Pathway-based Expression Profile for Breast Cancer Diagnoses

C. Cava, G. Bertoli, I. Castiglioni

***Abstract*— Microarray experiments have made possible to identify breast cancer marker gene signatures. However, gene expression-based signatures present limitations because they do not consider metabolic role of the genes and are affected by genetic heterogeneity across patient cohorts. Considering the activity of entire pathways rather than the expression levels of individual genes can be a way to exceed these limits. We evaluated and compared five methods of pathway-level aggregation of gene expression data. Our results confirmed the important role of pathway expression profile in breast cancer diagnostic classification (accuracy >90%). However, although assessed on a limited number of samples and datasets, this study shows that using dissimilarity representation among patients does not improve the classification of pathway-based expression profiles.**

I. Introduction

In the recent years, microarray gene expression experiments identified an increasing number of disease markers [1-3]. In Breast Cancer (BC) different gene signatures have been identified [4-8] but their reproducibility, and overlap is poor.

These limits can be explained by genetic heterogeneity across patients and by the fact that changes in expression of the few genes governing cancer development control several downstream effectors: these effectors are mainly found in different gene signatures [9] and could be involved in different pathways. Thus, the identified gene signature can only partially represent the genes involved in the process.

A way to overcome these limits is to focus on groups of genes that fall within common pathways, instead of individual genes. In turn, each sub-pathway is a distinct functional part within a larger interaction network. Thus, sub-pathways could be considered instead of full pathways in searching for key classification markers.

Two types of methods were used to provide pathway analysis: i) differentially expressed genes are identified and then their principal pathways are selected [e.g. 10]; ii) each pathway is examined to find pathways with differentially expressed common genes [e.g. 11]. We hypothesized that the pathway-based method has several advantages over the expression analysis of individual genes: i) the resulting sub-pathway provide models of the molecular mechanisms underlying cancer; ii) the results give an easy interpretation on the function of genes sets for biologist iii) a pathway expression profile may be consistent across the samples, while expression of individual genes in a pathway may differ considerable across samples [12]. There are different pathway-level aggregation methods [e.g. 13-16]. This study presents an evaluation of 5 pathway-level aggregation methods of gene expression data for BC diagnoses.

II. Materials and Methods

A. Microarray data set

We used one public BC microarray data set from the Gene Expression Omnibus (GEO) database (GSE39004), containing 94 BC samples: 47 samples of macro-dissected tumor tissue and 47 adjacent noncancerous tissue. The dataset came from the Affymetrix Gene Chip Human Gene 1.0 ST Arrays platform.

-Normalization

Probe cell intensity data was processed by the RMA algorithm [17].

B. Gene set enrichment analysis

With the purpose to identify a group of differentially expressed genes, enriched for a particular gene set, we used biological pathway-based analysis called Gene Set Enrichment Analysis (GSEA) [11]. We focused on 403 biological pathways derived from the KEGG and BioCarta pathway database [18]. All probe sets were pre-ranked using t-test with respect to their correlation with normal and tumor tissue. Then, GSEA analyzed ranked list of genes, and an enrichment score (ES) was calculated for a given gene set, which indicated if a gene set was found differentially expressed between normal and tumor tissue.

The nominal *p*-value estimated the statistical significance of the enrichment score for a single gene set.

Several pathways, found differentially expressed, were obtained.

C. Pathway expression profiles

This study compared five pathway-level aggregation methods of gene expression data. The five methods were grouped into three categories: mean-based of all genes, mean based of top SAM genes and dissimilarity distances.

Cava, G. Bertoli and I. Castiglioni are with the Institute of Molecular Bioimaging and Physiology of the National Research Council (IBFM-CNR), Milan, Italy (corresponding author e-mail: isabella.castiglioni@ibfm.cnr.it).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

-Mean-Based of all genes

The expression profiles of all the genes, grouped in a differentially expressed pathway, were combined by taking their mean value.

