Biomarkers for Epithelial Ovarian Cancers

Pritha Mahata
Pritha.Mahata@newcastle.edu.au
School of Electrical Engineering and Computer Science, University of Newcastle, University Drive Callaghan, Newcastle, NSW 2308, Australia.

Abstract

Epithelial carcinoma of the ovary is one of the most common gynecological malignancies and the fifth most frequent cause of cancer death in women. Currently blood test of advanced epithelial tumors are reflected in a high level of CA 125 antigen. However, it is not a good marker for early stage tumors, and may yield false positives. Clearly, there is a need for better understanding of the molecular pathogenesis of *epithelial ovarian cancer*, so that new drug targets or biomarkers that facilitate early detection can be identified. This work concentrates on finding genetic markers for three epithelial ovarian tumors, using a simple computational method.

We give a small set of genetic markers which are able to distinguish clear cell and mucinous ovarian cancers (13 and 26 genes respectively) from other epithelial ovarian tumors with 100% accuracy. We obtain the genes *HNF1-beta* (*TCF2*) and *GGT1* as the best markers for the clear cell and *CEACAM6 (NCA)* as the best marker for mucinous ovarian tumors. We employ a feature selection technique based on minimum probability of error for this purpose. We give a ranking of the important genes responsible for these tumors and validate the results using the leave-one-out cross-validation technique.

Using this method, we also agree with the common notion that *WT1* is one of the best genes to separate serous ovarian tumors from other epithelial ovarian tumors.

Keywords: epithelial ovarian cancer, gene selection, density estimation, minimum probability of classification error

1 Introduction

Although ovarian cancer (*OvCa*) is the most common gynecological malignancy with a relatively poor 5-yr survival record, the mechanism by which these tumors arise is not well understood. Epithelial tumors, which begin on the ovary surface, are the most common. On the basis of morphological criteria, there are four major types of primary epithelial ovarian adenocarcinomas (clear cell, mucinous, serous and endometroid), *EOC* in short. The clear-cell (*OC*) and mucinous ovarian (*OM*) tumors, each account for 10% of EOC's, while serous (*OS*) and endometroid ovarian (*OE*) tumors comprise the rest of the $50-60\%$ and $20-25\%$ of ovarian cancers, respectively. A number of studies have noted particularly unfavorable prognosis for the clear cell carcinomas. Also, the serous and endometroid cancers are mostly present in advanced stages.

Cancer antigen 125 (CA125) is the most widely used biomarker for ovarian cancer (see, e.g., [5, 11]). Although concentrations of CA125 are abnormal in about 80% of patients with advanced-stage disease, they are increased in only $40-50\%$ of patients in the early-stage ovarian cancer [11]. In fact, high serum CA125 is also seen in many benign gynecological diseases and other types of cancer [5]. There are also other works to find markers using proteomic approaches (e.g., see [15]). However, most works for finding proteomic markers concentrated on distinguishing between normal and tumorous epithelial ovarian cells. Even though this is an absolute requirement for diagnosis of ovarian cancer, it is not enough to find the cause of different EOCs with distinctive morphological differences. Thus, there is an evident need of biomarkers for analysing differences between various EOCs.

The application of DNA microarray technology has enabled the study of gene expression profiles of large numbers of tumor samples and has provided an opportunity to classify different diseases based on characteristic expression patterns (see [3] for a work on diffuse large B-cell lymphoma (DLBCL), [8] for colon cancer, and [19] for breast cancer). In fact, there were several studies to find responsible genes in different epithelial ovarian tumors as well, e.g., see the works of [6, 16, 17, 23]. However, excepting the work of [17], other studies do not have sufficient number of samples from all four morphologically different subtypes of EOCs. For instance, the work of [16] does not consider any mucinous or endometroid ovarian tumors. The work of [6] considers only the short-term and long-term survivors of the serous subtype. Finally, authors in [23] considered short-term and long-term survivors from all EOCs. For this reason, we analyse the dataset of [17], which contains samples of all four EOCs.

In this work, we find genetic markers for clear-cell, mucinous and serous ovarian tumors. The genetic markers provided in this work in fact have a better accuracy of classification than the genes mentioned in [17], while we used the same classifier. This is due to the fact that the gene selection algorithm is robust in presence of outliers unlike many other methods.

