Combining Nomogram and Microarray Data for Predicting Prostate Cancer Recurrence

Yijun Sun, Yunpeng Cai, and Steve Goodison

Abstract— **The derivation of molecular signatures indicative of disease status and behavior are required to facilitate the optimal choice of treatment for prostate cancer patients. We conducted a computational analysis of gene expression profile data obtained from 79 cases, 39 of which were classified as having disease recurrence, to investigate whether an advanced computational algorithm can derive more accurate prognostic signatures for prostate cancer. At the 90% sensitivity level, a newly derived genetic signature achieved 85% specificity. This is the first reported genetic signature to outperform a clinically used postoperative nomogram. Furthermore, a hybrid signature derived by combination of the nomogram and gene expression data significantly outperformed both genetic and clinical signatures, and achieved a specificity of 95%. Our study demonstrates the possibility of utilizing both genetic and clinical information for highly accurate prostate cancer prognosis beyond the current clinical systems, and shows that more advanced computational modeling of microarray and clinical data is warranted before clinical application of predictive signatures is considered.**

I. INTRODUCTION

Prostate cancer is the most common male cancer by incidence, and the second most common cause of male cancer death in the United States. In 2008, it is estimated that approximately 186,320 new cases will be diagnosed and 28,660 men will die from this disease. The mortality rate for prostate cancer is declining due to improvements in earlier detection and in local therapy strategies. However, the ability to predict the metastatic behavior of a patient's cancer, as well as to detect and eradicate disease recurrence remains some of the greatest clinical challenges in oncology. It is estimated that 25-40% of men undergoing radical prostatectomy will have disease relapse, often termed a biochemical recurrence, as the first clinical indication a rising serum level of prostate specific antigen (PSA) [1]. The accurate identification of patients at risk for relapse would greatly facilitate the rational application of adjuvant treatment strategies.

Accurate prediction models based on standard clinical variables already exist for prostate cancer recurrence after radical prostatectomy [2]. A postoperative nomogram [3] is one of the most frequently used tools in current clinical settings. It predicts prostate cancer progression by estimating 5 and 7-year progression-free probability (PFP) after radical prostatectomy based on serum PSA, Gleason grade, surgical margin status, and pathologic stage. Though well calibrated and repeatedly validated, the nomogram performs only slightly better than mid-way between a model with perfect discrimination and one with no discrimination. Yet, to date, no single biomarker, nor any prognostic molecular models based on high-throughput gene expression analysis, has been able to significantly improve upon the predictive accuracy of the postoperative nomogram [4], [5].

The advent of microarray gene expression technology has greatly enabled the search for predictive disease biomarkers in the past few years. Numerous exploratory studies have demonstrated the potential value of gene expression signatures in assessing the risk of post-surgical disease recurrence beyond the current clinical systems [6], [7], [8], [9], [11]. However, most of studies focused on breast cancer prognosis [11]. Moreover, many existing predictive models were derived using relatively simple computational algorithms, and the critical issue of whether existing gene signatures are ready for randomized, prospective clinical validation trials is still under debate in the oncology community [12]. The key to resolving the issue is the development of advanced computational algorithms, particularly feature selection algorithms that are capable of identifying relevant genes from tens of thousands genes on the basis of a limited number of patient tissue samples. However, feature selection for high-dimensional data still remains one of the major challenges in statistical machine learning [13]. This seriously undermines the performance of many currently used data analysis algorithms in terms of their speed and accuracy, and represents a major obstacle in translating predictive models established in exploratory studies to clinical applications. We have recently derived a new feature selection algorithm that addresses several major issues with prior work. The algorithm performs remarkably well in the presence of a huge number of irrelevant features. It allows one to process many thousands of features within a few minutes on a personal computer, yet maintaining a very high accuracy that is nearly insensitive to a growing number of irrelevant features [14].

In this study, we conducted a computational analysis to investigate whether the application of our computational algorithms can lead to the derivation of more accurate prognostic molecular signatures for predicting prostate cancer recurrence. We analyzed a prostate tissue gene expression dataset established at MSKCC [4], and used a rigorous experimental protocol to compare the prognostic performance

Manuscript received July 5, 2008. This work is supported in part by the Komen Breast Cancer Foundation under grant No. BCTR0707587.

