Practical Bioinformatics

Mark Voorhies

5/15/2019

Gotchas

We are currently using three different languages in Jupyter

- By default, code cells are in Python
- Cells that start with % are "Jupyter magic"

```
%cd
```

- Cells that start with ! are in Bash

```
!wget 'http://histo.ucsf.edu/BMS270/'
```

Gotchas

We are currently using three different languages in Jupyter

- By default, code cells are in Python
- Cells that start with % are "Jupyter magic"

```
%cd
```

- Cells that start with ! are in Bash

```
!wget 'http://histo.ucsf.edu/BMS270/'
```

- Jupyter will *sometimes* accept bash commands without the !, but *don't* make this a habit

Gotchas

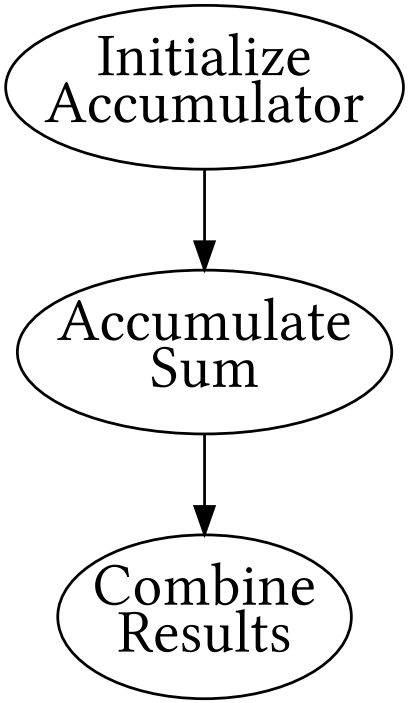
- Statements that precede code blocks (if, def, for, while, ...) end with a colon.

```
def mean(x):
    s = 0.0
    for i in x:
        s += i
    return s/len(x)
```

- You can use tab and shift-tab in Jupyter to indent/unindent blocks of code

Mean

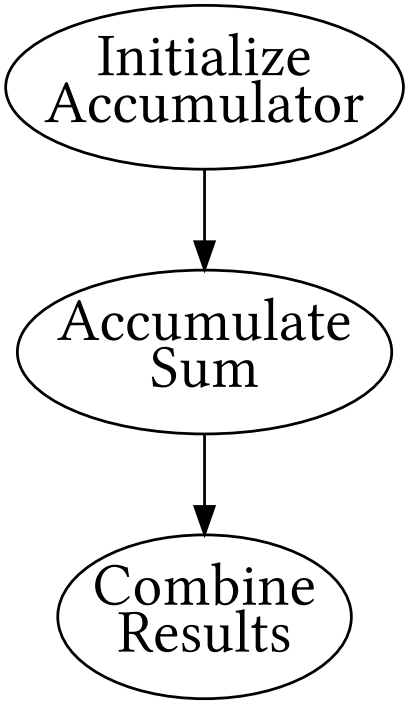
```
def mean(x):
    s = 0.0
    for i in x:
        s += i
    return s/len(x)
```



Mean

```
def mean(x):
    s = 0.0
    for i in x:
        s += i
    return s/len(x)

def mean(x):
    return sum(x)/float(len(x))
```



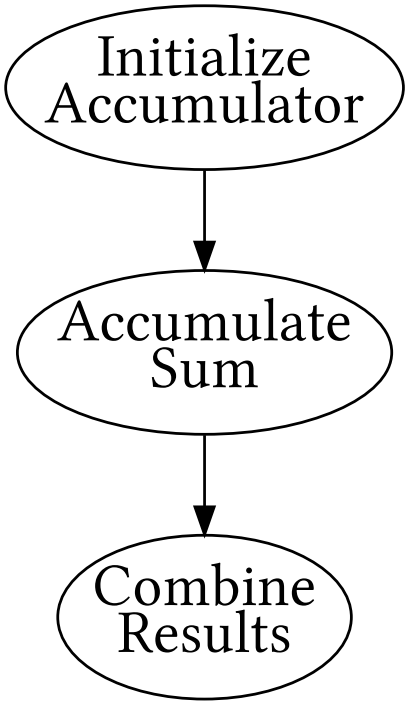
Standard Deviation

$$\sigma_x = \sqrt{\frac{\sum_i^N (x_i - \bar{x})^2}{N-1}}$$

Standard Deviation

$$\sigma_x = \sqrt{\frac{\sum_i^N (x_i - \bar{x})^2}{N-1}}$$

```
def stdev(x):
    m = mean(x)
    s = 0.0
    for i in x:
        s += (i - m)**2
    return (s/(len(x) - 1))**.5
```



