Introduction to Microarray Data Analysis and Gene Networks
Lecture 3 and practical

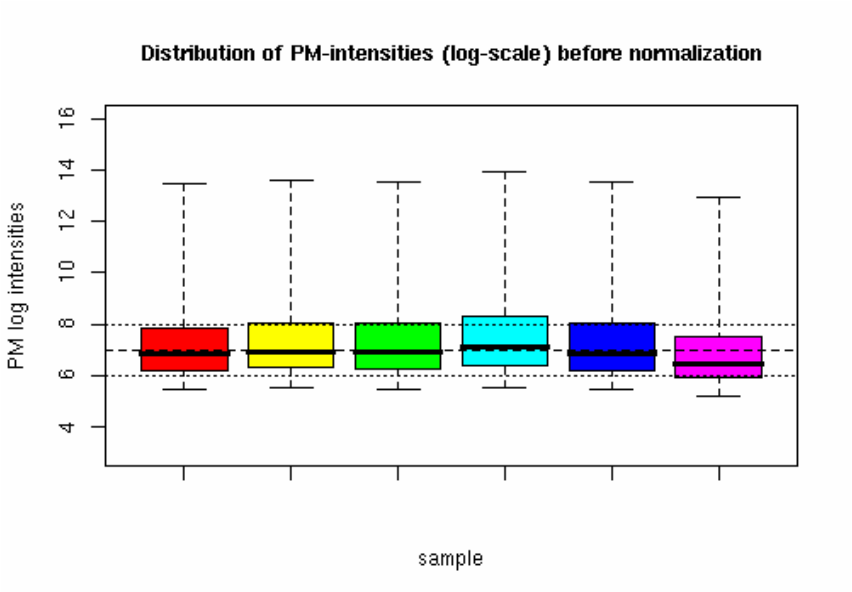
Alvis Brazma

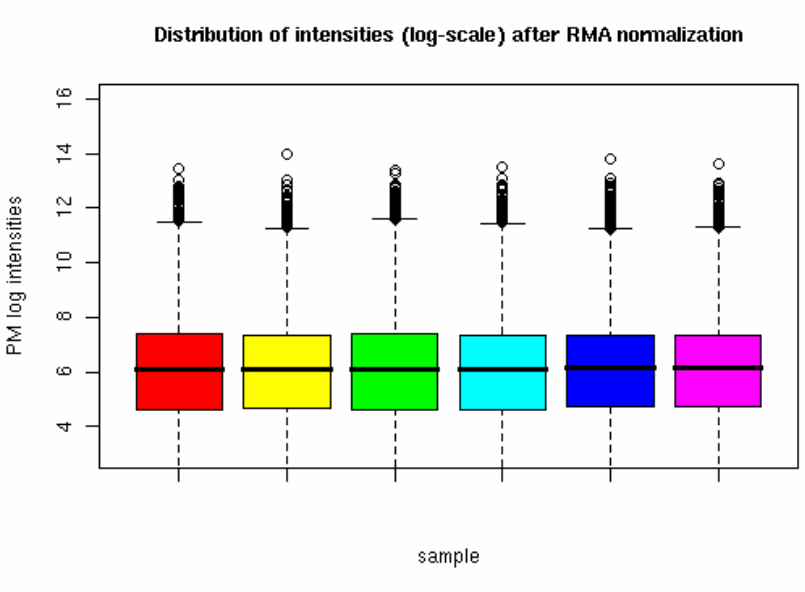
European Bioinformatics Institute

Robust **M**ulti-array **A**verage (RMA) normalisation

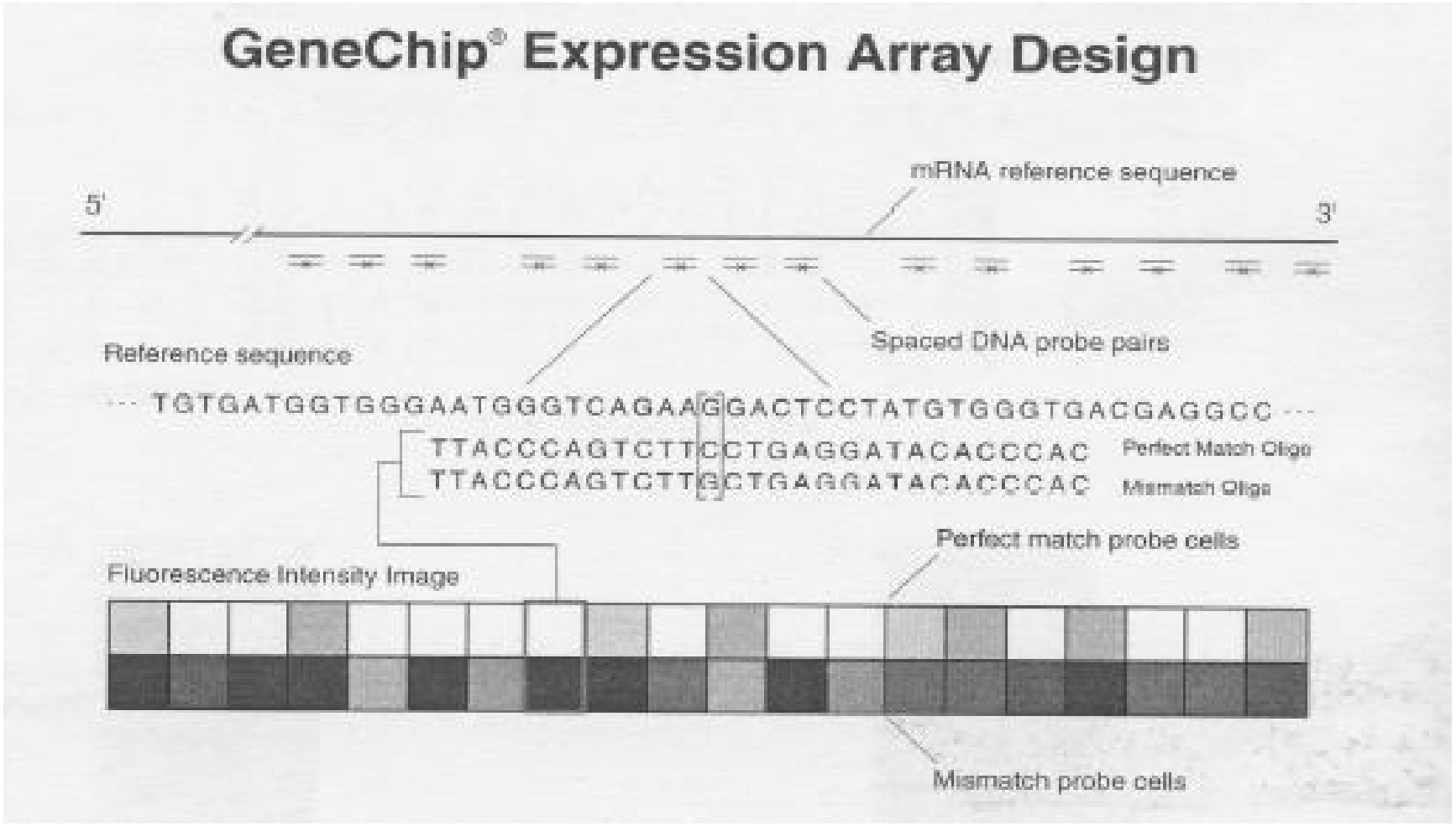
- Order each column of data (i.e. the points from each array) from highest to lowest expression value
- Calculate the mean of the highest expression value in each column
- Replace each highest value in the original array by that mean value
- Repeat the procedure using the second-highest value in each column, and continue until all values have been replaced by their respective means

