P,Q value

Created time: 2009.8.27
Updated time: 2010.04.28
Written by: Trysia Chen

p value

we calculate the p value using a hypergeometric distribution. If a whole genome has N total genes, among which M are involved in the term under investigation, and the set of genes has n total genes, among which m are involved in the same term, the p value for the term is calculated as follows:

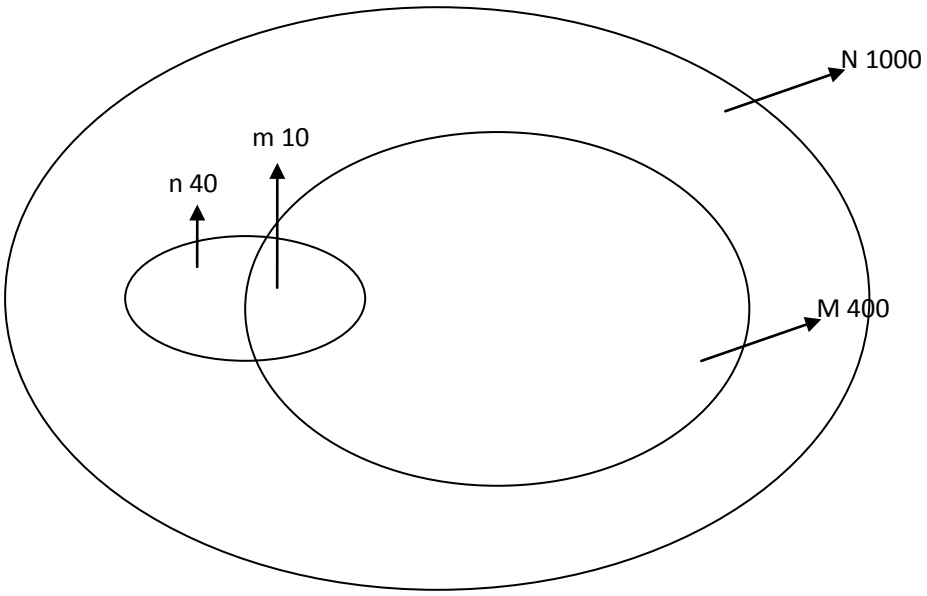
$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

N: total number of genes

M: total number of genes annotated with this term

n: number of genes in the list

k: number of genes in the list annotated with this term



The genes are enriched significantly in the pathway when P value< 0.05.
Use R code to calculate the p q value.

Q value

Q is Fasle Discovery Rate（FDR），Q default 0.05, the less the q is, the more

significant the genes(or proteins) enriched in the one pathway(or GO term), and the less FDR。

P vector is (P_1,P_2…P_m), v = rank p, return v(v_1,v_2,…v_m) to $\mathbf{q_i}$= eta0*m*p_i/v_i

Eta0 estimated adaptive(Benjamini and Hochberg (2000) J. Behav. Educ. Statist.)method

Reference

1 vol 21 no,19 2005,pages 3787-3793,Automated genome annotation and pathway identification using the KEGG Orthology(KO) as a controlled vocabulary,Xinzeng Mao,Tao Cai etal.

2 Johnson, N. L., Kotz, S., and Kemp, A. W. (1992) Univariate Discrete Distributions, Second Edition. New York: Wiley.

3 GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis，Qi Zheng1,2 and Xiu-Jie Wang1,* W358–W363 Nucleic Acids Research, 2008, Vol. 36, Web Server issue

4 Benjamini and Hochberg JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS.2000; 25: 60-83