Practical Bioinformatics

Mark Voorhies

4/6/2017

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
from importlib import reload
reload(my_module)
```

Pearson distances

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 distances

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 distances

Pearson similarity

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

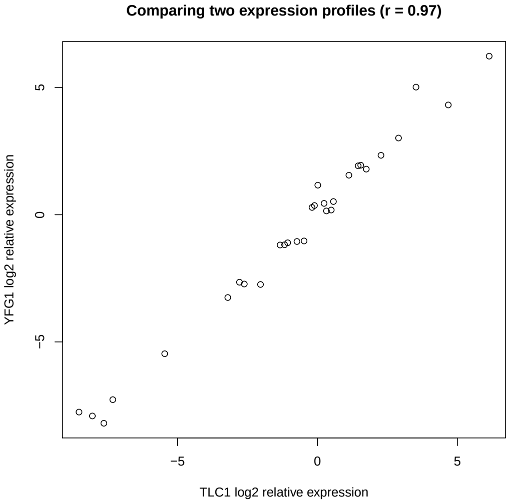
Pearson distance

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

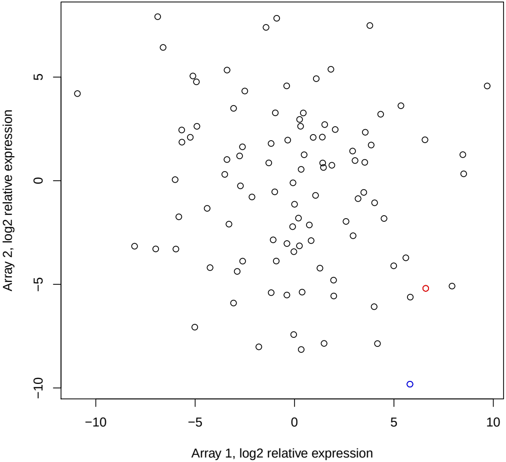
Euclidean distance

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

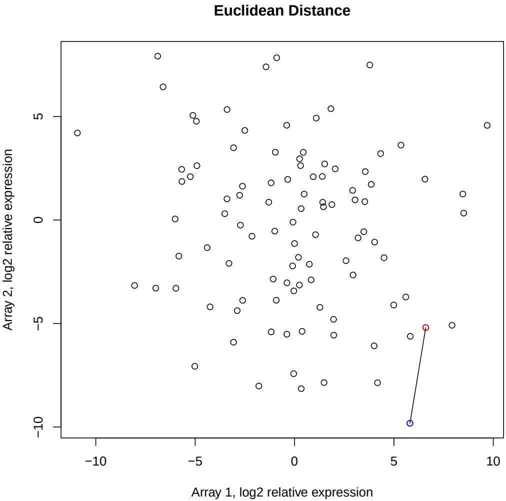
Comparing all measurements for two genes



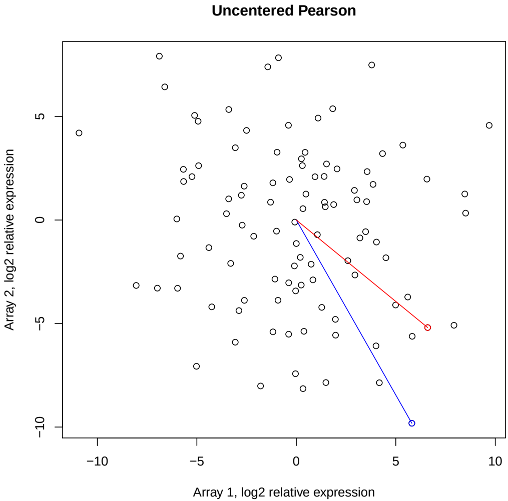
Comparing all genes for two measurements



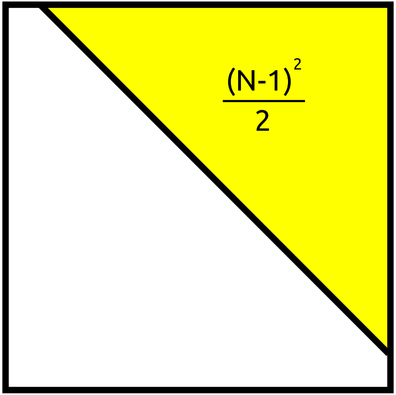
Comparing all genes for two measurements



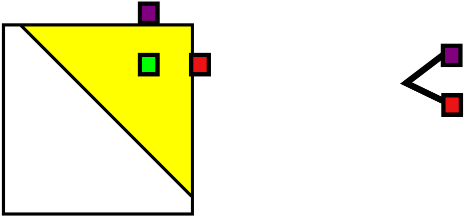
Comparing all genes for two measurements



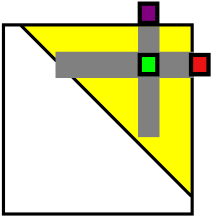
Measure all pairwise distances under distance metric

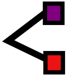


Hierarchical Clustering

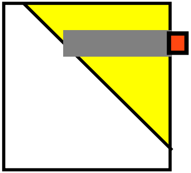


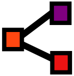
Hierarchical Clustering



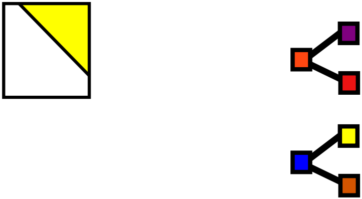


Hierarchical Clustering

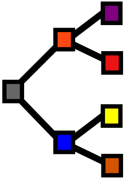




