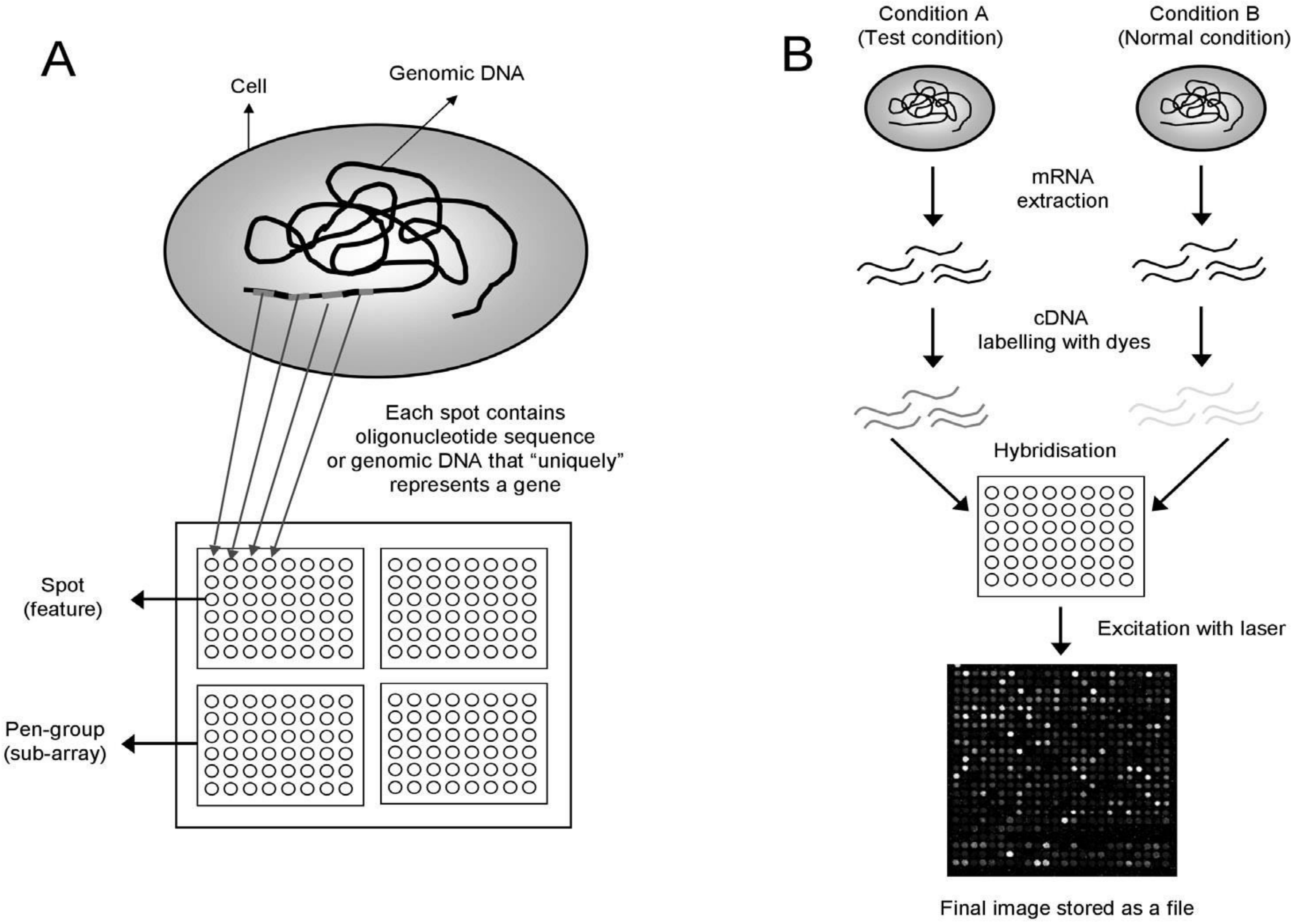
Microarrays

- **1. INTRODUCTION**
- Functional genomics involves the analysis of large datasets of information derived from various biological experiments. One such type of large-scale experiment involves monitoring the expression levels of thousands of genes simultaneously under a particular condition,

  called gene expression analysis.
- Microarray technology makes this possible and the quantity of data generated from each experiment is enormous, dwarfing the amount of data generated by genome sequencing projects


A
Cell
Genomic DNA
Each spot contains
oligonucleotide sequence
or genomic DNA that “uniquely”
represents a gene
Spot
(feature)
Pen-group
(sub-array)
B
Condition A
(Test condition)
Condition B
(Normal condition)
mRNA
extraction
cDNA
labelling with dyes
Hybridisation
Excitation with laser
Final image stored as a file

- If the gene was expressed to the same extent in both conditions, one would find the spot to be yellow, and if the gene was not expressed in both conditions, the spot would be black.
- Thus, what is seen at the end of the experimental stage is an image of the microarray, in which each spot that corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene.

OVERVIEW OF IMAGE PROCESSING, TRANSFORMATION AND NORMALIZATION

Image processing involves the following steps:

- 1. *Identification of the spots and distinguishing them from spurious signals.*
- 2. *Determination of the spot area to be surveyed, determination of the local region to estimate background hybridization.*
- 3. *Reporting summary statistics and assigning spot intensity after subtracting for background intensity.*

- Most approaches use the spot median value, with the background median value subtracted from it, as the metric to represent spot intensity.
- The median intensity is a value where half the measured pixels have intensities greater than this value and the other half of the measured pixels have intensities less than this value
- The other method is to use total intensity values, which has an
- advantage of being insensitive to misidentification of spots (as few more pixels with zero value in the background will not affect the total intensity), but has a disadvantage of being prone to be skewed by a few pixels with extreme intensity values

- Another consideration in image processing is the number of pixels to be included for
- measurement in the spot image. For many scanners, the default pixel size is 10µm. This means that an average spot of diameter of 200µm will have ~314 pixels.
- However, for a smaller spot diameter, it is better to use a smaller pixel size to ensure enough pixels are sampled. Most scanners now allow the pixel size of 5µm.

