Author's response to reviews

Title: Classification Between Tumor and Normal Tissues Based on The Pair-wise Gene Expression Ratio

Authors:

YeeLeng Yap (daniely@hkusua.hku.hk)
XueWu Zhang (xwzhang@hkucc.hku.hk)
Mt Ling (patling@hkucc.hku.hk)
XiangHong Wang (xhwang@hkucc.hku.hk)
Yc Wong (ycwong@hkucc.hku.hk)
Antoine Danchin (adanchin@pasteur.fr)

Version: 3 Date: 25 July 2004

Author's response to reviews: see over
Answer to Reviewer’s Report (In RED)
Reviewer’s report (In BLACK)
Title: Class Prediction Between Tumor and Normal Tissues Based on The Pair-wise Gene Expression Ratio
Version Date: 2 April 2004
Reviewer: Inge Jonassen
Reviewer’s report:
General

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)
The presentation is not very clear. A number of aspects need to be more detailed described. At the same time the manuscript is too verbose and containing too many figures and tables. A major revision is needed and it would need to be reviewed again in light of more detailed description of critical aspects of the methodology applied.

Statistics for revised version
Total words of the entire manuscript: 10 629 → 7479 words
Number of word in abstract: 536 → 340 words
Number of words in main text: 5781 → 4508 words
Number of references: 98 → 54 references
Number of figures: 12 → 9 figures
Number of tables: 9 → 8 tables

Some critical aspects:
- It is not clear whether feature selection is performed only using the training data or on the full data set before the leave one out cross validation is applied using the features (genes or gene ratios) selected by analyzing the full data set. See Ambroise and McLachlan PNAS 2002 for a discussion.

Answer: For this study, all tissues samples and all genes are recruited for determination of key cancer-related genes (features) based their individual discriminating efficiency PRIOR to leave-one-out validation. This, I aware, has some inherent limitations as discussed in the article you suggested (Ambroise and McLachlan PNAS 2002). I will include the reference in this manuscript to make sure the readers aware of the limitation of using FULL DATASET to derive the discriminating efficiency of markers. Therefore, there is a need to further explore a more robust validation methods.

Admittedly, we did tried to have a separate training set for deriving a set of reliable features (can be single gene expression and pair-wise gene expression ratio) and do the leave-one-out validation in the remaining dataset, but this leads to some unrelated features because of the small training set that we can afford to have, and the outcome of the cancer genes selection also depends on the subjective selection of training sets that will subsequently effect our end classification efficiency (effects discussed in Ambroise and McLachlan PNAS 2002). Because the objective is to select reliable features in order to compare classification efficiency based on either single gene/gene ratio, we decided to use FULL dataset instead.

Changes to the manuscript:
“Promising results have been reported, claiming near-perfect classification accuracy [1]. However, the usually small number of samples per class in most studies and the highly bias cross validation procedures have cast doubt on the classification accuracy in terms of their statistical significance [2].”

- it is not clear what method is used for class prediction
Answer: Recursive partitioning method [1] was used to determine the class of an unknown samples. I have rewritten part of the method section to make this clearer.

Changes to the manuscript:
“The first step to discriminate between the normal/tumor tissue samples using a specific feature i (single gene expression) is to determine the threshold value, \( T_i \) (Figure 2). There are two criteria for deriving a valid threshold value. First, it has to delineate correctly the one-dimensional region for either all the normal/tumor tissues using the full dataset (Figure 2). Secondly, it has to minimize the percentage of false prediction. Take gene #1659 for
example. To fulfill the two aforementioned criteria, it was determined that the region greater than 63.7 incorporates all the tumor tissue samples (Figure 3). It classifies correctly all tumor tissue samples with an overall false prediction of 13.9% in the tumor set. This is performed repeatedly for each feature (single gene expression or pair-wise gene ratio) until all the threshold values \(T_i \ldots \text{all features}\) are determined.

Now, to classify an unknown sample using 2-feature model classifier, the outcome of the class will be the joined combination of the prediction based on the derived threshold values of the two features. This will be repeated for all tissue samples to obtain the overall classification accuracy. In this manner, we recursively evaluate the classification of tissue samples based on different combinations of \(N\) genes and investigated the classifiers up to 10-feature model classifier.