This work employs a new gene ranking technique using minimum probability of classification error (*MPE*) as a metric for "goodness" of a gene. In this method, each gene is annotated with a rank considering the situation that only the gene in question is solely used for classification purpose. Then the genes are ranked according to the MPEs. The lower the MPE, the better is the gene. This method of gene selection technique uses the ideas from Bayesian decision theory and probability density estimation techniques. Unlike most other methods (e.g., correlation-based methods), this method performs well even when there are multiple modes of a gene within the same class. It turns out that this method successfully detects a small set of genes with very small probability of error. Many of these genes have been explored in many biological wet-lab experiments. Furthermore, in each case our genes give better accuracy than previous analysis of this dataset.

Our best markers for OC tumors agree with previous works. However, our ranking of genes reveals that there are two genes for clear-cell tumors, *HNF1-beta* and *GGT1* which are the best markers for OC. More importantly, we find a few genes (e.g., *CEACAM6*, or *NCA* and *CEACAM5*) for mucinous tumors, which have long been suspected to be important for this type. Notice that these genes were not discussed in the work of [17]. We also give a few genetic markers (e.g., *WT1*) for distinguishing OS tumors from other EOCs.

2 Materials and Methods

2.1 Materials

The dataset we analyse here is an Affymetrix dataset of oligoneucleotide microarrays, contributed by Schwartz *et al.* in [17]. It contains 113 samples from ovarian tumors with 9 clear-cell (OC), 11 mucinous (OM), 36 endometroid (OE) and 57 serous (OS) ovarian tumors. Notice that 1 in OC, 1 in OM, 3 in OE and 4 in OS are considered to have mixed histologies. The microarray data consists of a two-dimensional matrix where each row corresponds to the gene expressions of 113 ovarian carcinomas and each column corresponds to the gene expression for each tumor patient. We log-transformed the data. Then, each row data was centered at its mean and then divided by the standard deviation. This normalization procedure is a standard practice for working with Affymetrix datasets.

However for our work, we excluded the 9 samples with mixed histologies from the total dataset (as in [17]). We used the expressions of the total $7,069$ genes for the rest of the 104 samples for further analysis.

2.2 Methods

2.2.1 Gene Selection

To obtain a genetic signature for a particular class a of tumors, we need to find the genes which are most differentially expressed between the class a and the other tumors (called class b or rest, hereafter).

The metric used for differentiability of a gene is the probability of making a misclassification, if the gene under consideration is solely used for classification purpose. Smaller is the probability of misclassification (error), better is the gene for the classification purpose. However, the probability of error depends on the decision-making algorithm. The optimal Bayes decision-making algorithm minimizes the probability of error. Let g denote the random variable for a gene's expression. Denote the class-conditional probability density functions (pdf) of g for class a and class b by $p(\cdot)$ and $q(\cdot)$ respectively. Then, if the sample is in class a, then

$$F_a(x) := \text{Probability}\{g < x\} = \int_{l_a}^{x} p(y)\ \mathrm{d}y, \tag{1}$$

where l_a is the minimum possible value of g. Similarly, if the sample is in class b,

$$F_b(x) := \text{Probability}\{g < x\} = \int_{l_b}^{x} q(y)\ \mathrm{d}y, \tag{2}$$

where l_b is defined accordingly. The functions $F_a(x)$ and $F_b(x)$ are the class-conditional probability distribution functions for a and b, respectively.

Now, suppose in an experiment, g assumes a value z and the class of the sample is unknown. The optimal decision rule to infer the class label c for the sample would be

$$c = \begin{cases} a, & \text{if } p(z) \geq q(z) \\ b, & \text{otherwise.} \end{cases} \tag{3}$$

The decision rule (3) minimizes the probability of classification error, and the corresponding minimum value is given by

$$r = \beta \int_{S_b} p(y)\ \mathrm{d}y \quad + \quad (1-\beta) \int_{S_a} q(y)\ \mathrm{d}y, \tag{4}$$

where S_a is the set of all z such that $p(z) \geq q(z)$, S_b is the set of all z such that $p(z) < q(z)$, β is the prior probability of getting a sample from class a and $(1-\beta)$ is prior probability of getting a sample from class b. Generally $\beta = 0.5$ in a practical scenario, *i.e.* both classes are equally likely. We use the minimum probability of error (MPE) r as a measure of the goodness of the gene g.

We approximate $F_a(\cdot)$ and $F_b(\cdot)$ as piecewise linear functions between predefined quantile points. We use a simple quantile-based approach for estimating these functions. Given $F_a(\cdot)$ and $F_b(\cdot)$ and the quantile points, the approximation of the piecewise constant probability density functions $p(.)$ and $q(.)$ for the two classes a and b is straightforward from Equations 1 and 2. Also, evaluation of the integral (4) reduces to the computation of the areas of certain rectangles. The estimated density functions for the gene *TCF2* are shown in Figure 1, where the solid line is $p(\cdot)$ and the dotted line is $q(\cdot)$. Since the two functions do not intersect in case of *TCF2*, the MPE r for this gene turns out to be 0. In general, the area of intersection of both the curves will correspond to the MPE of the gene under consideration.

