Method Description of STC

STC(Series Test of Cluster) mainly study gene expression profiles when organism change in sequence or receive some external environment stimulation. In this article, we take biology gene expression profiles at different time points as an example. STC algorithm can determine which profiles have a statically significant higher number of genes assigned, Further above result can reveal the change rule of organism at different time points. The following sections will present detailed steps.

Define model profiles

In order to control the maximal rangeability of gene expression between two adjacent time points we define parameter c. For example, if $\mathrm{c} = 2$, gene expression in next time points can up or down 0 to 2 units. For the total number of time points is n, we can define $(2\mathrm{c}+1)^{\mathrm{n}-1}$ model profiles. For example, the time series (0, 2, 1, 2) and (0,-1,-3,-4) are different model profiles. When the total number of time points is large, we select the part of model profiles.

Assign the gene expression data of samples to model profiles

Logarithmic standardization is applied for the gene expression data of samples. In detail, every gene expression data transform into $\log_2\left(X_{ij}/X_{1j}\right)$, where X_{ij} is the original value of j gene i time point. We define the set of model profiles is M and the set of gene expression profiles is G, we assigns each gene expression profiles $\mathrm{g_i} \in \mathrm{G}$ passing the filtering criteria to the model profile $\mathrm{m_i} \in \mathrm{M}$ that most closely matches the gene's expression profile as determined by the distance. Gene expression profiles assigned to the model profiles are called hereinafter the genes in the model profiles. Where distance $\mathrm{d}(\mathrm{g_i}, \mathrm{m_i}) = 1 - \rho(\mathrm{g_i}, \mathrm{m_i})$ and $\rho(\mathrm{g_i}, \mathrm{m_i})$ is the correlation coefficient of $\mathrm{g_i}$ and $\mathrm{m_i}$.

Verify the model profile's significance

We defined the null hypothesis that the values of any two different time points are independent. If the number of genes in the model profile under the true ordering of time points is significantly more than the number of genes under the random ordering

of time points, the model profile which represented the gene expression profile of the organism is apparently off the null hypothesis. That is to say, the profile is characteristic in the change process of biological sample. The algorithm can determine which profiles have a statically significant higher number of genes assigned using a permutation test.

In detail, we let the number of time points is n, every gene have $n!$ permutation, for every permutation, we assigns each gene to the model profile that most closely matches the gene's expression profile. s_i^j denote the number of gene that is assigned in i model profile in j permutation. We let $S_i = \sum_j s_i^j$. If the data is generated under the null hypothesis, $E_i = S_i/(n!)$ is the predicted number of genes in the model profile. Note different model profiles have different number of genes, in general $E_i \neq |G|/m$. We assume the number of genes in the model profile obey the binomial distribution whose parameters are $|G|$ and $E_i/|G|$. We let $t(m_i)$ is the number of genes in the m_i th model profile. The p_value is $p(X \geq t(m_i))$, $X \sim Bin(|G|, E_i/|G|)$, so we can obtain the significant level of single model profile.

The control of type I error of the overall multiple comparison test

The individual significance test is implemented for m model profiles, so we need control of type I error of the multiple comparison test. We apply the Bonferroni correction, that is to say, we strongly control the FWER (family-wise error rate) by $p(X \geq t(m_i))/m$.

Reference

[1] Xiao S, Mo D, Wang Q, et al. Aberrant host immune response induced by highly virulent PRRSV identified by digital gene expression tag profiling[J]. BMC genomics, 2010, 11(1): 544.

GCBI