

Statistical Methods

Tumor proteomic patterns predict classification and tumor behavior in human non-small cell lung cancer

Yu Shyr

yu.shyr@vanderbilt.edu

The statistical analyses for the proteomic data were focused on the following steps: (1) selecting the important proteins that were differentially expressed among the histological groups, (2) using the class prediction model based upon the Weighted Flexible Compound Covariate Method (WFCCM)^{1,2,3} to verify if the proteins selected in step one have the statistically significant prediction power on the training samples, (3) applying the prediction model generated from step two to a set of blinded samples for examining the prediction power on the blinded samples, and (4) employing the agglomerative hierarchical clustering algorithm⁴ to investigate the pattern among the statistically significant discriminator proteins as well as the biologic status.

The selection of important proteins was based on Kruskal-Wallis test, Fisher's exact test (dichotomize the expression level as present or not), t-test, Significance Analysis of Microarrays (SAM)⁵, Weighted Gene Analysis (WGA)⁶ and the modified info score method⁷, the cutoff points for each method were $p < 0.0001$, $p < 0.0001$, $p < 0.0001$, 3.5, 2 and 0 respectively. The cutoff points were determined based on the significance as well as the prediction power of each method. The protein was on the final list if it met at least three of these six selection criteria.

The WFCCM was employed in the class-prediction model based on the selected proteins. This method was designed to combine the most significant proteins associated

with the biologic status from each analysis method, e.g., Kruskal-Wallis test, Fisher's exact test, permutation t-test, SAM, WGA, and modified Info Score. In other words, the WFCCM is an extension of the compound covariate method which allows considering more than one statistical analysis method into the compound covariate, and it reduces the dimensionality of the problem using a new covariate obtained as a weighted sum of the most important predictors. The WFCCM for tumor sample i is defined as $WFCCM(i) = \sum_j [\sum_k (ST_{jk})] [W_j] x_{ij}$, where x_{ij} is the log-ratio measured in tissue sample i for protein j . ST_{jk} is the standardized statistic, e.g., t-statistic, for statistical analysis method k . W_j is the weight of protein j which is defined as $W_j = [(\sum_k I_{jk} / K) (1 - \text{Info Score}_j)]$, where $I_{jk} = 1$, if the protein j is statistically significant in method k ; where $I_{jk} = 0$, if the protein j is not statistically significant in method k .

The class-prediction model was applied to determine whether the patterns of protein expression could be used to classify tissue samples into two classes according to the chosen parameter, e.g., normal tissue vs. tumor tissue. We estimated the misclassification rate using leave-one-out cross-validated class prediction method based on the WFCCM. This leave-one-out cross-validated method was processed in four steps. First, WFCCM was applied to calculate the single compound covariate for each tissue sample based on the significant proteins. Second, one tissue sample was selected and removed from the data set, and the distance between two tissue classes for the remaining tissue samples was calculated. Third, the removed tissue sample was classified based on the closeness of the distance of two tissue classes. Fourth, step 2 and 3 were repeated for each tissue sample. To determine whether the accuracy for predicting membership of tissue samples into the given classes (as measured by the number of correct

classifications) was better than the accuracy that could be attained for predicting membership into random grouping of the tissue samples, we created 5,000 random data sets by permuting class labels among the tissue samples. The cross-validated class prediction was performed on the resulting data sets and the percentage of permutations that resulted in as few or fewer misclassifications as for the original labeling of samples was reported. If less than 0.05 of the permutations resulted in as few or fewer misclassifications, the accuracy of prediction into the given classes was considered significant.

The prediction of the blinded samples was completed using the method described above. The blinded sample was classified based on the closeness of the distance of two tissue classes which was determined using the WFCCM.

The agglomerative hierarchical clustering algorithm was applied to investigate the pattern among the statistically significant discriminator proteins as well as the biologic status using M. Eisen's software⁸.

The similar analyses described above were applied to the survival data. The significant proteins were selected based on Log-rank test, Wilcoxon test, Likelihood Ratio test (viewed the expression level as present or not), and Cox proportional hazard model (viewed the expression level as a continuous variable). The cutoff points were $p < 0.0001$ for all these tests. The WFCCM method was then employed to calculate the summary score for each patient. Finally, a sensitivity analysis based on the concept of the *receiver operating characteristic (ROC)* curve method was applied to distinguish the survival patterns. The focus of the sensitive analysis was to maximize the odds ratios of the median

survivals and to minimize the degree of the imbalance of the sample sizes for each possible partition of the summary score.

Although the perfect or near perfect statistical models are reported in this study, there are some statistical limitations need be addressed. First, since the study sample size is small, a larger scale study to confirm our findings is necessary, and the number of the peaks reported in this paper was not based on the smallest number of the peaks that could discriminate the classes but based on the statistical evidence. The possibility of achieving similar misclassification rates based on different subsets of the peaks certainly exists. In addition, the near perfect discrimination obtained using the agglomerative hierarchical clustering is not surprising as it uses covariates that were themselves chosen to have maximal discriminating power.

Reference:

- 1 Tukey JW. Tightening the clinical trial. *Control Clin Trials* 1993; **14**(4): 266-285.
- 2 Shyr Y. Analysis and Interpretation of Array Data in Human Lung Cancer Using Statistical Class-Prediction Model. AACR annual meeting, San Francisco, CA. *Program/Proceeding Supplement* 2002; 41-42.
- 3 Shyr Y and Kim KM. Weighted Flexible Compound Covariate Method for Classifying Microarray Data. In: *A Practical Approach to Microarray Data Analysis* (Berrar, D., ed.), Kluwer Academic Publishers, Norwell, MA, USA. 2003; 186- 200.
- 4 Everitt BS. *Cluster Analysis*, Edward Arnold, New York. 1993.
- 5 Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci U S A* 2001; **98**(9): 5116-5121.
- 6 Hedenfalk I., Duggan D., Chen Y., Radmacher M., Bittner M., Simon R., Meltzer P., Gusterson B., Esteller M., Kallioniemi O.P., Wilfond B., Borg A., Trent J. Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 2001; **344**(8): 539-548.
- 7 Ben-Dor A, Friedman N, Yakhini Z. Scoring Genes for Relevance. Tech Report AGL-2000-13, Agilent Labs, Agilent Technologies. 2000.

8 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998; **95**(25): 14863-14868.