Complementary Gene Signature Integration in Multiplatform Microarray Experiments

Michalis E. Blazadonakis, Michalis Zervakis, and Dimitrios Kafetzopoulos

Abstract—The concept of gene signature overlap has been addressed previously in a number of research papers. A common conclusion is the absence of significant overlap. In this work we verify the above fact, but we also assess the issue of similarities not on the gene level but on the biology level hidden underneath a given signature. We proceed by taking into account the biological knowledge that exists among different signatures and use it as a means of integrating them and refining their statistical significance on the datasets. In this form, by integrating biological knowledge with information stemming from data distributions, we derive a unified signature that is significantly improved over its predecessors in terms of performance and robustness. Our motive behind this approach is to assess the problem of evaluating different signatures not in a competitive but rather in a complementary manner, where one is treated as a pool of knowledge contributing to a global and unified solution.

Index Terms—Gene Signature Integration, Gene Signature Overlap, Gene Signature Evaluation, Breast Cancer.

I. INTRODUCTION

Modern biological and biomedical research has been challenged by the release of the Human Genome Working Draft [1]. In turn, this has led to the development of high-throughput methods of genomic, proteomic and metabolomic analysis [2], such as DNA microarrays that allow the simultaneous measurement of the expression of every gene in a cellular genome.

An open problem in breast cancer, as in other types of diseases, relates to the decision of enrolling the patient in the appropriate treatment protocol. A question that a physician is often faced with is to whether chemotherapy is indeed necessary for the patient; this dilemma is a direct consequence of the fact that even though chemotherapy or hormonal therapy reduces the risk of distance metastasis by approximately 1/3, 70-80% of the patients receiving adjuvant therapy would have survived without it [3], [4]. However, it is generally agreed that patients with poor prognostic features benefit the most from adjuvant therapy [5], [6].

Additionally, insurmountable inconsistencies in histological grading forced the American Joint Committee on Cancer to exclude histological tumor grading from its staging criteria [7]. Hence, increasing the prognostic value through the use of stable and robust markers is more than a necessity; this is a direction towards which microarray technology is expected to contribute in the near future. Using, complementary DNA microarray (Agilent Hu25K gene chip) to analyze breast cancer tissues, Laura Van’t Veer et al. [8] derived a signature of 70 genes which in turn is able to distinguish between low and high risk patient groups, outperforming significantly the performance of the NIH and St. Gallen Criteria [9]. The test set that was used involved 295 new patients which also included 61 patients from the 78 training set of study [8]. The signature has been approved by FDA under the product name ‘MammaPrint’[10] as the first clear product that profiles genetic activity, measuring the likelihood of tumor recurrence. It may help doctors in planning appropriate therapy for a patient when used in accordance with other clinical criteria and laboratory tests. Additionally, the European Organization for Research and Treatment of Cancer (EORTC) [11] launched the MINDACT [12] (Microarray In Node Negative Disease may avoid ChemoTherapy) project opening the road for randomized trials on Van’t Veer’s et al. [8] result.

On similar grounds, but using Affymetrix Human U133 gene chip, Wang et al. [13] derived a 76-gene signature consisting of 60 genes for patients positive to estrogen receptors (ER) and 16 genes for ER-negative patients. The signature showed 93% (52/56) sensitivity with 48% (55/115) specificity and 0.69 area under the receiver operating characteristic curve (AUC), in an independent test set of 171 new patients. Using the data sets published by Laura Van’t Veer et al. [8] and using a linear filter, we derived a 57-gene signature published in [14] which demonstrated a similar survival prediction performance as with the 70 gene signature of Van’t Veer et al. in the 234 new cases published by Van De Vijver et al. [9]. We make a note that in order to increase the test-set power, we removed the 61 patients that were used both for the test and training purposes by Van De Vijver et al. [9].

Even though these studies refer to similar clinical problems with data acquired under similar patient conditions, the number of common genes across studies is minimal. L. Ein-Dor et al. [15] explain the almost zero amount of overlap at the gene level among different gene signature, by statistically demonstrating that there is not a unique gene set with the statistical characteristics of the reported gene signature, so that the
solution may not be unique. In addition, cross platform evaluation i.e., application and performance of one group’s signature in another group’s data is still an open issue in the field.

