Letter

How many genes are needed for early detection of breast cancer, based on gene expression patterns in peripheral blood cells?

Wuju Li

Center of Computational Biology, Beijing Institute of Basic Medical Sciences, Beijing, China

Corresponding author: Wuju Li, liwj@nic.bmi.ac.cn

Published: 29 July 2005
This article is online at http://breast-cancer-research.com/content/7/5/E5

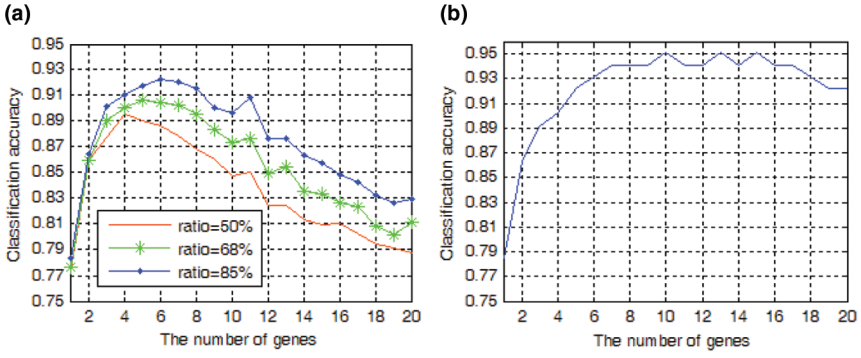
Breast Cancer Research 2005, **7**:E5 (DOI 10.1186/bcr1295)

See related research by Sharma *et al.*, http://breast-cancer-research.com/content/7/5/R634

In their recent report [1], Sharma and coworkers explore the early detection of breast cancer. They analyzed a gene expression data set (1368 genes in 62 normal and 40 tumour samples, including sample duplication in different batches) using the nearest shrunken centroid method. They identified a panel of 37 genes that permitted early detection, with the classification accuracy being about 82%. This is a typical problem with sample classification based on gene expression profiling. The objective is to achieve high prediction accuracy with as few genes as possible, and so feature selection plays an important role; examination of a large number of genes will increase the dimensionality, computational complexity, and clinical cost. According to our previous study of data sets from patients with colon cancer, leukaemia and breast cancer [2], we estimated that five or six genes – rather than 37 – would be sufficient for the early detection of beast cancer [1]. So how many genes are indeed needed? In order to address this question, we evaluated the data presented by Sharma and coworkers using the Tclass system [2].

In the Tclass system, Fisher's linear discriminant analysis and a step-wise optimization procedure for feature selection are used to analyze a batch adjusted data set [1] in two ways. The first is to take the prediction accuracy from the training set as the object function. The second way is to take the classification accuracy from the leave-one-out cross-validation as the object function. For the former, the selected optimal feature sets are evaluated by randomly dividing all tissue samples into a training set (e.g. 50%, 67%, or 85% of samples) and a test set 200 times. The relationship between the prediction accuracy and the number of genes is illustrated in Fig. 1, which shows that the greatest prediction accuracy was achieved using six genes (Fig. 1a); other peaks in accuracy occurred when 10, 13, or 15 genes were used (Fig. 1b). Furthermore, two genes – the 481th (BC009696) and the 801th (BC000514) – permitted classification accuracy as high as 86%, which is greater than the 82% achieved by Sharma and coworkers [1] with the selected 37 genes.

Figure 1



Number of genes examined and classification accuracy. The relationship between the number of genes and classification accuracy is shown for **(a)** different partition ratios and **(b)** leave-one-out cross-validation analysis.

In summary, we may draw the following conclusions. First, the number of genes needed for early detection of breast cancer is fewer than 10, based on the data set in the report by Sharma and coworkers [1]. Second, the classification accuracy will gradually decrease when the number of genes exceeds 6 (Fig. 1a) and 10 (Fig. 1b). Related details and information regarding the Tclass system are available upon request or from our website [3].

Competing interests

The authors(s) declare that they have no competing interests.

Acknowledgements

This work is supported by grant #5042021 from Beijing Natural Science Foundation and grant #30470411 from National Natural Science Foundation of China.

References

1. Sharma P, Sahni NS, Tibshirani R, Skaane P, Urdal P, Berghagen H, Jensen M, Kristiansen L, Moen C, Sharma P, *et al*.: **Early detection of breast cancer based on gene-expression patterns in peripheral blood cells.** *Breast Cancer Res* 2005, **7:** R634-R644.
2. Li WJ, Xiong MM: **Tclass: tumor classification system based on gene expression profile.** *Bioinformatics* 2002, **18:**325-326.
3. **How many genes are needed for early detection of breast cancer based on gene-expression patterns in peripheral blood cells?** [http://www.biosun.com.cn/tclass/]