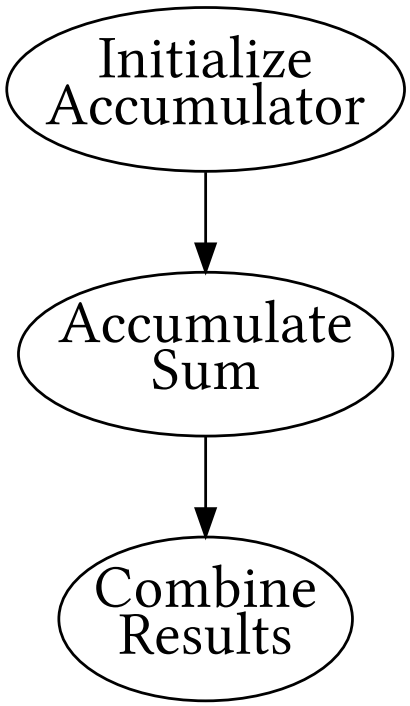
Pearson's Correlation Coefficient

$$r(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Pearson's Correlation Coefficient

```
def pearson(x,y):
    mx = mean(x)
    my = mean(y)
    sxy = 0.0
    ssx = 0.0
    ssy = 0.0
    for i in range(len(x)):
        dx = x[i] - mx
        dy = y[i] - my
        sxy += dx*dy
        ssx += dx**2
        ssy += dy**2
    return sxy/((ssx*ssy)**.5)
```

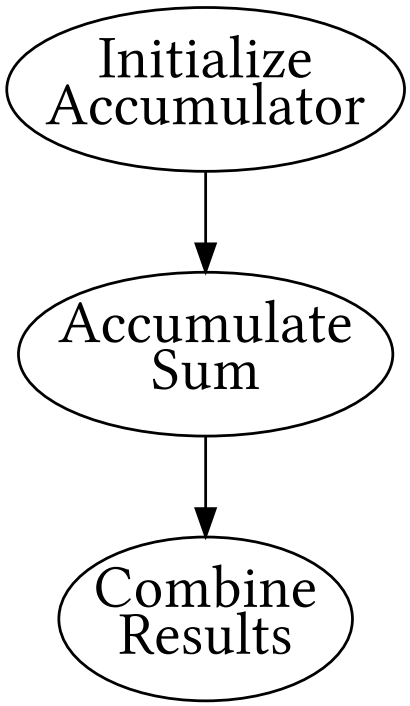
$$r(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$



Pearson's Correlation Coefficient

```
def pearson(x,y):
    mx = mean(x)
    my = mean(y)
    sxy = 0.0
    ssx = 0.0
    ssy = 0.0
    for i,j in zip(x,y):
        dx = i - mx
        dy = j - my
        sxy += dx*dy
        ssx += dx**2
        ssy += dy**2
    return sxy/((ssx*ssy)**.5)
```