In this research, we study the previously mentioned three gene signatures and verify that indeed they share negligible amount of common genes. We proceed then by searching for commonalities not on the gene level but on the biology level by means of Pathways and GOBPs (Gene Ontology Biological Processes), inspired by the work of Yu et al. [16]. At this level we discover a significant amount of overlap among the three signatures, hidden in the biology underlined by each one of them rather than in the genes themselves. Taking advantage and integrating appropriately this background knowledge, we proceed by deriving a unified 28-gene signature which demonstrates some interesting statistical characteristics: a) it gives 87% sensitivity (47/54), 58% specificity (105/180) and 0.73 AUC in the 234 new cases published by Van De Vijver et al. [9], outperforming significantly the initial three signatures b) it performs well in cross platform evaluation by distinguishing the two prognostic groups in a Kaplan-Meier survival analysis and c) it consists of a significantly less number of genes than its predecessor initial signatures.

II. BACKGROUND KNOWLEDGE
A. Gene Ontologies and Pathways
The Gene Ontology project [17] provides a controlled vocabulary to describe gene and gene product attributes in any organism. The Gene Ontology Biological Processes (GOBP) constitutes basic background knowledge organizing genes into ensembles according to the biological process they participate. Using GOBPs we can find a collection of genes that are involved in a specific biological process or find the biological processes that a specific gene is involved to. Notice that while there are uniquely identified biological processes, a specific gene may be involved in more than one process.

A pathway, on the other hand, is a series of biochemical reactions occurring within a cell. Such processes are usually rapid, lasting in the order of milliseconds in the case of ion flux, or minutes for the activation of protein, some can take hours and even days (as is the case with gene expression) to complete. Overall, GOBPs constitute a conceptual network of biological processes and interactions, dynamically adapted and updated with evolving knowledge, while a pathway refers to specific and rather strict biological functions accomplished through a series of biochemical reactions.

In this study we employ such biological background knowledge provided through GOBPs [17] and Pathways [18] to either validate, enhance or generalize derived results. More specifically, we employ the GOBPs associated with individual signatures and exploit common structures at this knowledge abstraction level to support the validity of the signatures. We also focus on the 20 pathways published by NetPath [18], 10 of which are related to immune system (immune signaling pathways) and 10 related to cancer (cancer signaling pathways).

The algorithmic tools that are utilized in our signature analysis and integration scheme are briefly discussed in the following sections.

B. The Hyper-Geometric Distribution Probability (HGDP)
The HGPD is a discrete probability distribution that describes the number of successes (number of genes belonging in a specific pathway), in a sequence of N draws corresponding to the N genes constituting a gene signature, from a finite population of M total genes without replacement. It is given as:

\[ p = f(x|K,M,N) = \binom{K}{x} \binom{M-K}{N-x} \]  (1)

where \( x \) is the number of genes in the signature belonging in the pathway, \( M \) is the total number of genes in data, \( K \) the number of the pathway genes that exist in the data and \( N \) is the number of genes in the examined gene signature.

Using the value of equation (1) we can assess the statistical significance of a GOBP or pathway as to whether its selection in the final gene signature is statistically significant or not. The lower the value of \( p \) the more significant the specific pathway or GOBP is, since it has a low probability of being a random guess. Thus, ranking the biological processes (pathways and GOBPs) that are pointed to by the genes of the signature according to their corresponding \( p \) values leads to a ranked list of process significance. We consider as significant those GOBPs or pathways with a \( p \)-value of less than 0.05.

