Practical Bioinformatics

Mark Voorhies

6/16/2010

Review

- Files
- Member functions
- dir()

Gotchas

- Locating files

```
from tkFileDialog import askopenfilename
print askopenfilename()
```

Gotchas

- Locating files

```
from tkFileDialog import askopenfilename
print askopenfilename()
```

- Forward slash (/) is the path separator, backslash (\) is the escape character
  - "/path/to/my/file"
  - "Text with tabs:\t newlines:\n and carriage-returns:\r"

Gotchas

- Locating files

```
from tkFileDialog import askopenfilename
print askopenfilename()
```

- Forward slash (/) is the path separator, backslash (\) is the escape character
  - "/path/to/my/file"
  - "Text with tabs:\t newlines:\n and carriage-returns:\r"
- Reading from a file moves the file pointer

Gotchas

- Locating files

```
from tkFileDialog import askopenfilename
print askopenfilename()
```

- Forward slash (/) is the path separator, backslash (\) is the escape character
  - "/path/to/my/file"
  - "Text with tabs:\t newlines:\n and carriage-returns:\r"
- Reading from a file moves the file pointer
- When indentation gets tricky in an interactive session, switch to using modules.

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]]
```

File exercises

1. Write a function to read a file formatted like supp2data.tdt and return a list of the lines in the file.
2. Change the function to return a list of lists of fields from the file (*i.e.*, the equivalent of the table that you would see in a spreadsheet).
3. Change the function to return [genes, annotations, ratios] where:
   - genes is a list corresponding to the first column of the file
   - annotations is a list corresponding to the second column of the file
   - ratios is a list of lists corresponding to the matrix of ratios

   and the header line is omitted.

Coercing data

Converting between data types:

```
number = float("3.5")
text = str(3.5)
```

Coercing data

Converting between data types:

```
number = float("3.5")
text = str(3.5)
```

Catching exceptions:

```
try:
    a = float("twenty")
except:
    a = None

if(i < 0):
    raise ValueError
```

More file exercises

1. Update your function to return the ratio matrix as floating point numbers rather than strings. Empty cells should be set to 0.0 or None (your choice).

Microarray workflow

- Experiment
- Data aquisition (optical)
- Pre-processing (image analysis and normalization)
  - *e.g.*, SpotReader, GenePix, ...
- Archival
  - *e.g.*, NOMAD, Acuity, MADAM, GEO, ...
- Analysis
  - Clustering, data reduction
    - *e.g.*, Cluster3, MeV, Acuity, R/Bioconductor, ...
  - Annotation, aggregation
  - Significance tests

Analysis workflow

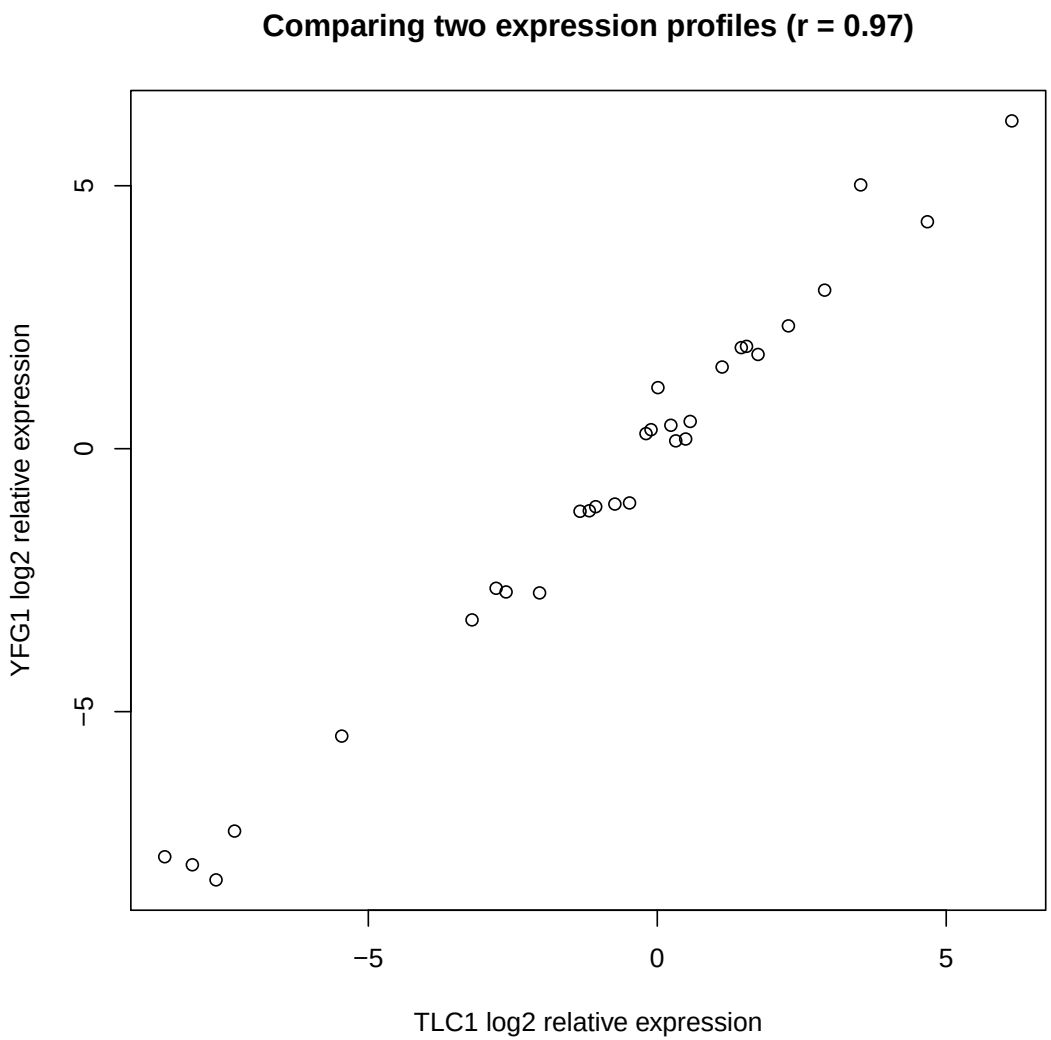
- Calculate pairwise distances
- Find a simplified representation of the distance matrix

Analysis workflow

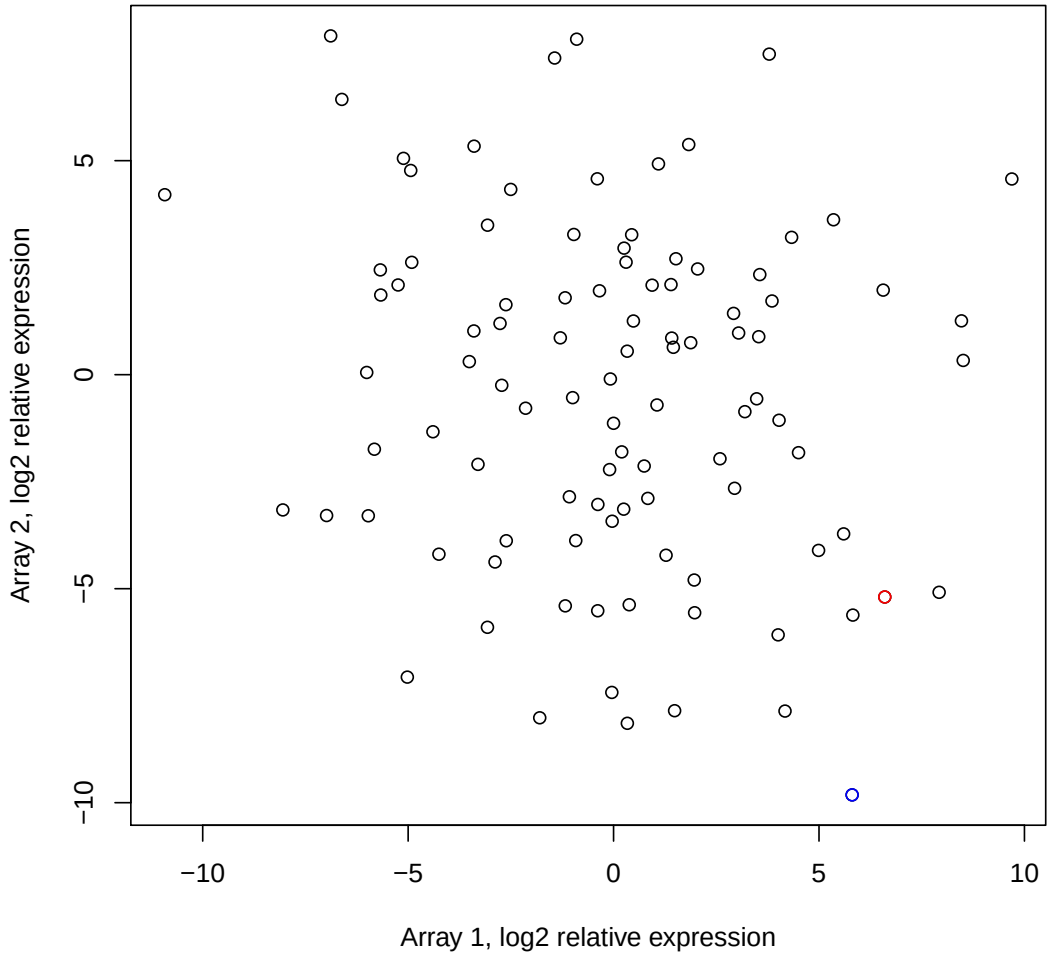
- Calculate pairwise distances
- Find a simplified representation of the distance matrix