Before and after RMA normalisation





GeneChip® Expression Array Design



RMA normalisation – steps from intensities to (pseudo) expression levels

1. Subtract the background intensity from each intensity value (if this has not already been done), in a way that ensures that all expression values are positive.
2. Take the log to base 2 of each expression value.
3. Normalise the log data as follows:
   a) Order each column of data (i.e. the points from each array) from highest to lowest expression value
   b) Calculate the mean of the highest expression value in each column
   c) Replace each highest value in the original array by that mean value
   d) Repeat the procedure using the second-highest value in each column, and continue until all values have been replaced by their respective means
4. The obtained ‘expression values’ will be gene specific

Practical part – find appropriate Affy dataset in ArrayExpress

- Browse ArrayExpress (www.ebi.ac.uk/arrayexpress) ‘Experiments’ (use Mozilla Firefox or Internet Explorer, not Safari)
- Filter on some Affymetrix array (e.g., U133A). Select an Affymetrix based experiment done on one array desing, with raw data present, consisting of about ~10 cel files (e.g., E-ATMX-10)
- Explore the experiment description, click on raw data and upload it in a directory on your PC

Open account in Expressi0n Profiler and load the data

- Open Expression Profiler in a browser (ie., go to www.ebi.ac.uk/expressionprofiler)
- Open an account, log in
- Go to Data import, Expression data
- Select Affymetrix and import the saved raw data
- Go to Normalisation, select RMA and click Execute
- Select 500 most variable genes and go to clustering

Distance measure

- Gene expression profiles can be considered vectors and the distance between them can be measured the same way as between vectors

Matrices and vectors

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$

X(1,1)	X(1,2)			X(1,5)
X(2,1)	X(2,2)			X(2,5)
X(3,1)	X(3,2)			X(3,5)
X(n,1)	X(n,2)			X(n,5)

The rows or columns of the matrix define *vectors* *A=(a1, …, ak)* (*e.g.*, *Ai=(xi1,…, xim)* for *i*-th row of the matrix and *Aj=(x1j,…,xnj)* for *j*-th column).