C. Fisher Ratio - Gene Ranking Criterion
Variations of Fisher’s ratio have been applied in a number of studies as a criterion of gene significance for population discrimination [8] [19]. The Fisher ratio is given by:

\[ f(g_i, \bar{g}_i, \sigma^+ g_i, \sigma^- g_i) = \frac{\bar{g}_i - \mu g_i}{\sigma g_i + \sigma g_i} \]  (2)

where, \( \mu g_i \), \( \mu g_i \), \( \sigma g_i \), \( \sigma g_i \) and \( \sigma g_i \) are the means and standard deviations of the expressions of gene \( g_i \) in positive and negative class respectively. According to this metric, it is obvious that genes which differentiate more their expression in the two situations are assigned higher weights than those which differentiate less between the two classes.

D. The Nearest Centroid Classifier
For classification purposes we use the nearest centroid prediction rule [20]. Each patient is classified according to the distance between his/her signature and the two average profiles; the predicted class is the one closer to examined profile, by means of the Euclidean distance. Such a classifier can be formulated as follows:

\[ f(x) = \text{sign}(x - c) \cdot w \]  (3)

where

\[ c = \frac{c_+ + c_-}{2} \]  (4)

\[ w = c_+ - c_- \]  (5)
and \( c_+ \) and \( c_- \) are the centroids of the positive and negative classes respectively. This method is similar to that used by Van’t Veer et al. [8].

III. A PROPOSAL FOR SIGNATURE INTEGRATION

For notation purposes we refer to our 57-gene signature [14] as S1, the 70-gene Van’t Veer’s et al. [8] signature as S2 and the 76-gene Wang’s et al. [13] signature as S3. We first notice that S1 and S2 have been derived from the same dataset [8], whereas S3 is derived from a different-center dataset [13] that also differs in terms of the gene-chip technology employed. Furthermore, the different signatures have been derived using different algorithmic approaches. We propose a methodology for signature integration, which can be formulated by a four level procedure as depicted in Fig. 1.

In the first level the three signatures are considered as pools of knowledge, by examining every gene in each signature and collecting all GOBPs and pathways that are associated with it. Hence, at this level we derive a total collection of GOBPs and pathways hidden behind the signatures.

In the second level, the GOBPs and pathways are ordered according to their significance, as measured by the HGDP presented in section II.B. We further focus only on the statistically important (HGDP \(<0.05\)) processes. Notice here that different microarray platforms use different technologies and different sets of genes, which in turn implies a different collection of significant processes for each platform. This means that a specific GOBP, for instance, could be statistically significant for one platform (gene chip) while it is probably not significant for another. Hence, the biological background of each signature may be strongly dependent on the application platform. In our study, signatures S1 and S2 were derived using Agilent Hu25K while S3 using an Affymetrix Human U133 gene chip. The overall significance results as measured on the corresponding derivation platforms are summarized in Table I, whereas details are presented in an appendix, which is accessible through the Web (endnote 1). The genes that are involved in all significant GOBPs and pathways form three larger pools of Background Knowledge (BK1, BK2 and BK3).

### Table I

**Number of significant GOBPs and pathways located in S1, S2 and S3**

<table>
<thead>
<tr>
<th>Background Knowledge</th>
<th># of significant GOBPs</th>
<th># of significant pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>BK1</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>BK2</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>BK3</td>
<td>32</td>
<td>1</td>
</tr>
</tbody>
</table>

In the third level we proceed to knowledge integration (Fig. 1) by combining the pools of genes resulting from each individual experiment. Since S1 and S2 use the same application platform, we treat the derived knowledge as complementary parts of one another and express this complementarily by taking the union of BK1 and BK2; note that this is actually a union of the genes contained in BK1 and BK2. Nevertheless, since S3 is derived from a different platform, its contribution to the other two signatures may be viewed only in terms of the common biological knowledge that may exist between the two different platforms. Thus, for the integration of this last signature to the previous two, we exploit the intersection operation and the overall integration process is expressed as BK1\(\cap\)BK2 \(\cap\)BK3 , resulting in a collection of genes consistent with the common knowledge between two different microarray platforms and among the three different signatures.

The fourth step exploits the derived set of genes that share biological significance from all platforms, datasets and/or algorithmic schemes that have been exploited for the derivation of the available signatures. It obtains the integrated set of genes and attempts to further refine it by exploring its statistical significance on one of the previous, or even on a completely independent dataset. It is worth noticing that the third-level information (set of genes) can be tested and statistically validated on various datasets (or data distributions), so that the integration scheme can be expanded in level 4 through the intersection of the resulting signatures.