Y. Sun is with the Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL 32611. sunyijun@biotech.ufl.edu

Y. Cai is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611. caiyp@u.edu

S. Goodison is with the Department of Surgery, University of Florida, Jacksonville, FL 32009. steve.goodison@jax.ufl.edu

of a newly identified genetic signature with those previously derived. Receiver operator characteristic (ROC) curves and survival data analyses demonstrate the superior performance of the new signature over previous work, and suggest that the application of this approach to large-scale cohort studies may lead to the derivation of prognostic prostate cancer signatures that are worthy of clinical validation trials. We further derived a hybrid prognostic signature by integrating gene expression data and clinical variables that significantly outperformed both the gene signature and nomogram. Our results demonstrate that advanced computational modeling can significantly improve the accuracy of prognostic signatures for prostate cancer, advocating the notion that more computational analysis of microarray and clinical data is warranted before clinical trials of predictive signatures are considered.

II. MATERIALS AND METHODS

A. Dataset

We analyzed the gene expression and clinical dataset used in the study published by [4]. The data set was built from tissue samples obtained from 79 patients with clinically localized prostate cancer treated by radical prostatectomy at MSKCC between 1993 and 1999. Thirty-nine cases had disease recurrence as classified by 3 consecutive increases in the serum level of PSA after radical prostatectomy, and forty samples were classified as non-recurrent samples by virtue of maintaining an undetectable PSA ($<$ 0.05 ng/mL) for at least 5 years after radical prostatectomy. No patient received any neo-adjuvant or adjuvant therapy before documented disease recurrence. The complete clinical characteristics of the 79 primary tumors are listed in [4]. Samples were snap frozen, examined histologically and enriched for neoplastic epithelium by macrodissection. Gene expression analysis was carried out using the Affymetrix U133A human gene array which has 22,283 features for individual gene/EST clusters, as per manufacturers instructions. Image processing was performed using Affymetrix Microarray Suite 5.0 to produce cel files which were used directly in our analyses.

B. Feature Selection Algorithm

High-throughput microarray technologies now routinely produce data sets with an unprecedented number of genes characterizing each patient sample, which greatly facilitates the search for predictive disease biomarkers through multivariate data analyses. However, it also poses a serious challenge to existing learning algorithms. With a limited number of patient samples, a learning algorithm can easily overfit training data, resulting in an over-optimistic or even zero training error, but with a poor generalization performance on unseen test data. A commonly used practice is to perform feature selection to identify a small fraction of genes that drive cancerous tumor growth and/or spread [15], [16]. Many existing feature selection algorithms rely on a heuristic combinatorial search (e.g., forward and backward selection), which has no guarantee of any optimality. In the presence of tens of thousands of irrelevant genes, computational complexity becomes a serious issue, and even a heuristic search becomes computationally not feasible. For this reason, many gene identification algorithms resort to filter methods that evaluate genes individually based on some information-theoretic measures such as Fisher score and p-value of t-test (see, for example, [6], [7]). Although filter methods work well for exploratory purposes, the obtained gene signatures are far from optimal for clinical applications.

We recently derived a new feature selection algorithm that addresses several major issues with prior work, including problems with computational complexity, solution accuracy, and capability to handle problems with extremely large data dimensionality [14]. The key idea of the algorithm is to decompose an complex model into a set of locally linear ones through local learning, and then estimate feature relevance globally within a large margin framework with ℓ_1 regularization. We below present a brief review of the algorithm. We start by defining the margin. Suppose we have a training dataset consisting of N samples, each represented by J features. Given a distance function, we find two nearest neighbors of each sample $\mathbf{x}_n$, one from the same class (called *nearest hit* or NH), and the other from the different class (called *nearest miss* or NM) [17]. The margin of $\mathbf{x}_n$ is then defined as $\rho_n = d(\mathbf{x}_n, \mathrm{NM}(\mathbf{x}_n)) - d(\mathbf{x}_n, \mathrm{NH}(\mathbf{x}_n))$, where $d(\cdot)$ is the distance function. For the purpose of this paper, we use the block distance to define a sample's margin and nearest neighbors, while other standard definitions may also be used. An intuitive interpretation of this margin is a measure as to how much $\mathbf{x}_n$ can "move" in the feature space before being misclassified. By the large margin theory [18], a classifier that minimizes a margin-based error function usually generalizes well on unseen test data. One natural idea then is to scale each feature, and thus obtain a weighted feature space, parameterized by a nonnegative vector $\mathbf{w}$, so that a margin-based criterion function in the *induced* feature space is maximized. The margin of $\mathbf{x}_n$, computed with respect to $\mathbf{w}$, is given by:

$$\rho_n(\mathbf{w}) = d(\mathbf{x}_n, \mathrm{NM}(\mathbf{x}_n)|\mathbf{w}) - d(\mathbf{x}_n, \mathrm{NH}(\mathbf{x}_n)|\mathbf{w}) \ . \quad (1)$$

By defining $\mathbf{z}_n = |\mathbf{x}_n - \mathrm{NM}(\mathbf{x}_n)| - |\mathbf{x}_n - \mathrm{NH}(\mathbf{x}_n)|$, where $|\cdot|$ is an element-wise absolute operator, $\rho_n(\mathbf{w})$ can be simplified as $\rho_n(\mathbf{w}) = \mathbf{w}^T\mathbf{z}_n$, which is a linear function of $\mathbf{w}$. By construction, the magnitude of each element of $\mathbf{w}$ reflects the relevance of the corresponding feature in a learning process. Note that the margin thus defined requires only information about the neighborhood of $\mathbf{x}_n$, while no assumption is made about the underlying data distribution. This implies that by local learning we can transform an arbitrary nonlinear problem into a set of locally linear ones.

The main problem with the above margin definition, however, is that the nearest neighbors of a given sample are unknown before learning. To account for the uncertainty in defining local information, we develop a probabilistic model where the nearest neighbors of each sample are treated as latent variables. We define a probability that sample $\mathbf{x}_i$ is the nearest hit or miss of $\mathbf{x}_n$, $P(\mathbf{x}_i{=}\mathrm{NH}(\mathbf{x}_n)|\mathbf{w})$ or $P(\mathbf{x}_i{=}\mathrm{NM}(\mathbf{x}_n)|\mathbf{w})$, depending on the class to which

$\mathbf{x}_i$ belongs. Following the principles of the expectation-maximization algorithm, we estimate the margin through taking the expectation of $\rho_n(\mathbf{w})$ by averaging out the latent variables:

$$\begin{aligned} \bar{\rho}_n(\mathbf{w}) &= \mathbf{w}^T \Big(\textstyle\sum_{i \in \mathcal{M}_n} P(\mathbf{x}_i{=}\text{NM}(\mathbf{x}_n)|\mathbf{w}) |\mathbf{x}_n - \mathbf{x}_i| - \\ &\qquad \textstyle\sum_{i \in \mathcal{H}_n} P(\mathbf{x}_i{=}\text{NH}(\mathbf{x}_n)|\mathbf{w}) |\mathbf{x}_n - \mathbf{x}_i| \Big) \\ &= \mathbf{w}^T \bar{\mathbf{z}}_n \ , \end{aligned} \tag{2}$$

where $\mathcal{M}_n$ contains all samples that have a different label from $\mathbf{x}_n$, and $\mathcal{H}_n$ contains all samples that have the same label as $\mathbf{x}_n$, excluding $\mathbf{x}_n$. The probabilities $P(\mathbf{x}_i{=}\text{NH}(\mathbf{x}_n)|\mathbf{w})$ and $P(\mathbf{x}_i{=}\text{NM}(\mathbf{x}_n)|\mathbf{w})$ are estimated through the standard kernel density estimation method:

$$P(\mathbf{x}_i{=}\text{NM}(\mathbf{x}_n)|\mathbf{w}) = \frac{k(\|\mathbf{x}_n - \mathbf{x}_i\|_\mathbf{w})}{\sum_{j \in \mathcal{M}_n} k(\|\mathbf{x}_n - \mathbf{x}_j\|_\mathbf{w})}, \forall i{\in}\mathcal{M}_n \tag{3}$$

$$P(\mathbf{x}_i{=}\text{NH}(\mathbf{x}_n)|\mathbf{w}) = \frac{k(\|\mathbf{x}_n - \mathbf{x}_i\|_\mathbf{w})}{\sum_{j \in \mathcal{H}_n} k(\|\mathbf{x}_n - \mathbf{x}_j\|_\mathbf{w})}, \forall i{\in}\mathcal{H}_n, \tag{4}$$

where $k(\cdot)$ is a kernel function. Specifically, we use exponential kernel $k(d){=}\exp(-d/\delta)$ where kernel width δ determines the resolution at which the data is locally analyzed.