Hierarchical Clustering



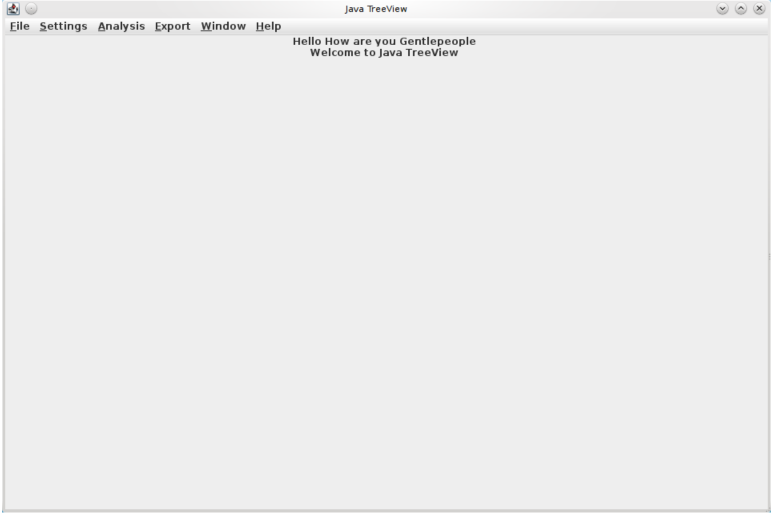
Hierarchical Clustering



> 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.

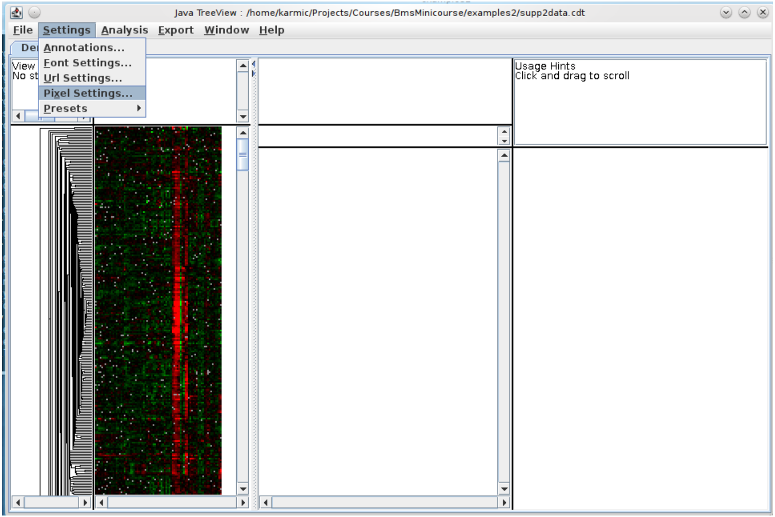
Bill Cleveland and Rick Becker
http://stat.bell-labs.com/project/trellis/interview.html

Using JavaTreeView

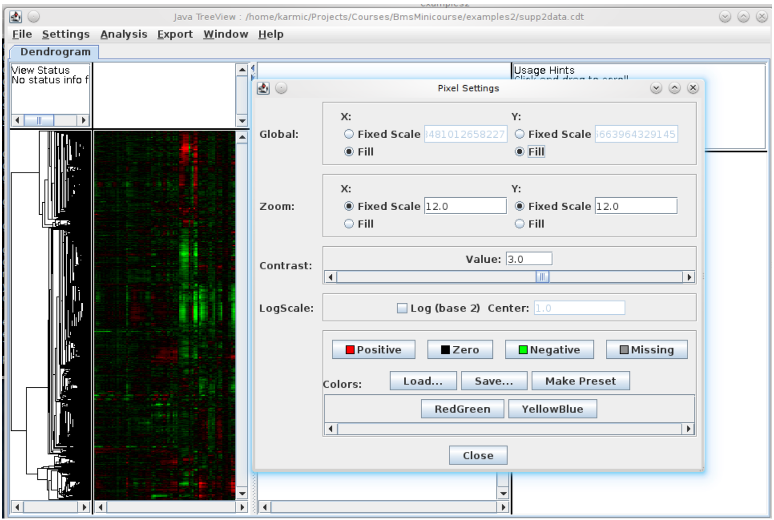


Hello How are you Gentlepeople
Welcome to Java TreeView

Adjust pixel settings for global view



Adjust pixel settings for global view



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

Pixel Settings

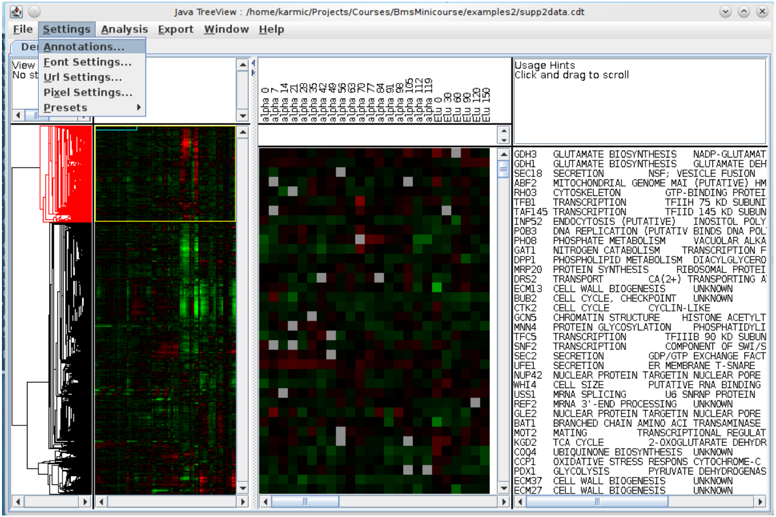
Global:
X: ○ Fixed Scale 481012658227 ● Fill
Y: ○ Fixed Scale 663964329145 ● Fill

Zoom:
X: ● Fixed Scale 12.0 ○ Fill
Y: ● Fixed Scale 12.0 ○ Fill

Contrast: Value: 3.0

LogScale: □ Log (base 2) Center: 1.0

Select annotation columns

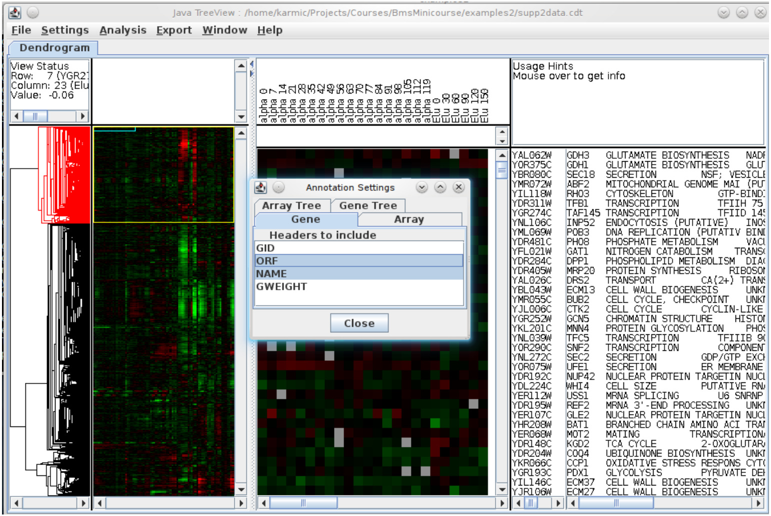


Java TreeView : /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt

Usage Hints
Click and drag to scroll

```
GDH3   GLUTAMATE BIOSYNTHESIS   NADP-GLUTAMAT
GDH1   GLUTAMATE BIOSYNTHESIS   GLUTAMATE DEH
SEC18  SECRETION           NSF; VESICLE FUSION
ABF2   MITOCHONDRIAL GENOME MAI (PUTATIVE) HM
RHO3   CYTOSKELETON          GTP-BINDING PROTEI
TFB1   TRANSCRIPTION         TFIIH 75 KD SUBUNI
