**Supplementary Materials**

**Genome-wide gene expression pattern analysis for predicting adenocarcinoma progression[[1]](#footnote-1)\***

**Fei Zhang 1, Shi-xiang Wang 1,** **Xing Chao2, Kai Song1\***

1 School of Chemical Engineering and Technology, Tianjin University, 300350 Tianjin, P.R. China

2 Department of Clinical Science, University of Texas Southwestern Medical Center, 75390, Dallas, Texas, USA

**Significance Analysis of Microarrays (SAM)**

SAM is a statistical technique for identifying whether changes in gene expression are statistically significant ([1](#_ENREF_1)). SAM identifies statistically significant genes by computing a statistic for each gene 𝑗. The score is used to measure the strength of the relationship between X-variable and a response variable (class membership). Genes with scores greater than a threshold are assumed to be significantly related to class membership, and the threshold can be adjusted to identify smaller or larger sets of genes. The proportion of genes incorrectly identified as significant is determined by the false discovery rate (FDR) ([2](#_ENREF_2), [3](#_ENREF_3)).

**Partial Least Squares (PLS)**

PLS developed by Wold et al. is a fast supervised dimension reduction method which can handle numerous of predictor variables ([4](#_ENREF_4)). PLS orthogonal components are constructed to maximize linear combinations of original predictor variables *X* correlate with the response values *Y* while accounting for as much variance in predictors as possible ([5](#_ENREF_5), [6](#_ENREF_6)). In the current study, we can suppose that variables *X* is amatrix of *n* samples and *p* genes (the raw data should be centralized to zero mean), and response values *Y* is a matrix (in the triple class classification model, q=3), and it indicates the classification of stage I, stage II and stage III. The linear relationship between the predictor matrixand the response values is shown below:

 (1)

whereis the regression coefficient matrix andis the residual matrix. PLS regression is on the basis of basic principal component decomposition:

 (2)

 (3)

whereis the latent variables matrix of *X*,  is the latent variables matrix of *Y*,andare matrices of coefficients,andare matrices of random errors,is the number of latent variables. The relationship between *U* and *T* is:

 (4)

whereis the *i*th latent variables matrix of *X*,  is the *i*th latent variables matrix of *Y*, *bi* is the coefficient determined by the minimum deviation *ri*.

Finally, the matrix *B* regression coefficients matrix is:

(5)



whereis a matrix of weights.

So  (the number of latent variables) is the only parameter in PLS which need to be decided, with the latent variables increase, the information of original data preserved is increasing, until reaching the maximal value, which is the rank of, all the information of original data is contained in latent variables.

**References**

1. Damle MT & Kshirsagar M (2012) Role of Permutations in Significance Analysis of Microarray and Clustering of Significant Microarray Gene list. *International Journal of Computer Science Issues(IJCSI)* 9(2).

2. Tusher VG, Tibshirani R, & Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response (vol 98, pg 5116, 2001). *Proceedings of the National Academy of Sciences of the United States of America* 98(18):10515-10515.

3. Benjamini Y (2010) Discovering the false discovery rate. *J R Stat Soc B* 72:405-416.

4. Boulesteix AL (2004) PLS dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology* 3:Article33.

5. Tan YX, Shi LB, Tong WD, Hwang GTG, & Wang C (2004) Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Comput Biol Chem* 28(3):235-244.

6. Nguyen DV & Rocke DM (2002) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* 18(9):1216-1226.

1. \*Address: School of Chemical Engineering and Technology, Tianjin University, Tianjin, China, 300350

   This work was supported by the National Natural Science Foundation of China (No. 31271351)

   Correspondence to: SONG Kai, Email: ksong@tju.edu.cn [↑](#footnote-ref-1)