Gene expression profiling of breast cancer survivability by pooled cDNA microarray analysis using logistic regression, artificial neural networks and decision trees

Chou, Hsiu-Ling,Yao, Chung-Tay,Su, Sui-Lun,Lee, Chia-Yi,胡光宇,Terng, Harn-Jing,Shih, Yun-Wen,Chang, Yu-Tien,Lu, Yu-Fen,Chang, Chi-Wen,Wahlqvist, Mark L,Wetter, Thomas,Chu, Chi-Ming
Bioinformatics
Computer Science and Informatics
kyhu@chu.edu.tw

Abstract

BACKGROUND: Microarray technology can acquire information about thousands of genes simultaneously. We analyzed published breast cancer microarray databases to predict five-year recurrence and compared the performance of three data mining algorithms of artificial neural networks (ANN), decision trees (DT) and logistic regression (LR) and two composite models of DT-ANN and DT-LR. The collection of microarray datasets from the Gene Expression Omnibus, four breast cancer datasets were pooled for predicting five-year breast cancer relapse. After data compilation, 757 subjects, 5 clinical variables and 13,452 genetic variables were aggregated. The bootstrap method, Mann--Whitney U test and 20-fold cross-validation were performed to investigate candidate genes with 100 most-significant p-values. The predictive powers of DT, LR and ANN models were assessed using accuracy and the area under ROC curve. The associated genes were evaluated using Cox regression. RESULTS: The DT models exhibited the lowest predictive power and the poorest extrapolation when applied to the test samples. The ANN models displayed the best predictive power and showed the best extrapolation. The 21 most-associated genes, as determined by integration of each model, were analyzed using Cox regression with a 3.53-fold (95% CI: 2.24-5.58) increased risk of breast cancer five-year recurrence... CONCLUSIONS: The 21 selected genes can predict breast cancer recurrence. Among these genes, CCNB1, PLK1 and TOP2A are in the cell cycle G2/M DNA damage checkpoint pathway. Oncologists can offer the genetic information for patients when understanding the gene expression profiles on breast cancer recurrence.

Keyword：Breast Cancer、Microarray、Artificial Neural Network、Logistic Regression 、Decision Tree