TAF145 TRANSCRIPTION         TFIID 145 KD SUBUN
INP52  ENDOCYTOSIS (PUTATIVE)   INOSITOL POLY
POB3   DNA REPLICATION (PUTATIV BINDS DNA POL
PHO8   PHOSPHATE METABOLISM     VACUOLAR ALKA
GAT1   NITROGEN CATABOLISM    TRANSCRIPTION F
DPP1   PHOSPHOLIPID METABOLISM  DIACYLGLYCERO
MRP20  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
TFC5   TRANSCRIPTION         TFIIIB 90 KD SUBUN
SNF2   TRANSCRIPTION         COMPONENT OF SWI/S
SEC2   SECRETION           GDP/GTP EXCHANGE FACT
UFE1   SECRETION           ER MEMBRANE T-SNARE
NUP42  NUCLEAR PROTEIN TARGETIN NUCLEAR PORE
WHI4   CELL SIZE           PUTATIVE RNA BINDING
USS1   MRNA SPLICING         U6 SNRNP PROTEIN
REF2   MRNA 3'-END PROCESSING   UNKNOWN
GLE2   NUCLEAR PROTEIN TARGETIN NUCLEAR PORE
BAT1   BRANCHED CHAIN AMINO ACI TRANSAMINASE
MOT2   MATING             TRANSCRIPTIONAL REGULAT
KGD2   TCA CYCLE           2-OXOGLUTARATE DEHYDR
COQ4   UBIQUINONE BIOSYNTHESIS  UNKNOWN
CCP1   OXIDATIVE STRESS RESPONS CYTOCHROME-C
PDX1   GLYCOLYSIS          PYRUVATE DEHYDROGENAS
ECM37  CELL WALL BIOGENESIS     UNKNOWN
ECM27  CELL WALL BIOGENESIS     UNKNOWN
```

Select annotation columns



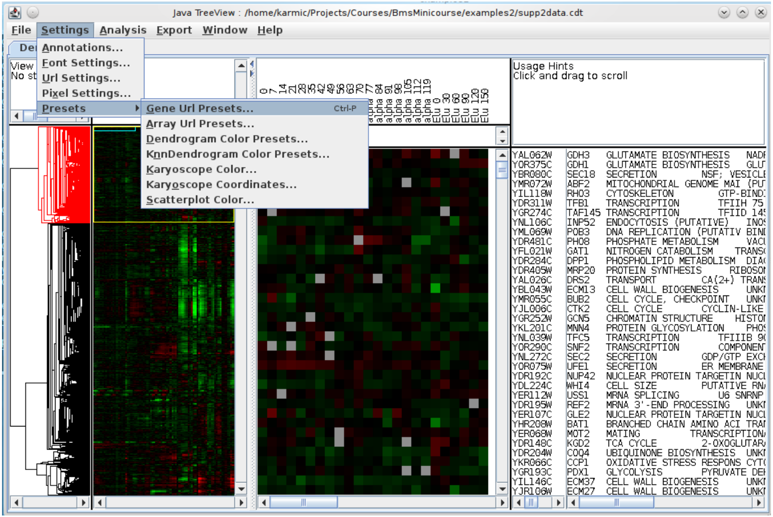
```
View Status
Row:    7 (YGR27
Column: 23 (Elu
Value:  -0.06
```

Headers to include

- GID
- ORF
- NAME
- GWEIGHT

```
YAL062W  GDH3   GLUTAMATE BIOSYNTHESIS   NADI
