# GROUP-BIOMARKERS IDENTIFICATION IN OVARIAN CARCINOMA

*[1]Alain B. Tchagang, [2]Ahmed H. Tewfik, [3]Amy P.N. Skubitz, and [4]Keith Skubitz*

Dept. of [1]Biomedical Engineering, [2]Electrical and Computer Engineering, [3]Lab. Pathology and Medicine, and [4]Medicine
University of Minnesota, 55455, Minneapolis, USA, Email: *{tcha0003, tewfik, skubi002, skubi001}@umn.edu*

## ABSTRACT

In this paper, we propose group-biomarkers as an alternative to the traditional single biomarkers used to date for the detection of ovarian cancer. Group-biomarkers are a set of genes that are used simultaneously for the diagnosis of early-stage and/or recurrent cancer. We describe a procedure for identifying such group-biomarkers from a data set of gene expression levels corresponding to normal and diseased ovarian tissue as well as tissue from other organs. The procedure starts with a list of potential single biomarkers. It then uses an order preserving biclustering step to identify other genes that are co-regulated with the candidate single biomarkers across the normal and diseased ovarian tissue and tissue from other organ. We present a statistical analysis that demonstrates that group-biomarkers have a much better detection performance than single biomarkers as exhibited by receiver operating characteristics curves.

***Index Terms—*** Biclustering, biomarkers, *DNA* microarray, ovarian cancer

## 1. INTRODUCTION

Each year in the United States, about 24,000 new cases of ovarian cancer are diagnosed and 14,000 deaths are attributed to it. Contributing to the poor prognosis is the lack of symptoms in the early stages of the disease. More than 75% of diagnoses are made in stage III and IV, after distant metastasis has occurred. The 5-year survival rate for women diagnosed with late-stage disease is 25%, compared to more than 90% for women diagnosed with stage 1 of the disease. The well-known CA-125 test is useful for tracking patients already diagnosed with ovarian cancer, but has not proven sensitive enough to be used as an early diagnosis test [1].

In recent years, large-scale gene expression analyses have been performed to identify differentially expressed genes in ovarian carcinoma. See, e.g., [1] and the references cited there in. A common goal of these studies was to identify potential tumor markers for the diagnosis of early-stage ovarian cancer, as well as to use these markers as targets for improved therapy and treatment of the disease during all stages. These earlier studies compared the gene expression profiles of tissues or cell lines derived from ovarian cancer samples, normal ovaries, other normal samples, and other types of tumors. The cumulative results of these gene expression studies reveal more than 150 potentially up-regulated genes that are associated with ovarian cancer.

So far, techniques for identifying potential biomarkers in ovarian carcinoma and cancer studies in general, have primarily focused on detecting single-biomarkers that is single gene that can be used for early detection and/or recurrent ovarian cancer. While these pioneering approaches have proven to be successful in addressing several challenges in ovarian cancer and cancer in general, they exhibit high false positive rates. Here, we introduce and develop a novel concept termed: group-biomarkers. Group-biomarkers are a group of co-regulated genes that can be jointly used for the diagnosis of early-stage and/or recurrent ovarian cancer. The quest for group-biomarkers is motivated by the fact that different groups of individuals exhibit different patterns of cancer onset. It is therefore hoped that group-biomarkers will identify the cancer onset pattern that a particular individual is likely to experience and therefore help select the most discriminative biomarker for the corresponding pattern. Group-biomarkers have many advantages over single-biomarkers. As we demonstrate in this paper, they provide a more reliable early detection of ovarian cancer. Furthermore, but they indicate what genes may be involved in the early development of ovarian cancer and how they interact with each other. As such, they also provide a target for ovarian cancer therapy.