We note by passing that there are more sophisticated techniques for estimating the probability density functions [18]. Such methods were employed in [22]. However, these methods require large number of samples to work properly.

2.2.2 Gene Selection by Correlation Method

We also use correlation-based method for the sake of comparison. In correlation-based method, first each gene is sorted within each class. Let there be n_a samples in class a and n_b samples in class b. Consider gene expression $g = [g_1, \ldots, g_{n_a}, g_{n_a+1}, \ldots, g_{n_a+n_b}]$ and $n = n_a + n_b$. Also, let $v = [v_a; v_b]$ where $v_a = l_1, \ldots, l_{n_a}$ and $v_b = m_1, \ldots, m_{n_b}$ with $l_1 = \cdots = l_{n_a} = 1$ and $m_1 = \cdots = m_{n_b} = -1$. Then Pearson's correlation coefficient ρ between g and v is given by

$$\rho_g = \frac{\sum_{k=1}^{n}(g_k - \bar{g})(v_k - \bar{v})}{\sqrt{\sum_{k=1}^{n}(g_k - \bar{g})^2 \sum_{k=1}^{n}(v_k - \bar{v})^2}}$$

with

$$\bar{g} = \frac{1}{n}\sum_{k=1}^{n} g_k, \quad \text{and} \quad \bar{v} = \frac{1}{n}\sum_{k=1}^{n} v_k.$$

Finally, genes are ranked according to the absolute values of their Pearson's correlation coefficient ρ_g, i.e, $\text{rank}_g = |\rho_g|$. Higher the value of ρ_g, better is the gene. Notice that this method is heavily affected by the presence of the outliers.

2.2.3 k-Nearest Neighbor Classifier

The authors of [17] used a k-nearest neighbor classifier with $k = 5$. For the sake of comparison, we do the same. KNN classifiers are based on finding the k nearest samples in the training set, and taking a majority vote among the classes of the k samples. We measured the 'nearest' by Euclidean distance. We used matlab's *knnclassify* routine for this purpose. In fact, we use a leave-one-out crossvalidation (LOOCV, in short) technique for establishing the fact that the high-ranked genes obtained by the above method are indeed important markers. In this technique, each time we omit one distinct sample; we use the rest of the 103 samples for training and then we classify the omitted one using KNN. Thus the procedure is run 104 times in total.

We also experimented with more sophisticated classifiers like support-vector machines [21], but it did not give any better accuracies than the KNN classifier.

3 Results

3.1 Markers for OC

Of all of the epithelial ovarian cancers, ovarian clear cell carcinoma (OC) of the ovary has the worst prognosis. We selected 30 top genes with minimum MPE for distinguishing between clear cell and other tumors. Interestingly, even the 30th gene has a very small MPE (0.053). Our method of ranking genes finds two genes *TCF2* and *GGT1* with MPE 0. The pdfs for *TCF2* is shown in Figure 1. Other highly ranked OC markers are *LAMB1*, *FXYD2*, *PVR*, *GLRX*, *DAPK1*, *MITF*, *TNFAIP1*, *NR1H4*, *RBP4*, *ANXA4*, *CYP2C18*, *FDPS*, *ESR1*, etc. For these genes, MPE values range from 0.007 to 0.033. Notice that some of the above genes are missed in previous biological experiments for OC tumors. Thus, further wet-lab experiments are required for them. However, some of the above genes are also known to be significant by the biologists.

TCF2: The significance of finding *TCF2* with 0 MPE is that recently *TCF2* (also known as *HNF-1Beta*) has been suggested to be an excellent OC-specific molecular marker [20]. This gene is expected to be a molecular target for therapy of OC. Authors of [20] report that *HNF-1 Beta* is up-regulated in OC, while it is downregulated in other epithelial ovarian cancers and the reduction of HNF-1 beta expression by RNA interference induced apoptotic cell death in OC cells. In the dataset of [17], in fact we found that not only OC has highest level of *HNF-1 Beta*, but also the fact that the mucinous samples have lower expression than OC, but higher expression than the rest of the tumors.

GGT1: The gene *GGT1* is considered in [9] where authors found that GGT1 is expressed in many common ovarian epithelial neoplasms. However, we find that in the considered dataset the expression of *GGT1* in OC is much higher than the others.

LAMB1: It is known that a sequence in the *LAMB1* (*Laminin beta 1*) is shown to have the capacity to inhibit metastasis.