YOR375C  GDH1   GLUTAMATE BIOSYNTHESIS   GLUT
YBR080C  SEC18  SECRETION          NSF; VESICLI
YMR072W  ABF2   MITOCHONDRIAL GENOME MAI (PUT
YIL118W  RHO3   CYTOSKELETON          GTP-BINDI
YDR311W  TFB1   TRANSCRIPTION         TFIIH 75
YGR274C  TAF145 TRANSCRIPTION         TFIID 14
YNL106C  INP52  ENDOCYTOSIS (PUTATIVE)   INOS
YML069W  POB3   DNA REPLICATION (PUTATIV BINI
YDR481C  PHO8   PHOSPHATE METABOLISM     VACI
YFL021W  GAT1   NITROGEN CATABOLISM    TRANS
YDR284C  DPP1   PHOSPHOLIPID METABOLISM  DIA
YDR405W  MRP20  PROTEIN SYNTHESIS      RIBOSON
YAL026C  DRS2   TRANSPORT          CA(2+) TRANS
YBL043W  ECM13  CELL WALL BIOGENESIS     UNKI
YMR055C  BUB2   CELL CYCLE, CHECKPOINT   UNKI
YJL006C  CTK2   CELL CYCLE         CYCLIN-LIKE
YGR252W  GCN5   CHROMATIN STRUCTURE    HISTOI
YKL201C  MNN4   PROTEIN GLYCOSYLATION    PHOS
YNL039W  TFC5   TRANSCRIPTION         TFIIIB 9
YOR290C  SNF2   TRANSCRIPTION         COMPONENT
YNL272C  SEC2   SECRETION          GDP/GTP EXCI
YOR075W  UFE1   SECRETION          ER MEMBRANE
YDR192C  NUP42  NUCLEAR PROTEIN TARGETIN NUCI
YDL224C  WHI4   CELL SIZE          PUTATIVE RN
YER112W  USS1   MRNA SPLICING         U6 SNRNP
YDR195W  REF2   MRNA 3'-END PROCESSING   UNKI
YER107C  GLE2   NUCLEAR PROTEIN TARGETIN NUCI
YHR208W  BAT1   BRANCHED CHAIN AMINO ACI TRAI
