Supplementary Method

The cosine coefficient and Pearson's correlation coefficient

The cosine coefficient between gene j and k denoted by c_{jk} is given as follows;

$$c_{jk} = \frac{\sum_{i=1}^{n} X_{ij} X_{ik}}{\left(\sum_{i=1}^{n} X_{ij}^{2}\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} X_{ik}^{2}\right)^{\frac{1}{2}}}. \qquad (1)$$

Here, X_{ij} denotes the expression level of gene j in i tissue. Pearson's correlation coefficient between gene j and k denoted by r_{jk} is given by the following equation.

$$r_{jk} = \frac{\sum_{i=1}^{n} \left(X_{ij} - \overline{X_j}\right)\left(X_{jk} - \overline{X_k}\right)}{\left\{\left[\sum_{i=1}^{n} \left(X_{ij} - \overline{X_j}\right)^{2}\right]\left[\sum_{i=1}^{n} \left(X_{ik} - \overline{X_k}\right)^{2}\right]\right\}^{\frac{1}{2}}}. \qquad (2)$$

$\overline{X_j}$ and $\overline{X_k}$ denote the average of expression level of gene j and k respectively, that is if the expression data is available among all the 10 tissues, the average should be 0.1(10%). These similarity correlation coefficients are defined on the range -1.0 to 1.0. A large positive value, close to 1, indicates the expression pattern retrieve is very similar to a query pattern; and a negative value, close to -1, indicates the expression pattern retrieve is dissimilar. However, note that since the expression level, X, is always 0 or positive, it can be easily verified from equation (1) that the range of c_{jk} should be 0 to 1. These values reflect the extent of differentiation of expression patterns of two genes that

are represented by 10 dimensional vectors. Intuitively, these values are considered a measurement in terms of the angle between the two vectors projecting from the origin. Note that in the case of Pearson's correlation coefficient, the end point for each vector is moved by subtracting the average (see equation (2)).

REFERENCE

CLUSTER ANALYSIS FOR RESEARCHERS, H. C. Romesburg, LULU PRESS, 2004