Huang *et al. BMC Bioinformatics* 2010, **11**(Suppl 4):P23
http://www.biomedcentral.com/1471-2105/11/S4/P23

BMC Bioinformatics

POSTER PRESENTATION Open Access

Application of Pearson correlation coefficient (PCC) and Kolmogorov-Smirnov distance (KSD) metrics to identify disease-specific biomarker genes

Hung-Chung Huang[1,2], Siyuan Zheng[1,2,3], Zhongming Zhao[1,2,3*]

From UT-ORNL-KBRIN Bioinformatics Summit 2010
Cadiz, KY, USA. 19-21 March 2010

Background

DNA microarrays have been widely applied in cancer research for better diagnosis and prediction of the disease states. Traditionally, most microarray studies aim to identify differentially expressed genes (DEGs) by comparing the average gene expression levels between two groups (e.g., the treated vs. control or disease vs. non-disease) based on statistical analysis such as t-test and Significance Analysis of Microarrays (SAM) [1,2].

Materials and methods

In this study, we defined the gene expression profile (GEP) of a gene as the distribution of the $\log_2$ values of its normalized expression signal intensities across the samples in the similarly studied microarrays. We hypothesized that the biomarker genes that distinguish disease samples from normal samples might form distinct GEPs between comparison groups. We applied Pearson Correlation Coefficient (PCC) and Kolmogorov-Smirnov Distance (KSD) metrics to identify disease-specific biomarkers by comparing GEPs between normal and disease states and then applied this technology to disease (e.g., cancer) related studies in order to discover some disease genes as biomarker candidates. These biomarkers' gene profiles in normal and disease samples might be used to diagnose or monitor patient's disease state via regular gene expression analysis.

Table 1 Top 10 gene pairs for top prediction accuracies on PCA diagnosis.

Down gene	Up gene	True positive	True negative	Accuracy
PCC sort*				
ACTA1	*CRISP3*	67/90	73/81	140/171
TGFB3	*BICD1*	72/90	68/81	140/171
ACTA1	*HPN*	76/90	63/81	139/171
MYL9	*CRISP3*	64/90	75/81	139/171
AL044599	*BICD1*	75/90	64/81	139/171
DMN	*CRISP3*	65/90	73/81	138/171
GJA1	*CRISP3*	70/90	68/81	138/171
AL036744	*CRISP3*	65/90	73/81	138/171
DMN	*BICD1*	69/90	69/81	138/171
ADH5	*BICD1*	71/90	67/81	138/171
KSD sort**				
GSTP1	*CRISP3*	68/90	72/81	140/171
AOC3	*CRISP3*	69/90	70/81	139/171
GSTP1	*UBE2C*	66/90	73/81	139/171
HLA-E	*RGS10*	71/90	68/81	139/171
GSTP1	*HPN*	70/90	68/81	138/171
DMN	*CRISP3*	65/90	73/81	138/171
GJA1	*CRISP3*	70/90	68/81	138/171
HLA-E	*UBE2C*	61/90	77/81	138/171
DMN	*BICD1*	69/90	69/81	138/171
PALLD	*BICD1*	66/90	72/81	138/171

[*] PCC sort: significant genes were separated into down- and up- regulated groups, then the top 20 genes (sorted by Pearson Correlation Coefficient in the cancer vs. normal GEPs for each gene) in each group were selected to generate pair-wise gene-pairs for the PCA prediction.

[**] KSD sort: significant genes were separated into down- and up- regulated groups, then the top 20 genes (sorted by Kolmogorov-Smirnov Distance in the cancer vs. normal GEPs for each gene) in each group were selected to generate pair-wise gene-pairs for the PCA prediction.

* Correspondence: zhongming.zhao@vanderbilt.edu
[1] Functional Genomics Shared Resource, Vanderbilt University Medical Center, Nashville, TN 37232, USA

Results and conclusion

We applied the PCC and KSD metrics to three prostate cancer related microarray datasets. They were generated from the same study and were available in the GEO database (a total of 81 normal samples and 90 prostate cancer samples) [3]. Using the cutoff values KSD > 0.4 and PCC < 0.7, we found 230 biomarker candidate genes. Our Gene Ontology (GO) analysis found that the top ranked biomarker candidate genes for prostate cancer were highly enriched in molecular functions such as "cytoskeletal protein binding" category. We used the top two ranked genes (*ACTA1*, encoding an actin subunit, and *HPN*, encoding hepsin) to demonstrate that prostate cancer might be diagnosed and monitored by marker genes. Furthermore, we picked top 20 significantly up-regulated and top 20 down-regulated genes based on PCC and KSD sorting. We found gene pairs comprising one up-regulated and another down-regulated had always best prediction performance (Table 1). Our study provided a promising tool to identify the potential biomarker genes for disease diagnosis and prognosis.

Author details

[1]Functional Genomics Shared Resource, Vanderbilt University Medical Center, Nashville, TN 37232, USA. [2]Bioinformatics Resource Center, Vanderbilt University Medical Center, Nashville, TN 37203, USA. [3]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA.

Published: 23 July 2010

References

1. Jafari P, Azuaje F: An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Med Inform Decis Mak* 2006, **6**:27.
2. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001, **98**:5116-5121.
3. Chandran UR, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, Liang W, Michalopoulos G, Becich M, Monzon FA: Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer* 2007, **7**:64.

doi:10.1186/1471-2105-11-S4-P23
Cite this article as: Huang *et al.*: Application of Pearson correlation coefficient (PCC) and Kolmogorov-Smirnov distance (KSD) metrics to identify disease-specific biomarker genes. *BMC Bioinformatics* 2010 **11** (Suppl 4):P23.