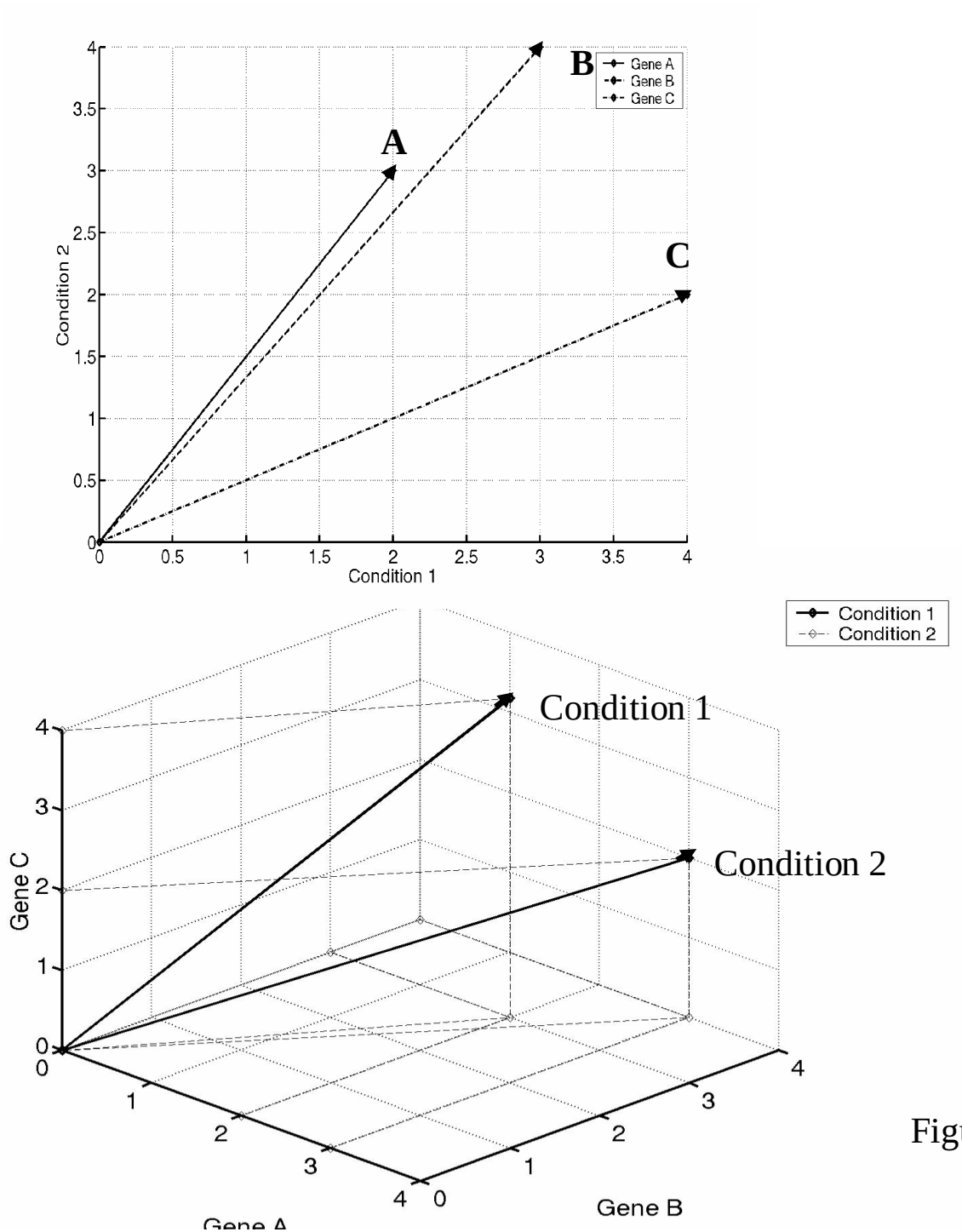

B
A
C
Gene A
Gene B
Gene C
Condition 2
Condition 1
0
0.5
1
1.5
2
2.5
3
3.5
4
Condition 1
Condition 2
Condition 1
Condition 2
Gene C
Gene A
Gene B

Figure 4.2

The length of a vector

Given a vector *A=(a1, …, ak)*, we define its length |*A*| as

$$\left|A\right| = \sqrt{a_1^2 + \ldots + a_k^2}$$

Distance measures

A distance measure D(*A,B*) is said to be *metric*, if it satisfies the following properties:

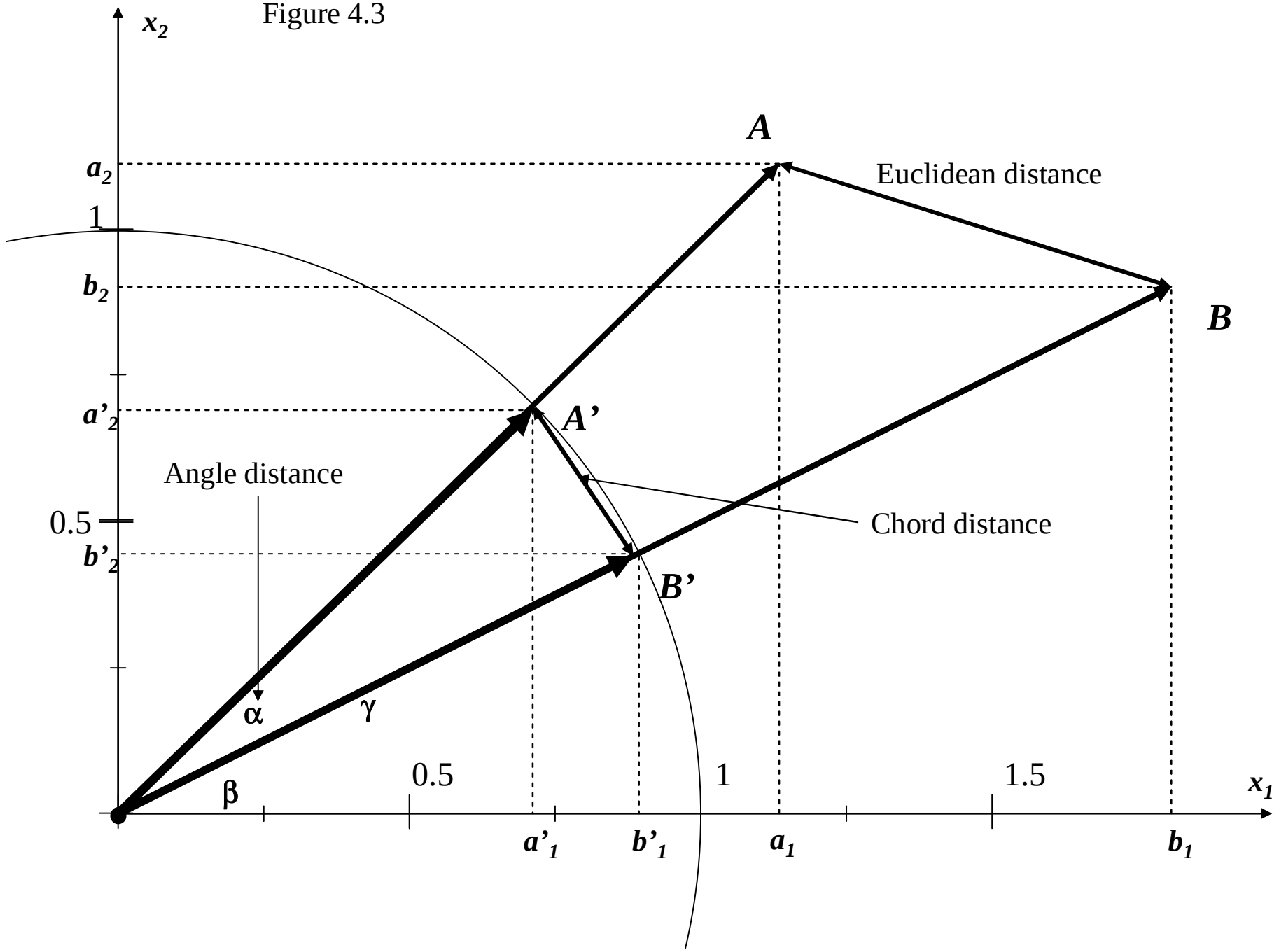
- if $A=B$, then D(*A,B*) = 0, *i.e*., the distance of an object to itself is 0;
- if $A \neq B$, then D(*A,B*) $\geq$ 0, *i.e*., the distance is always nonnegative;
- D(*A,B*) = D(*B,A*), *i.e*., it does not matter in which order we measure the distance;
- D(*A,B*) + D(*B,C*) $\geq$ D(*A,C*), *i.e*., given three objects, the length of a direct path from the first to the third objects cannot be greater than the length of the path through the second object.