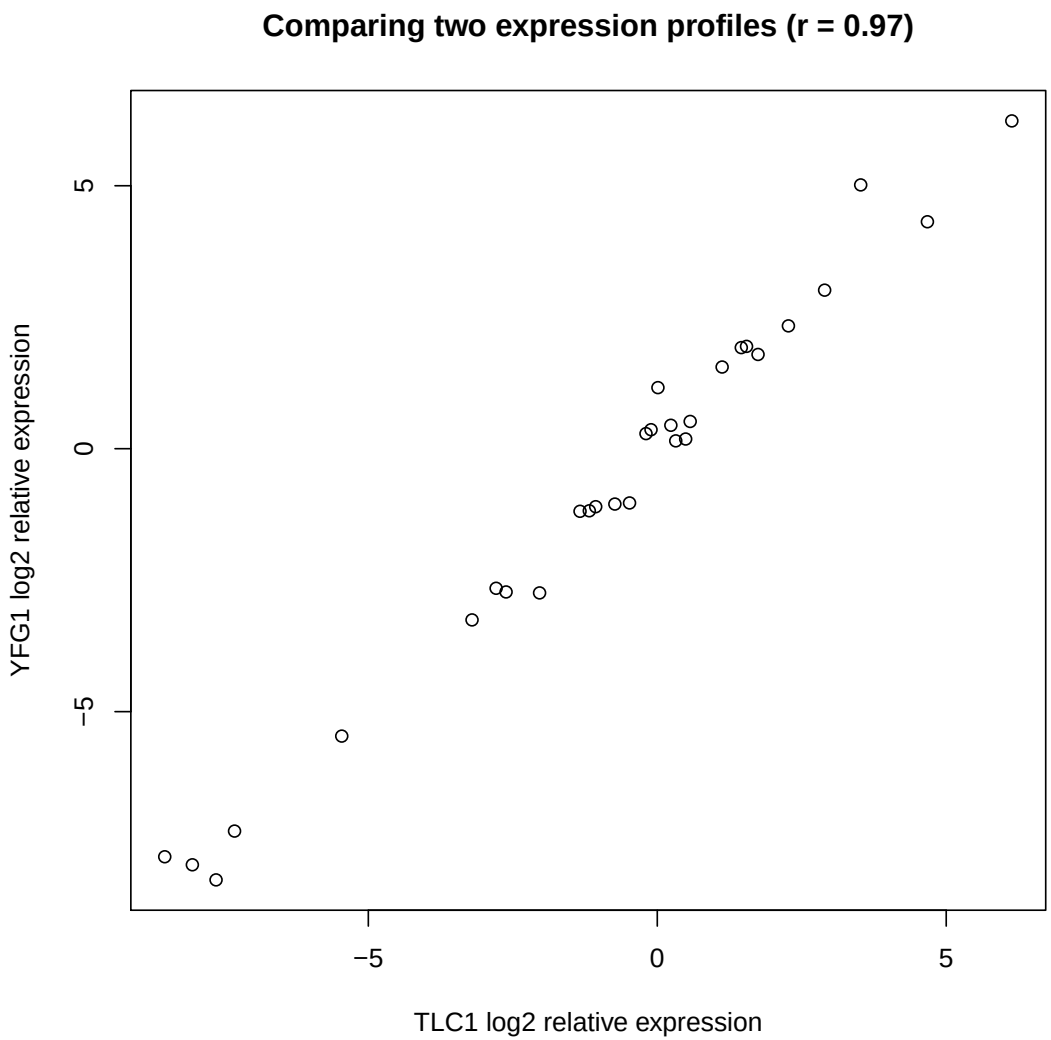
$$r(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$



