SAS 系统算法原理

SAS 系统统计学运算部分采用 R-software(The R Project for Statistical Computing, http://www.r-project.org/)中的 R-package 中的函数。

1、 统计学分析

a) DiffGene 计算，*t* 检验(t-test)和 SAM 检验(Significant Analysis of Microarray) ：
核心函数：T-test: t.test(R default)[1]， SAM: samr (R default)[2]

b) 一维方差分析，one way ANOVA:
核心函数：aov (R default)[3]

c) 二维方差分析，two way ANOVA:
核心函数：aov (R default)[3]

d) 主成份分析，Principal Component Analysis， PCA:
核心函数：cmdscale (R default)[4]

e) 相关性分析，Correlation analysis:
利用皮尔森关联算法[5,6](Pearson correlation)计算实验数据两两之间的相关性，检验样品之间的相似/异度。

f) 聚类分析，Hierarchical Clustering:
核心函数：hclust (R default, complete linkage)[7]

SAS 系统功能注释结合 R-software 和 7 大公共数据库，对差异基因进行富集度计算和功能注释。

2、 功能注释

g) 差异基因基本信息注释：
数据库来源：NCBI Entrez Gene 数据库 http://www.ncbi.nlm.nih.gov/gene/

h) 差异基因的 GO 富集分析，GO enrichment analysis:
富集度 *p* 值算法：R-package Fisher’s Exact Test [8, 9, 10],
富集度 *q* 法：R-package John Storey’s method [11, 12, 13, 14]
数据库来源：Gene Ontology 数据库 http://www.geneontology.org/

i) 差异基因的 Pathway 富集分析，Pathway enrichment analysis:
富集度 *p* 值算法：R-package Fisher’s Exact Test [8, 9, 10],
富集度 *q* 法：R-package John Storey’s method [11, 12, 13, 14]
数据库来源：KEGG 数据库 http://www.genome.jp/kegg/,
Biocarta 数据库 http://www.biocarta.com/

j) 差异基因编码的蛋白质相互作用关联查询：
数据库来源：人类蛋白质相互作用数据库 (HPRD) Human Protein Reference Database
http://www.hprd.org/
分子相互作用数据库 (MINT) a Molecular Interaction database
http://mint.bio.uniroma2.it/mint/

k) 差异基因对应的小分子 RNA(microRNA)关联查询：
数据库来源：Sanger microRNA 数据库 http://www.mirbase.org/

3、 参考文献

1． George Casella, Roger L.Berger, Statistical Inference, chapter 2

2． Tusher, V., Tibshirani, R. and Chu, G. (2001): Significance analysis of microarrays applied to the ionizing radiation response PNAS 2001 98:5116-5121

3． Chambers, J. M., Freeny, A and Heiberger, R. M. (1992) Analysis of variance; designed experiments. Chapter 5

4． Cox, T. F. and Cox, M. A. A. (1994) Multidimensional scaling. Chapman and Hall

5． J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1):59-66, Feb 1988.

6． Stigler, Stephen M. (1989). Francis Galton’s Account of the Invention of Correlation. *Statistical Science 4 (2).*

7． Anderberg, M. R. (1973).Cluster Analysis for Applications. Academic Press: New York

8． Fisher, R. A. (1922). “On the interpretation of χ^2 from contingency tables, and the calculation of P”. Journal of the Royal Statistical Society85 (1)：87-94. doi:10.2307/2340521. JSTOR 2340521.

9． Fisher, R.A. (1954). Statistical Methods for Research Workers. Oliver and Boyd.

10. Agresti, Alan (1992). “A Survey of Exact Inference for Contingency Tables”. Statistical Science 7 (1):131-153.

11. Storey JD. (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, 64: 479-498.

12. Storey JD and Tibshirani R. (2003) Statistical significance for genome-wide experiments. Proceedings of the National Academy of Sciences, 100: 9440-9445.

13. Storey JD. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Annals of Statistics, 31: 2013-2035.

14. Storey JD, Taylor JE, and Siegmund D. (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. Journal of the Royal Statistical Society, Society, Series B, 66:187-205.

上海伯豪生物技术有限公司
生物信息部
2015 年 7 月