Once the margins are defined, the problem of learning feature weights can be directly solved within the large margin framework. For computational convenience, we perform the estimation within the logistic regression formulation [19]. In molecular classification, we expect that most of genes are irrelevant. To encourage the sparseness, one commonly used strategy is to add ℓ_1 penalty of $\mathbf{w}$ to an objective function, which leads to the following optimization problem:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} \log\left(1 + \exp(-\mathbf{w}^T \bar{\mathbf{z}}_n)\right) + \lambda \|\mathbf{w}\|_1 \ , \tag{5}$$

subject to $w_j \geq 0, 1 \leq j \leq J$, where λ is a parameter that controls the sparseness of the solution. The nonnegative constraint on $\mathbf{w}$ can be absorbed into the objective function by setting $w_j = v_j^2$, which yields an unconstrained optimization problem, whose solution can be readily found through gradient descent with a simple update rule:

$$v_j \leftarrow v_j - \eta \left(\lambda - \sum_{n=1}^{N} \frac{\exp(-\sum_j {v_j}^2 \bar{z}_n(j))}{1 + \exp(-\sum_j {v_j}^2 \bar{z}_n(j))} \bar{z}_n(j) \right) v_j \ , \tag{6}$$

where η is the learning rate determined by the standard line search. It can be shown that, for fixed $\bar{\mathbf{z}}_n$, the solution obtained when the gradient vanishes is a global minimizer, given a nonzero initial point $v_j^{(0)}$.

Since $\bar{\mathbf{z}}_n$ implicitly depends on $\mathbf{w}$ through the probabilities $P(\mathbf{x}_i{=}\text{NH}(\mathbf{x}_n)|\mathbf{w})$ and $P(\mathbf{x}_i{=}\text{NM}(\mathbf{x}_n)|\mathbf{w})$, we use a fixed-point recursion method that alternatively refines the estimates of the probabilities and feature weights until convergence. By using the Banach fixed point theorem, it can be proved that our algorithm converges to a *unique* solution for any nonnegative initial feature weights, under a loose condition that a kernel width is sufficiently large [14].

Compared with existing methods, our algorithm has several nice properties. First, it avoids any heuristic combinatorial search, and allows one to process many thousands of features within a few minute on a personal computer. Second, the algorithm has two levels of regularization, i.e., the implicit leave-one-out and explicit ℓ_1 regularization, thereby ensuring a good generalization capability of the classifier constructed using selected features on unseen test samples. Third, unlike many existing methods, ours has a strong theoretical foundation. Our theoretical analysis suggested that the algorithm have a logarithmical sample complexity with respect to the input data dimensionality. That is, the number of samples needed for maintaining the same level of learning accuracy grows only *logarithmically* with the data dimensionality. This property makes the algorithm very suitable for microarray data analysis. We have conducted a large-scale experiment on a wide variety of synthetic and real-world data sets that demonstrated that the algorithm can achieve close-to-optimal solutions in the presence of many thousands of irrelevant features. Due to space limitation, many technical details are omitted. Interested reader may refer to [14] for detailed discussion of the algorithm.