2.2 Expression ratios: the primary comparison

- We saw that the relative expression level for a gene can be measured as the amount of red or green light emitted after excitation. The most common metric used to relate this information is called expression ratio. It is denoted here as *Tk and defined as:*

$$T_k = \frac{R_k}{G_k}$$

- For each gene *k on the array, where Rk represents the spot intensity metric for the test* sample and *Gk represents the spot intensity metric for the reference sample. As mentioned* above, the spot intensity metric for each gene can be represented as a total intensity value or a background subtracted median value. If we choose the median pixel value, then the median expression ratio for a given spot is:

$$T_{median} = \frac{R_{median}^{spot} - R_{median}^{background}}{G_{median}^{spot} - G_{median}^{background}}$$

- where R_{median}^{spot} *and* $R_{median}^{background}$ *are the median intensity values for the spot and background* respectively, for the test sample.

2.3 Transformations of the expression ratio

- The expression ratio is a relevant way of representing expression differences in a very intuitive manner.
- For example, genes that do not differ in their expression level will have an expression ratio of 1.
- However, this representation may be unhelpful when one has to represent up-regulation and down-regulation. For example, a gene that is up-regulated by a factor of 4 has an expression ratio of 4 (R/G = 4G/G = 4). However, for the case where a gene is down regulated by a factor of 4, the expression ratio becomes 0.25 (R/G = R/4R = 1/4). Thus up-regulation is blown up and mapped between 1 and infinity, whereas down-regulation is compressed and mapped between 0 and 1.

$$Up-regulation \xrightarrow{mapped} [1,\infty]$$

$$Down-regulation \xrightarrow{mapped} [0,1]$$

- To eliminate this inconsistency in the mapping interval, one can perform two kinds of transformations of the expression ratio, namely, inverse transformation and logarithmic transformation.

Inverse or reciprocal transformation

However, this method also has a problem in that the mapping space is
discontinuous between
–1 and +1 and hence becomes a problem in most mathematical analyses
downstream of
this step.

$$\text{Fold change} = \begin{cases} T_k & \text{if } T_k \geq 1 \\ -1/T_k & \text{if } T_k < 1 \end{cases} \qquad \text{e.g.:} \quad \text{Fold change} = \begin{cases} 4 & \text{when } T_k = 4 \\ -4 & \text{when } T_k = 0.25 \end{cases}$$

Logarithmic transformation

- A better transformation procedure is to take the logarithm base 2 value of the expression ratio (*i.e. log2 (expression ratio)).*
- *This has the major advantage that it treats differential* up-regulation and down-regulation equally, and also has a continuous mapping space.
- For example, if the expression ratio is 1, then log2 (1) equals 0 represents no change in expression. If the expression ratio is 4, then log2 (4) equals +2 and for expression ratio of log2 (1/4) equals -2. Thus, in this transformation the mapping space is continuous and upregulation and down-regulation are comparable.