We present a procedure for identifying all group-biomarkers from a given, properly selected set of gene expression data. Our methodology is based on using the ability of a modified biclustering technique combined with sensitivity analysis of gene expression levels to identify all potential single-biomarkers found by prior studies as well as many more candidates that had been missed in the literature. We then use an order preserving modified biclustering technique to identify genes that are co-regulated with the candidate single-biomarkers. Each set of genes produced by this approach is a candidate group-biomarker that is then validated using additional analysis such as chemical analysis. Statistical analysis of group-biomarkers shows that

1-4244-0728-1/07/$20.00 ©2007 IEEE

they provide better early detection of ovarian cancer than single-biomarkers.

The rest of this paper is organized as follows. In paragraph 2, we present the data set used in this study. In Paragraph 3, we perform the analysis of potential biomarkers that can be used for early detection and/or recurrent ovarian cancer. Finally, we conclude in paragraph 4.

## 2. MATERIALS

The gene expression data that we used in this study corresponds to 44 normal ovaries, 10 borderline ovarian cancer tissues, 17 serous papillary ovarian carcinoma tumors, and 20 metastases of serous papillary ovarian carcinoma to the omentum. For comparison purposes, we have also used 21 others tissue sets that encompassed 372 different tissue samples: 25 normal adipose tissues, 4 normal breast (from which adipose tissue was removed), 20 normal cervix, 32 normal colon, 11 normal kidney, 12 normal liver, 24 normal lung, 45 normal myometrium, 9 normal omentum, 30 normal skeletal muscle, 17 normal skin, 15 normal small intestine, 69 normal thymus, 9 tonsils with lymphoid hyperplasia, 2 endometrial hyperplasia, 4 squamous cell carcinoma of the cervix, 3 colon adenocarcinoma, 6 endometrial adenocarcinoma, 8 kidney cell carcinoma, 6 lung adenocarcinoma, 10 squamous carcinoma of the lung, 22 gall bladder with chronic inflammation. The tissues were provided by the University of Minnesota Cancer Center's Tissue Procurement Facility. Bulk tumor and normal tissues were identified, dissected, and snap-frozen in liquid nitrogen within 15 to 30 minutes of resection from the patient. Tissue sections were made from each sample, stained with hematoxylin and eosin (H&E), and examined independently by two pathologists to confirm the pathological state of each sample. All tissue samples underwent stringent quality control measures to verify the integrity of the *RNA* before use in gene array experiments.

The gene expression was determined by Gene Logic Inc. using *Affymetrix HG_U95A* arrays containing about ~12,626 genes. The gene expression matrix was normalized using *Affymetrix* (*M.A.S. 4.0.1*), and the *log-floor* data transform with a *floor* value of *1* was performed. Because of missing values, 5 metastases of serous papillary ovarian carcinoma tissues were removed, and about 74 genes were eliminated because they all had missing values. Thus the final gene expression matrix used for simulation contained: 44 normal ovaries, 10 borderlines, 17 serous papillary ovarian carcinoma tumors, and 15 metastases of serous papillary ovarian carcinoma to the omentum, about 12626 genes among which 12000 are known genes. The data was organized in a *12626 x 86* matrix where the rows represents the 12626 genes, the columns the 86 conditions among which conditions 1 to 44 correspond to the 44 normal ovaries, conditions 45 to 54 the 10 borderline, 55 to 71 to the 17 serous papillary ovarian carcinoma, and conditions 72 to 86 to the 15 metastases of serous papillary ovarian carcinoma to the omentum. The borderlines are classified here as stage I of ovarian cancer and the other ones as late stage II and stage III of ovarian cancer.

## 3. POTENTIAL BIOMARKERS IDENTIFICATION

As mentioned earlier, we propose to identify group-biomarkers by starting from a list of single biomarkers and using an order preserving biclustering approach to find potential group-biomarkers. We describe each step of the process in this section.