C. Experimental Procedure

To avoid possible overfitting of a computational model to training data, we used an experimental protocol with the leave-one-out cross validation (LOOCV) method to estimate classifier parameters and prediction performance. The experimental protocol consists of inner and outer loops. In the inner loop, LOOCV is performed to estimate the optimal classifier parameters based on the training data provided by the outer loop, and in the outer loop, a held-out sample is classified using the best parameters from the inner loop. The experiment is repeated until each sample has been tested. The held-out testing sample is not involved in any stage of the training process. The classification parameters that need to be specified in the inner loop include the kernel width and regularization parameter of the feature selection algorithm, as well as the structural parameters of a classifier, which leads to a multi-dimensional parameter search. To make the experiment computationally feasible, we adopted some heuristic simplifications. Linear discriminant analysis (LDA) was used to estimate classification performances and tune the input parameters. One major advantage of LDA, compared to other classifiers (e.g., SVM and neural networks), is that LDA has no structural parameters. We predefined the kernel width as 5, and estimated the regularization parameter through LOOCV in the inner loop. In our simulation study, we found that the choice of the kernel width is not critical, and the algorithm yields nearly identical prediction performance for a large range of values for this parameter. We comment that a comprehensive parameter searching may lead to a more accurate prediction performance but with a much higher computational complexity.

Kaplan-Meier survival plots and log-rank tests were used to assess the predictive values of different prognostic approaches. The Mantel-Cox estimation of hazard ratio was performed to quantify the relative risk of biochemical recurrence in the bad-prognosis group compared with the good-prognosis group. A hazard ratio above 1.0 indicates that the patients assigned to the bad-prognosis group have a higher probability to develop disease recurrence than those in the good-prognosis group. In most microarray data analyses, the numbers of available patient samples are usually quite small, and some performance measurements (e.g., hazard ratios) are heavily influenced by the choice of a decision threshold. A receiver operating characteristic (ROC) curve obtained by varying a decision threshold provides a direct view on how a predictive approach performs at the different sensitivity and specificity levels. The specificity is defined as the probability that a patient who did not experience disease recurrence was assigned to the good-prognosis group, and the sensitivity is the probability that a patient who developed disease recurrence was in the bad-prognosis group. The most frequently used criterion for comparing multiple ROC curves is the area under a ROC curve, commonly denoted as AUC, which can range from 0.5 (no discrimination) to 1.0 (perfect ability to discriminate). MedCalc version 8.0 (MedCalc Software, Mariakerke, Belgium) was used to perform the ROC curve analysis. A p-value of 0.05 is considered statistically significant.

III. RESULTS

We developed two computational models to predict the biochemical recurrence of prostate cancer. The first model is based exclusively on gene expression data obtained from tissue samples, and the second combines the predictive information of both genetic and clinical variables. Specifically, in the latter combination (or hybrid) model we used as clinical variable the 7-year probability of disease recurrence estimated by the postoperative nomogram.

ROC curve analysis was performed to compare the prediction performance of the two novel prognosis models and the nomogram (Fig. 1). The nomogram performed reasonably well, consistent with multiple studies reported in the literature [3], but the genetic model predicted disease recurrence more accurately than the nomogram, specifically in the high specificity region. At the 90% sensitivity level, the genetic signature correctly classified 69 out of 79 samples (87%), including 34 non-recurrent and 35 recurrent tumors. To our knowledge, this is the first reported genetic signature in the literature that outperforms the clinically used predictive nomogram. Furthermore, a hybrid signature derived by combining the gene expression data with clinical information outperformed both the nomogram and the genetic signature. At the 90% sensitivity level, the hybrid signature improved the specificities of the genetic model and nomogram by about 10% and 20%, respectively (Table I). It correctly classified 74 out of 79 samples (94%), including 38 non-recurrent and 36 recurrent tumors. Statistical analysis of the ROC curves using MedCalc Software revealed the predictive accuracy of the hybrid signature to be significantly superior to that of the postoperative nomogram ($p\text{-value} < 0.0001$) and the gene-expression model ($p\text{-value} < 0.05$). The odds ratio (OR) of the hybrid and genetic models, reported in Table I, show that the patients assigned to the bad-prognosis group are 18.2 (95% CI: 5.9- 56.2) and 16.5 (95% CI: 5.4 - 51.0) times more likely to develop disease recurrence than those assigned to the good-prognosis group, respectively, which is much higher than that of the nomogram (8.4, 95% CI: 2.9 - 24.6).