IV. EXPERIMENTAL RESULTS

A. Data Sets

For our research we used the following well known and publicly available breast cancer data sets:

1) Van’t Veer’s Data Set

This dataset is published by Van’t Veer et al. [8] and consists of 78 cases of 24481 genes each, and 78 patients. Among those 44 cases are characterized negative and correspond to patients that remain disease free for a period of at least five years, whereas the
remaining 34 are characterized positive and correspond to patients that developed a relapse within a period of five years. An additional set consisting of 19 cases is given (7 negative and 12 positive). The 78 sample set of the data is used for training purposes while the 19 additional samples are used for independent test set evaluation. 293 genes expressing missing information for all 78 patients were removed and the remaining 13604 missing values were substituted using Expectation Maximization (EM) imputation [21].

2) Van’t Veer’s Data Set
This data set consists of 234 cases and it is basically a physical continuation of Van’t Veer’s data set presented in the previous section, it consists of 24496 genes and 234 patients, 180 negative and 54 positive, note that we have excluded from the initial provided data the 61 samples that were both used for training and testing purposes by Van De Vijver et al. [9].

3) Wang’s Data Set
In addition, the data published by Wang et al. [13] is also employed for further validation purposes. It consists of 286 patients and 22283 genes; the whole data set is used for validation purposes, and for testing the cross platform prediction ability of a signature derived based on different experimental design and application platform. Based on the available data, we analyze our results according to disease recurrence. Hence, those patients for which no recurrence occurred are characterized as the good prognosis group and belong to the negative class, while those for which a reoccurrence occurred belong to the positive class.

B. Gene Overlap among Signatures
In this section we present some interesting results on the common characteristics of the studied signatures. S1 and S2 share 5 genes in common (5% overlap = 5/(57+70-5)), but there are no common genes between any of these signatures and S3. We observe only a small 5% overlap between S1 and S2, even though they are derived from a common microarray platform along with a similar gene selection philosophy (variations of the Fisher’s ratio criterion) [8], [14]. On the other hand, S3 is derived from a different microarray platform along with a different philosophy on the gene selection method based on the Cox regression model [13]. Due to platform differences, only 37 of the 76 genes of S3 are located in the microarray platform of S1 and S2.

C. The Gene Selection Method
Towards supporting different signatures with biological knowledge, we first consider the two signatures S1 and S2, each associated with its own platform. For each one we derived the significant GOBP and pathways and compiled the set of significant genes BK1 and BK2. By statistically evaluating these sets on the original dataset, we derive the biologically supported signature S1BK and S2BK, as refined versions of the statistically supported signatures S1 and S2, respectively. The BK1 collection enumerates a count of 3535 genes that were then used as candidates for deriving the 70-gene S1BK signature. In a similar manner BK2 collection enumerated a count of 3854 genes that were used in turn for deriving the 70-gene signature S2BK. Subsequently, in order to evaluate the potential of signature integration on the same dataset, we preceded one step further and considered the collection of BK1 BK2 genes to derive the integrated signature S1S2BK consisted of 69 genes that are both statistically and biologically supported.

To derive S1BK, S2BK and S1S2BK we used a recursive filter elimination approach, where genes are ranked in decreasing order according to their Fisher’s score as given in equation (2). At each step, the lower-rank genes are eliminated and the process continues recursively until the gene list is empty. In each recursion we proceed by cutting off so many genes so that the remaining ones form a power of two, up to 1024 genes; then we eliminate a set of 124 genes up to the point of 900 remaining genes; from this point on we cut sets of 100 genes up to 100 surviving ones and, finally, we proceed by eliminating one gene at a time up to the end of the process. For training and validation purposes we used the 78 and 19 patient sets of Van’t Veer (see section A.1) respectively. At the end of the process the set of signatures that are closer to 70 and achieve the highest classification accuracy on the independent test set of 19 samples are selected as the final gene signature. We use the criterion of 70 genes to be compliant with the 70-gene cutoff point in Van’t Veer’s et al. [8] signature.