### 3.1. Single-Biomarkers Identification

In [5], we presented an exhaustive method for the identification of all potential single-biomarkers. The paper also makes several important contributions. It applies a novel high-profile set of biclustering techniques recently developed by Tewfik and Tchagang [2]-[3] combined with a sensitivity analysis to the above unique and comprehensive set of gene expression data generated by Gene Logic Inc. from tissues collected at the University of Minnesota by Skubitz et *al* [1]. Since the approach of [2]-[3] can find all biclusters in a given set of data, the paper reproduces the discoveries of prior studies and identifies several additional more promising single-biomarkers candidates. More significantly, unlike most prior studies, it identifies genes that are down-regulated in ovarian carcinoma, indicating the lack of a suppressor function.

By combining the biclustering technique of [2]-[3] with a sensitivity analysis of the results by varying the thresholds used for data quantization, in [5] we did identify 481 genes upregulated in ovarian cancer tissues compared only to normal ovarian tissue. This set included all 150 genes found to be upregulated in ovarian cancer tissues in previous studies [1]. After the filtering process, we identified 55 upregulated in ovarian cancer tissues compared to normal ovarian tissue and the other 372 non-ovarian tissues. This set included all 40 candidate single-biomarkers listed in [1]. Using the same methodology, we had also identified 127 genes downregulated in ovarian cancer tissues compared only to normal ovarian tissue. We refer the reader to [5] for more development.

### 3.2. Group-Biomarkers Identification

Our procedure for identifying group-biomarkers relies on the observation that such sets of genes exhibit coherent behavior across a sub-group of patients. By acting simultaneously in a coherent manner, such genes could trigger the onset cancer. To uncover such patterns from the *DNA* microarray data, we treat each tissue as a separate condition, and seek sets of genes that are coregulated with a potential single-biomarker across as many conditions or tissues as possible. Since the ordering of the individual tissue is immaterial, we use the order preserving biclustering approach of [4] to find such sets of genes. The resulting biclusters of group-biomarkers implicitly segment the human population into non-overlapping groups that exhibit distinct patterns of cancer onset or recurrence.

The Order Preserving Submatrix Problems (*OPSM*) was introduced in [6] by Ben-Dor et *al* as a way of discovering local structure in a gene expression data *i.e*: biclusters with coherent evolutions. In [3]-[4], we also did develop a biclustering technique with low complexity to address the *OPSM* issue.

Typically, given a gene expression data matrix $A = [a_{nm}]$, with rows corresponding to the set of genes $G = \{g_1, ..., g_N\}$, columns to the set of experimental conditions $C = \{c_1, ..., c_M\}$, and $a_{nm}$ a real number that represents the expression level of the gene corresponding to row $n$ under the specific condition corresponding to column $m$, a submatrix $B$ of $A$ is order preserving if there is a permutation of its columns under which the sequence of values in every row is strictly increasing. More precisely, in the case of expression data, such a submatrix corresponds to a group of genes whose expression levels induce some linear order across a subset of the conditions.

Such patterns might arise, for example, if the experiments in $C$ represent distinct stages in the progression of a disease or in a cellular process, and the expression levels of all genes in $G$ vary across the stages in the same way. For example, in expression data that comes from a population of patients, such as in Bittner et *al*. [7], it is reasonable to expect that each individual is in a particular stage of the disease. There is a set of genes that are coexpressed with this progression, and we therefore expect the data to contain a set of genes and a set of patients such that the genes are identically ordered on this set of patients. The same situation occurs when considering data from nominally identical exposure to the environmental effects, data from drug treatment, data representing some temporal progression, etc. In many cases, the data contains more than one such pattern. For example, in cancer data, patients can be staged according to the disease progression (as in this study), as well as according to the extent of genetic abnormalities. These two orders on some subset of tissues are not necessarily correlated. Therefore, even in data where some nominal order is given a priori, we are seeking related or unrelated hidden orders and the sets of genes that support them.

