Inter-platform Concordance of Gene Expression Data for the Prediction of Chemical Mode of Action

Chathura Siriwardhana, Susmita Datta, Somnath Datta*

Department of Bioinformatics and Biostatistics,University of Louisville, KY 40202, USA

1. Introduction

The purpose of this study is two fold: (i) develop a classifier that has high accuracy in both microarray and RNA-seq platforms and (ii) study the concordance of multiple standard classifiers in the two platforms. We use seven standard classifiers and an adaptive ensemble classifier built around them to achieve these goals. The dataset for our study is resulted from a Rat liver experiment conducted by the FDA SEQC consortium to assess the performance of modern gene transcript expression profiling methods and released as part of 2015 Critical Assessment of Massive Data Analysis (CAMDA) challenges. The Rat liver experiment was designed for developing predictive models to predict the chemical Mode of Action (MOA). A previous comprehensive analysis (Wang et al. 2014) of the above gnomic data suggested weak classification accuracies for a set of classifiers applied to multiple platforms.

2. Data

The dataset consists of gene expression responses profiled by Affymetrix microarray and Illumina RNA-seq in rat liver tissues from 105 male Sprague-Dawley Rats, which were exposed to 27 different chemicals represented by 9 different MOAs. Microarray and RNA-seq platforms contain gene expression measurements of nearly 31,000 and 46,000 genes, respectively. In the original experiment, a training set is formed with 45 rats, which were treated with 15 chemicals corresponding to MOAs of "PPARA",

*somnath.datta@louisville.edu

"CAR/PXR", "AhR", "Cytotoxic", "DNA damage", and 18 controls. Test set contains data on 36 rats which were treated with 12 chemicals corresponding to "PPARA", "CAR/PXR", "ER", "HMGCOA" and 6 controls. We noticed that two MOAs, "ER" and "HMGCOA", are presented only in the test set. Due to duplication and removal of some initial samples, the data set profiled by RNA-seq contains 116 samples, which causes imbalance between training sets among platforms. We further noticed, approximately 22,253 average expressions per sample in RNA-seq data were recorded as "NA", where it indicates an insufficient number of reads mapped onto the gene to provide a reliable gene expression estimate. As a result, around 16,133 expression measurements remained, once all "NA"s were removed.

3. Methodology

Support Vector Machine (SVM), Random Forest (RF), Neural Network (NN), Linear and Quadric Discernment Analysis (LDA, QDA) are some examples of standard techniques widely applied in classification problems. For high dimensional data, these classifiers are often combined with dimension reduction, variable selection, or penalization techniques such as Partial Least Squares (PLS), Principle Component Analysis (PCA), Random Forest (RF) based importance measures, L_1 regularization, etc., for greater applicability and improved prediction accuracy (Boulesteix, 2004, Dai, 2006). However, the accuracies of these individual classifiers are highly variable and dependent on the true underlying data structures of various classes. Datta et al. (2010) described an optimal adaptive ensemble classifier via bagging and rank aggregation to offer a classification solution that has good performance across multitude of data structures. The ensemble classifier we used is developed with a set of seven standard classifiers, namely, SVM, RF, LDA, PLS+RF (Random Forest using the PLS terms), PLS+LDA (linear discriminant analysis using the PLS terms) , PCA+RF (Random Forest using the principal components), PCA+LDA (LDA using the principal components), and Recursive Partitioning (RPART).

We conducted three different analyses to study the performances of our classifiers in classifying the MOAs: (1) Classifiers trained and tested on each individual platforms; (2) Classifiers trained in one platform and tested on the other platform; (3) Classifiers trained on the perturbed training set with permuted gene expressions for each platform followed by accuracy calculation for identification of important variables (genes).