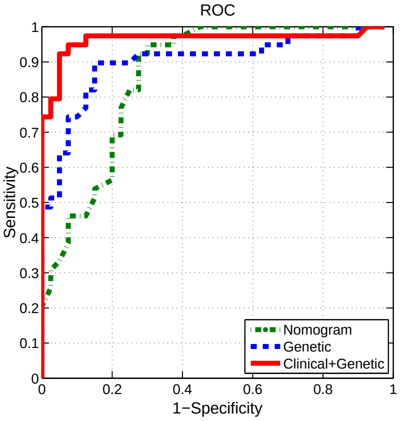

Fig. 1. Receiver operating characteristic (ROC) plot comparing the prediction performance of the nomogram, genetic and hybrid (combination of nomogram and genetic) models.

To further demonstrate the predictive value of the three approaches in assessing the risk of biochemical recurrence in prostate cancer patients, survival data analyses were performed. The Kaplan-Meier curve of the hybrid model, plotted in Fig. 2, shows a significant difference in the probability of remaining free of disease recurrence in patients with good or bad prognosis ($p\text{-value} < 0.001$). The Mantel-Cox estimate of hazard ratio of biochemical recurrence of prostate cancer within five years for the hybrid model is 29.1 (95% CI: 8.3 - 102.1), which is much larger than those of either the nomogram (11.9, 95% CI: 3.8 - 36.9) or the genetic model (18.0, 95% CI: 5.9 - 54.6). At the 5-year end point, all three approaches have similar low relapse rates in patients with good prognosis, but the patients assigned to the bad-prognosis group by the hybrid model have a much lower probability of remaining free of disease recurrence (0.21, 95% CI: 0.12 - 0.40) than that determined by the nomogram (0.35, 95% CI: 0.22 - 0.50) .

We also performed an experiment to compare the prediction performance of our algorithm with those obtained by using SVM-RFE [20] and ℓ_1 regularized logistic regression [21]. Both algorithms can perform classification directly. The results, reported in Fig. 3, show that the two competing algorithms perform worse than our algorithm. However, the results suggest that combining the nomogram with genetic information can indeed improve the prediction performance.

With a small sample size, in each iteration in LOOCV, the derived computational model may generate a different prognostic signature since the training data used is different. In the genetic modeling approach, a 5, 6, 7 and 8-gene model was developed in 7, 43, 24 and 5 iterations, respectively. A total of 11 genes were identified in the genetic prognostic

TABLE I

Prediction results of the nomogram, genetic and hybrid predictive models. The specificity was computed at the 90% sensitivity level.

Methods	AUC(95% CI)	Specificity	Odd Ratio (95% CI)	Hazard Ratio (HR) HR (95% CI)	 p-value
Nomogram	0.86 (0.77 - 0.93)	73%	8.4 (2.9 - 24.6)	11.9 (3.8-36.9)	< 0.001
Genetic	0.90 (0.81 - 0.96)	85%	16.5 (5.4 - 51.0)	18.0 (5.9 - 54.6)	< 0.001
Hybrid	0.96 (0.90 - 0.99)	95%	18.2 (5.9 - 56.2)	29.1 (8.3 - 102.1)	< 0.001

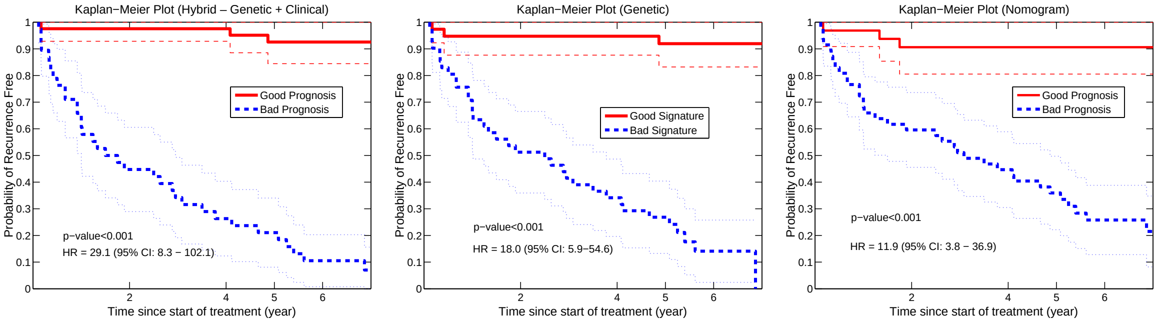

Fig. 2. Kaplan-Meier estimation of the probabilities of remaining free of disease recurrence for patients with good or bad prognosis. The "Genetic" signature was derived from gene expression data, the "Nomogram" is an existing clinical prediction model, and the "Hybrid" signature was the combination of both. The p-values were computed by log-rank test.

