

# Distance Measures for Gene Expression (and other high-dimensional) Data

Utah State University – Spring 2014  
STAT 5570: Statistical Bioinformatics  
Notes 2.2

1

## References

- Chapter 12 of Bioconductor Monograph (course text)

2

## ALL Data (to be used extensively)

- A public data set looking at Acute Lymphoblastic Leukemia
- RMA-preprocessed data
  - Chiaretti et al., Blood (2004) 103(7)
  - 12625 genes (hgu95av2 Affymetrix GeneChip)
  - 128 samples (arrays)
  - phenotypic data on all 128 patients, including type of disease:
    - 95 B-cell cancer
    - 33 T-cell cancer

3

## What to do with expression measures?

- Summarize
  - combine estimates across arrays
- Compare
  - between sample types
    - Graphically – to get visual representation  
scatterplots, tree diagrams, etc.
    - Numerically – to assess significance

4

## Comparison across arrays

- What are we looking for?
  - Genes that – change expression between sample types
  - Genes that – have similar ... expression profiles
- How do we know when we've found it?
  - graphical and statistical checks
  - best for motivating / supporting hypotheses
- What do we do once we've found it?
  - exploit changes / similarities – drug discovery, pathways, e.g.

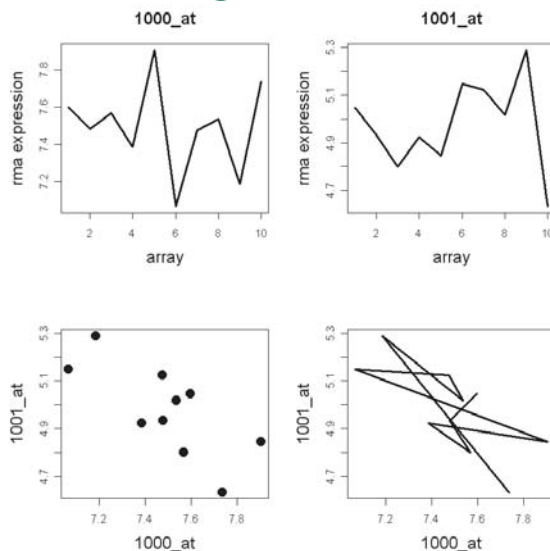
5

## Before graphical comparisons

- Need to have some measure of distance:
  - how “far apart” are genes’ expression “profiles”
    - (expression levels across arrays)
- How to define this?
  - consider just 2 genes at a time

6

## Visualizing similarity or distance



look at 2  
genes on  
10 arrays

7

## R code – visualizing similarity or distance

```
library(affy); library(ALL)
data(ALL) # ALL is an ExpressionSet object (as from rma)
gn <- featureNames(ALL)
all.exprs <- exprs(ALL)
par(mfrow=c(2,2))
plot(all.exprs[1,c(1:10)],main=gn[1],type='l',lwd=2,
      xlab='array',ylab='rma expression',
      cex.main=1.5,cex.lab=1.5)
plot(all.exprs[2,c(1:10)],main=gn[2],type='l',lwd=2,
      xlab='array',ylab='rma expression',
      cex.main=1.5,cex.lab=1.5)
plot(all.exprs[1,c(1:10)],all.exprs[2,c(1:10)],
      xlab=gn[1],ylab=gn[2],pch=16,cex=2,
      cex.main=1.5,cex.lab=1.5)
plot(all.exprs[1,c(1:10)],all.exprs[2,c(1:10)],
      xlab=gn[1],ylab=gn[2],type='l',lwd=2,
      cex.main=1.5,cex.lab=1.5)
```

8

## Mathematical Properties of “Distance”

For vectors  $x$  and  $y$ , suppose function  $d(x,y)$  satisfies:

□ Non-negativity  $d(x, y) \geq 0$ , with equality iff  $x = y$

□ Symmetry  $d(x, y) = d(y, x)$

□ Triangle Inequality  $d(x, y) + d(y, z) \geq d(x, z)$

Then  $d$  is called a metric

- special case of distance, where definiteness and triangle inequality are not necessary

9

## Familiar Distances

- Euclidean Metric -

$$d_{\text{euc}}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Pearson Sample Correlation Distance -

$$d_{\text{cor}}(x, y) = 1 - \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = 1 - r(x, y) = 0 \text{ for “zero distance” or “perfect correlation”}$$

- Relation for “standardized” vectors:

$$d_{\text{euc}}(x, y) = \sqrt{2n(1 - r(x, y))} \quad (\text{vectors have mean 0 and “length” 1})$$

10

## Mahalanobis Distance

- Consider vectors  $x$  and  $y$  from a multivariate distribution with mean  $\mu$  and covariance  $\Sigma$ .

- Let  $d(x,y)$  be

$$(x - y)' \Sigma^{-1} (x - y)$$

- What does this mean?  
data are more variable in some directions
- When will this be equivalent to Euclidean distance (or square of Euc. dist.)?  
When  $\Sigma$  is the identity matrix (so the pairs are iid)

11

## Effect of transformations on distances

- Suppose we transform the data:  
 $(x,y) \rightarrow (g(x),g(y)) = (x',y')$
- Then  $d(x,y) = d(g^{-1}(g(x)),g^{-1}(g(y))) = d'(x',y')$  ;  $d'$  is a new metric
- The distance between two points will depend on the transformation (or scale) of the data –

Usually:

- raw intensities
- log intensities
- glog intensities

12

## Distances between distributions

- Instead of looking at “profiles” across samples, consider gene’s expression levels as indep. sample from a population

- Are two genes’ distributions similar?
- Which samples have the same distributions?