- Mean based of top SAM genes

In this method mean expression of a pathway is represented by the mean expression of key genes found by Significance Analysis of Microarray (SAM) [19] for each pathway. Our aim was to select significant genes based on differential expression between normal and tumor tissue. The genes were considered up/down-regulated if their mean expression in tumor samples were significantly higher/lower (FDR, q value <0.01) than in normal samples. The genes, as found up or down regulated in expression, were identified by submitting IDs probes from the HGU133Array to Affymetrix through the Netaffxtool (www.affymetrix.com/analysis/index.affx).

In the next phase, we want to define a sub-pathway obtained differentially expressed by the GSEA algorithm.

To find a sub-pathway, we identified key genes that satisfy the following two criteria: 1) genes are included in SAM analysis, and 2) genes belong to the same differentially expressed pathway.

- Dissimilarity distances

Dissimilarity distances have been proved useful in many application fields. Recent studies [20,21] used with success dissimilarity representation among patients, considering the expression of individual genes. To our knowledge, dissimilarity representation is not used in pathway-based expression profiles. Our goal is to give a *dissimilarity representation,* which can express, through a function $D(x,y)$, the dissimilarity between the mean expression levels of altered genes in a pathway for the pair of patients x and y. The following ordinary distances (from the R bioDistance package [22] were considered: i) Euclidean distance, ii) Manhattan distance, iii) Kendall's τ-distance.

D. *Validation*

To evaluate the performance of the pathway level aggregation methods we used a machine learning algorithm, trained on the identified pathway-based expression profile and tested on the ability to differentiate normal and BC tissues with respect to the pathway expression profile.

-Machine learning

A Rapid Miner (RM) *workflow* (WF) [23] was designed.

The RM workflow implemented standard Support Vector Machine (SVM) algorithm. The main issues of this workflow were characterized by the following processes:

a) SVM Parameter Optimization. We optimized the inference accuracy over a space of given SVM feasible learning parameters. The following values were used: kernel.γ - from 0 to 5, step 30; kernel.C - from 0 to 5, step 30; kernel.type ∈ {ANOVA, DOT, RADIAL}.

b) Cross Validation. The SVM was validated by a k-fold cross-validation process. We used k=5, k=10 and k=15.

The performance of the classification was obtained in terms of Accuracy and Balanced Accuracy for the following case-control study: normal and tumor samples. Cross validation of the classifier was performed for two different breast datasets: GSE39004, as previously described, and GSE10797 containing 15 breast samples: 10 samples of invasive BC and 5 normal breast tissue. This last database was used to avoid cohort specific bias.

III. RESULTS

A. *Gene set enrichment analysis*

We found 2 up-regulated gene sets (p-value <0.01): Biocarta G2-cell cycle, Kegg DNA replication, and 5 down-regulated gene sets (p-value <0.05): Kegg Adipocytokine Signaling, Kegg Fatty Acid Metabolism, Biocarta PPARA, Kegg WNT Signaling and Biocarta GPCR.

Table I shows the pathways with the number of genes presented.

TABLE I. DIFFERENTIALLY EXPRESSED PATHWAYS

UP REGULATED PATHWAY		
Pathway	N° genes	p-value
I: Biocarta G2-cell cycle	24	<0.01
II: Kegg DNA replication	36	<0.01
DOWN REGULATED PATHWAY		
III: Kegg Adipocytokine Signaling	67	<0.05
IV: Kegg Fatty Acid Metabolism	42	<0.05
V: Biocarta PPARA	58	<0.05
VI: Kegg WNT Signaling	151	<0.05
VII: Biocarta GPCR	37	<0.05

B. *Pathway-based expression profile*

We obtained a pathway-based expression profiles for the seven differentially expressed pathways using the five considered pathway-level aggregation methods.

-Mean-Based of all genes

We obtained pathway expression profiles representing the mean expression values of the 24 genes for the I pathway, 36 genes for the II pathway, 67 genes for the III pathway, 42 genes for the IV pathway, 58 genes for the V pathway, 151 genes for the VI pathway and 37 genes for the VII pathway.

- Mean based of top SAM genes

SAM analysis identified 1974 up-regulated and 1933 down -regulated genes between normal and tumor samples.

