Mining Significant Gene to Metabolite Correlations in NCI-60 Dataset

Gang Su[1], Chris Beecher[2], Manhong Dai[3], Brian Athey[1,3], Fan Meng[1,3]

[1]University of Michigan, National Center for Integrative Biomedical Informatics

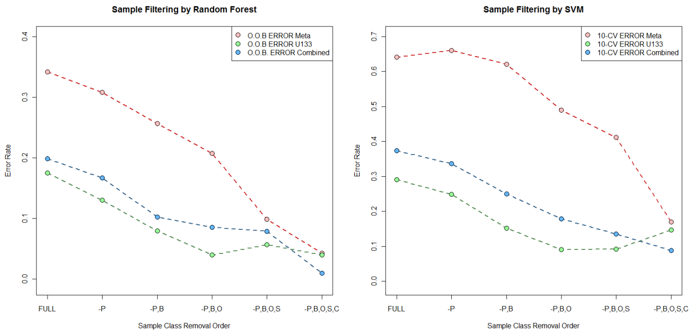
[2] Michigan Center for Translational Pathology

[3] Molecular & Behavioral Neuroscience Institute and Psychiatry Department

Background: Rapid development of Metabolomics profiling has generated high quality datasets which capture various downstream biological states, processes and fluctuations of cell metabolism. The study of Metabolomics not only opened new frontiers in molecular interaction and pathway analysis, but also promise enormous potential of revealing novel biological inferences when integrating with other high dimensional 'omics' data, such as Proteomics and Transcriptomics. The NCI-60 data, commissioned by NCI and publicly available online, provides a suite of comprehensive measurement on mRNA, metabolite, protein and epigenetics over 9 different classes of cancer cell lines. This dataset is ideal material for mining underlying relationships between 'omics' datasets.

Data Screening and Sample Classification: We performed two-way data screening to produce a confident Metabolomics dataset for subsequent correlation analysis. The first way is probe-wise screening: when cross validating NCI-60 Metabolomics dataset with a reference copy, we discovered that this dataset contains approximately 30% imputed missing values. Also this dataset contains more outliers than the corresponding microarray data. To create a subset of 'cleaner' data, we first removed the metabolite profiles with less than 20 valid values, then filtered out profiles which had Pearson Correlation Coefficient less than 0.7 between the NCI-60 data and our reference data. This yields 220 metabolite profiles. The second way is sample-wise screening: The initial classification error using all the 220 metabolite profile over 9 cancer classes is ~0.5, which is unacceptable. Some samples may contain higher than average noise level which will consequently deteriorate the follow-up correlation analysis. We evaluated the quality of each cancer class by comparing step-wise classification errors from Metabolomics and Gene microarray data. The best classifiers are selected at each step using a backwards elimination method based on random forest. At each step, the cancer class contributes most to the classification error was removed. Both the O.O.B. error rate from Random Forest and 10-fold cross validation error rate from Support Vector Machine (SVM) were investigated. The process was repeated 20 times and the mean error rate at each step is plotted in figure 1. Cancer cell line classes: Breast(B), CNS(S), Colon(C), Leukemia(L), Melanoma(M), Non-Small(N), Cell(C), Lung(L) Ovarian(O), Prostate(P), Renal(R).

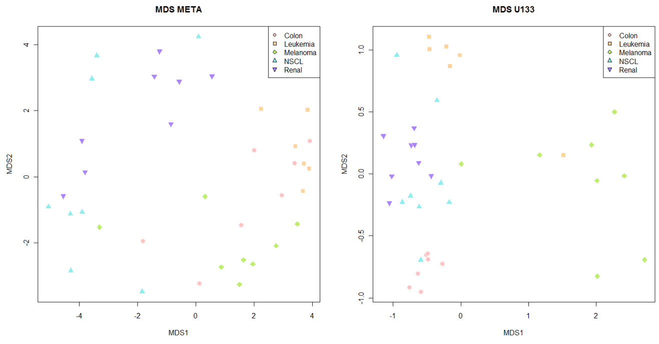
Figure 1 Classification Errors



From Figure 1.a we can see that classification error rate decreases while we remove the most mis-classified samples, which is expected. It should be noted that when removed five most misclassified cancer classes, for both Random Forest and Support Vector Machine, the combined best classifiers from gene expression and metabolite profile outperform either one alone. This do suggests that Metabolomics data capture significant amount of information from the cell despite the fact that they are more noisy and heterogeneous; after cleaning and processing, Metabolomics classifiers can achieve comparable classification performance with microarray. We chose five cancer classes (removed P,B,O,S) for correlation analysis, where the O.O.D error of Random Forest is around 0.1 and the largest classification error for cancer class is smaller than 0.3.