- What would this mean biologically?

Do the genes behave similarly across samples?  
- i.e., are they functionally similar

13

## Kullback-Leibler Information

- Suppose gene k expression levels are a vector from distribution with density  $f_k$ ; compare genes 1 & 2:

$$KLI(f_1, f_2) = E_1 \left[ \log \frac{f_1(X)}{f_2(X)} \right] = \int \log \frac{f_1(x)}{f_2(x)} f_1(x) dx \quad = 0 \text{ if } f_1 = f_2$$

$$KLD(f_1, f_2) = (KLI(f_1, f_2) + KLI(f_2, f_1)) / 2 \quad \text{(to make KLI symmetric)}$$

- Maybe assume  $f_k$  is  $N(\mu_k, \Sigma_k)$  and  $\Sigma_k$  is diagonal  
 estimate from expr. levels  $\xrightarrow{\quad}$   
 estimate from SE of expr. levels  $\xrightarrow{\quad}$  ↑ independent samples

14

## Mutual information

- Let  $f = f(X, Y)$  be the joint density of X & Y
- Let  $g = f_1(X) f_2(Y)$  be the product of marginal densities

$$MI(f_1, f_2) = E_f \left[ \log \frac{f(X, Y)}{g(X, Y)} \right]$$

- When will  $MI = 0$ ?  
When X & Y are independent  
(then  $f=g$ )

15

## Interpreting “distances” between distributions

- What does KLD measure?  
distance between distributions (shape, etc.)
- What does MI measure?  
distance from independence
- How are they related?  $KLI(f, g) = MI(f_1, f_2)$
- Assuming normality and equality of covariance matrices:  
KLD is a Mahalanobis distance

16

## How to implement these?

- Can assume normality (or some other distribution)

But need to verify assumption is appropriate

- Could group (bin) data and treat as discrete

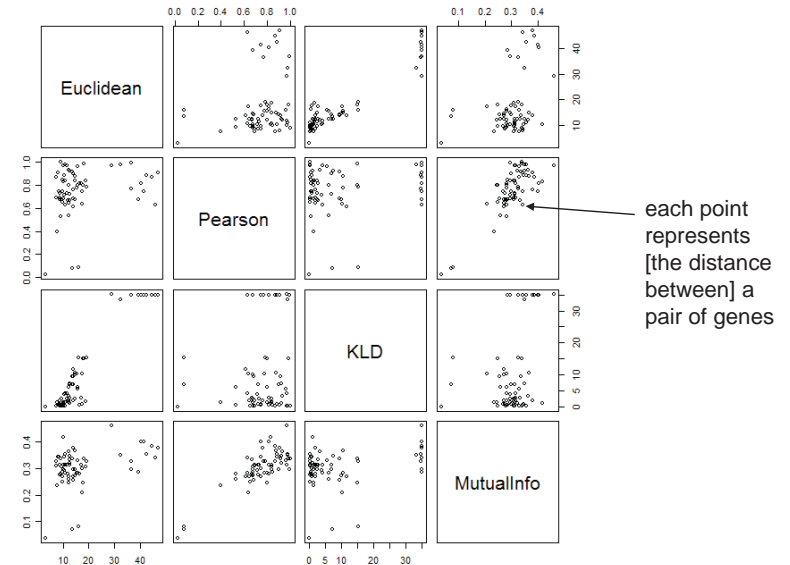
But problems arise when distributions are not similar locally

- Could apply density estimation with numerical integration

This is computationally expensive – but can be done

17

## Comparison of distances for 12 genes



18

## Comparison of distances in Bioconductor

```
library(affy); library(ALL); library(bioDist)
data(ALL); gn <- featureNames(ALL)

gn.list <-
  c("36199_at", "39020_at", "2031_s_at", "39723_at",
    "1635_at", "1636_g_at", "39730_at", "34740_at",
    "41763_g_at", "38050_at", "1000_at", "1001_at")
t <- is.element(gn, gn.list)
small.eset <- exprs(ALL[t,])
```

19

## Comparison of distances in Bioconductor

```
d.euc <- euc(small.eset) # creates a distance matrix
d.cor <- cor.dist(small.eset)
d.KLD <- KLdist.matrix(small.eset)
d.MI <- MI.dist(small.eset)

dist.frame <- data.frame(as.vector(d.euc),
  as.vector(d.cor), as.vector(d.KLD),
  as.vector(d.MI))
names(dist.frame) <-
  c("Euclidean", "Pearson", "KLD", "MutualInfo")
plot(dist.frame)
```

20

## Summary

---

- Compare expression levels
    - Between genes
    - Between samples
  - Look at distances
    - Different definitions: Euclidean, Correlation, Kullback-Leibler, Mutual
    - Different interpretations
  - Look at graphical representations – more coming up
-