Missing values management

One of the problems customarily encountered when analyzing gene expression data (especially with microarrays) is the occurence of missing values.

To deal with them, we use a methodology suggested by Kaufman and Rousseeuw (Kaufman, L. and Rousseeuw, P.J. (1990) *Finding groups in data: an introduction to cluster analysis.* Wiley, New York – see page 14-15). The mathematical description of this method is given below. In general, the operations where missing values have to be taken into consideration in Step 1 and Step 2 are the calculations of Euclidean distances and the calculation of mean expression profiles.

Suppose that $A=\{g_i(g_i^1,g_i^2,\ldots, g_i^E), i=1,\ldots,Q\}$ is a set of Q gene expression profiles g_i where E is the number of measurements for each gene (subscript indicates the gene number; superscript indicates measurement number). Suppose that the measurement numbers of the missing values for gene expression profile g_i are given by the set $P_i=\{p_{i,m}\}_{m=1,\ldots,Mi}$, where Mi is the number of missing values in g_i. For example, suppose that E=7 and g1={1, 3, –9, *, 5, *, 0} (* indicates a missing value), then $P_1=\{4,6\}$ ($p_{1,1}=4$; $p_{1,2}=6$; $M_1=2$).

If we want to calculate the Euclidean distance $d(g_k,g_l)$ between g_k and g_l, we have to take their missing values into account. Suppose that $\#(P_k \cup P_l) < E$, otherwise $d(g_k,g_l)$ is undefined (# means number of elements in a given set). We define $d(g_k,g_l)$ as:

$$d(g_k, g_l) = \sqrt{\frac{E}{E - \#(P_k \cup P_l)} \sum_{i \notin (P_k \cup P_l)} \left(g_k^i - g_l^i\right)^2}$$

For example if E=7 and g_1={1, *, *, -7, 9, 0, -1} and g_2={*, 2, *, 5, 1, *, *} then P_1={2,3}, P_2={1,3,6,7}, P_1UP_2={1,2,3,6,7} and #(P_1UP_2)=5. The distance d(g_1,g_2) is given by:

$$d(g_1, g_2) = \sqrt{\frac{7}{7-5}\left[(-7-5)^2 + (9-1)^2\right]} = 26.98$$

If we want to calculate the mean expression profile g_{av} of A, we also have to take the missing values into account. The j-th measurement of g_{av} (g_{av}^j) is defined as follows (Note that * . 0 = 0):

$$g_{av}^j = \begin{cases} \dfrac{1}{N(j)} \displaystyle\sum_{p=1}^{Q} (g_p^j . D(p, j)) & \text{if} \quad N(j) \neq 0 \\ \text{missing} & \text{if} \quad N(j) = 0 \end{cases}$$

where

$$D(p, j) = \begin{cases} 1 & \text{if} \quad j \notin P_p \\ 0 & \text{if} \quad j \in P_p \end{cases}$$

$$N(j) = \sum_{p=1}^{Q} D(p, j)$$

and

For example if E=7 and A={g_1,g_2,g_3} (Q=3) where

g_1={1, *, *, -7, 9, 0, -1}

g_2={*, 2, *, 5, 1, *, *}

g_3={2, 3, *, -9, *, 6, *}

then

$$g_{av} = \left\{\frac{1+2}{2}, \frac{2+3}{2}, *, \frac{-7+5-9}{3}, \frac{9+1}{2}, \frac{6+0}{2}, \frac{-1}{1}\right\} = \{1.5, 2.5, *, 3.667, 5, 3, -1\}$$

Note that the approach described above discards the need to replace the missing values by fictive values (e.g., by using interpolation).