2.4 Data normalization

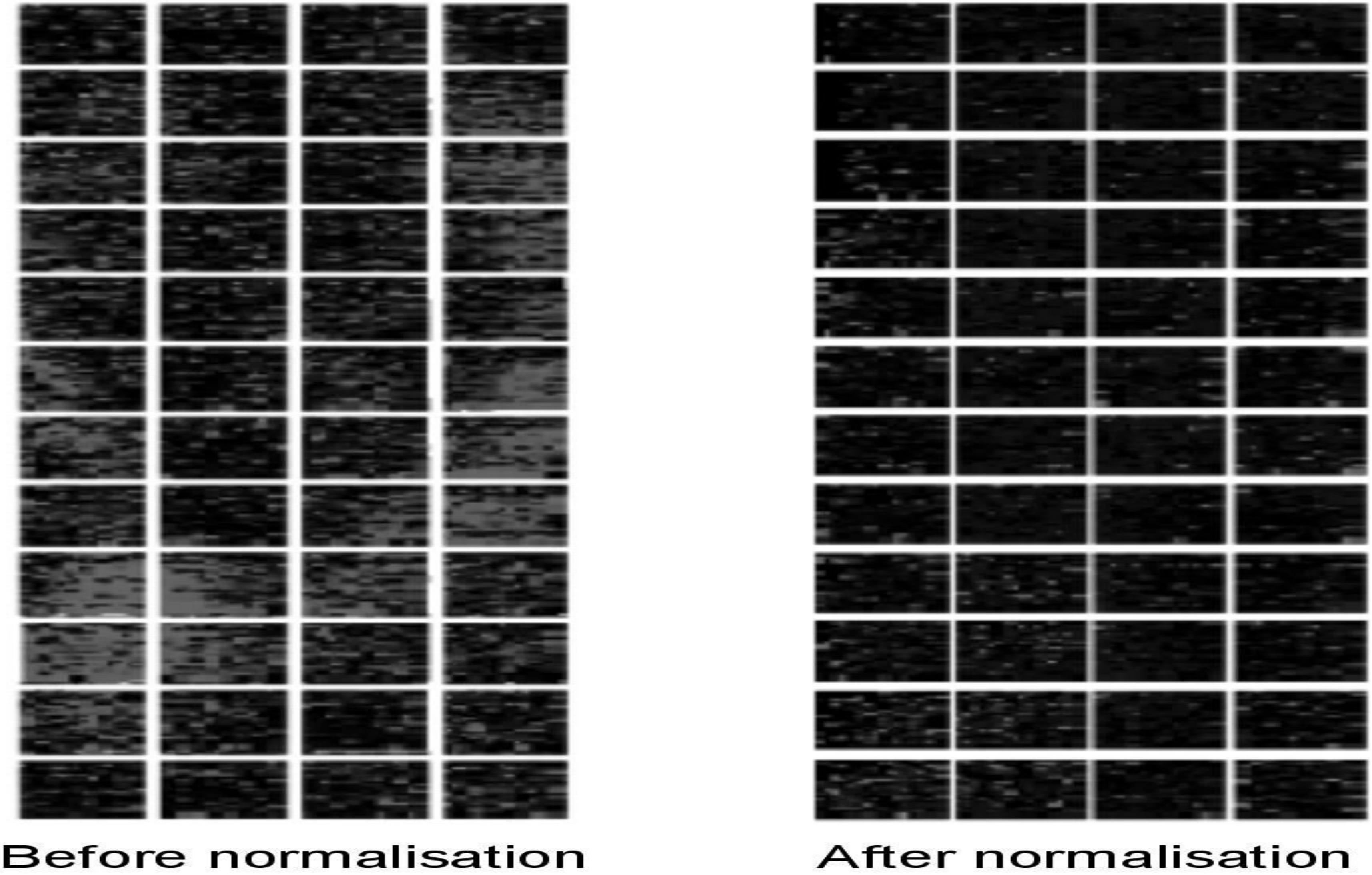
2.4 Data normalization

- when one compares the expression levels of genes that should not change in the two conditions (say, housekeeping genes), what one quite often finds is that an average expression ratio of such genes deviates from 1
- In the case of microarray experiments, as for any large-scale experiments, there are many sources of systematic variation that affect measurements of gene expression levels.

- Normalization is a term that is used to describe the process of eliminating such variations to allow appropriate comparison of data obtained from the two samples.

- The first step in a normalization procedure is to choose a gene-set (which consists of genes for which expression levels should not change under the conditions studied, that is the expression ratio for all genes in the gene-set is expected to be 1.
- From that set, a *normalization factor, which is a number that accounts for the variability seen in the gene set,* is calculated. It is then applied to the other genes in the microarray experiment.

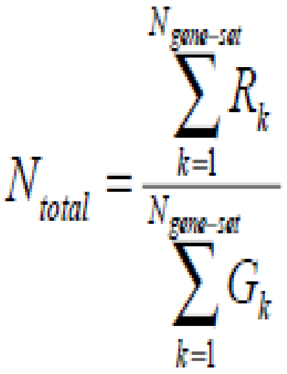
Figure 3. Gene expression data before and after the normalization procedure. Note that before normalization the image had many spots of different intensities, but after normalization only spots that are really different light up. This image was kindly provided by N. Luscombe*. Colour figure at:* *http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/*.



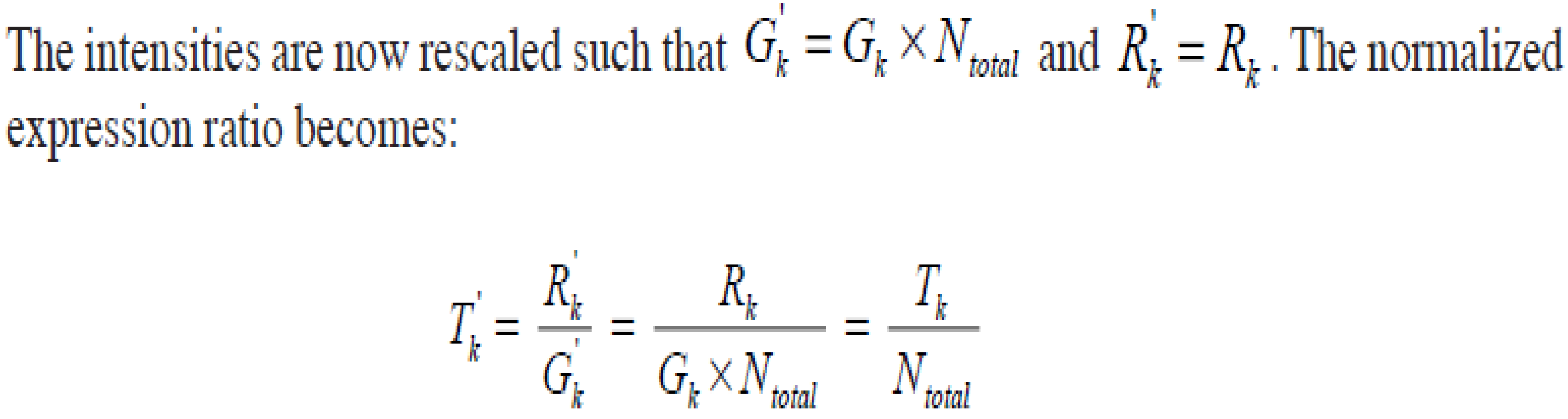
Total intensity normalization

- The basic assumption in a total intensity normalization is that the total quantity of RNA for the two samples is the same.
- Also assuming that the same number of molecules of RNA from both samples hybridize to the microarray, the total hybridization intensities for the gene-sets should be equal.

- So, a normalization factor can be calculated as:



$$N_{total} = \frac{\sum_{k=1}^{N_{gene-set}} R_k}{\sum_{k=1}^{N_{gene-set}} G_k}$$



The intensities are now rescaled such that $G_k^{'} = G_k \times N_{total}$ and $R_k^{'} = R_k$. The normalized expression ratio becomes:

$$T_k^{'} = \frac{R_k^{'}}{G_k^{'}} = \frac{R_k}{G_k \times N_{total}} = \frac{T_k}{N_{total}}$$

Which is equivalent to:

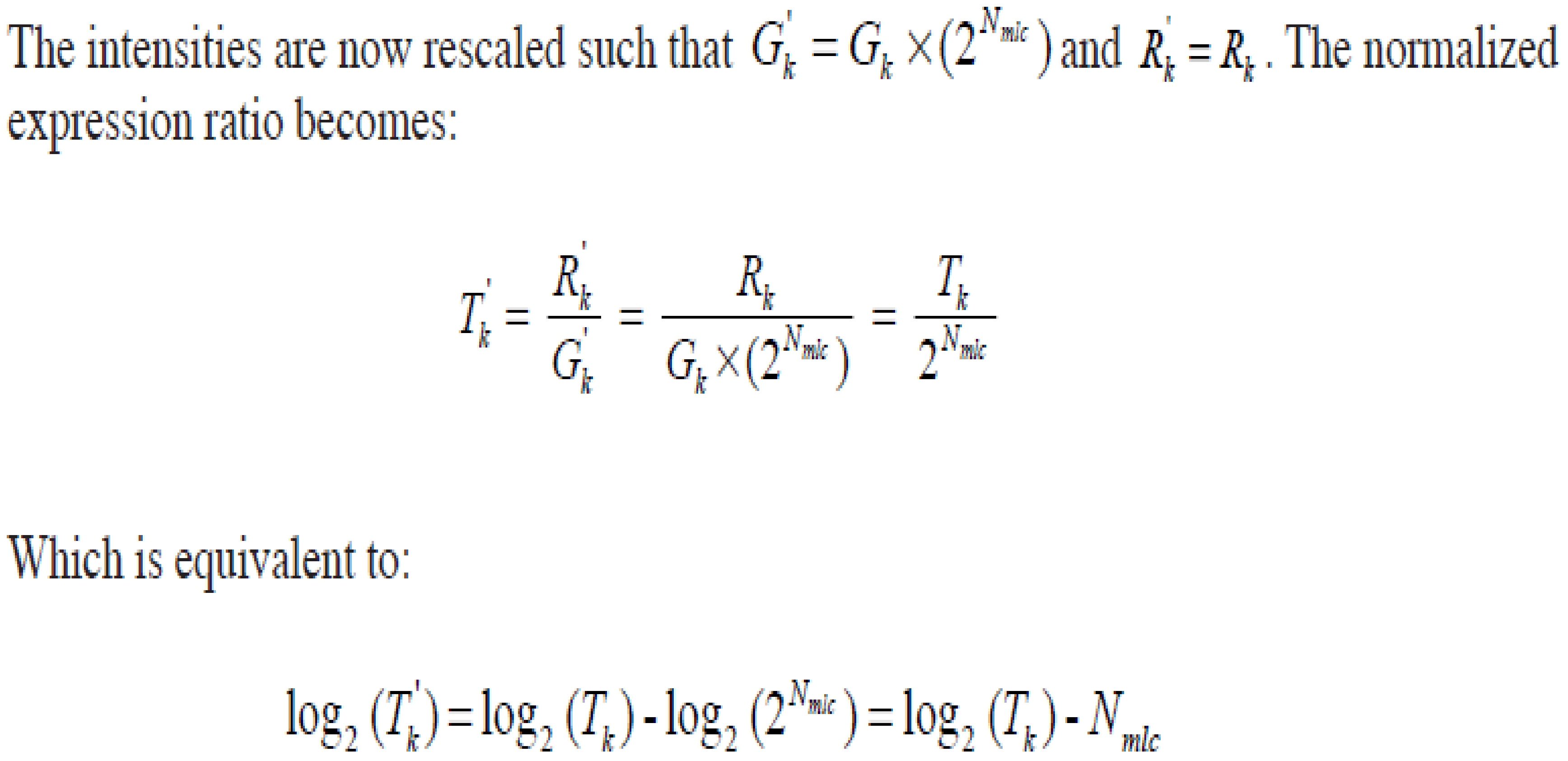
$$\log_2 (T_k^{'}) = \log_2 (T_k) - \log_2 (N_{total})$$

This now adjusts the ratio such that the mean ratio for the gene set is equal to 1.

Mean log centring

In this method, the basic assumption is that the mean $\log_2$ (expression ratio) should be equal to 0 for the gene-set. In this case, the normalization factor can be calculated as:

$$N_{mlc} = \frac{\sum_{k=1}^{N_{gene-set}} \log_2\left(\frac{R_k}{G_k}\right)}{N_{gene-set}}$$



The intensities are now rescaled such that $G_k^{'} = G_k \times (2^{N_{mlc}})$ and $R_k^{'} = R_k$. The normalized expression ratio becomes:

$$T_k^{'} = \frac{R_k^{'}}{G_k^{'}} = \frac{R_k}{G_k \times (2^{N_{mlc}})} = \frac{T_k}{2^{N_{mlc}}}$$

Which is equivalent to:

$$\log_2(T_k^{'}) = \log_2(T_k) - \log_2(2^{N_{mlc}}) = \log_2(T_k) - N_{mlc}$$

- This adjusts the ratio such that the mean log2 (expression ratio) for the gene-set is equal to 0.

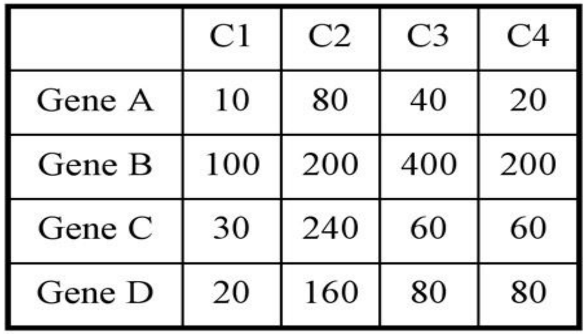
- Other normalization methods include: linear regression, Chen’s ratio statistics and Lowess normalization.
- The next step following the normalization procedure is to filter low intensity data using specific threshold or relative threshold imposed according to the background intensity.

3. ANALYSIS OF GENE EXPRESSION DATA

- The processed data, after the normalization procedure, can then be represented in the form of a matrix, often called gene expression matrix (Table 1A).
- Each row in the matrix corresponds to a particular gene and each column could either correspond to an experimental condition or a specific time point at which expression of the genes has been measured.
- The expression levels for a gene across different experimental conditions are cumulatively called the gene expression profile, and the expression levels for all genes under an experimental condition are cumulatively called the sample expression profile.

