研究论文

不相交主成分分析(PCA)和遗传算法(GA)用于差异表达基因的识别

苏振强[1,3], HONG Hui-Xiao[2], TONG Wei-Da[3], PERKINS Roger[2], 邵学广[4], 蔡文生[1,4]

1. 中国科学技术大学化学系, 合肥 230026;
2. Division of Bioinformatics, Z-Tech at FDA's National Center for Toxicological Research, Jefferson, AR 72079, USA;
3. Center for Toxicoinformatics, National Center for Toxicological Research (NCTR), US Food and Drug Administration(FDA), Jefferson, AR 72079, USA;
4. 南开大学化学系, 天津 300071

摘要　　建立了一种基于不相交主成分分析(Disjoint PCA)和遗传算法(GA)的特征变量选择方法, 并用于从基因表达谱(Gene expression profiles)数据中识别差异表达的基因. 在该方法中, 用不相交主成分分析评估基因组在区分两类不同样品时的区分能力; 用GA寻找区分能力最强的基因组; 所识别基因的偶然相关性用统计方法评估. 由于该方法考虑了基因间的协同作用更接近于基因的生物过程, 从而使所识别的基因具有更好的差异表达能力. 将该方法应用于肝细胞癌(HCC)样品的基因芯片数据分析, 结果表明, 所识别的基因具有较强的区分能力, 优于常用的基因芯片显著性分析(Significance analysis of microarrays, SAM)方法.

关键词　　基因芯片　主成分分析(PCA)　遗传算法(GA)　基因芯片显著性分析(SAM)　偶然相关

分类号　0652

Identification of Differentially Expressed Genes Using Disjoint Principal Component Analysis Coupled with Genetic Algorithm

SU Zhen-Qiang[1,3], HONG Hui-Xiao[2], TONG Wei-Da[3]*, PERKINS Roger[2], SHAO Xue-Guang[4], CAI Wen-Sheng[1,4]*

1. Department of Chemistry, University of Science and Technology of China, Hefei 230026, China;
2. Division of Bioinformatics, Z-Tech at FDA's National Center for Toxicological Research, Jefferson, AR 72079, USA;
3. Center for Toxicoinformatics, National Center for Toxicological Research(NCTR), US Food and Drug Administration(FDA), Jefferson, AR 72079, USA;
4. Department of Chemistry, Nankai University, Tianjin 300071, China

Abstract A new method for the feature selection using disjoint principal component analysis(PCA) coupled with genetic algorithm(GA) was proposed and was used to identify differentially expressed genes based on microarray gene expression profiles. The discriminatory power of combination of genes is assessed with using disjoint PCA, the combinatorial optimization problem of genes is solved by using GA, and the chance correlation of genes is assessed by a statistic method. Due to considering the cooperation between genes which is a way to approximate the synergistic regulation by genes during the biological processes, the genes identified by our method are capable of powerful ability to express the differences. This method has been applied to analyze the gene microarray data of hepatocellular caricinoma(HCC). It is found tha

t the genes identified by the proposed method has more discriminatory power in distinguishing two-class samples than those identified by SAM(significance analysis of microarrays), which is very popular in the analysis of microarray data.

Key words Microarray Principal component analysis(PCA) Genetic algorithm(GA) Significance analysis of microarrays(SAM) Chance correlation