## 4. EXPERIMENTAL VALIDATION

### 4.1. Results

Since the diagnostic of cancer will be performed using blood analysis, we are looking for patterns that are specific to ovarian cancer. Therefore, we are only interested in genes that are highly expressed in ovarian cancer tissues compared to normal ovarian tissues and non-ovarian tissues. In other terms, using each one of the 55 single-biomarker uncovered in [5], as reference, we have applied the above procedures to the set of microarray data used in our study and described in Section 2. We then obtained an initial group of 55 potential group-biomarkers. After removing overlapping groups, we ended up with a group of 22 potential group-biomarkers that can be used for early detection of ovarian cancer.

Figure (1) shows a subgroup of genes that are highly coregulated across a subset of tissues when the expression level of *Integrin beta 8* is taken as reference. This group-biomarker comprises *Integrin beta 8*, *Cadherin 6*, Kallikrein 8, Forkhead box J1, and Bone morphogenetic protein 7, all single-biomarkers discovered in [5] and some by previous studies [1]. It does show that those single-biomarkers work together in this subgroup of patient. The other genes in this group-biomarker represent the genes that exhibit coherent behavior with the above mentioned single-biomarkers. They are highly expressed in ovarian cancer tissues compared to normal ovarian tissues and non-ovarian tissues but, they also show some expression in about 10% of the non-ovarian tissues.

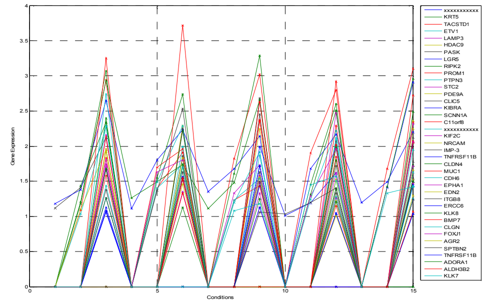

**Figure 1**: Example of potential group-biomarker composed of 39 genes. This group-biomarker comprises *Integrin beta 8*, *Cadherin 6*, Kallikrein 8, Forkhead box J1, and Bone morphogenetic protein 7 which are highly and only expressed in ovarian cancer tissues.

### 4.2. Statistical Analysis

We can compare the detection performance of group-biomarkers to that of individual biomarkers as follows. Let $a$ be the number of healthy tissues that screen positive, $b$ the number of diseased tissues that screen positive, $c$ the number of healthy tissues that screen negative and $d$ the number of diseased tissues that screen negative. We define the following parameters:

- The *sensitivity of a biomarker* (*Se*): number of diseased tissues that screen positive divided by the total number of diseased tissues: equation (1).

$$Se = \frac{b}{b+d} \qquad (1)$$

- The *specificity of a biomarker* (*Sp*): number of healthy tissues that screen negative divided by the total number of healthy tissues: equation (2).

$$Sp = \frac{c}{c+a} \qquad (2)$$

Using the above parameters, we plotted the Receiver Operating Characteristics (*ROC*) curve of the potential single-biomarkers and group-biomarkers: equation (3).

$$\textit{sensitivity} = f(1\textit{-specificity}) \qquad (3)$$

Figure (2) for example represents the *ROC* curve of the group-biomarker represented by figure (1) above and each potential single-biomarkers that are part of it: *(Integrin beta 8, Cadherin 6*, Kallikrein 8, Forkhead box J1, and Bone morphogenetic protein 7*)*. The red line represents the *ROC* of the group-biomarker when at least 70% of its genes are expressed, the green line the *ROC* of the group-biomarker when 100% of its genes are expressed, the blue line represents the *ROC* curve of *Integrin beta 8*, the yellow line the *ROC* curve of *Cadherin 6*, the magenta line the *ROC* curve of *Forkhead box J1*, the black line the *ROC* curve of *Bone morphogenetic protein 7*, the cyan line the *ROC* curve of *Kallikrein 8*. We can easily see that in any case group-biomarker either expressed at 70% or 100% performs better than each single-biomarker. The analysis of the other 22 group-biomarkers confirmed this conclusion.