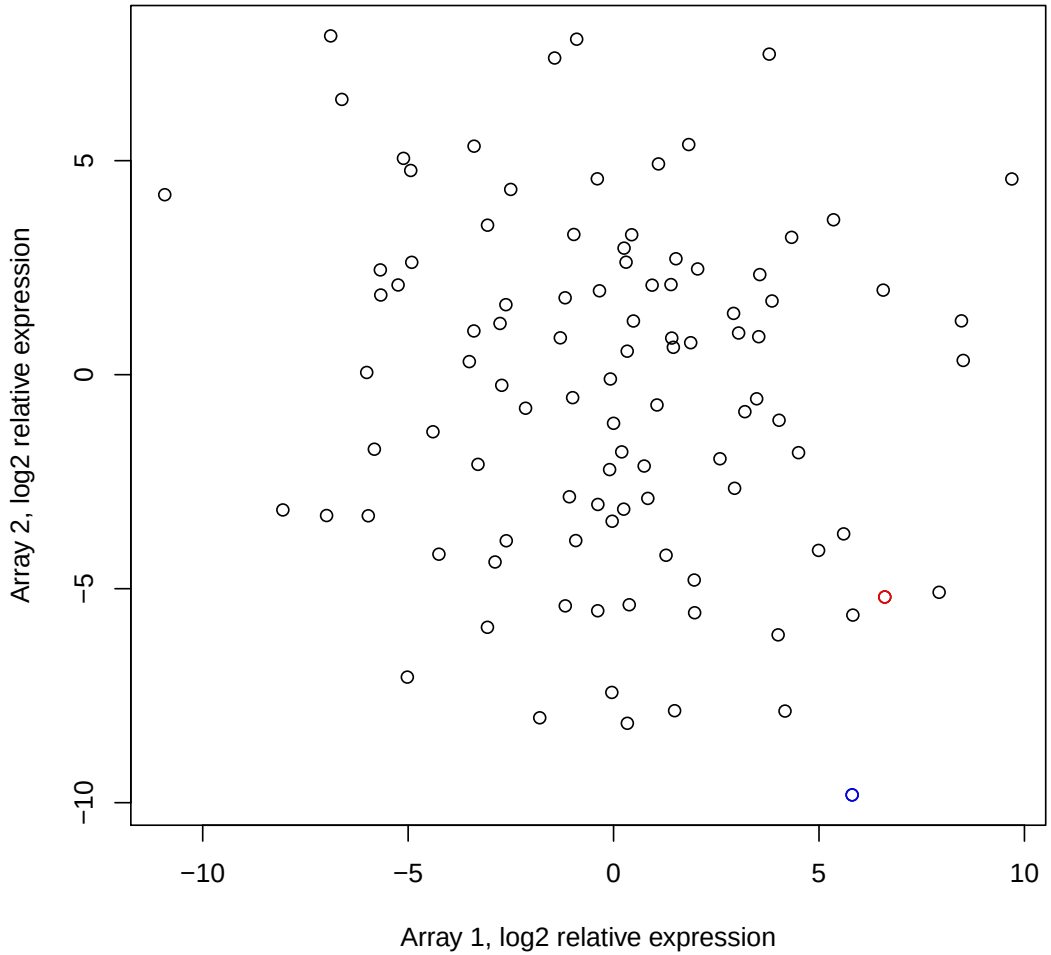
[T]he relational graphic – in its barest form, the scatterplot and its variants – is the greatest of all graphical designs. It links at least two variables, encouraging and even imploring the viewer to assess the possible causal relationship between the plotted variables.

–Edward Tufte

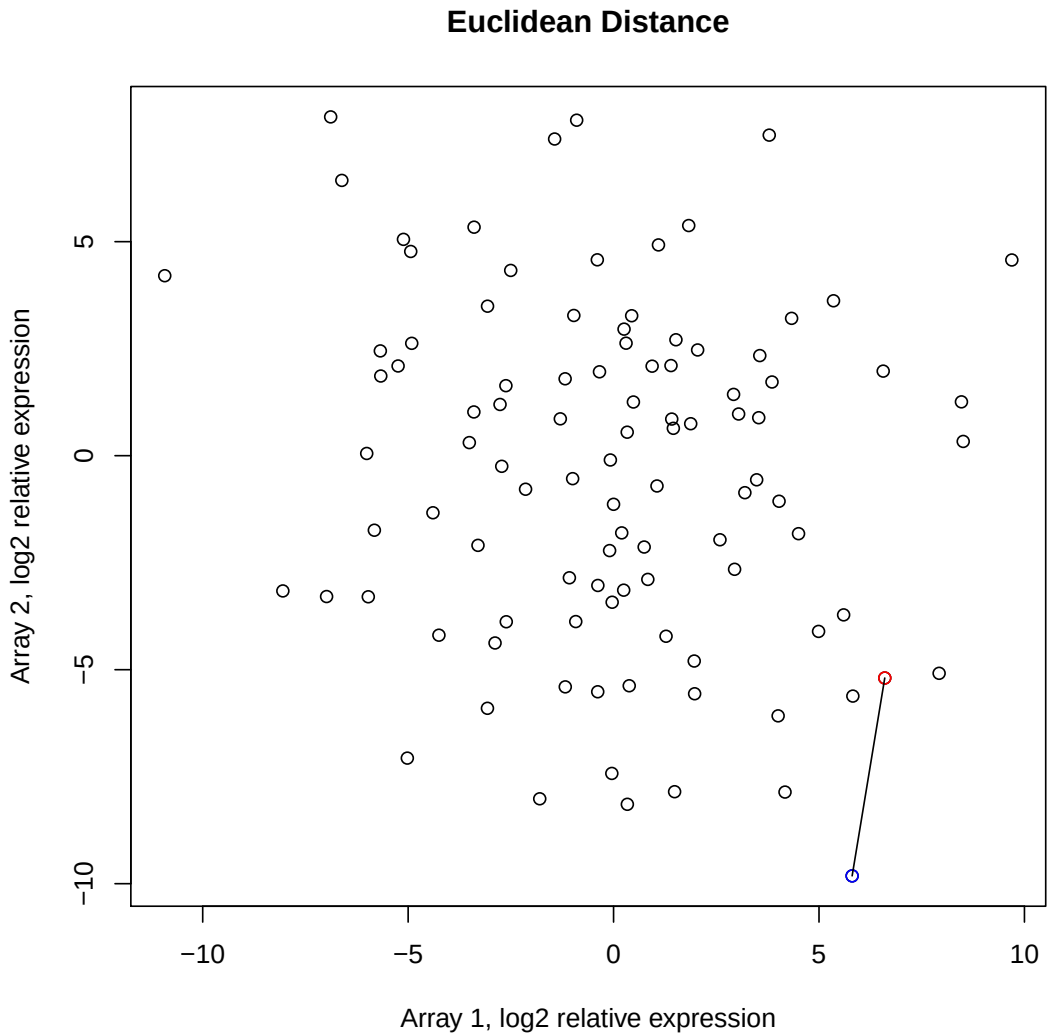
Comparing all measurements for two genes



