BIO508: Lab Session 8

Common mistakes in HW4 & HW5

- `reFASTA` function. Common mistakes: `reFASTA( "> i dee " )` returns `" i dee"` or `"i dee "`. `reFASTA( "> ATGGC" )` returns `None`.

```
def reFASTA( strLine ):
  """
>>> reFASTA( "> id" )
'id'

>>> reFASTA( ">id id" )
'id id'

>>> reFASTA( "> i dee " )
'i dee'

>>> reFASTA( "ATGGC" )
''
"""

mtch = re.search( r'^>\s*(.*?)\s*$', strLine )
return ( mtch.group( 1 ) if mtch else "" )
```

- `rePOS` function. Common mistakes: `rePOS( "compliementary" )` returns `"None"` or `[]`.

Different distance metrics

A visual example for the differences between Pearson correlation, Spearman correlation, and Euclidean distance:

Euclidean Distance

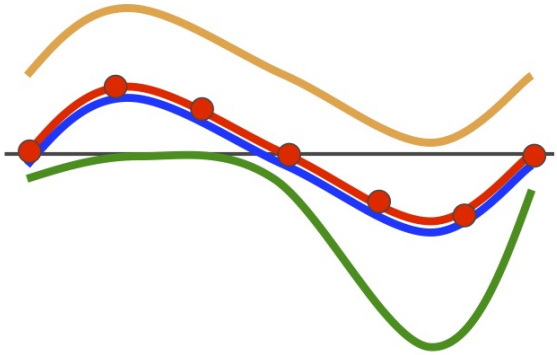
$$L_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Spearman Correlation

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (rank[x_i] - rank[y_i])^2}{n(n^2 - 1)}$$

Pearson Correlation

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

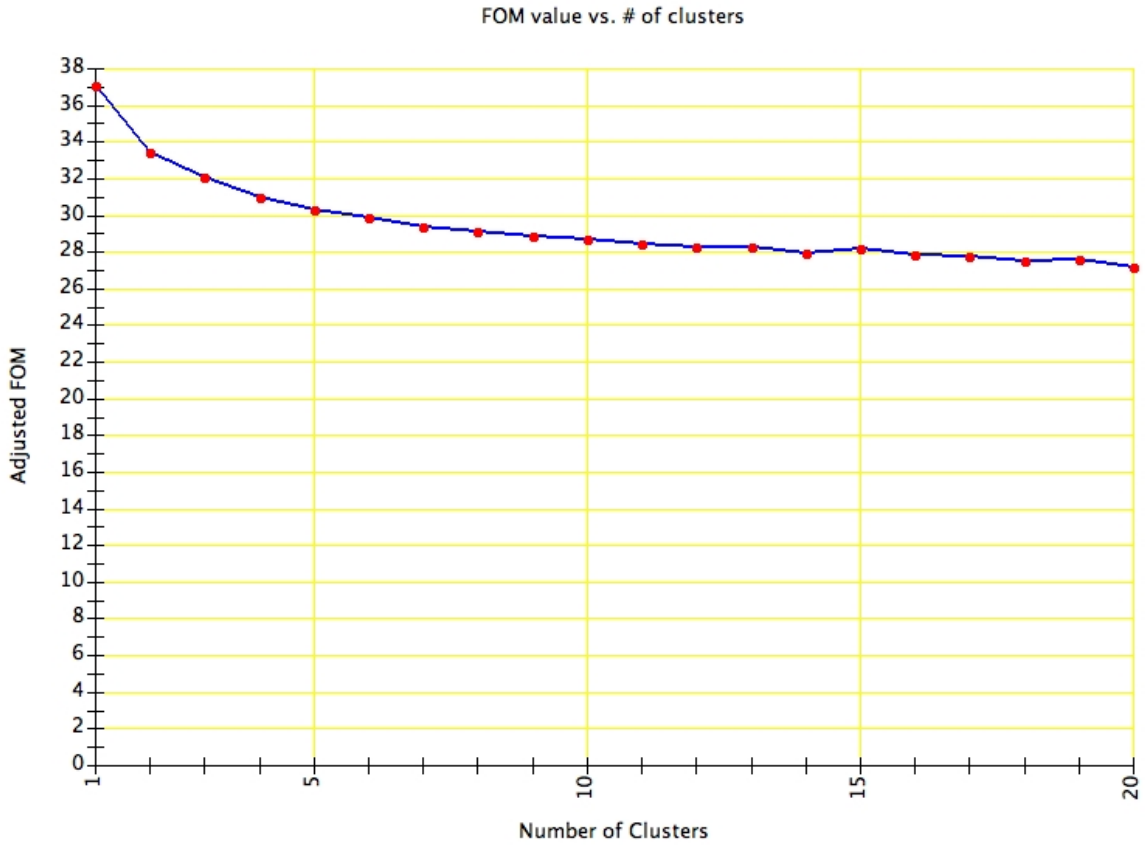


- Euclidean distance is also much more sensitive to outliers.
- Pearson correlation only tries to capture linear correlations.

```
> x <- c( -3, -2, -1, 0, 1, 2, 3 )
> y <- c( 9, 4, 1, 0, 1, 4, 9 )
> cor( x, y, method='pearson' )
[1] 0
```

Figure of merit

The original paper could be found at `http://bioinformatics.oxfordjournals.org/content/17/4/309.long`. Here is an example result from MeV:



This is the explanaiton in the original paper: "Intuitively, a clustering has possible biological significance if genes in the same cluster tend to have similar expression levels in additional experiments that were not used to form the clusters. We estimate this predictive power by removing one experiment, E, from the data set, clustering genes based on the remaining data, and then measuring the within-cluster similarity of expression values in experiment E."