Assessing Histological Grades of Primary Breast Cancer Using Gene Signatures

Priyanka Arora, Lee Carraher,Ben Landis, Li Jing, Samuel Schmidt

December 10, 2013

Abstract

This paper is a re-Analysis of the Sotiriou et. al. paper on using tumor histological grade expressed gene signatures for breast cancer prognosis [1]. The principle assumptions of this process are that the classification of histological grade and tumor progression has a strong correlation with the length of survival time. Furthermore, we suggest grade 2 tumors are misclassified grade 1 or grade 3 tumors. The result of developing better gene signatures for classifying tumor grades can further assist in diagnosis and treatment options for breast cancer patients.

1 Background

2 Purpose and Hypothesis

3 Methods and Materials

3.1 Data Acquisition and Sampling

Five gene expression datasets were used in this study which were obtained by microarray analysis (using Affymetrix U133A Genechips) of tumor samples from 661 patients with primary breast cancer.
They are enumerated below:

- The training set KJX64.
- The validation set KJ125.
- The National Cancer Institute (NCI) dataset from Sotiriou et al. [2]
- The Stanford/Norway (STNO) dataset from Sorlie et al. [3]
- The Nederlands Kanker Instituut (NKI) 2 dataset from Van de Vijver et al. [4]

Histologic tumor grade in these datasets was based on the Elston – Ellis grading system [5] and the standardized mean difference of Hedges and Olkin [6] was used to rank genes by their differential expression. The probe sets of the Affymetrix U133A Genechips were mapped to other microarray platforms by matching the Unigene identifiers (version 180), following the technique mentioned in Praz et al. [7] For gene expression analysis, RNA was isolated by utilization of TRIzol, which is a monophasic solution of phenol and guanidinium isothiocyanate, used for solubilizing biological material and denaturation of proteins. Agilent Bioanalyzer was used for optimization and quality control of the RNA obtained from all those tumor samples and only the ones with good quality of RNA were acknowledged for further examination. For the purpose of our analysis, while selecting the grade-associated genes, we only considered the ER-positive tumors and excluded the ones with ER-negative and NA status on account of the relationship between the ER status and histologic grade. As mentioned in the Sotiriou et al. paper [1], majority of the ER-negative tumors are grouped as either intermediate (Grade 2) or high (Grade 3) histologic grade, assuming that we had utilized all histologic grade 1 and 3 tumors despite their ER status during our analysis, we might have ended up choosing ER-identified genes that were falsely connected with grade.

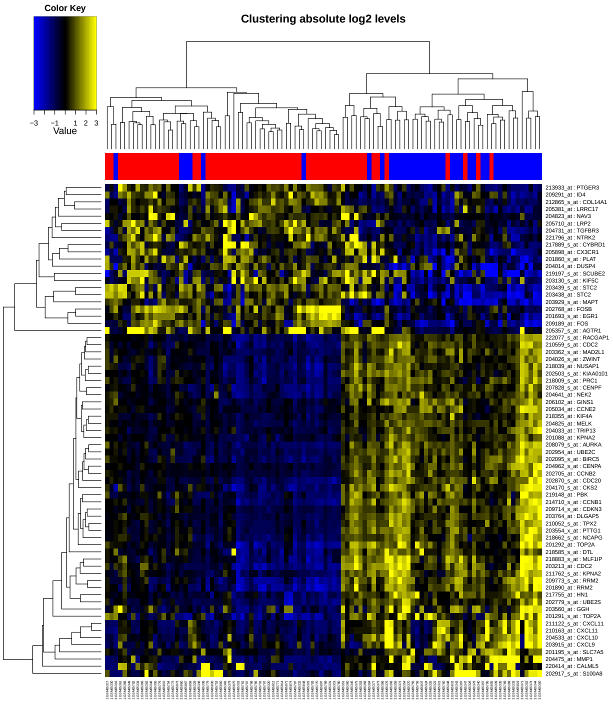

Figure 1: Grade 1 and Grade 3 Differentially Expressed

$$Centroid(Grade_X) = \left[\sum_{x \in X} \frac{x_1}{|X|}, \sum_{x \in X} \frac{x_2}{|X|}, \ldots, \sum_{x \in X} \frac{x_{|genesignature|}}{|X|} \right]$$