DAPK1: Death-associated protein kinase 1 (*DAPK1*) is a positive mediator of gamma-interferon induced programmed cell death. It is a tumor suppressor candidate.

ANXA4: *ANXA4* is almost exclusively expressed in epithelial cells.

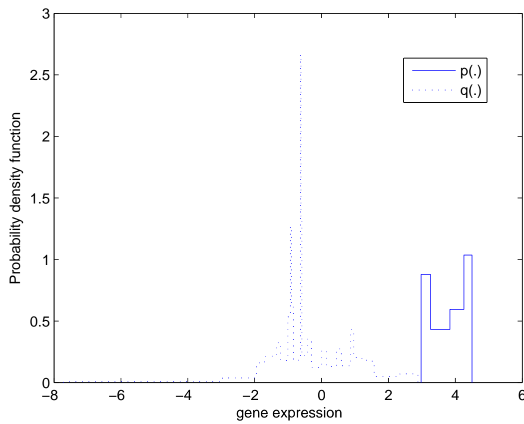

Figure 1: p (q) is the probability density functions of the gene *TCF2* for OC (rest, resp.). Notice that the functions $p(.)$ and $q(.)$ do not have any intersection at all. Thus, the MPE for the gene *TCF2* is 0.

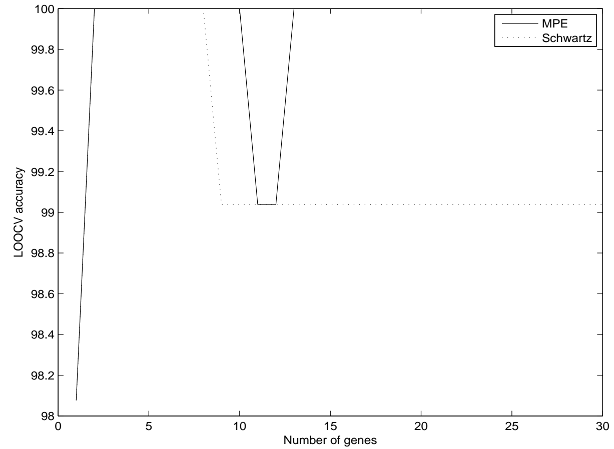

Figure 2: The accuracy of classification using LOOCV in OC Vs the rest signature of 30 genes. The solid (dashed) line gives the accuracy for genes selected by MPE (from the genes in Schwartz *et al.*)

CYP2C18 and ESR1: *CYP2C18* and *ESR1* are suspected to have roles in ovarian cancer.

For the sake of comparison, we order the 73 genes found by [17] using our ranking method. Notice that the best 7 genes in the above list are also found in [17]. However, 14 of the genes in this list are missing in the ranked list of first 30 genes of [17]. Many of our genes not only have a low MPE, but also are biologically significant. For instance, *DAPK1*, *CYP2C18*, *ESR1*, etc have low MPE for distinguishing clear cell from others, but these genes are missed when just an average fold-change across different EOCs is considered in [17].

In correlation-based method, the highest ranked gene is *FXYD2* with $\text{rank}_{FXYD2} = 0.76$. Next best gene was *RBP4* with $\text{rank}_{RBP4} = 0.72$. Even though we find these genes in our list, correlation-based method fails to obtain *TCF2* and *GGT1* as good markers for OC.

3.1.1 Validation

Figure 2 shows the result of leave-one-out cross-validation (LOOCV) with the genes from clear cell vs rest signature. In this figure, the solid line represents the accuracy of classification with respect to the first 30 genes obtained by our method from the original 7,069 genes. The dashed line represents the LOOCV accuracy with respect to the first 30 of the 73 genes provided by [17]. Notice that we ranked these 73 genes by MPE for the sake of a fair comparison. Since, taking large number of genes deteriorates the accuracies, we restrict our comparisons upto 30 genes.

It turns out that our accuracy exactly matches with that of [17] using 7 genes, since those genes are the same in both cases. However, accuracy of [17] drops from our accuracy, when larger number of genes are considered. Notice that in that case the intersection of genes is smaller (9 out of 23 precisely). Interestingly, if we include the samples with the mixed histologies, the above accuracies in Figure 2 remain intact using the genes obtained by our method. Also, notice that the work of [17] reports that they always miss one clear-cell sample out of the 8 samples, when they used a total of 158 genes which are important for all EOCs. This scenario does not occur in our case, which points to the possibility that our gene set is possibly better than the earlier published markers for OC. Also, for finding out a single type of EOC, it is wiser to use a small set of genetic markers and do a binary classification with a group vs. the rest, rather than use a large number of genes and do a multi-class classification.