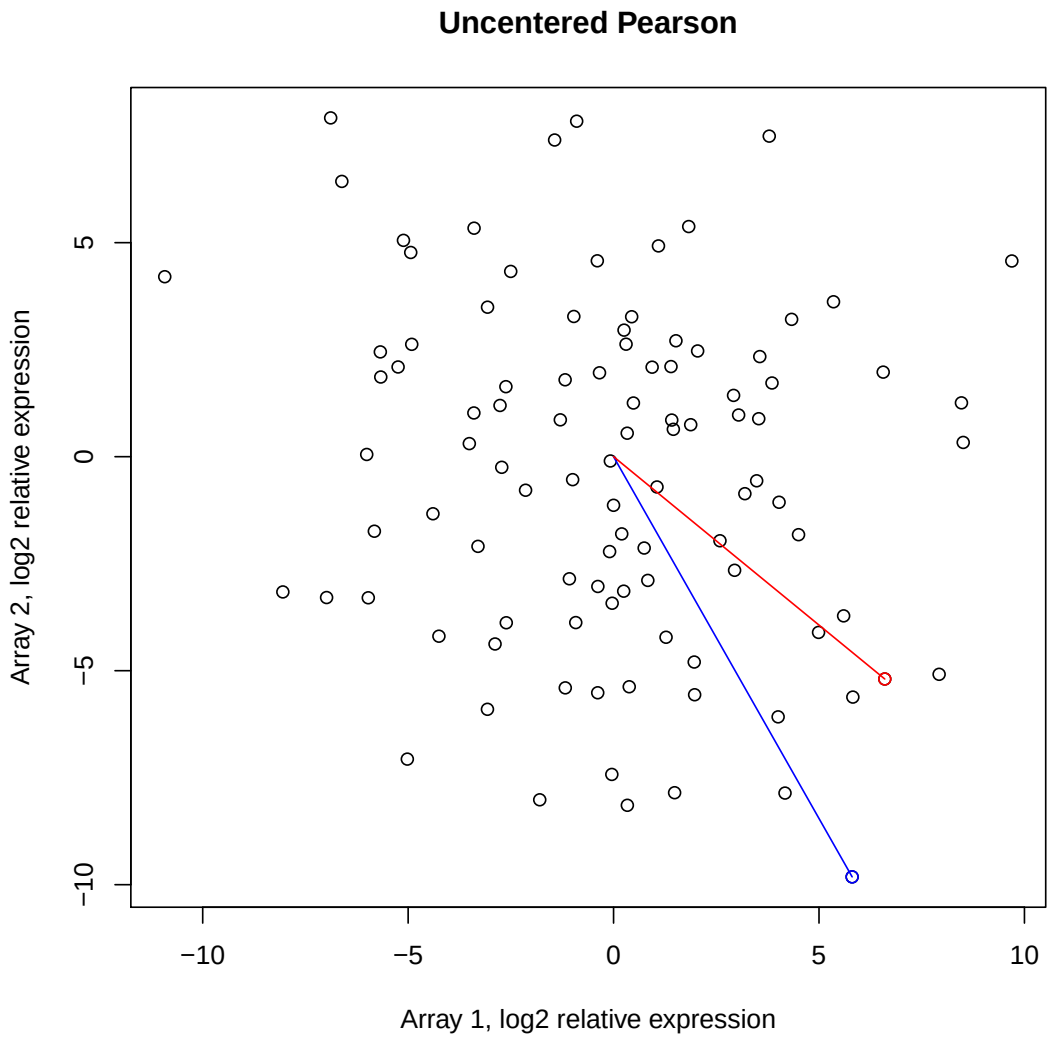
Comparing all genes for two measurements