Figure 2: Grade Gene Expression Centroid Generation

3.2 Differential Expression Modeling Methods

Our method of assessing histological grade of primary breast cancer consists of three phases. The first phase is a semi-supervised ranking of genes and grade 1 and grade 3 tumors based on maximum differential expression. The second phase consists of selecting a set of genes that are up-regulated with grade 3 tumors and down regulated in grade 3. In the third phase, we generated two groups of tumor samples. Classification of the grade 2 tumors is then performed using these groups as models. Below we expand upon our process in detail.

The gene signature identification process consisted of assessing the Sotiriou breast cancer data set with Genomics Portals (genomicsportals.org). Genomics portals performs differential analysis on genes and samples using the CLEAN algorithm [8]. CLEAN co-clusters the genes and sample grades hierarchically, using data from the dataset and a database of known gene functional groups. Highly differentially expressed genes were identified with a p-value of .001, and fold change over expression level of 2. The results of this analysis were inspected in genomics portal's Treeview navigator [9], and the visually most cohesive up-regulated set of genes corresponding to grade 3 tumors was selected as our gene signature1. Inspection with Treeview failed to suggested a visually significant cluster of down-regulated genes for grade 3 tumors, therefore only up-regulated ones were included in our model. We acquired a set of 49 differentially expressed up-regulated grade 3 tumor genes from this analysis.

Using the selected set of 49 genes as a signature we supposed a |gene signature|−dimensional subspace to build our model. Withing the subspace, we used the clinically provided grade labels to generate centroids for the two , 1 and 3 - grade, groups following the expression (Figure 2) . Noise suppression refinement was performed by removing samples that deviated from the group cluster centers by more than 2 standard deviations of the cluster's mean variance.

For visual assessment of our analysis, we queried the grade 2 tumors against the grade 1 and 3

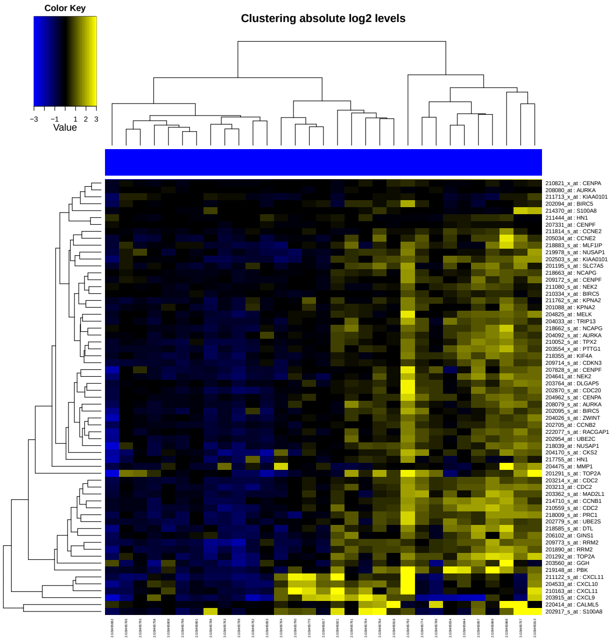

Figure 3: Grade 2 Differentially Expressed Genes from Grade 1 and Grade 3

$$Grade(X) = \underset{C \in Centroids}{\text{argmin}} \left(\sqrt{\sum_{g \in genelist} (X_g - C_g)^2} \right)$$

Figure 4: Nearest Centroid Classification

differentially expressed genes. The resulting chart shows a distinct clustering between the grade 1-like and grade 3-like tumors that follow approximately the grade 3 and grade 1 distributions. Assuming the dataset comprises the actual distribution of grade 1 to grade 3 tumors, approximately 55:64 for grade 1 and grade 3 respectively (figure 3) we reassert our hypothesis of the lack of a grade 2 class distinction.