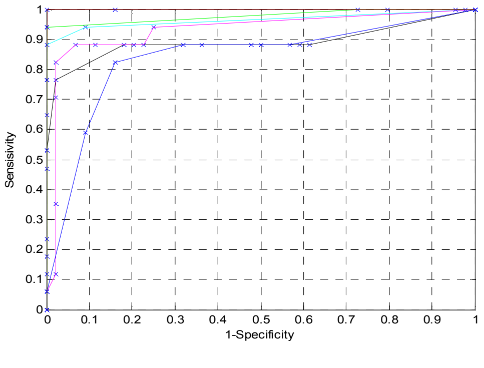

**Figure 2**: *ROC* curve, group-biomarker of figure (1) and each one of its single-biomarker. The red line represents the *ROC* of the group-biomarker when at least 70% of its genes are expressed, the green line the *ROC* of the group-biomarker when 100% of its genes are expressed, the blue line represents the *ROC* curve of *Integrin beta 8*, the yellow line the *ROC* curve of *Cadherin 6*, the magenta line the *ROC* curve of *Forkhead box J1*, the black line the *ROC* curve of *Bone morphogenetic protein 7*, the cyan line the *ROC* curve of *Kallikrein 8*.

## 5. CONCLUSION

In this study, we proposed and develop a novel concept termed group-biomarkers that can be used for early detection and/or recurrent ovarian cancer. Statistical analysis of the potential 22 group-biomarkers obtained shows that they do perform better than single-biomarkers.

The well separated histograms of their gene expression patterns in normal and non-ovarian tissues and cancerous ovarian tissues of the group-biomarkers identified make them more promising and robust biomarkers for early detection and/or recurrent ovarian cancer using blood diagnostic. Immunohistochemistry analysis and reverse transcriptase polymerase chain reaction screening of all group-biomarkers are currently in progress and will allow their biological validation. Also, we are currently performing some biological search to see if there exists a biological correlation among the genes that belong to the same group-biomarker. The aim here is to see how those genes interact with each other during the early stage of the disease.

## 6. REFERENCES

[1] Hibbs, K., Skubitz, K.M. Pambuccian, S., Casey, R.C., Burleson, K.M, Oegema, T., Jr., Thiele, J.J., Grindle, S.M., Bliss, R., and Skubitz, A.P.N. (2004) Differential gene expression in ovarian carcinoma: Identification of potential biomarkers. *American Journal of Pathology* 165(2):397-414

[2] Alain B. Tchagang and Ahmed H. Tewfik, *DNA Microarray Data Analysis: A Novel Biclustering Algorithm Approach*, EURASIP Journal on Applied Signal Processing 2006 (2006), Article ID 59809, 12 pages

[3] A. H. Tewfik, A. B. Tchagang, and L. Vertatschitsch "Parallel Identification of Gene Biclusters with Coherent Evolution", *IEEE Transaction on Signal Processing, Special Issue on Genomics Signal Processing*, Vol. 54, no. 6, pp. 2408-2417, June 2006.

[4] A. B. Tchagang, A. H. Tewfik, and A. P.N. Skubitz "Analysis of Order Preserving Genes Biclusters", *Proc. of IEEE International Workshop on Genomic Signal Processing and Statistics*, The College Station, Texas, May 28 - 30, 2006, GENSIPS 2006.

[5] A. B. Tchagang, A. H. Tewfik, A. P.N. Skubitz, and K. M. Skubitz "Uncovering Potential Biomarkers in Ovarian Carcinoma via Biclustering of *DNA* Microarray Data", *Proc. of IEEE International Workshop on Genomic Signal Processing and Statistics*, The College Station, Texas, May 28 - 30, 2006, GENSIPS 2006

[6] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem," *Proc. Sixth Int'l Conf. Computational Biology (RECOMB '02)*, pp. 49-57, 2002.

[7] M. Bittner, P. Meltzer, Y. Chen, et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling", *Nature*, 406 (6795):536-40, 2000.