write("\t1.0")
        for j in self.ratios[i]:
            if(j == None):
                fp.write("\t"+"")
            else:
                fp.write("\t%f" % j)
        fp.write("\n")
```

Clustering exercises

- 1 Cluster `supp2data.tdt` and explore the results in `JavaTreeView`. Can you identify the clusters from figure 2 of the Eisen paper. Click on gene names to open the corresponding SGD annotations in your web browser. Are the current annotations consistent with those in `supp2data.tdt`? Are they consistent with the clustering pattern?
- 2 Write you pairwise distance matrix to a CDT file (in this case, the rows and columns are *both* genes) and visualize it in `JavaTreeView`.
- 3 Add functions to `TdtRatios` to reproduce the shuffling controls in figure 3 of the Eisen paper (removing correlations among genes and/or arrays).
- 4 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.