Classification of grade 2 tumors was then computed by nearest centroid under the euclidean norm following the method in figure 4. The new classifications were then used to recompute the KM-Survival analysis on tumor grades 2A (or grade 1-like) and 2b (or grade 3-like).

3.3 Other Unsupervised Learning Methods

In addition to the use of genomics portals to identify differentially expressed genes, two common unsupervised machine learning techniques were used to classify histological grades based on genetic expression. The first method used was the K-Means method. We will forgo an exhaustive explanation of K-Means here as it is a long known method with many descriptions and analysis available online. K-Means is an unsupervised clustering method that attempts to minimize the inter-cluster distance on a set of N points and K cluster assignments. The method is greedy and iterative, with few guarantees on optimality, however combined with random starts, the overall performance of K-Means is actually quite good for spherical clusters [10]. For our testing, following our aforementioned hypothesis of the existence of only two genetically identifiable tumor grades (low and high), we performed 100 tests using K-Means on 50:50 training:test data comprising grade 1 and grade 3 tumors. The reclassification accuracy results are given in (3.3).

- Mean 0.8382
- Mode 0.84
- Min 0.58

- Max 0.94
- Var 0.00315

From the results it is clear there is significant error in the upper and lower bounds for accuracy. This though partially due to the K-Means algorithm itself, may have resulted from a badly conditioned training and test dataset split. As 50% grade 1 and 50% grade 3 would suggest the optimal split, we assert that the actual min accuracy performance is slightly lower than the real performance of K-Means, while the max is no better than could ideally be achieved from random 50:50 splitting.

The second unsupervised method, Projection to Latent Structures Regression (or Partial Least Squares Regression, *PLSR*) [11] proved to be slightly less accurate in its ability to correctly classify data. Despite its accuracy performance shortcomings, it did however perform well in its ability to identify highly differentially expressed genes.

4 Results and Discussion

5 Conclusion

90% of the time it works all the time!

References

[1] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262–272, 2006.

[2] C. Sotiriou, S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu, "Breast cancer classification and prognosis based on gene expression profiles from a population-based study," *Proceedings of the National Academy of Sciences*, vol. 100, no. 18, pp. 10393–10398, 2003.

[3] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A.-L. Børresen-Dale, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10869–10874, 2001.

[4] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *N. Engl. J. Med.*, vol. 347, pp. 1999–2009, Dec 2002.

[5] C. Elston and I. Ellis, "Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. cw elston & io ellis. histopathology 1991; 19; 403–410," *Histopathology*, vol. 41, no. 3a, pp. 151–151, 2002.

[6] L. V. Hedges, I. Olkin, M. Statistiker, I. Olkin, and I. Olkin, "Statistical methods for meta-analysis," 1985.

[7] V. Praz, V. Jagannathan, and P. Bucher, "CleanEx: a database of heterogeneous gene expression data based on a consistent gene nomenclature," *Nucleic Acids Res.*, vol. 32, pp. D542–547, Jan 2004.

[8] J. Freudenberg, V. Joshi, Z. Hu, and M. Medvedovic, "Clean: Clustering enrichment analysis," *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–15, 2009.

[9] K. Shinde, M. Phatak, F. Johannes, J. Chen, Q. Li, J. Vineet, Z. Hu, K. Ghosh, J. Meller, and M. Medvedovic, “Genomics portals: integrative web-platform for mining genomics data,” *BMC Genomics*, vol. 11, no. 1, pp. 1–10, 2010.

[10] A. Jain, “Data clustering: 50 years beyond k-means,” in *Machine Learning and Knowledge Discovery in Databases* (W. Daelemans, B. Goethals, and K. Morik, eds.), vol. 5211 of *Lecture Notes in Computer Science*, pp. 3–4, Springer Berlin Heidelberg, 2008.

[11] S. Wold, M. Sjöström, and L. Eriksson, “Pls-regression: a basic tool of chemometrics,” *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109 – 130, 2001. ¡ce:title¿PLS Methods¡/ce:title¿.