趋势分析方法学说明

趋势分析主要研究生物体按照一定顺序发生变化或者受到外界环境刺激时基因表达变化所呈现趋势特征，本文以生物体在不同时间点基因表达的变化为例。趋势分析主要是通过统计学的方法得到时间序列样本中与时间模式相关的某些趋势模型以及与趋势模型紧密相关的基因群，从而揭示生物样本在时间变化过程中所特有的变化规律，其中趋势模型的定义可见于步骤描述。

趋势分析的具体步骤如下：

一、定义趋势模型

首先，通过定义参数c来控制基因在两个时间点之间最大的变化幅度，例如$c = 2$时，在两个时间点之间，基因表达量可以向上变化 1 个或 2 个单位的幅度，也可以无变化，或者向下变化 1 个或 2 个单位的幅度。对于n个时间点，于是就可以得到$(2c + 1)^{n-1}$个不同的趋势。这样在 4 个时间节点的情况下，时间序列(0,2,1,2)与序列(0,-1,-3,-4)都是定义的趋势模型，我们称之为 Profile。对于节点数较多的情况下，我们会对部分趋势予以优化选择。

二、建立样本基因表达数据与趋势模型的关系

首先对序列样本的基因表达数据进行对数标准化，方法是求序列样本中每个时间点的基因表达数据与第一个时间点的比值，再取以 2 为底的对数。定义趋势模型的集合为 M，基因表达样本序列的集合为 G，根据基因表达样本序列与趋势集合 M 中哪一个趋势的距离最小，把每一个基因表达样本序列$g_i \in G$分配到趋势模型$m_i \in M$中，下文简称趋势中的基因，其中距离$d(g_i, m_i)$的定义为$d(g_i, m_i) = 1 - \rho(g_i, m_i)$，$\rho(g_i, m_i)$表示$g_i$与$m_i$的相关系数。

三、检验上述关系模型的显著性

定义零假设：任一个时间点上的值与过去和将来的时间点的值都是独立的。如果实际分配的基因个数显著大于随机状态下分配的基因个数，则表明代表生物功能的趋势模型显然偏离了零假设，也即该趋势是生物样本变化过程中所特有的，而不是随机发生的。于是通过模拟置换的方法，改变基因表达的样本顺序，计算分配到每个趋势模型中的基因数目，来检验某个趋势模型是否具有统计显著性意义。设有 n 个样本点，于是每个基因都有 n！次置换，对于每一次置换，我们将基因都分配到与它最接近的趋势模型中。用s_i^j表示在置换 j（j 表示 n!次置换中的一次）中，分配到趋势模型 i 中的基因数量。我们设 $S_i = \sum_j s_i^j$，如果数据是在零假设下产生的，则$E_i = S_i/(n!)$就是每个趋势模型的基因预测数目。要注意的是不同的趋势模型含有基因数目也不同，所以一般$E_i \neq |G|/m$。于是可以假设每个趋势中基因数量是关于参数|G|和$E_i/|G|$的二项分布，设$t(m_i)$为分配到趋势模型m_i中基因的数目， $t(m_i)$个基因分配到趋势模型m_i中的 p 值为$p(X \geq t(m_i))$，$X \sim Bin(|G|, E_i/|G|)$，于是我们可以得到单个趋势模型的显著性水平。

四、多重比较检验整体第一类错误的控制

我们对 m 个趋势模型进行了单个显著性检验，所以需要进行多重比较检验的第一类错误的控制。这里我们采用 Bonferroni 校正，即通过$p(X \geq t(m_i))/m$时，来强控制 FWER(family-wise error rate)。

参考文献

[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.