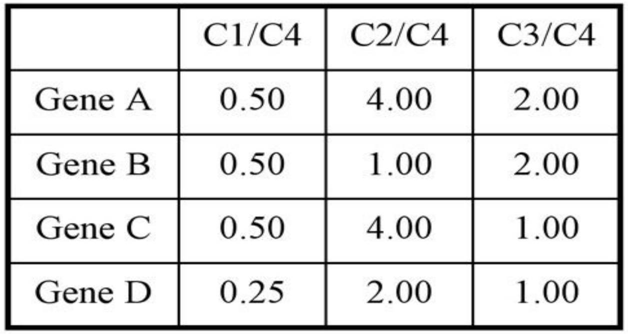
Table 1. A: Gene expression matrix that contains rows representing genes and columns representing particular conditions. Each cell contains a value, given in arbitrary units, that reflects the expression level of a gene under a corresponding condition. B: Condition C4 is used as a reference and all other conditions are normalized with respect to C4 to obtain expression ratios. C: In this table all expression ratios were converted into the log2 (expression ratio) values. This representation has an advantage of treating up-regulation and down-regulation on comparable scales. D: Discrete values for the elements in Table 1.C. Genes with log2 (expression ratio) values greater than 1 were changed to 1, genes with values less than –1 were changed to –1. Any value between –1 and 1 was changed to 0.

Table 1.A: Absolute measurement



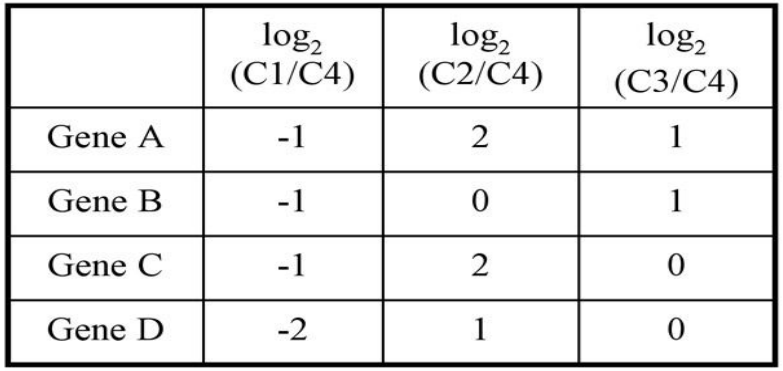
	C1	C2	C3	C4
Gene A	10	80	40	20
Gene B	100	200	400	200
Gene C	30	240	60	60
Gene D	20	160	80	80

Table 1.B: Relative measurement



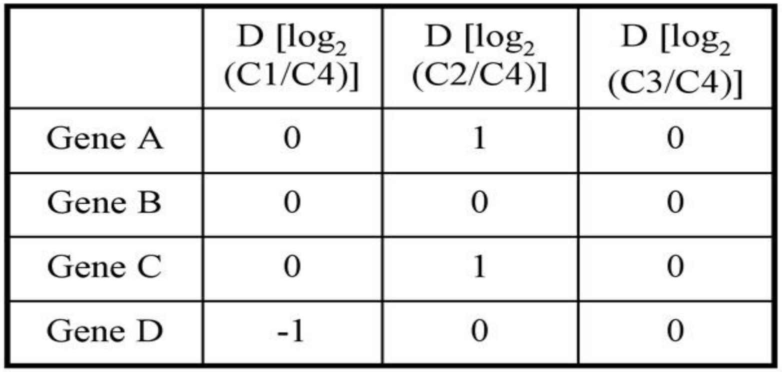
	C1/C4	C2/C4	C3/C4
Gene A	0.50	4.00	2.00
Gene B	0.50	1.00	2.00
Gene C	0.50	4.00	1.00
Gene D	0.25	2.00	1.00

Table 1.C: $\log_2$(relative measurement)



	$\log_2$ (C1/C4)	$\log_2$ (C2/C4)	$\log_2$ (C3/C4)
Gene A	-1	2	1
Gene B	-1	0	1
Gene C	-1	2	0
Gene D	-2	1	0

Table 1.D: Discrete values



	D [$\log_2$ (C1/C4)]	D [$\log_2$ (C2/C4)]	D [$\log_2$ (C3/C4)]
Gene A	0	1	0
Gene B	0	0	0
Gene C	0	1	0
Gene D	-1	0	0

- Once we have obtained the gene expression
- matrix (Table 1A), additional levels of annotation can be added either to the gene or to the sample. For example, the function of the genes can be provided, or the additional details on the biology of the sample may be provided, such as disease state or normal state’.

supervised learning

- Depending on whether the annotation is used or not, analysis of gene expression data can be classified into two different types, namely supervised or unsupervised learning.
- In the case of a supervised learning, we do use the annotation of either the gene or the sample, and create clusters of genes or samples in order to identify patterns that are characteristic
- for the cluster. For example, we could separate sample expression profiles into ‘disease state’ and ‘normal state’ groups, and then look for patterns that separate the sample profile of the ‘disease state’ from the sample profile of the ‘normal state’.

unsupervised learning

- In the case of an unsupervised learning, the expression data is analysed to identify patterns that can group genes or samples into clusters without the use of any form of annotation.
- For example, genes with similar expression profiles can be clustered together without the use of any annotation. However, annotation information may be taken into account at a later stage to make meaningful biological inferences.

3.1 Representation of gene expression data

- Expression data can be represented in five
- d *Absolute measurement* In the case of an absolute measurement, each cell in the matrix will represent the expression level of the gene in abstract units. Note that it is not meaningful to compare expression levels of genes across two different conditions in absolute units, because the starting amounts of mRNA could be different.
- Table 1A shows a sample gene expression matrix with each cell containing the expression level in abstract units. different ways, which are described below:

- *Relative measurement or expression ratio*
- In the case of a relative measurement or representations involving expression ratio, the
- expression level of a gene in abstract units is normalized with respect to its expression in a
- reference condition. This gives the expression ratio of the gene in relative units. Note that
- in such cases, a ratio of 4000/100 will lead to the same result as 40/10. Thus any information on
- absolute measurement will be lost in such a representation, but now meaningful comparison
- across different conditions can be made as long as the same reference condition is used to get
- the expression ratio.(table 1B)

- *log2(expression ratio)*
- In the case of tables representing the log2 (expression ratio) values, information on upregulation
- and down-regulation is captured and is mapped in a symmetric manner. For
- example, 4-fold up-regulation maps to log2 (4) = 2 and a 4-fold down-regulation maps to
- log2 (1/4) = -2. Thus, from this table the fold-change for a differentially regulated gene
- under any condition can be easily recognised. Table 1C shows the log2 (expression ratio)
- values of the genes under different conditions.