We found 104 unique key genes included in SAM analysis and belonging to the considered differentially expressed pathways. Pathways obtained by GSEA analysis contain now a reduced number of genes, as reported in table II.

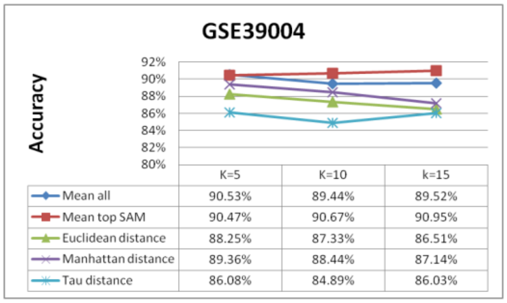
TABLE II. PATHWAY-KEY GENES

UP REGULATED PATHWAY		
Pathway	Key Genes	p-value
I: Biocarta G2-cell cycle	10	<0.01
II: Kegg DNA replication	18	<0.01
DOWN REGULATED PATHWAY		
III: Kegg Adipocytokine Signaling	18	<0.05
IV: Kegg Fatty Acid Metabolism	18	<0.05
V: Biocarta PPARA	16	<0.05
VI: Kegg WNT Signaling	29	<0.05
VII: Biocarta GPCR	8	<0.05

We obtained pathway expression profiles representing the mean expression values of these key-genes.

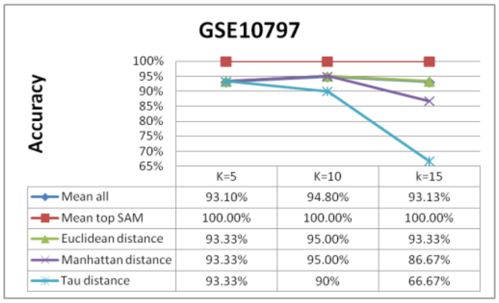
- Dissimilarity distances

The expression profiles of genes in a pathway were combined by taking their mean and for N samples in each data sets we obtained NxN dissimilarities matrix.



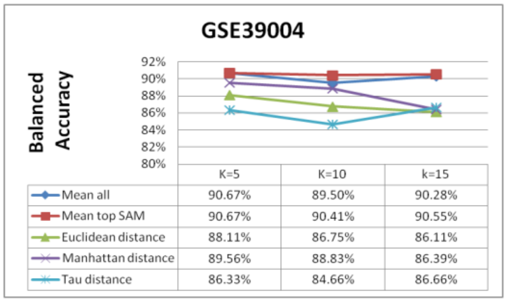
	K=5	K=10	k=15
Mean all	90.53%	89.44%	89.52%
Mean top SAM	90.47%	90.67%	90.95%
Euclidean distance	88.25%	87.33%	86.51%
Manhattan distance	89.36%	88.44%	87.14%
Tau distance	86.08%	84.89%	86.03%

Figure 1. Accuracy of each method is plotted on varying of k-cross-validation for GSE39004.



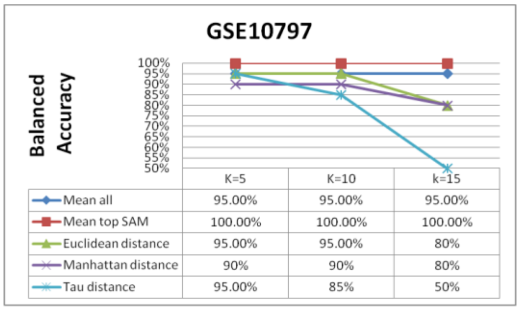
	K=5	K=10	k=15
Mean all	93.10%	94.80%	93.13%
Mean top SAM	100.00%	100.00%	100.00%
Euclidean distance	93.33%	95.00%	93.33%
Manhattan distance	93.33%	95.00%	86.67%
Tau distance	93.33%	90%	66.67%

Figure 2. Accuracy of each method is plotted on varying of k-cross-validation for GSE10797.