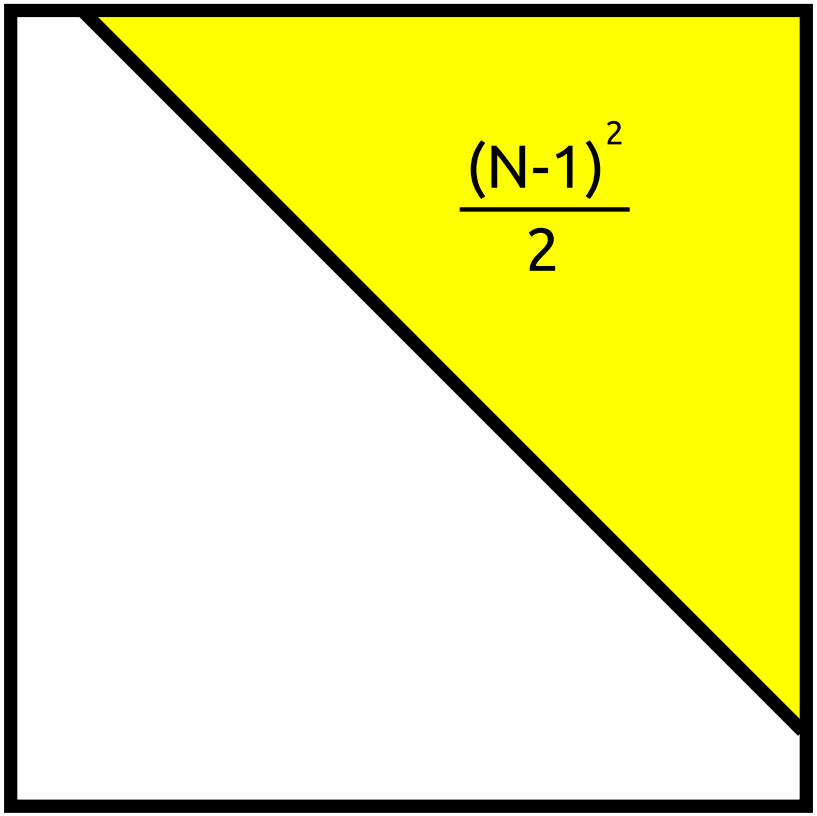
Comparing all genes for two measurements



Comparing all genes for two measurements



Measure all pairwise distances under distance metric



List tricks

Adding data to a list:

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

List tricks

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]]
```

Homework

1. Download and install JavaTreeView
2. Write a function to calculate all pairwise Pearson correlations for the first N rows of the macrophage expression profiles.
   - Start with $N = 10$, then work up to $N = 4000$
   - The intrepid may choose to work up to the full data set, but note that this may lock up both your guest and host machine due to running out of RAM.
3. *Optional:* Read PNAS 95:14863
4. *Optional:* Repeat problem 2, replacing the Pearson correlation with the distance metric from the PNAS paper or with one of the distance metrics from the Cluster3 manual.