- *Discrete values*
- Another way of representing information is to convert to discrete numbers the values in the tables mentioned above.
- In the case of converting the absolute measurement to discrete numbers, a binary expression matrix of 1 and 0 can be used, where 1 means that the gene is expressed above a user defined threshold, and 0 means that the gene is expressed below this threshold.
- In the case of making the relative expression tables or log2 (expression ratio) tables discrete, values can be divided into 3 classes, +1, 0 and –1, where +1 represents a gene that is positively regulated, 0 represents a gene that is not differentially regulated and –1 represents a gene that is repressed

- .The process of making the values discrete loses a lot of information, but is useful to analyse expression profile les using algorithms that cannot handle real value expression matrices, for example algorithms calculating mutual information between genes or samples.
- Table 1D shows discrete values for the log2 (expression ratio) table.

- *Representation of expression profiles as vectors*
- So far we have seen how individual cells in the gene expression matrix can be represented.
- Similarly, an expression profile (of a gene or a sample) can be thought of as a vector and can be represented in vector space.
- For example, an expression profile of a gene can be considered as a vector in *n dimensional space (where n is the number of conditions), and an* expression profile of a sample with *m genes can be considered as a vector in m dimensional* space (where *m is the number of genes).*

- In the example given below, the gene expression matrix X with *m genes across n conditions is considered to be an m x n matrix, where the* expression value for gene *i in condition j is denoted as xij:*

$$X = \begin{bmatrix} x_{11} & x_{12} & .. & x_{1n} \\ x_{21} & x_{22} & .. & x_{2n} \\ .. & .. & .. & .. \\ x_{m1} & x_{m2} & .. & x_{mn} \end{bmatrix}$$

The expression profile of a gene *i* can be represented as a row vector:

$$G_i = \begin{bmatrix} x_{i1}, & x_{i2}, & x_{i3}, & ..., & x_{in} \end{bmatrix}$$

The expression profile of a sample *j* can be represented as a column vector:

$$G_i = \begin{bmatrix} x_{1j} \\ x_{2j} \\ .. \\ x_{mj} \end{bmatrix}$$

3.2 Distance measures

- Analysis of gene expression data is primarily based on comparison of gene expression profiles or sample expression profiles.
- In order to compare expression profiles, we need a measure to quantify how similar or dissimilar are the objects that are being considered

Euclidean distance

- Euclidean distance is one of the common distance measures used to calculate similarity between expression profiles.
- The Euclidean distance between two vectors of dimension 2, say *A=[a1, a2] and B=[b1, b2] can be calculated as:*

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

- For instance two genes with expression profiles in two conditions *G1=[1,2] and G2=[2,3],*

the Euclidean distance can be calculated as:

$$D_{Euc}(G_1,G_2)=\sqrt{(1\ -2\ )^2+(2\ -3\ )^2}=\sqrt{2}$$

- Thus for genes with expression data available for n conditions, represented as *A=[a1, .., an] and B=[b1, .., bn], Euclidean distance can be calculated as:*

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

- In other words, the Euclidean distance between two genes is the square root of the sum of the squares of the distances between the values in each condition (dimension).

Pearson correlation coefficient

- One of the most commonly used metrics to measure similarity between expression profiles is the Pearson correlation coefficient (PCC) (Eisen *et al. 1998).*
- *Given the expression* ratios for two genes under three conditions *A=[a1, a2, a3] and B=[b1, b2, b3], PCC can be* computed as follows:

$$\overline{a} = \frac{a_1 + a_2 + a_3}{3} \quad and \quad \overline{b} = \frac{b_1 + b_2 + b_3}{3}$$

Step2: “Mean centre” expression profiles

$$\overline{A} = (a_1 - \overline{a},\ a_2 - \overline{a},\ a_3 - \overline{a}) \quad and \quad \overline{B} = (b_1 - \overline{b},\ b_2 - \overline{b},\ b_3 - \overline{b})$$

Step3: Calculate PCC as the cosine of the angle between the mean-centred profiles

$$PCC = \frac{\overline{A} \circ \overline{B}}{\left|\overline{A}\right| \left|\overline{B}\right|}$$

Where,

$$\overline{A} \circ \overline{B} = \sum_{i=1}^{n} (a_i - \overline{a}) \times (b_i - \overline{b})$$

$$\left|\overline{A}\right| = \sqrt{\sum_{i=1}^{n} (a_i - \overline{a})^2} \quad \text{and} \quad \left|\overline{B}\right| = \sqrt{\sum_{i=1}^{n} (b_i - \overline{b})^2}$$

- The reason why we “mean centre” the expression profiles is to make sure that we compare shapes of the expression profiles and not their magnitude.
- Mean centering maintains the shape of the profile, but it changes the magnitude of the profile as shown in Figure 4.
- A PCC value of 1 essentially means that the two genes have similar expression profiles and a value of –1 means that the two genes have exactly opposite expression profiles. A value of 0 means that no relationship can be inferred between the expression profiles of genes.
- In reality, PCC values range from –1 to +1. A PCC value ≥ 0.7 suggests that the genes behave similarly and a PCC value ≤ -0.7 suggests that the genes have opposite behavior. The value of 0.7 is an arbitrary cut-off, and in real cases this value can be chosen depending on the dataset used. An example calculation is shown below:

Rank correlation coefficient

- Rank correlation coefficient (RCC) is a distance measure that does not take into account the actual magnitude of the expression ratio in each condition, but takes into account the rank of the expression ratio. For example, consider two genes *A = [2, 3, 9, 15, 8]*
- *and B =* [*2, 7, 15, 25, 13*]*. When we consider the rank of the values for different conditions for* gene A, we get the following:

- <2 (rank = 1) < 3 (rank = 2) < 8 (rank = 3) < 9 (rank = 4) < 15 (rank = 5) which is equivalent to *A = [1, 2, 4, 5, 3].*
- Similarly, for gene *B, we get the ranks for the values for the different conditions as:*
- 2 (rank = 1) < 7 (rank = 2) < 13 (rank = 3) < 15 (rank = 4) < 25 (rank = 5), which is equivalent to *B = [1, 2, 4, 5, 3].*

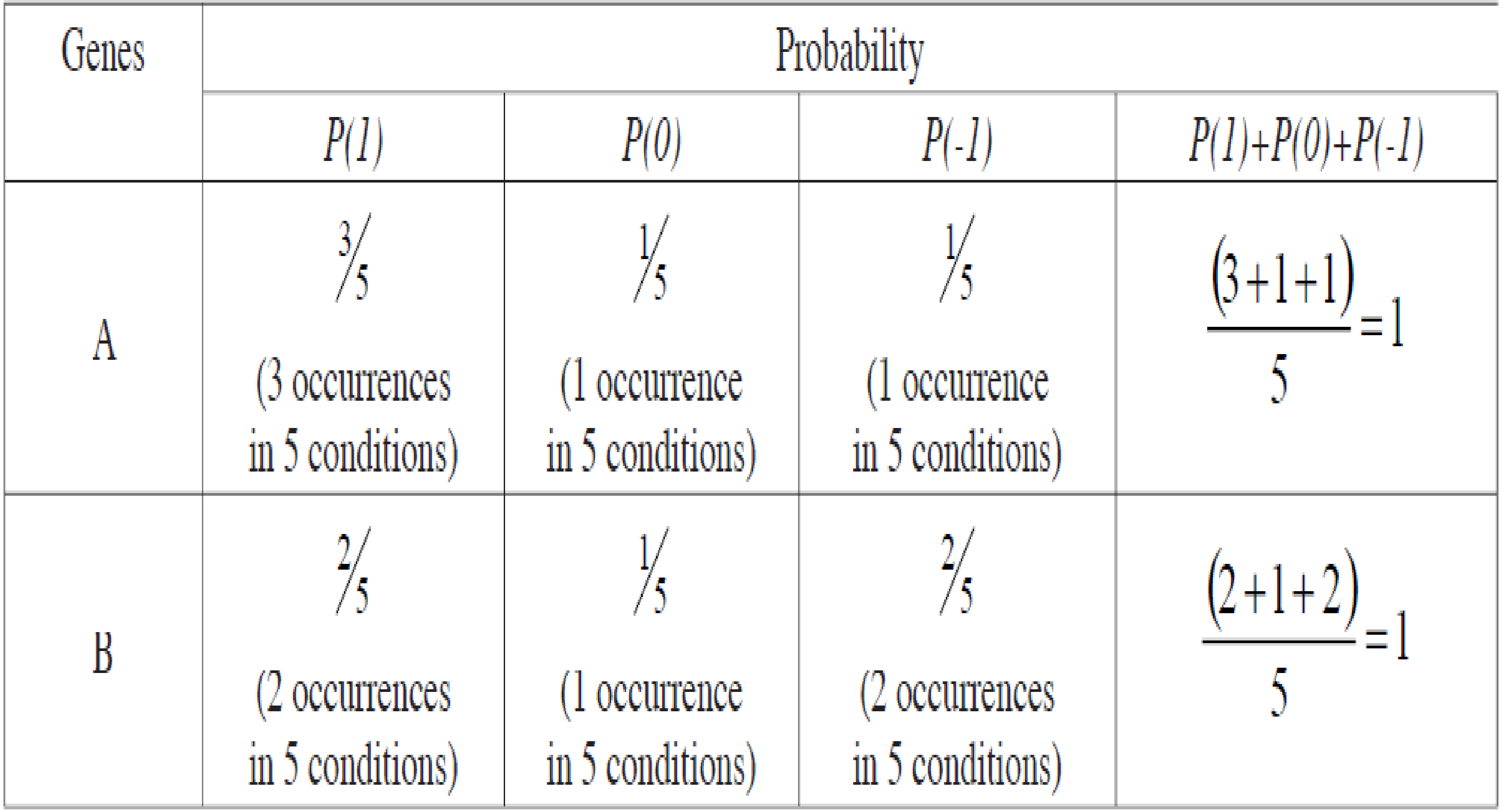
- Rank correlation coefficient is the PCC calculated on the expression profiles converted into their rank profiles. In the above case the two genes have exactly the same rank profile, thus rank correlation coefficient becomes 1.
- However, PCC is not applicable when two values within a rank profile are repeated. In this case, the rank correlation coefficient can be directly computed as:

$$D_{rank}(A,B) = 1 - 6 \times \sum_{i=1}^{n} \frac{d_i^{\,2}}{n(n^2 - 1)}$$

- Where *n is the number of conditions (dimension of the profile) and di is the difference* between ranks for the two genes at condition *i. An advantage of RCC is that it is not* sensitive to outliers in the data.

