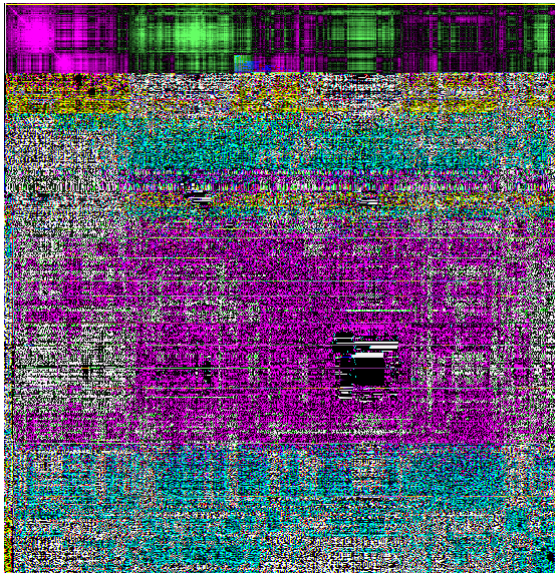


# Practical Bioinformatics

Mark Voorhies

5/24/2013

# Clustering exercises { Visualizing the distance matrix



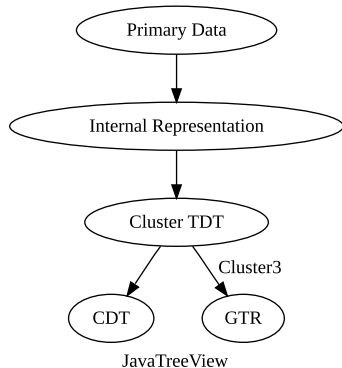
- Running Cluster3 from the command line
  - /Applications/Cluster.app/Contents/MacOS/Cluster
  - /Program Files/Stanford University/Cluster3/Cluster.com
- Command-line programs are like functions
- `\man program"` is like `\help(function)"`
- Use the `subprocess` module to run command-line programs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

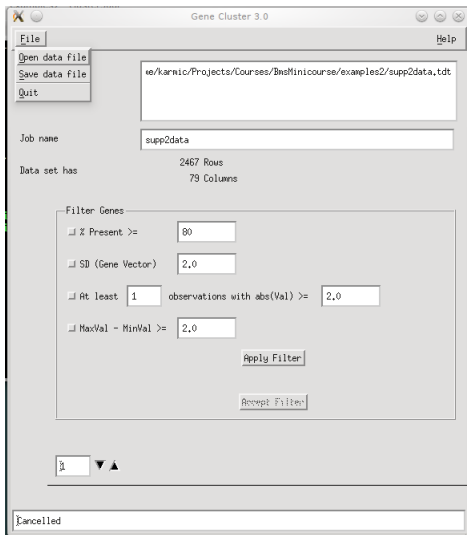
Apply Filter

Accept Filter

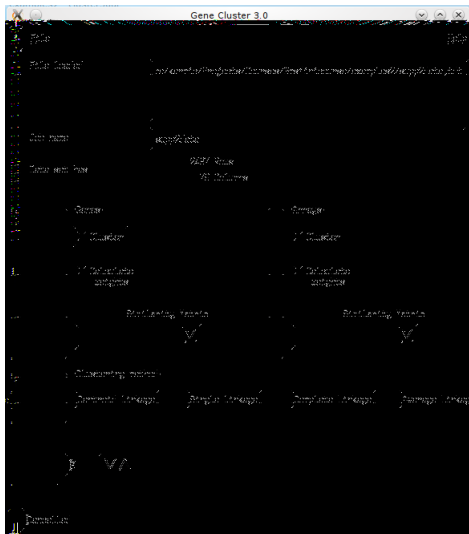
1

1

# Load your data



# Choose distance function





# Choose linking method



# Activate and detach annotation window

The screenshot shows the Java TreeView application window titled "java TreeView : /home/karmic/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 main window is divided into three panes:

- Left Pane:** A dendrogram showing hierarchical clustering of samples. A red vertical bar highlights a specific cluster.
- Center Pane:** A heatmap visualization where rows represent genes and columns represent samples. The color scale ranges from green (low expression) to red (high expression).
- Right Pane:** A table of gene annotations. The first column lists gene IDs, and the second column lists their associated biological processes.

**Usage Hints:** Click and drag to scroll

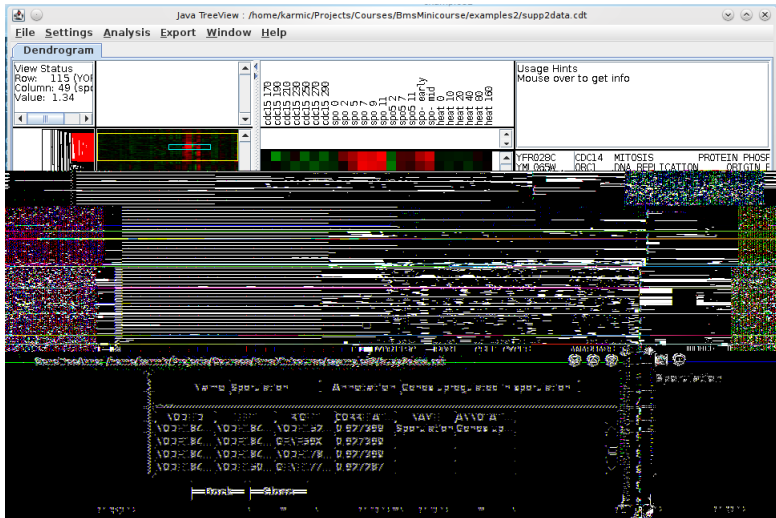
Gene ID	Annotation
YAL063W	GLUTAMATE BIOSYNTHESIS NADF
YOR375C	GLUTAMATE BIOSYNTHESIS GLU
YBR080C	SECRETION NSF; VESICLI
YMR072W	MITOCHONDRIAL GENOME MAI (PU
YDR311W	RHO3 CYTOSKELETON GTP-BIND
YOR274C	TFB1 TRANSCRIPTION TFIID 75
YML106C	INP52 ENDOCYTOSIS (PUTATIVE) INO
YML069W	POB3 DNA REPLICATION (PUTATIV BINI
YDR481C	PHO8 PHOSPHATE METABOLISM VACU
YFL021W	GAT1 NITROGEN CATABOLISM TRANS
YDR284C	DPP1 PHOSPHOLIPID METABOLISM DIA
YDR495W	MRF20 PROTEIN SYNTHESIS RIBOSOM
YAL029C	DPS2 TRANSPORT CA(2+) TRANS
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 COMPONENT
YML272C	SEC2 SECRETION GDP/GTP EXO
YOR075W	UFEL SECRETION ER MEMBRANE
YDR192C	NUP42 NUCLEAR PROTEIN TARGETIN NUCL
YDL224C	WH14 CELL SIZE PUTATIVE RN
YER112W	USJ1 MRNA SPLICING U6 SNRNP
YOR185W	REF2 MRNA 3'-END PROCESSING UNK
YER107C	GLE2 NUCLEAR PROTEIN TARGETIN NUCL
YHR208W	BAT1 BRANCHED CHAIN AMINO ACI TRAI
YER066W	MOT2 MATING TRANSCRIPTION
YDR148C	KGD2 TCA CYCLE 2-OXOGLUTAR
YDR204W	COQ4 UBIQUINONE BIOSYNTHESIS UNK

# 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 displays a table with columns: "NODEID", "LEFT", "RIGHT", "CORRELAT...", "NAME", and "ANNOTATI...". The table contains 25 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



# 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 { 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)))
```