In general, there is no established criteria to define prediction for an unknown class that was not represented in the training data. Thus, we performed the 1st analysis after removing all test samples belonging to two classes of "ER" and "HMGCOA". However, for the 2nd and the 3rd analyses we were able to retain all classes and data since in effect the the classifiers were trained on the union of training and testing data in each platform. We used normalized expression levels that came from microarray data using Robust Multi-Array Average (RMA) expression measurements (Irizarry et al., 2003), whereas data obtained for RNA-seq was already normalized via the Magic normalization. We felt that it would be more meaningful to perform an analysis with a common set of genes represented in both platforms for a comparative study. To that end the expression data for 8336 unique common genes were used in building our classifiers.

In the first analysis, we developed a set of classifiers directly using the training data with different classification algorithms and made predictions on the given test dataset in the same platform. However, since the classifier needed to run on both platforms for the 2nd analysis, each gene expression measurement was standardized, separately for both platforms, prior to the 2nd analysis. We performed a 10-fold cross validation for each individual classifier to select the number of components for PLS and PCA methods, separately for two platforms. We employed the same number of components to build the ensemble classifier. For the third analysis, we permute the expression of a single gene in the training set and fit a classifier on the modified training set followed by accuracy calculation on the test set. This was done for each of the gene common to both platforms. The reduction in accuracy as compared to the original (unperturbed) training set is a measure of importance of a given gene in the classification process. In order to reduce the computational burden, we did not use the ensemble classifier for this purpose. Instead the component classifier PLS+LDA which had an overall accuracy close to that of the ensemble classifier was used. The genes are then ranked according to their importance for both platforms.

4. Results

The results of analyses 1 and 2 are summarized in Figure 1. The left panel shows that the performance of each classifier is similar in both platforms since all the data points are fairly close to the diagonal line (Pearson's r =0.92). The accuracy of individual classifier varies from 17% to 75%, and as to be expected, the performance of the ensemble classifier is the best in both platforms. The overall accuracy of the optimal classification method is slightly better in microarray compared to RNA-seq (75% vs 67%). On the other hand, we observe a lower prediction accuracy for the class "PPARA" in RNA-seq (55.56%), compared to the microarray (88.87%) platform (not shown in the figure). Results of the second analysis summarized on the right hand plot shows even greater agreement between the prediction accuracies of the classifiers trained on a bigger training set in one platform and used to predict using the bigger test data on the other platform (Pearson's r =0.99). Remarkably, the ensemble classifier was able to provide 100% accurate predictions for both cases, regardless of the additional complexity caused by 8 varieties of classes. In this analysis, the component classifier PLS+LDA also performed similarly to the ensemble classifier in both cases yielding 100% accurate class predictions. Clearly, between the two types of dimension reduction methods, PLS performs better