We now turn our attention to the significant amount of overlap that exists between the signatures when the underlying background knowledge is considered. Specifically, between BK1 and BK2 there are 2531 genes in common leading to 52% overlap and implying that there is a lot of common biological structure, which however is hindered at the gene level. Considering S1 and S3, which share no common genes at all, at the biological process level there are 3385 genes in common between BK1 (3535 genes) and BK3 (7402 genes) or 45% overlap when the background biological knowledge is considered. Similarly, there are 3620 common genes (47% overlap) between BK1 and BK3.

D. Evaluating Initial Signatures with Background knowledge
In our first step of evaluation we measured the accuracy performance of S1 and S2 using the nearest centroid classifier trained by the 78 patient set of Van’t Veer et al. [8] (section A.1) and tested on the 234 new cases of Van’t Vijver data set.

**Table II**

<table>
<thead>
<tr>
<th>Signature</th>
<th>AUC</th>
<th>SEN</th>
<th>SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.66</td>
<td>74%</td>
<td>59%</td>
</tr>
<tr>
<td>S1BK (70-GS)</td>
<td>0.69</td>
<td>80%</td>
<td>59%</td>
</tr>
<tr>
<td>S2</td>
<td>0.69</td>
<td>76%</td>
<td>62%</td>
</tr>
<tr>
<td>S2BK (70-GS)</td>
<td>0.71</td>
<td>78%</td>
<td>63%</td>
</tr>
<tr>
<td>S1S2BK (69-GS)</td>
<td>0.73</td>
<td>81%</td>
<td>64%</td>
</tr>
</tbody>
</table>

**Table III**

<table>
<thead>
<tr>
<th>Signatures</th>
<th>Statistical performance of S* and S** on the independent test sets of 19 and 234 new patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>S1BK (70-GS)</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>S2BK (70-GS)</td>
<td></td>
</tr>
<tr>
<td>S1S2BK (69-GS)</td>
<td></td>
</tr>
</tbody>
</table>
Ind. Test Set (19 Patients) | Ind. Test Set (234 Patients)
--- | ---

<table>
<thead>
<tr>
<th>Signature</th>
<th>AUC</th>
<th>SEN</th>
<th>SPE</th>
<th>AUC</th>
<th>SEN</th>
<th>SPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S^*$</td>
<td>0.89</td>
<td>92%</td>
<td>86%</td>
<td>0.73</td>
<td>87%</td>
<td>58%</td>
</tr>
<tr>
<td>$S^{**}$</td>
<td>0.82</td>
<td>92%</td>
<td>71%</td>
<td>0.73</td>
<td>85%</td>
<td>60%</td>
</tr>
</tbody>
</table>

Fig. 2. ROC Curves (A), (B) of $S^*$ and $S^{**}$ respectively on the independent test of 19 (dotted curve) and 234 (solid curve) patients. Gene expression patterns (C), (E) and survival prediction (D), (F) of $S^*$ and $S^{**}$. The figures include the metrics of signature significance according to the log rank test for the survival distributions ($P_{LR}$) and the mean relapse time ($P$).

S1 derives a 0.66 AUC while S2 a 0.69, with sensitivity rate 74% and 76%, respectively. Furthermore, these two signatures were able to discriminate significantly between the two prognostic groups in a Kaplan-Meier survival analysis as reported [8], [14]. One interesting point associated with the proposed integration scheme is the significant increase in terms of accuracy when biological knowledge is used to derive signatures $S1BK$ and $S2BK$, respectively (rows 2, 4 of TABLE II).

A 6 unit increase in sensitivity and a 3 unit increase in AUC is achieved between S1 and S1BK, while a similar behavior, but with a lower degree of increase, is observed in the S2 case. These results are supporting the coupling of statistical approaches with biological information as a meaningful gene-selection procedure. Furthermore, notice that the S1S2BK signature improves even further the statistics of S1BK and S2BK (TABLE II) by giving a 0.73 AUC along with an 81%
sensitivity (true positive) and 64% in specificity (true negative). We stress out the increase in sensitivity due to its significant role in clinical implications and the fact that it enables doctors to decide more accurately on the therapeutic protocol, which is very important in cancer as to minimize the number of patients that receive unnecessary toxic treatment.