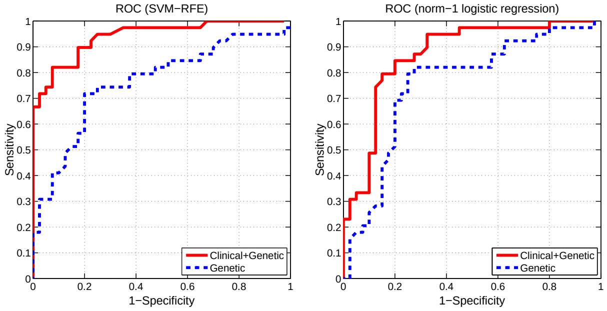

Fig. 3. ROC curves obtained by using SVM-RFE and norm-1 regularized logistical regression.

signature (Table II). The observed pattern (under- or over-expressed) in the recurrent cases for each gene, and the frequency of occurrence of each gene over 79 algorithm iterations, are listed in Table II. A high occurrence rate is an indication of the importance of the corresponding gene for predicting disease recurrence. In the hybrid modeling approach, the nomogram output was selected in all 79 iterations, and 4, 5 and 6 genes were identified in 69, 9, and 1 iteration(s), respectively. A total of 5 different genes were included in the hybrid models. Notably, all of these genes were also present in the genetic model, and three genes (PAK3, RPL23, and EI24) occurred at a high frequency in both the genetic and hybrid models (Table II).

IV. Discussion

The application of our feature selection algorithm to the MSKCC dataset enabled us to derive a genetic signature that predicts disease recurrence after radical prostatectomy with 87% overall accuracy. Furthermore, a hybrid signature derived by combining the gene expression data with the 7-year PFP score outperformed both the nomogram and the genetic signature, correctly classifying 74 out of 79 samples. Statistical analyses also clearly demonstrated the superiority of the hybrid signature over a prognostic system that uses only genetic or clinical markers. These data confirm the previous finding that the nomogram and gene expression models can provide complementary information for predicting biochemical recurrence of prostate cancer [4]. Though the nomogram performs very well when the estimated 7-year disease prognosis-free probability is larger than 90%, it assigns a significant number of non-recurrence patients to the bad prognosis group. It is evident in Fig. 4 that microarray data provides additional information to stratify these patients. While it is clear that the hybrid signature performs very well thus far, we should emphasize that in many cases clinical data is not available, or is not consistent across institutions, and thus it is important that the optimal genetic signatures are also pursued.

Three genes that were most highly weighted in both the genetic and hybrid signatures were RPL23, EI24, and PAK3. RPL23 is a member of the ribosomal protein family that acts to stabilize rRNA structure, regulate catalytic function, and integrate translation with other cellular processes, but recent studies have shown that many ribosomal proteins (RPs) have extra-ribosomal cellular functions independent of protein biosynthesis. A potential role for RPs in carcinogenesis and tumor progression is being founded on studies that have implicated ribosomal proteins not only as targets of tumor suppressors or proto-oncogenes, but also as more direct mediators of aspects of tumor progression [24]. RPL23 has been shown by Dai et al. [25] to be part of a multi-protein

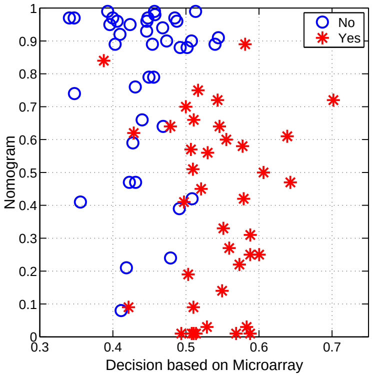

Fig. 4. Scatter plot of the prediction results obtained by using the nomogram and the genetic model. This demonstrates that the genetic and clinical markers contain complementary information in assessing the risk of a patient developing biochemical disease recurrence.

