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: 7 Date: 27 August 2004

Author's response to reviews: see over
Reviewer’s report
Title: Classification Between Tumor and Normal Tissues Based on The Pair-wise Gene Expression Ratio
Version: Date: 23 August 2004
Reviewer: Inge Jonassen
Reviewer’s report:
The revised manuscript is in my view much improved. Although I still find it verbose and think it would benefit from a shortening, I realize that this is partly a matter of taste and leave this decision to the editor.

Corrections highlighted in RED in revised manuscript.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

1) The recursive nature of the classification procedure used (Zhang et al) is not very well described. This manuscript simply says that the result of the test on the different features is combined - this should be clarified.

I have added explanation of the exact procedures taken to classify the unknown tissue samples. The manuscript had been revised accordingly to be clearer, instead of just referring the methodology to a published article. The revised version for the section is as below:

**Feature Partitioning Method [4] for Classification of Normal/Tumor Tissues using Single gene Expression.**

Regarding the Feature Partitioning Method (FPM), in order to discriminate between the normal/tumor tissues based on specific feature \(i\) (single gene expression), the first step is to determine the threshold value, \(T_i\), that can optimally splits all the tissue samples into tumor and normal tissue. The FPM algorithm has a recursive version [4], in which a decision tree depicting the classification rules for tissue samples was generated recursively. Both methods differ in the way \(T_i\)s are derived. Nonetheless, they are very intuitive and non-parametric in nature.

Also, they restrict no priori distribution patterns for features used. We adopted the simple FPM for tissue classification where each feature was treated individually. There are two criteria for deriving a valid threshold value \(T_i\) for each feature. First, it has to delineate correctly (discriminating efficiency=100%) the one-dimensional region (\(R_{\text{feature}_i}\)) for either all the normal/tumor tissues using all tissue samples (Figure 2).

Secondly, it has to minimize the percentage of false prediction for the other tissue type. Take gene #1659 for example. To fulfill the two aforementioned criteria, it was determined that the region greater than 63.7 (\(R_{\#1659}\)) incorporates all the tumor samples (Figure 4). It classified correctly all tumors (discriminating efficiency=100%) with an overall false prediction of 13.9% in the normal set. This was performed repeatedly for all features until all the threshold values \((T_i...all\ features)\) were determined.

Now, to classify an unknown sample using 2-feature model classifier, a combination of any two features and their corresponding pre-determined threshold values \(T_i\)s (selected from \(T_i...all\ features\) for each dataset) were recruited. The outcome of the tissue class will be determined depending on whether one/both the expression values of the unknown sample fall completely in either the normal/cancer region (\(R_{\text{feature}_i}\)). This is to say that if any of the two features from the unknown sample meets the criteria (\(R_{\text{feature}_i}\)) to be either normal/tumor tissue type (based on our definition, \(R_{\text{feature}_i}\) is a region with 100% discriminating efficiency for a specific tissue type), the unknown sample will be assigned to be normal/tumor respectively. This is repeated exhaustively for all possible combinations constituting of any two features. The procedure will be repeated for all tissue samples to evaluate the overall classification accuracy for 2-feature model classifier. In total, we evaluated the classification of tissue samples based on different combinations of \(N\) genes and investigated the classifiers up to 10-feature model classifier.
2) It is unclear what stops the authors from performing a more conservative test of prediction accuracy where not all examples are used in the feature selection step. The combination of including all features in feature selection and using a classifier method that may easily overfit, leaves the reader with an unclear impression. Do the high prediction accuracy arise because of the added value of using gene combinations or from the system's "information leakage"?

I agree with the Dr Jonassen that performing a more conservative test of prediction accuracy where NOT ALL samples are used to derive the threshold value, \( T_i \), in the feature selection step IS feasible. However, to adjust our methodology so that our results are unbiased (when comparing the outcome from single gene and pair-wise gene ratio) and in-line with our objective – that is to compare the classification efficiency between single gene and pair-wise gene ratio, we resorted to using ALL samples to derived threshold value, \( T_i \), in the feature selection step for both studies.

The main barrier in using ONLY part of the samples in the feature selection step is: how to select an unbiased training dataset (given that we already have small dataset, and also an unbalanced samples of normal/cancer in Colon dataset). Admittedly, we had a tough time deciding which sample should be included if we ONLY use part of the samples, and also have to worry about if the selected samples are a representable population for the entire dataset. Eventually, we ran into ambiguous results when a different population of samples was selected. Furthermore, we might miss important features because of the biased training dataset. In our recent study of our own microarray on lung adenocarcinomas, we have only 57 arrays of clinical samples (8 normal + 49). In this case, in order to derive the most meaningful pair-wise gene ratio for further qPCR verification (which we had done), we had resorted to using ALL samples as well… below please find our qPCR verification.

By including ALL samples for both studies (single gene and pair-wise gene ratio), we aimed to derive the MOST reliable threshold values and classified samples based on them (to biologists, identification of the most reliable feature will minimized waste of resources). Since the methodology is COMMON between both studies, and it includes all samples, the comparison of classification efficiency is valid and will reflect how well each feature (single gene and pair-wise gene ratio) can be used to delineate samples. On the other hand, there is already an obvious indication that showed improvement in feature-to-tissue_type correlation.
I also agree with Dr. Jonassen that the high prediction accuracy might arise from data overfit because we used the ALL SAMPLES in the feature selection step. However, because both studies (single gene and pair-wise gene ratio) undertook the same procedures, the outcome will still reflect the improved classification efficiency in the case of pair-wise gene ratio. With our follow-up studies (to be published soon), there are obviously improvements in cancer signals when genes are considered in pairs, as also discussed in your previous studies.

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
The order of figures had been re-organized. One figure (Figure 2 in the original document) had been deleted.
One mistakes in label has been corrected as follow:
1) Table 4. Prostate cancer: The coefficient of variation (CV) for the original dataset and transformed dataset according to their rank.
This table shows 10 data with lowest coefficient of variation, the complete table can be downloaded at http://web.hku.hk/~daniely/microarray.

Discretionary Revisions (which the author can choose to ignore)
The What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions
Level of interest: An article whose findings are important to those with closely related research
Quality of written English: Needs some language corrections before being published
Statistical review: No