E. Towards an Incremental Integration of Signatures

Proceeding in an incremental fashion, we utilize signature S3 stemming from yet a different platform and apply the rule BK1UBK2 ∩ BK3 as proposed in section III. This operation results in a collection of 1242 genes, which encode biological significance in both Van’t Veer and Wang platforms. By applying the gene selection method (section II.C) on the 78 cases of Van’t Veer’s data set, we derive a 28 gene signature denoted as S*, which achieves an 0.89 AUC (Fig. 2-part A-dotted curve, only two patients missed) on the independent test set of 19 patients (section A.I) while it gives a 0.73 AUC (Fig. 2-part A-solid curve) and with 87% sensitivity and 58% specificity on the 234 new patients of Van’t Vĳër’s cohort (TABLE III, 1st row). Notice the 6 unit increase in sensitivity in comparison to S1S2BK (TABLE II).

We once again stress out the increase in sensitivity as a crucial factor in medical decision making. Thus, among the 54 true positive cases, 47 are indeed classified as true positives and 7 as false negatives. On the other hand, among the 180 true negative cases, 104 are classified as true negative and 76 as false positive. This means that on a total of 111 cases classified as negative (104 true negative + 7 false negative) the 104 of them are indeed true negative, implying that a doctor can take the decision of “no chemoTherapy” with a high certainty of 94% (104/111); this is a positive side-effect of the increased sensitivity rate.

In Fig. 2 (part C) we present the gene expression analysis of the 28-gene signature S* based on the classification result depicted through a labeled hierarchical tree analysis. Within this procedure the class label is used as a value in the analysis, so that the two classes do not intermingle with each other, while at the same time the differences in the expression levels between the two groups are revealed. Notice the significant difference in the gene expression levels between the two prognostic groups, where the green sub-tree corresponds to the good prognosis and the red to the poor. In a similar manner the survival prediction analysis in Fig. 2 (part D), discriminates significantly between the two prognostic groups with the green and red lines corresponding to the good and poor prognosis respectively. In terms of the distance between the survival curves, the log rank test yield PLR < 0.001, which indicates the high statistical significance of the derived result. In addition we assess the statistical significance of the relapse time between the two curves and evaluate the null hypothesis $H_0: \bar{X} = \bar{X}_+$, where $\bar{X}_-, \bar{X}_+$ represent the mean relapse time between the two groups. Indeed, the null hypothesis is statistically rejected with $P < 0.05$, indicating that there exists significance difference of the mean relapse time between the two groups.

Notice that the signature $S^*$ includes biological knowledge consistent with the Wang platform, but its statistical significance is still assessed on the Van’t Veer platform. In this sense, it would be interesting to evaluate the intersection of $S^*$ and S1S2BK. Even though both signatures are statistically assessed on the same platform, $S^*$ can be thought of as a more “informative” extension of S1S2BK through the incorporation of knowledge from a different platform. Based on this intersection we derive a new signature (denoted by $S^{**}$) of 21 genes, which gives 0.82 AUC on the independent test set of 19 patients (Fig. 2-part B-dotted curve, 3 patients missed) while it still gives a 0.73 AUC (Fig. 2-part B-solid curve) on the test cohort of 234 new patients (TABLE III, 2nd row). Notice that we have a decrease in sensitivity by two units, which corresponds to one patient compared to the $S^*$ signature, while we have two units increase in specificity. Nevertheless, we consider this finding significant in terms of the robustness of the proposed signature $S^*$, since with even a fewer number of genes it still retains a high sensitivity rate coupled with a significant AUC rate. The performance of $S^{**}$ in terms of gene expression profile and survival prediction remains in the same level as $S^*$ is depicted in Fig. 2 (E and F parts). A detailed list of the genes of $S^*$ and $S^{**}$ is presented in the Web Appendix (endnote).