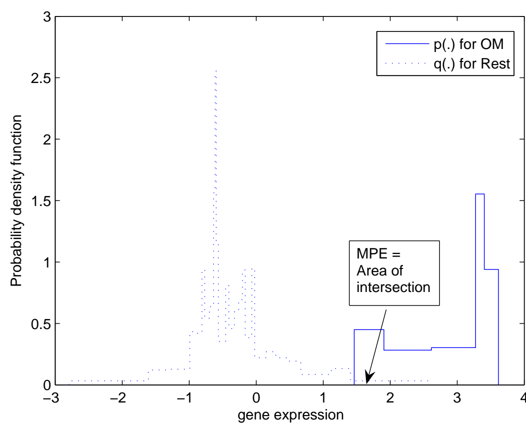

Figure 3: The probability density function $p(.)$ corresponding to the gene *CEACAM6* of the OM samples' is shown in solid line; and the same for the rest of the samples is shown in dashed line.

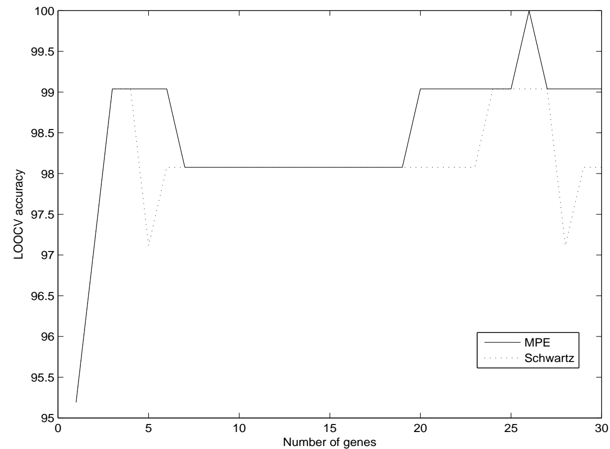

Figure 4: The accuracy of classification using LOOCV for OM vs. rest. The solid (dashed) line gives the accuracy for genes selected by MPE (from [17], genes were ranked by MPE).

Remark: The work of [16] also considers a signature for OC vs. other EOCs (mainly serous ovarian tumors) and has only 4 genes (*GLRX*, *NDRG1*, *NP* and ANXA4) common with the 73 genes in [17]. Our work has *GLRX*, *NDRG1*, *ANXA4*, and *ESR1* in common with that of [16]. However, the work in [16] does not obtain our best genes *TCF2* and *GGT1*.

3.2 Markers for OM

Next, we find marker genes which distinguish ovarian mucinous tumors (OM) from the rest of the ovarian tumors. The following genes are considered to be highly-ranked by MPE: *CEACAM6*, *CEACAM5*, *STS*, *TRADD*, *COL17A1*, *IQGAP2*, *SPINK1*, *TFF1*, *FUT3*, *CTSE*, etc. The minimum and maximum MPEs are 0.018 and 0.101, respectively for the above genes. Again, biological experiments were not done for all of the above. Thus, we find new markers to work on for OM tumors. In the following, we discuss the biological significance of some of the above OM markers.

(1) The best genes to distinguish OM tumors from the rest are CEACAM6 (alias *NCA*) and *CEACAM5*, which are called *carcinoembryonic antigen-related cell adhesion molecule 6* and *5* respectively. It was established in [10] that OM tumors contain considerable amounts of *CEACAM5* and *NCA*, whereas serous type neoplasms show negligible amounts or lack of these antigens. Immunological data indicate that ovarian mucinous and serous adenocarcinomas derive from separate lineages of epithelium. In fact, we find that *NCA* is found in smaller amount also in the OC and OE tumors. In Figure 3, we show how the probability density of *CEACAM6* for OM samples, $p(.)$ differs from the same function $q(.)$ for the rest of the samples. However, other feature selection methods missed to find it as the best marker for OM tumors.

(2) In [13], the gene *STS* was found to be overexpressed by immunohistochemically staining in 70% of OC, 33.3% of serous adenocarcinoma and 50.0% of mucinous adenocarcinoma and was localized in the cytoplasm of neoplastic epithelial cells. However, in the considered dataset, OM samples have higher expressions of *STS* than others.

(3) The results of [2] suggest significant level of expression of *COL17A1* (*BPAG2*) in ovarian carcinoma. In fact, in this dataset *Col17A1* is overexpressed in OM, but underexpressed in the rest.