Euclidean distance

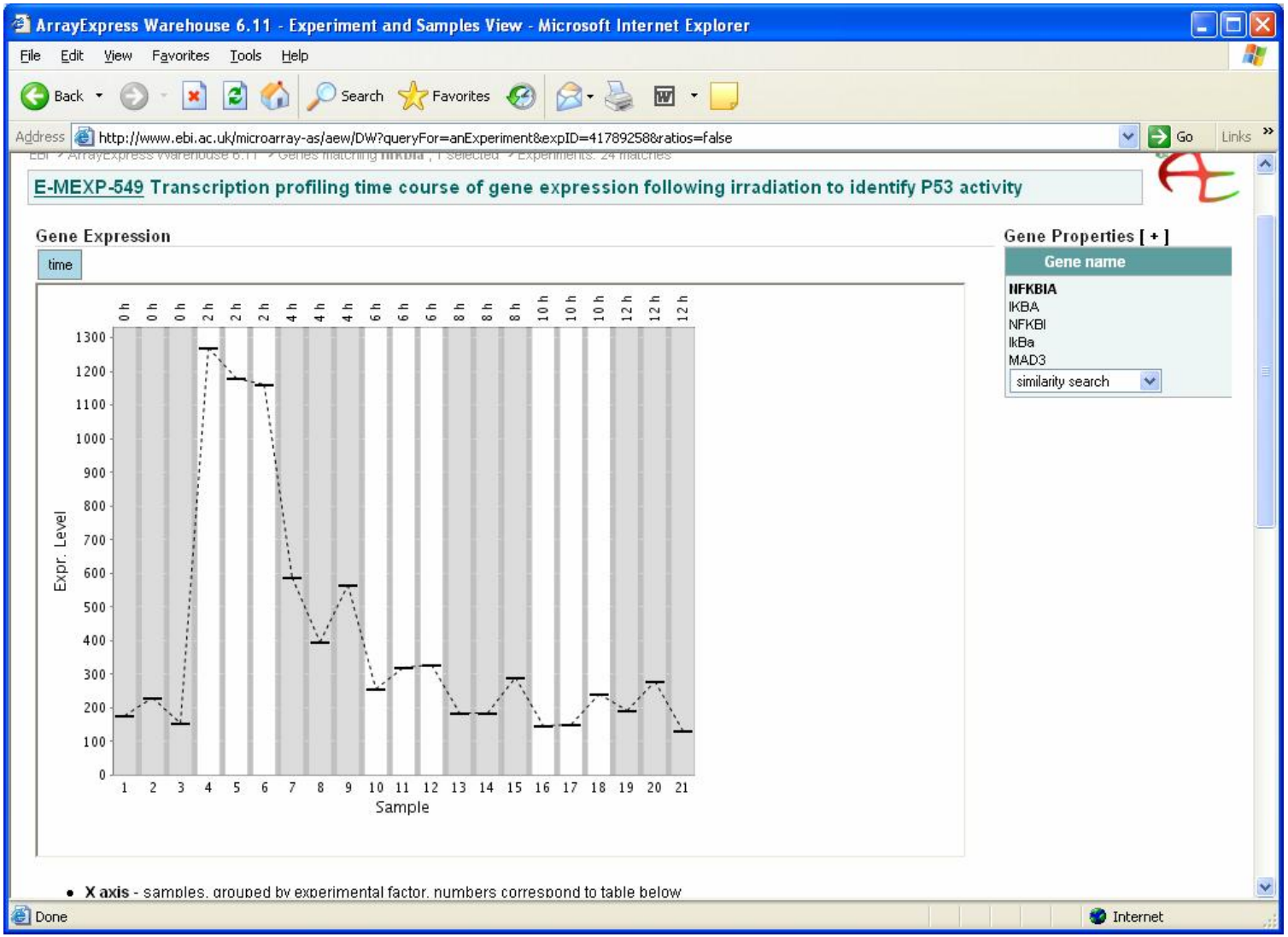
$$D_{Eucl}(A,B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