	K=5	K=10	k=15
Mean all	90.67%	89.50%	90.28%
Mean top SAM	90.67%	90.41%	90.55%
Euclidean distance	88.11%	86.75%	86.11%
Manhattan distance	89.56%	88.83%	86.39%
Tau distance	86.33%	84.66%	86.66%

Figure 3. Balanced accuracy of each method is plotted on varying of k-cross- validation for GSE39004.



	K=5	K=10	k=15
Mean all	95.00%	95.00%	95.00%
Mean top SAM	100.00%	100.00%	100.00%
Euclidean distance	95.00%	95.00%	80%
Manhattan distance	90%	90%	80%
Tau distance	95.00%	85%	50%

Figure 4. Balanced accuracy of each method is plotted on varying of k-cross- validation for GSE10797.

D. Validation

Results of accuracy of the SVM classification are shown in Figure 1-2, for GSE39004 and GSE10797, respectively, for the considered case-control study: normal vs tumor.

Results of balanced accuracy of the SVM classification are shown in Figure 3-4, for GSE39004 and GSE10797, respectively. The results were shown on varying of k-cross-validation.

Both mean of all genes and mean based of top SAM genes achieved good results. Mean based of top genes SAM slightly improved the performance with respect to mean of all genes. Dissimilarity distances don't seem to improve the performance of the classifier. Manhattan and Euclidean distances don't show considerable differences. Tau distances report the worst behavior.

IV. CONCLUSIONS

In this study, we evaluated and compared five methods of pathway-level aggregation of gene expression data.

The evaluation was performed with respect to accuracy and balanced accuracy. The best performances were obtained when SAM analysis was applied in a differentially expressed pathway. We demonstrated that incorporating pathway information into expression gene analysis-based BC diagnosis can provide good biological model.

Among the possible affected pathways, in the upregulated BC gene group we found cell cycle and DNA replication genes. DNA is constantly subjected to a number of surveillance mechanisms that constantly monitor its integrity, and control cell cycle progression. In the presence of DNA damage, cells activate pathways that lead to cell cycle checkpoints activation, DNA repair mechanisms, apoptosis and transcription. In cancer, these control mechanisms are altered, mainly for mutation in critical proteins. For this feature, cancer cells can stimulate their own growth, resist to apoptosis, multiply forever and stimulate angiogenesis [24].

Our pathway analysis revealed that downregulated genes belong to i) adipocytokine signaling pathway, ii) Fatty acid metabolism; iii) Peroxisome proliferator-activated receptor-alpha (PPARA) signaling pathway, iv) WNT signaling pathways; v) G-protein-coupled receptor (GPCR) signaling pathway. The relation between cancer and metabolic disorders was recognized several decades ago. In the last years, many groups have been studying systemic adipose tissue markers in cancer patients, revealing that high body mass index (BMI) values are strongly associated with increased incidence of several types of cancer and also premalignant lesions [25]. In particular, it has been suggested that adipose tissue may support tumor cell growth [26]. In fact, adipose tissue-derived factors, as adipocytokine, have been shown to influence the behavior of tumor cells, i.e. by promoting their proliferation in tridimensional structures [27]. It is quite unexpected, thus, to find in our analysis BC downregulated genes belonging to fatty acid metabolism pathway (ii), to adypocytokine signaling molecules (i), to PPARA signaling pathway (iii), whose activity increase fatty acid oxidation and decrease cytokine levels [28]. Nevertheless, it is still possible that, being PPAR pathway involved in the control of ERBB2-positive stem cells [29], the tumor tissue population analyzed in the databases contains mainly ERBB2-negative cells or non-stem cells, as demonstrated by the finding of the downregulation of WNT pathway genes (iv). Finally, GPCR family groups several proteins involved in cell proliferation control, promoting tumor cell invasion and metastasis, endothelial cell migration, and tumor angiogenesis [30]. Our analysis finds GPCR pathway genes to be down-regulated. It is possible that the samples considered in the databases are mainly non metastatic, while this pathway expression in human BC correlates with higher tumor grade and metastatic potential [31].

V. References

[1] Alizadeh, A. A., Eisen, M. B., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503-511.

[2] Golub, T. R., Slonim, D. K., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, *286*(5439), 531-537.