```
def cluster(distances):
    # Initialize tree with leaf nodes
    while(len(distances) > 1):
        # Find closest pair
        # Link pair in the tree
        # Update distances based on linking rule
```

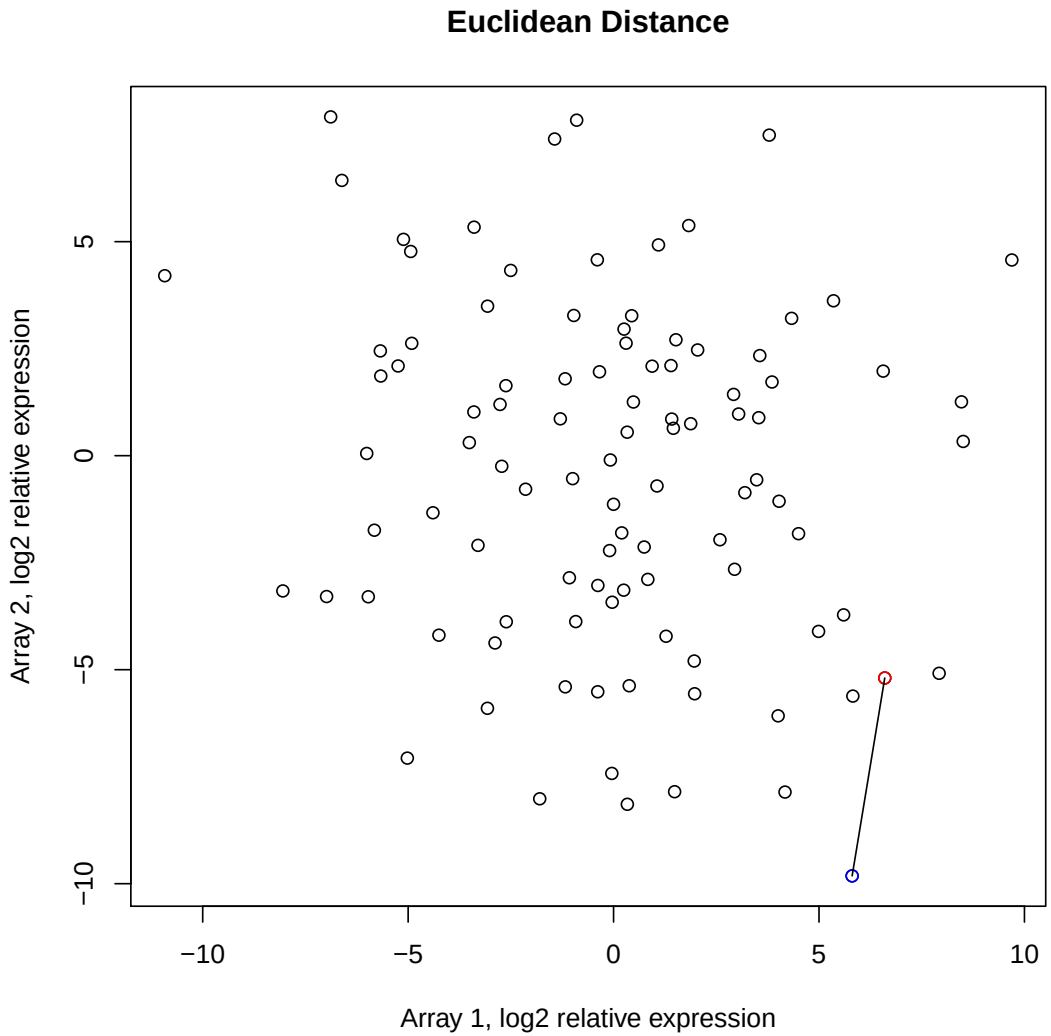
Comparing all measurements for two genes



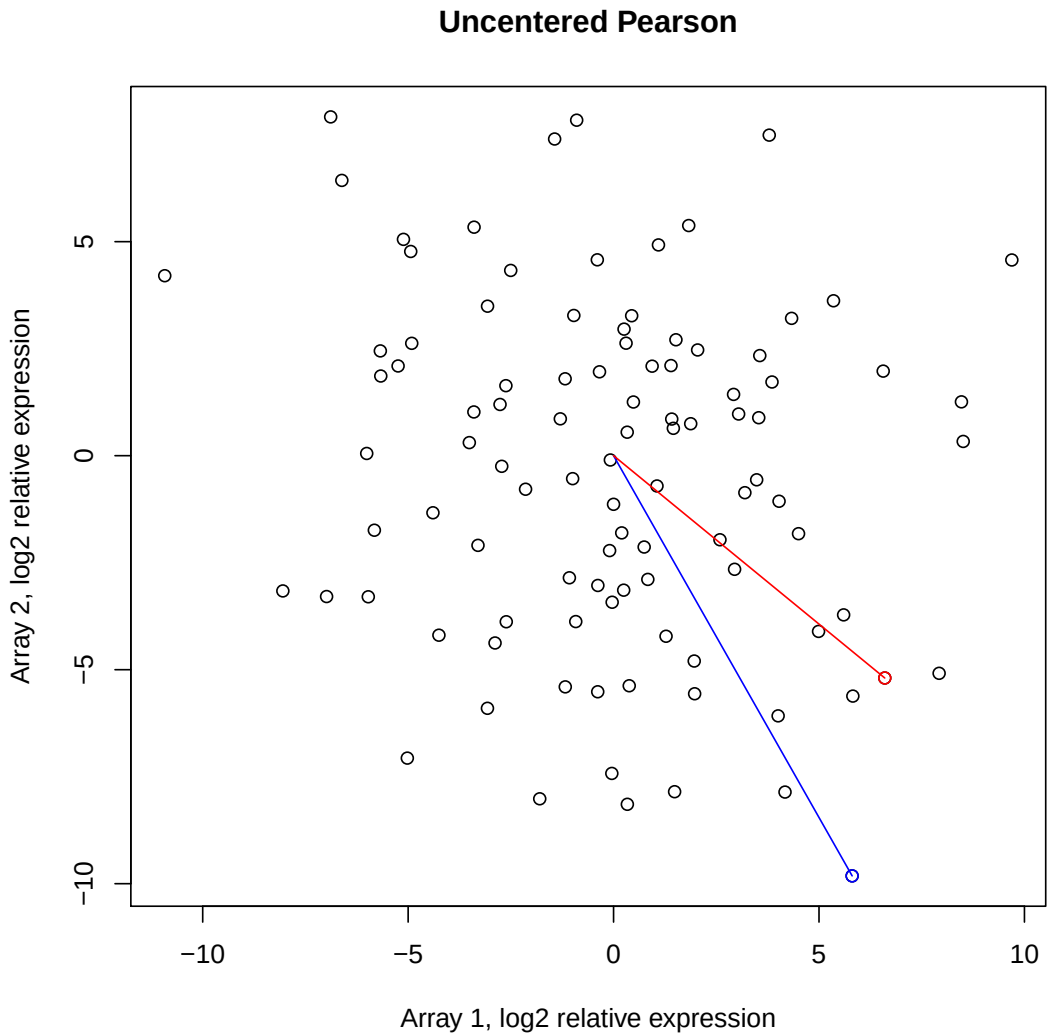
Comparing all genes for two measurements



Comparing all genes for two measurements



Comparing all genes for two measurements



Pearson distances

Pearson similarity

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

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

Pearson distances

Pearson similarity

$$s(x,y) = \sum_{i}^{N} \left(\frac{x_i - x_{offset}}{\phi_x} \right) \left(\frac{y_i - y_{offset}}{\phi_y} \right) \tag{3}$$

$$\phi_G = \sqrt{\sum_{i}^{N} (G_i - G_{offset})^2} \tag{4}$$

Pearson distances

Pearson similarity