YER068W  MOT2   MATING           TRANSCRIPTION
YDR148C  KGD2   TCA CYCLE          2-OXOGLUTAR
YDR204W  COQ4   UBIQUINONE BIOSYNTHESIS  UNKI
YKR066C  CCP1   OXIDATIVE STRESS RESPONS CYT
YGR193C  PDX1   GLYCOLYSIS         PYRUVATE DEI
YIL146C  ECM37  CELL WALL BIOGENESIS     UNKI
YJR106W  ECM27  CELL WALL BIOGENESIS     UNKI
```

Select URL for gene annotations

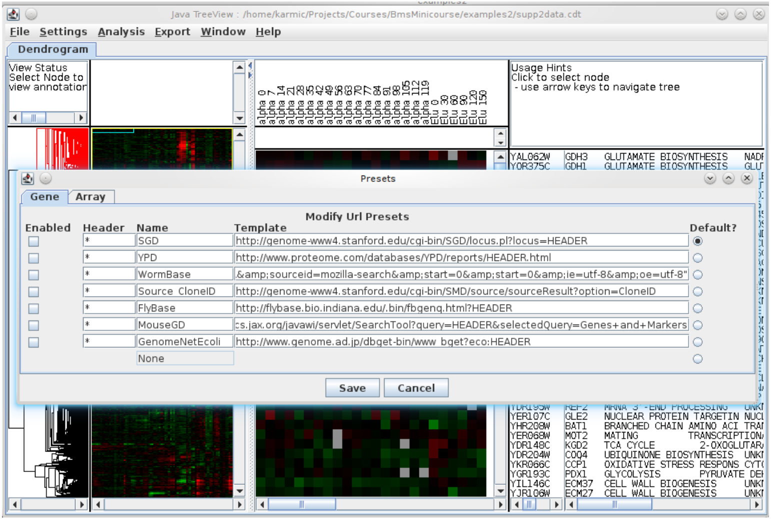


Java TreeView : /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt

Usage Hints
Click and drag to scroll

```
YAL062W  GDH3   GLUTAMATE BIOSYNTHESIS   NADF
YOR375C  GDH1   GLUTAMATE BIOSYNTHESIS   GLU
YBR080C  SEC18  SECRETION            NSF; VESICLE
YMR072W  ABF2   MITOCHONDRIAL GENOME MAI (PU
YIL118W  RHO3   CYTOSKELETON           GTP-BINDI
YDR311W  TFB1   TRANSCRIPTION          TFIIH 75
YGR274C  TAF145 TRANSCRIPTION          TFIID 145
YNL106C  INP52  ENDOCYTOSIS (PUTATIVE)   INOS
YML069W  POB3   DNA REPLICATION (PUTATIV BIND
YDR481C  PHO8   PHOSPHATE METABOLISM     VACU
YFL021W  GAT1   NITROGEN CATABOLISM    TRANSC
YDR284C  DPP1   PHOSPHOLIPID METABOLISM  DIAC
YDR405W  MRP20  PROTEIN SYNTHESIS      RIBOSON
YAL026C  DRS2   TRANSPORT            CA(2+) TRANS
YBL043W  ECM13  CELL WALL BIOGENESIS     UNKN
YMR055C  BUB2   CELL CYCLE, CHECKPOINT   UNKN
YJL006C  CTK2   CELL CYCLE           CYCLIN-LIKE