- it is claimed that the gene ratio approach alleviates the need for normalization - the argument for this is not convincing. The method is validated on one data set only. This aspect should be addressed in a more careful manner.

Answer: I have temperate my claim. It should not be perceived as replacing normalization because here in this study, I am proposing another way to discover cancer-related signals existed in the form of pair-wise RATIO instead of single gene aberration. The rationale is that genes are rarely alone, but always act to regulate or to be regulated by some other genes.

Changes in the manuscript:

“Here, instead of resolving to single gene expression, that depends heavily on normalization, for tissue classification, we presented a transformation method that uses pair-wise gene expression ratios within the same experiment as the discriminating axes. By doing so, we aimed to minimize the influence of different normalization methods considering that an experiment is self-consistent with the same factors affecting all genes in the same fashion. The rationale is that even when the normalization methods differ between two array experiments, their pair-wise gene expression ratios within the same experiment will remain relatively stable. If reliable cancer-related marker, exist in the form of pair-wise gene expression ratio, we are indeed discovered successfully, they will be relatively independent from the normalization method used on the dataset.”

We had carried out the similar analysis on a new dataset that we have produced in collaboration with Stanford university on lung adenocarcinomas and normal lung tissues, and verified also the single and pair-wise ratio of cancer gene marker quantitatively using qrt-PCR. The so-called housekeeping genes are found to be changing between normal cancer states, therefore causing problem in normalization. But, in some case, pair-wise ratio (that is like normalizing with the most correlated gene) produced favorable classification of tissue samples. We verified it using qrt-PCR. The result will be published shortly. I have attached some findings on the efficiency of classification of unknown cell samples in qrt-PCR results based on single gene and pair-wise gene ratio.

- the manuscript fails to refer to and discuss relationship to other papers considering pairs of genes or combinations of genes in the supervised analysis of gene expression data. For example Bø and Jonassen (Genome Biology 2002) proposed a method to consider pairs of genes in combination effectively able to discover pairs of genes whose ratios is informative about (e.g.) clinical outcome (that approach is not limited to ratios)

Answer: I have included this important paper in my revised manuscript. And had added discussion on some issues raised inside that paper.

Changes in the manuscript:

“Thirdly, although it has been recently established that genes segregate into clusters of interacting networks [3] instead of acting as one single entity, most cancer DNA-array studies have only investigated single gene aberration (up/down-regulated) when comparing tumor expression profiles to their corresponding controls. Interestingly, Bø and Jonassen investigated genes in pairs and demonstrated that gene pair can be used to discriminate efficiently the different tissue classes [4]. This idea about studying gene in pairs, or even higher order relationship, should be explored further to reveal new insight on multiple gene regulation and the formation of cancer cell.”

- in the analysis of gene ratios, not all possible gene ratios are considered since this is claimed to be too computationally demanding. I can hardly see that this is the case as for example Bø and Jonassen (see above) consider all gene pairs explicitly.
In Bø and Jonassen’s paper, the single gene expressions for the selected gene pairs were used as the discriminating axes for classifying tissue phenotypes. They studied how well different combinations of gene pairs separated between two tissue phenotypes in a multivariate way. In that study, they acknowledged that gene study in ALL-PAIRs procedure is indeed computationally expensive. If 100 genes were retained for classification purpose, there are essentially $100C_2 = 4950$ gene pairs to be considered.

In current study, we attempted to classify tissue phenotypes using single gene expression as well as pair-wise gene expression ratio as discriminating axes, and explore the classification potential of N-feature classifier up to order N=10 based on these discriminating axes. If, for instance, 100 single gene expression or pair-wise gene expression ratio were determined to be informative in separating tissue phenotypes (discriminating axes), in order to derive the performance of 10-feature model, we need to compute $100C_{10} = 1.731030945644000 \times 10^{13}$ difference combinations. This is significant computations unable for my computer to handle at this moment.

- The language and presentation should be improved
Answer: I have revised some grammatical mistakes and corrected some unclear phrases. Those corrected are in RED.

- The reference list is in my view too long and includes many very remotely relevant papers.
Answer: I have shortened the reference list.
Number of references: 100 → 55 references

References