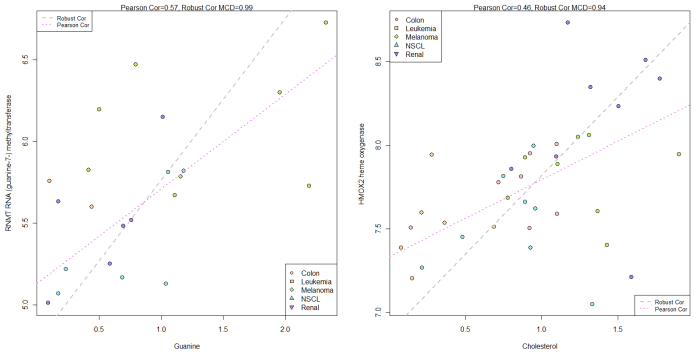
Sample Feature Comparison: We compared the structures of classification by super-imposing the 2D projection of Euclidian distance calculated on the best classifiers from Metabolomics and Microarray data respectively , transformed by 2-D Multi-dimensional scaling, as shown in figure 2. Notably, both the 2-D projections over five cancer classes have reasonable separation, and visually one profile can be rotated and approximately super-imposed on the other. The significance of such a match using Procrustes analysis assessed by Protest procedure over 1000 permutations give a p-value of 0.01. This suggests not only Metabolomics data can achieve equivalent classification performance as microarray, but also the cryptic structural information in both datasets resemble each other. This provides a theoretical basis for correlation analysis between Metabolomics and Microarray profiles.

Figure 2: MDS of Cell line distances



Robust Estimate of Correlations: Because of 1. Gene-metabolite correlation is worse than Gene-Gene correlation 2. Out Metabolomics dataset is not complete 3. Outliers, the classical Pearson, Spearman and Kendal correlation fail. We applied robust pair-wise correlation estimates, such as FMCD and pairwiseQK to tackle the 'contamination' of outliers in the data. In addition, we computed first-order robust partial correlation by recursive formula on all the Gene to Metabolite/(Gene/Metabolite) combinations to improve estimation of significant direct Gene to Metabolite correlations. We also performed Liquid-Association on all the Gene-Gene/Metabolite combinations to search for metabolite expression fluctuations that may significantly alter Gene to Gene expression correlations. In each case, the p-value is estimated by 5000 permutations. All the permutations were done on Depression Center Cluster. Figure 3 shows an example of robust correlation estimates. The left demonstrate s the correlation of Guanine to RNMT, which is ranked top on MCD robust estimate of correlations. The right one demonstrates cholesterol v.s. HMOX2, a gene documented to have involvement in cholesterol biosynthesis. Further analysis are proposed to combine literature mining with the short lists of significant Gene to Metabolite correlations for validation and inferences.

Figure 3: Robust estimation of Correlation examples



Summary and Conclusions: Based on our preliminary analysis, we produced a shortlist of best Metabolomics classifiers which can achieve equivalent performance to those from microarray, and the combined classifiers outperform either set of classifiers alone under some conditions. Furthermore, by implementing robust correlation estimates, we successfully circumvented the issue of missing values, noise and outliers and identified significant correlations which can be validated from literature. The direct relationships from partial correlation and subtle controlling relationships from liquid association can also reveal multi-fold of hidden knowledge in addition to direct correlation estimation. It would also be interesting to integrate proteomics and epi-genetics data to construct a association network structure to help us better understand the panoramic picture underlying 'omics'.

Acknowledgements: This work was supported by NIH Grant #U54 DA021519