$$s(x,y) = \sum_{i}^{N} \left(\frac{x_i - x_{offset}}{\sqrt{\sum_i^N (x_i - x_{offset})^2}} \right) \left(\frac{y_i - y_{offset}}{\sqrt{\sum_i^N (y_i - y_{offset})^2}} \right) \quad (5)$$

Pearson distances

Pearson similarity

$$s(x,y) = \frac{\sum_i^N (x_i - x_{offset})(y_i - y_{offset})}{\sqrt{\sum_i^N (x_i - x_{offset})^2}\sqrt{\sum_i^N (y_i - y_{offset})^2}} \quad (6)$$

Pearson distances

Pearson similarity

$$s(x,y) = \frac{\sum_i^N (x_i - x_{offset})(y_i - y_{offset})}{\sqrt{\sum_i^N (x_i - x_{offset})^2}\sqrt{\sum_i^N (y_i - y_{offset})^2}} \tag{6}$$

Pearson distance

$$d_{uncentered}(x,y) = 1 - s(x,y) \tag{7}$$

Distance functions

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

$$d(x,y) = 1 - \frac{\sum_i^N (x_i - x_{offset})(y_i - y_{offset})}{\sqrt{\sum_i^N (x_i - x_{offset})^2}\sqrt{\sum_i^N (y_i - y_{offset})^2}} \quad (8)$$

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.
4. Save the results of your pairwise distance calculation in the CDT format described in the JavaTreeView manual.