Mutual information

- A distance measure to compare genes whose profiles have been made discrete can be calculated using an entropy notion, called Shannon's entropy.
- This measure gives us a metric that is indicative of how much information from the expression profile of one gene can be obtained to predict the behaviour of the other gene.
- Consider the discrete expression profiles for two genes, *A = [1, 1, 0, 1, -1] and B = [1,-1, 0, 1, -1]. We know that at any condition, the values that have been made discrete can* be 1, 0 or –1. Thus, the probability for each state to occur in the profile for the two genes can be computed as follows:



Genes	Probability			
	$P(1)$	$P(0)$	$P(-1)$	$P(1)+P(0)+P(-1)$
A	$^{3}/_{5}$ (3 occurrences in 5 conditions)	$^{1}/_{5}$ (1 occurrence in 5 conditions)	$^{1}/_{5}$ (1 occurrence in 5 conditions)	$\frac{(3+1+1)}{5}=1$
B	$^{2}/_{5}$ (2 occurrences in 5 conditions)	$^{1}/_{5}$ (1 occurrence in 5 conditions)	$^{2}/_{5}$ (2 occurrences in 5 conditions)	$\frac{(2+1+2)}{5}=1$

From this table, the Shannon’s entropy for the genes can be calculated as:

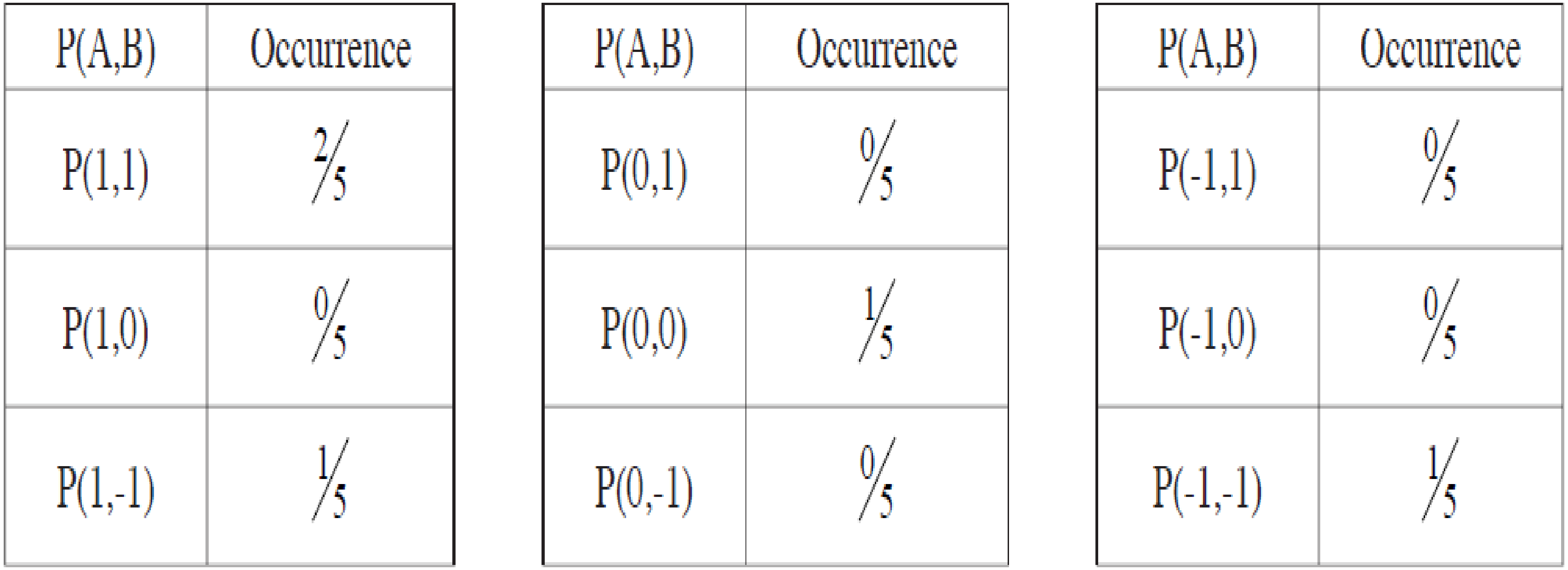
$$H(gene) = -\sum_{i=1}^{3} P_i \times \log_2 P_i$$

Note that *i* runs from 1 to 3 because there are three possible states (1, 0 and –1).

$$H(A) = -1\times(\tfrac{3}{5}\times \log_2 \tfrac{3}{5} + \tfrac{1}{5}\times\log_2 \tfrac{1}{5} + \tfrac{1}{5}\times\log_2 \tfrac{1}{5}) = 1.371$$

$$H(B) = -1\times(\tfrac{2}{5}\times \log_2 \tfrac{2}{5} + \tfrac{1}{5}\times\log_2 \tfrac{1}{5} + \tfrac{2}{5}\times\log_2 \tfrac{2}{5}) = 1.522$$

- The next step in our calculation is to consider how often gene A and gene B have the same state (1, 0, or -1) across given conditions.
- There are 9 possible pairwise combinations of states, and they are calculated for our example in the following manner:



P(A,B)	Occurrence
P(1,1)	2/5
P(1,0)	0/5
P(1,-1)	1/5

P(A,B)	Occurrence
P(0,1)	0/5
P(0,0)	1/5
P(0,-1)	0/5

P(A,B)	Occurrence
P(-1,1)	0/5
P(-1,0)	0/5
P(-1,-1)	1/5

The number of conditions in which both gene A and gene B have their values equal to 1 over all conditions is 2 out of 5 conditions, and so on. Another parameter we will need to calculate mutual information is joint entropy H(A,B):

$$H(A,B) = -\sum_{i,j=1}^{3} P_{ij} \times \log_2 P_{ij}$$

when both *i* and *j* independently run from 1 to 3, corresponding to the three states (1, 0 and –1).

$$H(A,B) = -1\times\left(\tfrac{2}{5}\times \log_2 \tfrac{2}{5} + \tfrac{1}{5}\times\log_2 \tfrac{1}{5} + \tfrac{1}{5}\times\log_2 \tfrac{1}{5} + \tfrac{1}{5}\times\log_2 \tfrac{1}{5}\right) = 1.923$$

For the above example, the mutual information between the two expression profiles, which provides a measure of the similarity between the two genes can be calculated as:

$$M(A,B) = H(A) + H(B) - H(A,B) = 1.371 + 1.522 - 1.923 = 0.970$$

- In general, the higher the mutual information score, the more similar are the two profiles.
- However, precise state and consequently, interpretation of the observed score would

depend on the number of conditions for which measurements were available.

- For our case of 5 conditions, the obtained score of 0.97 is high.