ANALYSIS OF GENE EXPRESSION DATA

MBI401-High throughput Data Generation & analysis

Mamta Sagar

Department of Bioinformatics

University Institute of Engineering & Technology, CSJM University, Kanpur

- 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 profi le.

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 refl ects 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 profi les into ‘disease state’ and ‘normal state’ groups, and then look for patterns that separate the sample profi le of the ‘disease state’ from the sample profi le 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 profi les 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 fi ve
- 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 profi les as vectors

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

$$G_i = [x_{i1}, \quad x_{i2}, \quad x_{i3}, \quad .., \quad x_{in}]$$

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

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

References

Lecture was prepared using following study material

- Roger Bumgarner, DNA microarrays: Types, Applications and their future Curr Protoc Mol Biol. 2013 January ; 0 22: Unit–22.1.. doi:10.1002/0471142727.mb2201s101.
- **Madan Babu, M., 2015. *An Introduction to Microarray Data Analysis*. [online] Mrc-lmb.cam.ac.uk. Available at: <http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/> [Accessed 3 December 2015].**