YGR252W  GCN5   CHROMATIN STRUCTURE    HISTON
YKL201C  MNN4   PROTEIN GLYCOSYLATION    PHOS
YNL039W  TFC5   TRANSCRIPTION          TFIIIB 90
YOR290C  SNF2   TRANSCRIPTION          COMPONENT
YNL272C  SEC2   SECRETION            GDP/GTP EXCH
YOR075W  UFE1   SECRETION            ER MEMBRANE
YDR192C  NUP42  NUCLEAR PROTEIN TARGETIN NUCL
YDL224C  WHI4   CELL SIZE            PUTATIVE RNA
YER112W  USS1   MRNA SPLICING          U6 SNRNP
YDR195W  REF2   MRNA 3'-END PROCESSING   UNKN
YER107C  GLE2   NUCLEAR PROTEIN TARGETIN NUCL
YHR208W  BAT1   BRANCHED CHAIN AMINO ACI TRAN
YER068W  MOT2   MATING             TRANSCRIPTION
YDR148C  KGD2   TCA CYCLE            2-OXOGLUTARA
YDR204W  COQ4   UBIQUINONE BIOSYNTHESIS  UNKN
YKR066C  CCP1   OXIDATIVE STRESS RESPONS CYTC
YGR193C  PDX1   GLYCOLYSIS           PYRUVATE DEH
YIL146C  ECM37  CELL WALL BIOGENESIS     UNKN
YJR106W  ECM27  CELL WALL BIOGENESIS     UNKN
```

Select URL for gene annotations



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

View Status
Select Node to
view annotation

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

```
YAL062W   GDH3    GLUTAMATE BIOSYNTHESIS  NADP
YOR375C   GDH1    GLUTAMATE BIOSYNTHESIS  GLU
```

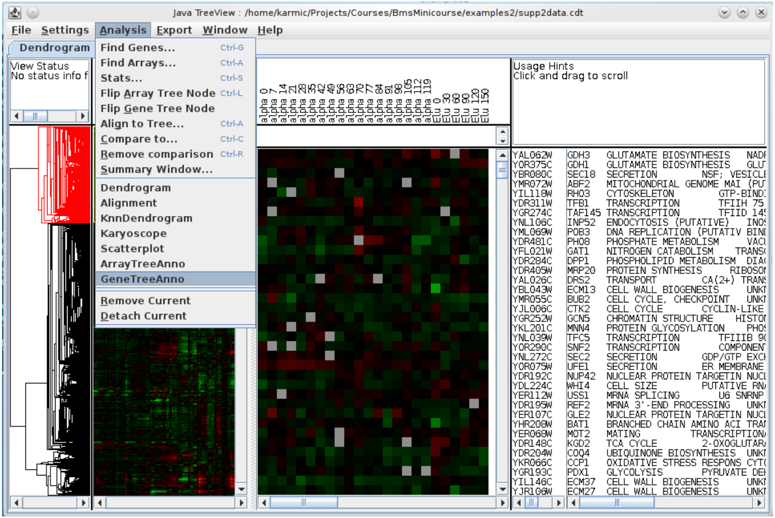
Presets

Modify Url Presets

Enabled	Header	Name	Template	Default?