F. Cross Platform Evaluation

An important step in our evaluation scenario is to assess one group’s signature on another group’s data, based on a 9-year survival prediction analysis. Towards this direction and starting from the initial signatures (S1, S2 and S3), we evaluate their ability to reveal the two prognostic groups when applied to a different dataset, produced with a different experimental protocol along with a different microarray platform. To accomplish this task we evaluate signatures S1 and S2 on Wang’s data set which was derived from an Affymetrix Human U133 gene chip, while S3 is tested on Van’t Vĳër’s data set derived from an Agilent Hu25K gene chip. We manage to locate 44/57 genes of S1 and 52/70 genes of S2 in the U133 gene chip, while 37/76 genes of S3 were located in Hu25 gene chip. Based on the expression profiles of the commonly identified genes in the cross platform datasets, we performed a SOM clustering step and assess the survival prediction in each of the two generated clusters in terms of Kaplan-Meier survival analysis. The 9-year results are depicted in Fig. 3, where we observe that S2 and S3 appear to be statistically significant according to the log rank test ($P_{LR} < 0.05$ for parts B and C of Fig. 3). Note that the log rank test compares the survival distributions of the two groups by measuring the area between the two curves; the wider the area, the more likely the trial reaches a statistical significance.

A deeper understanding of the cross platform potential of these signatures is obtained by measuring the statistical significance in terms of the relapse time between the two curves and evaluate the null hypothesis $H_0: \bar{X}_- = \bar{X}_+$, where $\bar{X}_-, \bar{X}_+$ represent the mean relapse time between the good and the poor prognosis groups; we expect $H_0$ to be rejected. This test indicates that the null hypothesis cannot be rejected for S1
and S3, since P > 0.05. Hence, the cross-platform evaluation of these signatures is inconclusive in terms of their statistical significance on other datasets. The performance of S2 on the other hand is statistically accepted since both p-values are < 0.05. Of special interest is the performance of S3 on survival analysis on the Van’t Vijver platform, where even though there exists a significance area gap between the two groups in Fig. 3 (C), verified also by the log rank test, this is not translated into a significant relapse-time difference.

By assessing the performance of S* and S** (again on Wang’s data set) we derive the result depicted in Fig. 4. Note that both S* and S** succeeded to improve significantly their performance over their predecessor signatures (S1, S2 and S3) in terms of the Kaplan-Meier survival analysis.

![Fig. 3. Cross platform evaluation through a Kaplan-Meier survival analysis of the initial signatures S1, S2, S3 in parts A, B and C respectively, one’s group signature is evaluated on another’s group data. The figures include the metrics of signature significance according to the log rank test for the survival distributions (P_LR) and the mean relapse time (P).](image)

![Fig. 4. Gene expression levels and Kaplan-Meier survival analysis in the cross platform evaluation performance of S* and S**; green and red curves correspond to the good and poor prognosis group respectively; for gene expression levels rows represent patients and columns genes; the gene expression levels of the good prognosis group are located below the yellow line while the poor expression cluster is above it. The figures include the metrics of signature significance according to the log rank test for the survival distributions (P_LR) and the mean relapse time (P).](image)

Both P_LR < 0.001 and P < 0.001 measures are significantly improved. They now fall below the 0.05 p-value (95% level of significance) and hence, they indicate statistically significant signatures. Notice also that the gene expression levels of the two signatures (Fig. 4) demonstrate a significant differentiation on their expression levels between the two prognostic groups. The cluster corresponding to the good prognosis is located below the yellow line, while the poor prognosis group being above it.

G. Randomness of the Derived Results

In this section we assess the question of randomness and consider if the result produced is a random one or not. To assess the randomness of the result we perform three additional computational experiments.