(4) In [14], it is claimed that *SPINK1*, alias *TATI* was very successful in diagnosing mucinous carcinomata of the ovaries; the rate of true positive findings was 64% versus 50% for CA 125. The authors of this work conclude that CA 125 remains the single tumor marker of choice in the diagnosis of malignant epithelial ovarian cancer, while *TATI* appears to be a valuable complementory marker with a higher sensitivity in cases of poorly differentiated and mucinous carcinoma. In this work, *SPINK1* has MPE 0.061.

(5) CTSE was found to be expressed in other mucinous tumors like mucin-producing adenoma (see [4]).

(6) The three trefoil factors (*TFF1*, *TFF2*, and *TFF3*) are small peptides believed to cross-link mucous glycoproteins and to play a role in the maintenance and repair of the gastrointestinal mucosa. Aberrant expression of *TFF1*, *TFF2*, and *TFF3* was frequently detected in gastric cell lines. Both *TFF1* and *TFF3* were found among the first 30 genes as markers for OM tumors.

Among all the above genes, the work of [17] only mentions *TFF1* as a good marker for differentiating OM against rest. The correlation-based method also has the highest value of $\text{rank}_{TFF1} = 0.76$. This method also finds the next 4 genes among the top genes: *CEACAM6*, *CEACAM5* and *CTSE*. Thus, the importance of *CEACAM6* is also verified by this method. Finally, *ANXA4* is the 12th gene as OC marker and the 18th gene as OM marker. In fact, this is a bi-modal gene and the expression levels in OC and OM have two distinctive levels. Thus, our method's novelty lies in finding multi-modal genes.

3.2.1 Validation

Following the same methodology as in case of OC tumors, we obtain accuracy curves with respect to the genes selected by MPE and the genes from [17] (shown in Figure 4). Our genes are ranked from the whole set of 7,069 genes whereas we rank 82 of genes given by [17], for mucinous vs. rest. Notice that we reach 100% accuracy when we consider the best 26 genes according to MPE. However, there are only 17 genes which are common among the two sets of 30 genes in our markers and those of [17]. Thus, it turns out that it is a bit harder to separate OM from other EOCs, as it is evident from the minimum MPE value 0.018 for *CEACAM6*.

3.3 Markers for OS

Next we discuss a few genetic markers for distinguishing serous ovarian tumors from other tumorous ovarian samples. However, as noticed by other works, serous tumors may have further subgroups. The best MPE we get is 0.137 for the Wilms tumor suppressor 1 gene *WT1* and this value is much higher than the previous two cases (see the probability density functions in Figure 5). Overall highly-ranked genes are *WT1*, *PRLR* (prolactin receptor), *S100A1*, *KLK7*, *PIK3CA*, *PTGS1*, *GAS1*, *PLAB*, *LU*, etc. The MPEs range from 0.137 to 0.188 for the above genes. We discuss a few of the above genes for which some biological findings exist.

WT1: Recent studies [12] reveal that expression of *WT1* may be indicative of an unfavorable prognosis in patients with advanced serous epithelial ovarian carcinoma. The work of [17] also finds *WT1* as a marker for OS tumors.

WIT-1: *WIT-1* has the same temporal and cell-restricted expression pattern, as *WT1*. Furthermore, *WIT-1* is located upstream of the *WT1* gene and these two genes are bi-directionally transcribed from the same promoter region.

KLK7: *KLK7* has been considered in [7] for differentiating between ovarian cancer cell line OVCAR-3 and PEO1 and normal epithelial cells. This gene is expressed in serous and not in the rest of the tumors.

PRLR: *PRLR* is found to be significantly related to breast cancer and this gene should also be experimentally investigated as a OS marker.

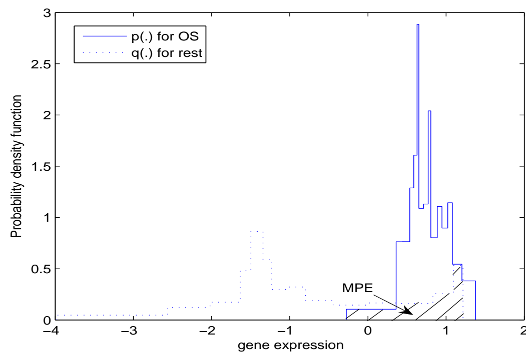

Figure 5: The probability density function $p(.)$ corresponding to the gene *WT1* of the OS samples' is shown in solid line; and the same for the rest of the samples is shown in dashed line.