☐	*	SGD	http://genome-www4.stanford.edu/cgi-bin/SGD/locus.pl?locus=HEADER	●
☐	*	YPD	http://www.proteome.com/databases/YPD/reports/HEADER.html	○
☐	*	WormBase	&amp;sourceid=mozilla-search&amp;start=0&amp;start=0&amp;ie=utf-8&amp;oe=utf-8"	○
☐	*	Source CloneID	http://genome-www4.stanford.edu/cgi-bin/SMD/source/sourceResult?option=CloneID	○
☐	*	FlyBase	http://flybase.bio.indiana.edu/.bin/fbgenq.html?HEADER	○
☐	*	MouseGD	cs.jax.org/javawi/servlet/SearchTool?query=HEADER&selectedQuery=Genes+and+Markers	○
☐	*	GenomeNetEcoli	http://www.genome.ad.jp/dbget-bin/www_bget?eco:HEADER	○
		None		○

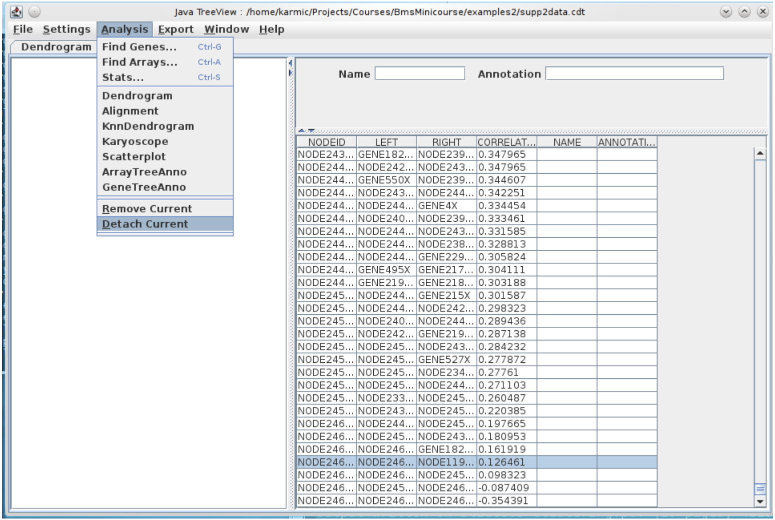
```
YDR195W   REF2    MRNA 3'-END PROCESSING  UNKN
YER107C   GLE2    NUCLEAR PROTEIN TARGETIN NUCL
YHR208W   BAT1    BRANCHED CHAIN AMINO ACI TRAN
YER068W   MOT2    MATING           TRANSCRIPTION
YDR148C   KGD2    TCA CYCLE          2-OXOGLUTAR
YDR204W   COQ4    UBIQUINONE BIOSYNTHESIS  UNKN
YKR066C   CCP1    OXIDATIVE STRESS RESPONS CYTO
YGR193C   PDX1    GLYCOLYSIS         PYRUVATE DEH
YIL146C   ECM37   CELL WALL BIOGENESIS     UNKN
YJR106W   ECM27   CELL WALL BIOGENESIS     UNKN
```

Activate and detach annotation window



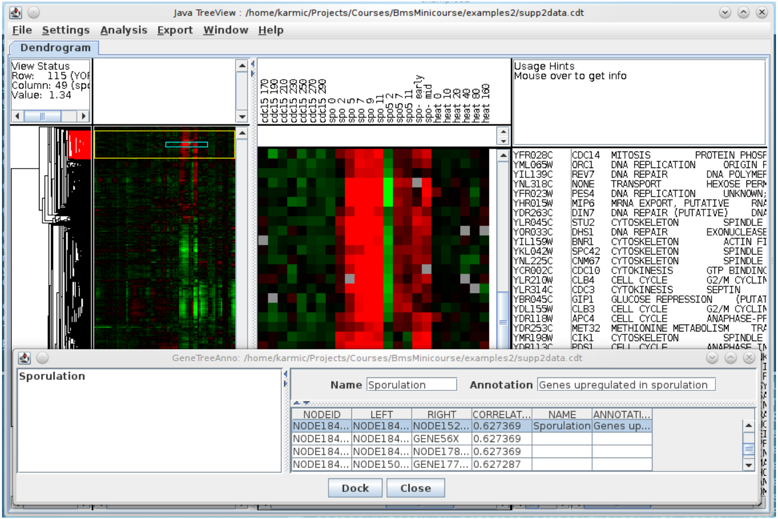
```
YAL062W  GDH3    GLUTAMATE BIOSYNTHESIS   NADF
YOR375C  GDH1    GLUTAMATE BIOSYNTHESIS   GLU
YBR080C  SEC18   SECRETION           NSF; VESICLE
YMR072W  ABF2    MITOCHONDRIAL GENOME MAI (PU
YIL118W  RHO3    CYTOSKELETON           GTP-BINDI
YDR311W  TFB1    TRANSCRIPTION          TFIIH 75
YGR274C  TAF145  TRANSCRIPTION          TFIID 145
YNL106C  INP52   ENDOCYTOSIS (PUTATIVE)   INOS
YML069W  POB3    DNA REPLICATION (PUTATIV BIND
YDR481C  PHO8    PHOSPHATE METABOLISM     VACU
YFL021W  GAT1    NITROGEN CATABOLISM    TRANSC
YDR284C  DPP1    PHOSPHOLIPID METABOLISM  DIAC
YDR405W  MRP20   PROTEIN SYNTHESIS      RIBOSON
YAL026C  DRS2    TRANSPORT           CA(2+) TRANS
YBL043W  ECM13   CELL WALL BIOGENESIS     UNKN
YMR055C  BUB2    CELL CYCLE, CHECKPOINT   UNKN
YJL006C  CTK2    CELL CYCLE          CYCLIN-LIKE
YGR252W  GCN5    CHROMATIN STRUCTURE    HISTON