In the first experiment we focus on the total of 24188 genes contained in the data set of 78 patients (section A.1) that was used as training set. We randomly select a set of 28 genes and measure its performance on the 234 new-patient cohort (Van’t Vijver’s data set), with this process being repeated 100 times. Then, we assess the statistical significance of achieving a 0.73 AUC_1 or more (matching the performance of S* and S**), while at the same time achieving an AUC_2 of 0.89 or more on the independent test set of 19 patients. Testing the null hypothesis

$$H_0 : AUC_1 \geq 0.73 \& AUC_2 \geq 0.89$$

at the 95% confidence level, we derive a p-value < 0.001, indicating that the result of S*
is very unlikely to have happened by chance. To be even more conservative, in the second experiment we performed the same scenario but focusing only on the 1242 gene set derived by the application of \( \text{BK1} \cap \text{BK2} \cap \text{BK3} \) operation. Once again, the null hypothesis \( H_0: \text{AUC}_1 \geq 0.73 \& \text{AUC}_2 \geq 0.89 \) was rejected with a p-value < 0.001.

In the third computational scenario we assess the likelihood of deriving by chance the 1242 gene set of biological knowledge extracted through the application of \( \text{BK1} \cap \text{BK2} \cap \text{BK3} \) operation. Note that even though we utilize biological knowledge from two different platforms, we use Van’t Veer’s data set to statistically refine this knowledge and derive a unified gene signature. Hence, we use the Van’t Veer data set (section IV.A.1) as our starting point. We performed 100 iterations on the original Van’t Veer 24188 gene set, where in each iteration we randomly selected a sample of 1242 genes. From each run we derive a 28-gene signature by applying the gene selection method and measure its AUC performance on both the independent test set of 19 samples (AUC\(_1\)) and the new-patient set of 234 patients (AUC\(_2\)). We subsequently assess the null hypothesis: \( H_0: \text{AUC}_1 \geq 0.73 \& \text{AUC}_2 \geq 0.89 \), which is rejected with a p-value < 0.001. Indeed, among 100 runs of 1242 genes each, \( H_0 \) was not satisfied even once, indicating that the 28-genes of S\(^*\) are tightly bound with the 1242 gene set of biological knowledge.

Overall, the statistical tests performed enhance the confidence that the produced signatures are unlikely to have happened by chance.

**V. CONCLUSIONS**

In this paper we address the problem of multiplatform studies, including the small overlap between derived gene signatures and the low efficiency of each signature on different platform datasets. The proposed integration approach considers the complementarily of different signatures, which is revealed at the level of biological information carried by the signature genes and aims at a signature evolution strategy that interferes between statistical and knowledge-based evaluation. We begin with the assumption that even though different gene signatures may share minor or no overlap, they are probably part of a more global solution, each one focusing on a different, complementary or overlapping part of the biological aspect of the problem. Experimental evidence supports this assumption, at a level of biological processes associated with the signature genes. Based on such biological knowledge, we propose an integration approach that groups together the biological processes triggered by different signatures and statistically evaluates the associated genes on a testing platform (which may be independent of the original datasets of the initial signatures). Such integration of gene signatures at the level of biological knowledge produces a statistically more significant signature, improving its performance over the initial ones and demonstrating reliable clinical prediction outcome on a different dataset from different experimental design and microarray platform. This further enhances the value of the integrated signature, implying that it does not entirely depend on the platform or experimental design, but it is most likely associated directly with the biological origin of the disease itself.

We notice at this end that the proposed signature integration scheme makes full use of biological knowledge form multiple (readily extends to more than two) platforms, but its statistical significance in a new integrated signature extraction process is validated on one platform (Van’t Veer). Nevertheless, it could be similarly validated on the other platform (Wang), or even on a completely independent platform, possibly resulting in different signatures. As an extension to our methodology, we currently explore the possibility of using the integration of such signatures at the final step of Fig. 1, resulting in a fully integrated scheme acquiring knowledge and statistical significance from all datasets.

**ACKNOWLEDGMENT**

Present work was supported by the Hellenic Ministry of Education.

**REFERENCES**


To access the Appendix, please visit [www.display.tuc.gr](http://www.display.tuc.gr) and follow the links: Publications/Browse by Category/Miscellaneous/Details/Related Downloads.