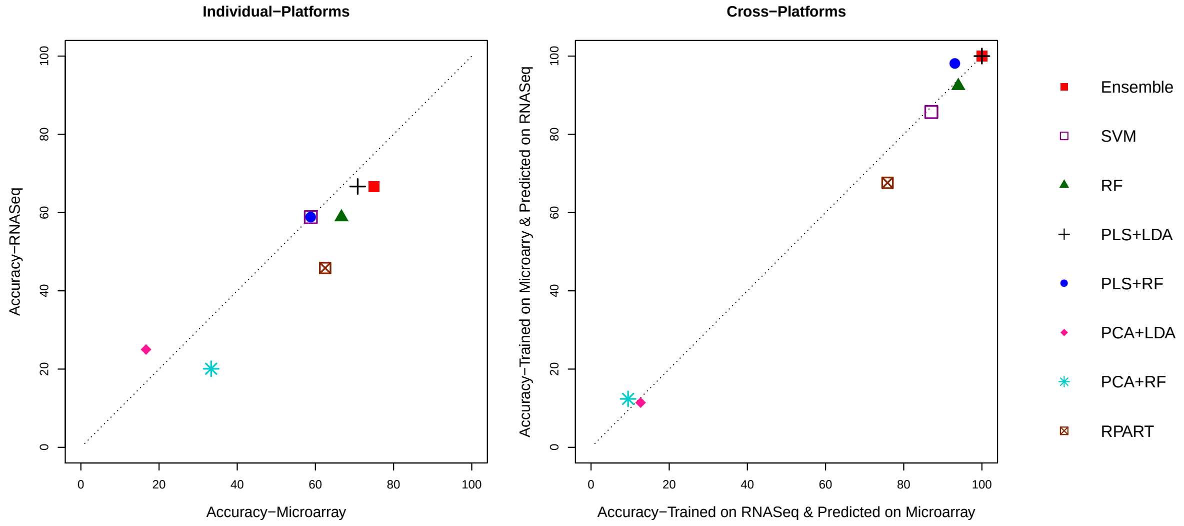

Figure 1: Plots between predication accuracies of RNAseq vs microarray test sets, by eight different classification techniques, for classifiers trained and predicted on individual platforms and cross platforms.

than PCA throughout this study. The performances of classifiers integrated with PCA are clearly the weakest among all individual classifiers in each scenario.

From the third analysis, we observed that five of ten most important genes for classification (Cyp1a1, Fam111a, Ugt2b, Akr1b8, and Hbb) were common between the two platforms. From literature search we found that CYP1A1 encodes a member of the cytochrome P450 superfamily of enzymes which catalyze many reactions involved in drug metabolism. Likewise, Ugt2b belongs to a large family of proteins capable of detoxifying a wide variety of both endogenous and exogenous substrates such as biogenic amines, steroids, bile acids, phenolic compounds and various other pharmacologically relevant compounds, including numerous carcinogens, toxic environmental pollutants and prescription drugs. Mutations in Hbb have been implicated in a number of blood disorders.

5. Discussion

In this study, we developed an ensemble classifier built on a set of standard classifiers to predict MOAs in Rat liver experiment data profiled by microarray and RNA-seq. The newly constructed ensemble classifier performed reasonably well in both platforms separately; we observed comparable overall predictability of MOAs in both test sets with 75% and 67% accuracies for microarray and RNA-seq, respectively. In an earlier classification approach applied on the same data, Wang et al. (2014) reported averaged overall accuracies of 58% and 61% for microarray and RNA-seq, suggesting a slightly better predictability in RNA-seq. However outcomes of these two studies are somewhat incomparable due to the differences in test data sets used. For example, we omit two unknown classes present in original test sets after including controls as a separate class, whereas in their analysis, two unknowns were considered as another class while discarding controls. Interestingly, once we trained classifiers to make predictions on cross platforms, the ensemble classifier provided 100% accurate predictions for all 8 classes presented in the whole experiment. This result exhibits a perfect cross platform concordance in view of classification. Clearly, throughout the whole analysis, none of the individual classifiers outperformed the ensemble classifier with respect to the overall accuracy. However, PLS+LDA performs equally well in many cases. We observe widely different classification performances among standard classifiers, which reflects the unreliability of restricting to a single classifier in case of high dimensional classification problems. On the other hand, this also proves the utility of the adaptive ensemble classifier which is expected to perform as good or better than the individual classifiers.

References

1. Boulesteix, A. (2004), "PLS dimension reduction for classification", *Statistical Applications in Genetics and Molecular Biology with Microarray Data*, Vol 3(1), pp. 1-30.

2. Dai, J.J., Lieu L., Rocke, D., (2006), "Dimension reduction for classification with gene expression microarray data", *Statistical Applications in Genetics and Molecular Biology with Microarray Data*, Vol 5(1), 1544-6115.

3. Datta, S., Datta, S., Pihur, V., (2010) "An adaptive optimal sensemble classifier via bagging and rank aggregation with application to high dimensional data", *BMC Bioinformatics*, 11:427.

4. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., Speed, T.P., (2003), "Summaries of Affymetrix GeneChip probe level data", *Nucleic Acids Research*, Vol 31, No. 4 e15.

5. Wang, C., Gong, B., Bushel, P.R., et al. (2014), "The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance", *Nature Biotechnology* 32, pp. 926-932.