$$D_{Eucl}(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

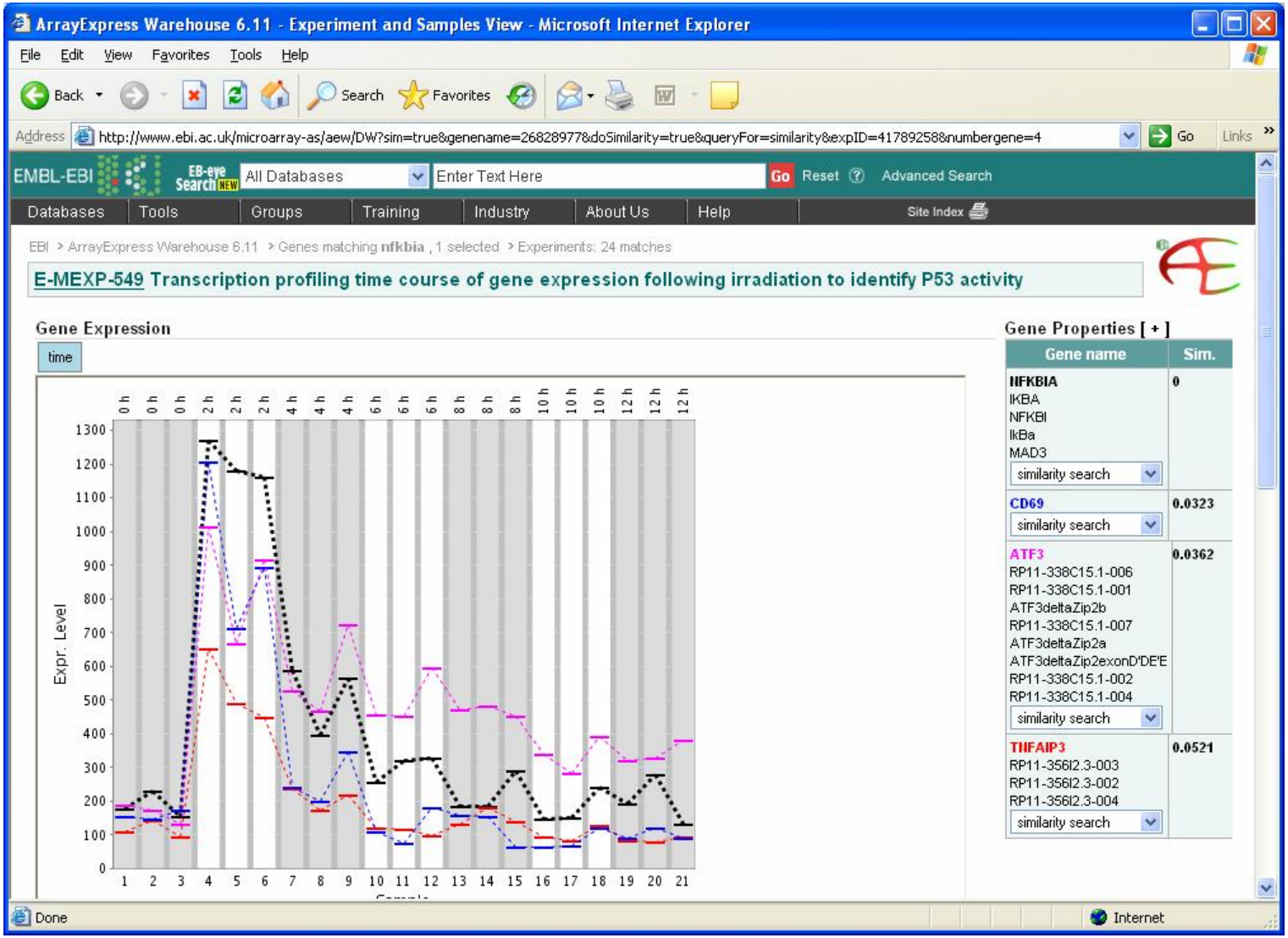
Figure 4.3


x_2
a_2
1
b_2
a'_2
0.5
b'_2
A
B
A'
B'
Euclidean distance
Angle distance
Chord distance
α
β
γ
0.5
1
1.5
x_1
a'_1
b'_1
a_1
b_1

Gene expression profile



Find genes with similar expression profiles



Practical

- Find in ArrayExpress experiment E-MEXP-57
- Go to View detailed data retrieval page
- Select normalised data, DB:genedb, and reporter name, click on Export
- Upload data in to Expression Profiler
- Go to transformations – apply Ratio -> Log ratio transformation, to Data selection, observe the distributions
- Go to transformations, perform KNN missing data imputation
- Select 400 most variable genes, do various clusterings