YKL201C  MNN4    PROTEIN GLYCOSYLATION    PHOS
YNL039W  TFC5    TRANSCRIPTION          TFIIIB 90
YOR290C  SNF2    TRANSCRIPTION          COMPONENT
YNL272C  SEC2    SECRETION           GDP/GTP EXCH
YOR075W  UFE1    SECRETION           ER MEMBRANE
YDR192C  NUP42   NUCLEAR PROTEIN TARGETIN NUCL
YDL224C  WHI4    CELL SIZE           PUTATIVE RNA
YER112W  USS1    MRNA SPLICING          U6 SNRNP
YDR195W  REF2    MRNA 3'-END PROCESSING   UNKN
YER107C  GLE2    NUCLEAR PROTEIN TARGETIN NUCL
YHR208W  BAT1    BRANCHED CHAIN AMINO ACI TRAN
YER068W  MOT2    MATING            TRANSCRIPTIONA
YDR148C  KGD2    TCA CYCLE           2-OXOGLUTARA
YDR204W  COQ4    UBIQUINONE BIOSYNTHESIS  UNKN
YKR066C  CCP1    OXIDATIVE STRESS RESPONS CYTC
YGR193C  PDX1    GLYCOLYSIS          PYRUVATE DEH
YIL146C  ECM37   CELL WALL BIOGENESIS     UNKN
YJR106W  ECM27   CELL WALL BIOGENESIS     UNKN
```

Activate and detach annotation window



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



View Status
Row: 115 (YOF
Column: 49 (spo
Value: 1.34

Usage Hints
Mouse over to get info

YFR028C	CDC14	MITOSIS	PROTEIN PHOSF
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	DIN7	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
YCR002C	CDC10	CYTOKINESIS	GTP BINDING
YLR210W	CLB4	CELL CYCLE	G2/M CYCLIN
YLR314C	CDC3	CYTOKINESIS	SEPTIN
YBR045C	GIP1	GLUCOSE REPRESSION	(PUTAT
YDL155W	CLB3	CELL CYCLE	G2/M CYCLIN
YDR118W	APC4	CELL CYCLE	ANAPHASE-PF
YDR253C	MET32	METHIONINE METABOLISM	TRA
YMR198W	CIK1	CYTOSKELETON	SPINDLE
YDR113C	PDS1	CELL CYCLE	ANAPHASE IN

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		

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 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).

```python
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])
```

Homework

1. Explore different clustering methods and/or distance methods
2. Try additional shufflings of the data: how do they affect your ability to cluster the data? *C.f. figure 3 the Eisen paper*
    - Permute the columns
    - Independently permute the columns of each row