Distance Measures

MBI401-High throughput Data Generation & analysis

Mamta Sagar
Department of Bioinformatics
University Institute of Engineering & Technology, CSJM University, Kanpur

Distance measures

- Analysis of gene expression data is primarily based on comparison of gene expression

profi les or sample expression profi les.

- In order to compare expression profi les, 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 profi les.
- 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 profi les 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 coeffi cient

- One of the most commonly used metrics to measure similarity between expression profi les is the Pearson correlation coeffi cient (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:

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

Step2: "Mean centre" expression profiles

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

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

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

Where,

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

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

- The reason why we “mean centre” the expression profi les is to make sure that we compare shapes of the expression profi les and not their magnitude.
- Mean centering maintains the shape of the profi le, but it changes the magnitude of the profi le as shown in Figure 4.
- A PCC value of 1 essentially means that the two genes have similar expression profi les 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 profi les 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 profi les 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 profi le for the two genes can be computed as follows:

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

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

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

- 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)	$\frac{2}{5}$
P(1,0)	$\frac{0}{5}$
P(1,-1)	$\frac{1}{5}$

P(A,B)	Occurrence
P(0,1)	$\frac{0}{5}$
P(0,0)	$\frac{1}{5}$
P(0,-1)	$\frac{0}{5}$

P(A,B)	Occurrence
P(-1,1)	$\frac{0}{5}$
P(-1,0)	$\frac{0}{5}$
P(-1,-1)	$\frac{1}{5}$

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).

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

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.

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].**