[3] Ramaswamy, S., Ross, et al. (2003). A molecular signature of metastasis in primary solid tumors. *Nature genetics*, *33*(1), 49-54.

[4] Van't Veer, L. J., Dai, H., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, *415*(6871), 530-536.

[5] Paik, S., Shak, S., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, *351*(27), 2817-2826.

[6] Sotiriou, C., Wirapati, P., et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, *98*(4), 262-272.

[7] Ivshina, A. V., George, J., et al. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, *66*(21), 10292-10301.

[8] Cava, C., Zoppis, I., et al. (2013, July). Combination of gene expression and genome copy number alteration has a prognostic value for breast cancer. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (pp. 608-611). IEEE.

[9] Chuang, H. Y., Lee, E., et al. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, *3*(1).

[10] Huang, D. W., Sherman, B. T., et al. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, *4*(1), 44-57.

[11] Subramanian, A., Tamayo, P., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545-15550.

[12] Hwang, S. (2012). Comparison and evaluation of pathway-level aggregation methods of gene expression data. *BMC genomics*, *13*(Suppl 7), S26.

[13] Edelman, E., Porrello, A., et al. (2006). Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics*, *22*(14), e108-e116.

[14] Guo, Z., Zhang, T., et al. (2005). Towards precise classification of cancers based on robust gene functional expression profiles. *BMC bioinformatics*, *6*(1), 58.

[15] Azuaje, F., Zheng, H., et al. (2011). Systems-based biological concordance and predictive reproducibility of gene set discovery methods in cardiovascular disease. *Journal of biomedical informatics*, *44*(4), 637-647.

[16] Lee, E., Chuang, H. Y., et al. (2008). Inferring pathway activity toward precise disease classification. *PLoS computational biology*, *4*(11), e1000217.

[17] Irizarry, R. A., Hobbs, B., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, *4*(2), 249-264.

[18] Kanehisa, M., Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *28*(1), 27-30.

[19] V. G. Tusher, et al. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, 98(9):5116{5121}, April 2001

[20] Cava, C., Zoppis, I., et al. (2013). Copy–Number Alterations for Tumor Progression Inference. In *Artificial Intelligence in Medicine* (pp. 104-109). Springer Berlin Heidelberg.

[21] Cava, C., Zoppis, I., et al. (2014). Combined analysis of chromosomal instabilities and gene expression for colon cancer progression inference. *Journal of clinical bioinformatics*, *4*(1), 2.

[22] Ding, B., Gentleman, R., et al. (2011). bioDist: Different distance measures. Bioconductor Software http://bioconductor.wustl.edu/bioc

[23] Ingo Mierswa, et al. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, KDD '06: Proc. Of the 12th ACM SIGKDD int. conf. on Know. disc. and data mining, pages 935–940, 2006.

[24] Hanahan, D., et al. (2000). The hallmarks of cancer. *cell*, *100*(1), 57-70.

[25] Renehan, A. G., et al. (2008). Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *The Lancet*, *371*(9612), 569-578.

[26] Currie, E., et al. (2013). Cellular fatty acid metabolism and cancer. *Cell metabolism*, *18*(2), 153-161.

[27] Vona-Davis, L., et al. (2007). Adipokines as endocrine, paracrine, and autocrine factors in breast cancer risk and progression. *Endocrine-related cancer*, *14*(2), 189-206.

[28] Golembesky, A. K., et al. (2008). Peroxisome proliferator-activated receptor-alpha (PPARA) genetic polymorphisms and breast cancer risk: a Long Island ancillary study. *Carcinogenesis*, *29*(10), 1944-1949.

[29] Wang, X., et al. (2013). PPARγ maintains ERBB2-positive breast cancer stem cells. *Oncogene*. 32(49):5512-21

[30] Lu, Y. Y., et al. (2013). Prometastatic GPCR CD97 Is a Direct Target of Tumor Suppressor microRNA-126. *ACS chemical biology*.

[31] Zajac, M., et al. (2011). GPR54 (KISS1R) transactivates EGFR to promote breast cancer cell invasiveness. *PloS one*, *6*(6), e21599.