生物物理学报(2005) 21(3), 187-194

PDF文档

结合基因功能分类体系Gene Ontology筛选聚类特征基因

徐建震[1]、郭政*[1,2,3]、李霞[1,2,3]、李永进[1]、刘帅[1]、屠康[1]
1 哈尔滨医科大学生物信息学系
2 哈尔滨工业大学计算机科学与技术学院
3 同济大学生命科学与技术学院

使用两套基因表达谱数据，按各基因的表达值方差，选择表达变异基因对样本聚类，发现一般使用方差较大的前10%的基因作为特征基因，就可以较好地对疾病样本聚类；对不同的疾病，包含聚类信息的特征基因有不同的分布特点。在此基础上，我们提出结合基因功能分类体系Gene Ontology，进一步筛选聚类特征基因的方法。通过检验在Gene Ontology(GO)中的每个功能类中的表达变异基因是否非随机地聚集，寻找疾病相关功能类，再根据相关功能类中的表达变异基因进行聚类分析。实验结果显示：结合功能体系GO进一步筛选表达变异基因作为聚类特征基因，可以保持或提高聚类准确性，并使得聚类结果具有明确的生物学意义。另外，本实验发现了一些可能和淋巴瘤和白血病相关的基因。

Feature Selection for Clustering Disease Samples Based on Gene Ontology

By analyzing two microarray datasets of leukemia and lymphoma, we demonstrate that the disease subtypes can be well clustered based on the top 10% genes expressed with the highest variations across disease samples. The feature genes, including strong clustering information, have different distribution characteristics in the two disease datasets. Based on this observation, we propose a new method to select feature genes for disease clustering based on gene expression profiles and gene functional knowledge. After annotating each individual gene to functional classes defined in Gene Ontology, we identify the disease relevant functional classes significantly enriched with differentially expressed genes, and then cluster disease samples by the differentially expressed genes contained in these identified functional classes. Our experiment results showed that the new clustering procedure performs better than that with traditional procedures. Besides, biological function comprehension can be achieved directly with this new approach. Two feature genes sets, which may functionally relevant to leukemia and lymphoma respectively, are extracted.

关键词