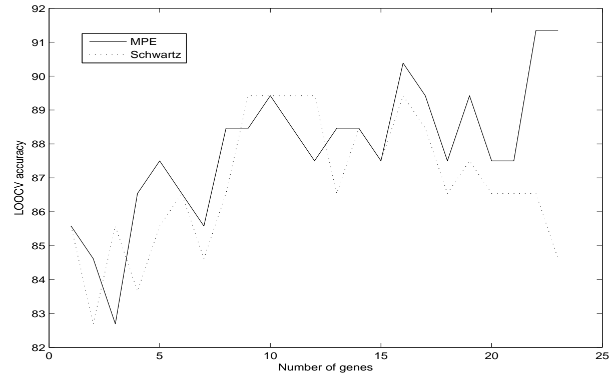

Figure 6: The accuracy of classification using LOOCV for OS Vs. rest. The solid (dashed) line gives the accuracy for genes selected by MPE (from [17]).

In Figure 6, we show the comparison of the genes selected by MPE and those from [17] using LOOCV for OS vs. rest. Notice that the maximum accuracy of 91.35% in this case is lower than the previous two cases. This is in tune with the belief that there may be some serous tumors which are similar to mucinous and clear cell tumors, while there is another set of serous tumors which are very similar to OE tumors. This is also reflected by the hierarchical clustering performed on this dataset (data not shown). However, among the 23 genes considered in each case, we found 9 genes which are found by us and [17]. Finally, by the correlation-based method as well, *WT1* was selected as the best gene for OS tumors with $\text{rank}_{WT1} = 0.67$. Here, it is interesting to note that among the genes selected by MPE, there are two genes *KLK7* and *LU*, which are also considered as markers to distinguish serous ovarian tumors from the normal ovarian samples (see [1]). Both these genes are relatively underexpressed in OC and OM, but overexpressed in both OS and OE tumors.

4 Discussion

For a long time, pathologists have attempted to classify tumors into biologically and clinically meaningful categories. This has been particularly challenging for ovarian tumors, given their morphological heterogeneity.

The contribution of this paper is in ranking the genetic markers specific for clear cell, mucinous and serous ovarian tumors. We show that (a) *HNF1-beta* and *GGT1* are the best genes for classifying OC tumors; (b) *CEACAM6* is the best gene for distinguishing OM from other OECs; and (c) *WT1* is best for OS tumors. Other feature selection methods could not rank *HNF1-beta* and *GGT1* as the best markers for OC. Also, only our method found *CEACAM6* to be the best marker for OM.

Notice that the tumors which are not classified as OC, OM or OS, by a classifier, will be considered as OE tumors. However, a similar gene selection for OE tumors results in a list where the best gene for the separation of OE from other EOCs has an MPE of 0.211. Related works also noted the fact that OE tumors are highly heterogeneous and contains further subtypes. So, we considered an unsupervised clustering of the 104 samples, where we found that OS and OE tumors are further subdivided in two parts, whereas OC and OM tumors mostly tend to be clustered together (data not shown).

Thus, our study offers additional support to the fact that ovarian cancer represents a group of distinct, but related diseases. Finally, further research is necessary for reliable integration of datasets from various microarray platforms to get a statistically significant study of all EOCs.

References

[1] Adib, T.R., Henderson, S., Perrett, C., Hewitt, D., Bourmpoulia, D., Ledermann, L., and Boshoff, C., Predicting biomarkers for ovarian cancer using gene-expression microarrays, *Br. J. Cancer*, 90(3):686–692, 2004.

[2] Aho, S. and J.Uitto, 180-kd bullous pemphigoid antigen/type xvii collagen: Tissue-specific expression and molecular interactions with keratin 18, *J. Cell. Biochem.*, 72(3):356–367, 1999.

[3] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yanq, L., Marti, G.E., Moore, T., Hudson, J. Jr, Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warne, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., and Staudt, L.M., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403(6769):503–511, 2000.

[4] Azuma, T., Hirai, M., Ito, S., Yamamoto, K., Taggart, R.T., Matsuba, T., Yasukawa, K., Uno, K., Hayakumo, T., and Nakajima, M., Expression of cathepsin e in pancreas: A possible tumor marker for pancreas, a preliminary report, *Int. J. Cancer*, 67(4):492–497, 1996.

[5] Bast, R.C. Jr, Klug, T.L., St John, E., Jenison, E., Niloff, J.M., Lazarus, H., Berkowitz, R.S., Leavitt, T., Griffiths, C.T., Parker, L., Zurawski, V.R. Jr, and Knapp, R.C., A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer, *N. Engl. J. Med.*, 309(15):883–887, 1983.

[6] Berchuck, A., Iversen, E.S., Lancaster, J.M., Pittman, J., Luo, J., Lee, P., Murphy, S., Dressman, H.K., Febbo, P.G., West, M., Nevins, J.R., and Marks, J.R., Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers, *Clin. Cancer Res.*, 11(10):3686–3696, 2005.

[7] Dong, Y., Kaushal, A., Brattsand, M., Nicklin, J., and Clements, J.A., Differential splicing of klk5 and klk7 in epithelial ovarian cancer produces novel variants with potential as cancer biomarkers, *Clin. Cancer Res.*, 9(5):1710–1720, 2003.

[8] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloompeld, C.D., and Lander, E.S., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, 286(5439):531–537, 1999.

[9] Hanigan, M.H., Frierson, H.F. Jr, Brown, J.E., Lovell, M.A., and Taylor, P.T., Human ovarian tumors express gamma-glutamyl transpeptidase, *Cancer Res.*, 54(1):286–290, 1994.

[10] Harlozinska-Szmyrka, A., Slesak, B., Richter, R., Rabczynski, J., and Cislo, M., Immunological reactivity of mucinous and serous ovarian adenocarcinomas, *Cancer Detect. Prev.*, 9(3–4):347–358, 1986.

[11] Menon, U. and Jekob, I.J., Ovarian cancer screening in the general population: Current status, *Int. J. Gynecol. Cancer*, 11:3–6, 2001.

[12] Netinatsunthorn, W., Hanprasertpong, J., Dechsukhum, C., Leetanaporn, R., and Geater, A., Wt1 gene expression as a prognostic marker in advanced serous epithelial ovarian carcinoma: An immunohistochemical study, *BMC Cancer*, 6(90), 2006.

[13] Okuda, T., Saito, H., Sekizawa, A., Shimizu, Y., Akamatsu, T., Kushima, M., Yanaihara, T., Okai, T., and Farina, A., Steroid sulfatase expression in ovarian clear cell adenocarcinoma: Immunohistochemical study, *Gynecol. Oncol.*, 82(3):427–434, 2001.

[14] Peters-Engl, C., Medl, M., Ogris, E., and Leodolter, S., Tumor-associated trypsin inhibitor (tati) and cancer antigen 125 (ca125) in patients with epithelial ovarian cancer, *Anticancer Res.*, 15(6B):2727–2730, 1995.

[15] Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., and Liotta, L.A., Use of proteomic patterns in serum to identify ovarian cancer, *Lancet North Am. Ed.*, 359(9306):572–577, 2002.

[16] Schaner, M.E., Ross, D.T., Ciaravino, G., Sørlie, T., Troyanskaya, O., Diehn, M., Wang, Y.C., Duran, G.E., Sikic, T.L., Caldeira, S., Skomedal, H., Tu, I.P., Hernandez-Boussard, T., Johnson, S.W., O'Dwyer, P.J., Fero, M.J., Kristensen, G.B., Børesen-Dale, A., Hastie, T., Tibshirani, R., van de Rijn, M., Teng, N.N., Longacre, T.A., Botstein, D., Brown, P.O., and Sikic, B.I., Gene expression patterns in ovarian carcinomas, *Mol. Biol. Cell*, 14(11):4376–4386, 2003.

[17] Schwartz, D.R., Kardia, S.L.R., Shedden, K.A., Kuick, R., Michailidis, G., Taylor, J.M.G., Misek, D.E., Wu, R., Zhai, Y., Darrah, D.M., Reed, H., Ellenson, L.H., Giordano, T.J., Fearon, E.R., Hanash, S.M., and Cho, K.R., Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas, *Cancer Res.*, 62(16):4722–4729, 2002.

[18] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, CRC Press, 1986.

[19] Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, H., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Eystein Lonning, P., and Børresen-Dale, A.L., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proc. Natl. Acad. Sci. USA*, 98(19):10869–10874, 2001.

[20] Tsuchiya, A., Sakamoto, M., Yasuda, J., Chuma, M., Ohta, T., Ohki, M., Yasugi, T., Taketani, Y., and Hirohashi, S., Expression profiling in ovarian clear cell carcinoma: identification of hepatocyte nuclear factor-1 beta as a molecular marker and a possible molecular target for therapy of ovarian clear cell carcinoma, *Am. J. Pathol.*, 163(6):2503–2512, 2003.

[21] Vapnik, V.N., *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[22] Vêncio, R.Z.N., Patrão, D.F.C., Baptista, C.S., Pereira, C.A.B., and Zingales, B., Bayboots: A model-free Bayesian tool to identify class markers from gene expression data, *Genet. Mol. Res.*, 5(1):138–142, 2006.

[23] Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J., Lockhart, D.J., Burger, R.A., and Hampt, G.M., Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer, *Proc. Natl. Acad. Sci. USA*, 98(3):1176–1181, 2001.