complex that regulates the activity of the oncoprotein HDM2 (human MDM2), a protein that is frequently over-expressed in various human carcinomas, soft tissue sarcomas, and other cancers [26]. HDM2 interacts with several growth suppressors and other proteins, including the tumor suppressor p53, the retinoblastoma susceptibility gene product Rb, and the growth suppressor p14, so any shift in the availability of HDM2 could lead to significant alterations of cellular phenotype. Etoposide induced gene 24 (EI24) is a p53-induced gene (PIG) that is located in chromosomal region 11q23-24 shown to be often mutated or deleted in solid tumors, including prostate [27]. EI24/PIG8 is localized in the endoplasmic reticulum (ER), and by virtue of its binding Bcl-2, has been linked with the modulation of apoptosis [28]. EI24 is a direct target of p53 transcriptional activation and is thought to involve in the formation of reactive oxygen species [29]. Perturbation of either of these mechanisms by changes in EI24 expression may contribute to prostate cancer progression. PAK3 is a Group I member of the p21-activated kinase (Pak) family serine/threonine protein kinases that bind to and modulate the activity of the small GTPases, Cdc42 and Rac. GTPase signaling controls many aspects of cellular response to the environment, and through these interactions, PAKs have been shown to be involved in the regulation of cellular processes such as gene transcription, cell morphology, motility, and apoptosis [30]. Interestingly, it has been revealed that one PAK family member is able to inhibit androgen receptor (AR) responsiveness, a critical function in prostate cells, by regulating nuclear translocation of the AR and thus preventing specific transcriptional responses [31]. There is growing evidence for a pivotal role of GTPases in tumor progression [32], [33], and is noteworthy that another of the 11 genes in the genetic prognostic signature is a GTPase-activating protein, named RICS, that also acts on Cdc42 and Rac [34]. The potential roles of these genes in prostate cancer progression deserve further investigation.

As well as an impact on clinical decision-making, it is hoped that microarray data will advance our understanding of cancer biology, which in turn will inform the development of new and effective therapies. The fact that diagnostic and prognostic signatures reported to date have been composed of tens or hundreds of genes means that the choosing of genes to study functionally remains difficult and somewhat arbitrary. A major advantage of our deriving accurate prognostic signatures comprising just a few genes greatly facilitates the task of functional investigation. The number of genes was further reduced to 5 in our clinical/genetic hybrid signature, and it is notable that all 5 genes were also amongst the 11 genes comprising the genetic signature. This was not necessarily to be expected, because the analysis used to derive the hybrid signature was not in any way informed by the genetic signature analysis. While they used the same raw data, the two signatures were derived entirely independently.

The derivation of disease-associated molecular signatures is necessarily an ongoing, dynamic process, in which, with the inclusion of more patient samples with consistent clinical information, a prognostic signature will be continuously refined [11]. Due to biological and technical limitations, tissue-based microarray analysis may not be able to achieve 100% accuracy. Yet, the application of our advanced feature selection algorithm has brought us close to optimality in this dataset. The ROC curves of our analyses depicted in Fig. 1 show that, in this cohort, there is now very little room for improvement, suggesting that the application of this approach to large-scale cohort studies may lead to the derivation of prognostic prostate cancer signatures that are worthy of clinical validation trials.

References

[1] M. Han, A. W. Partin, C. R. Pound, J. I. Epstein, and P. C. Walsh, “Long-term biochemical disease-free and cancer-specific survival following anatomic radical retropubic prostatectomy. The 15-year Johns Hopkins experience,” *Urol. Clin. North Am.*, vol. 28, no. 3, pp. 555–565, 2001.

[2] M. L. Blute, E. J. Bergstralh, A. Iocca, B. Scherer, and H. Zincke, “Use of gleason score, prostate specific antigen, seminal vesicle and margin status to predict biochemical failure after radical prostatectomy,” *J. Urol.*, vol. 165, no. 1, pp. 119–125, 2001.

[3] M. Kattan, T. Wheeler, and P. Scardino, “Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer,” *J. Clin. Oncol.*, vol. 17, no. 5, pp. 1499–1507, 1999.

[4] A. J. Stephenson, A. Smith, M. W. Kattan, *et. al*, “Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy,” *Cancer*, vol. 104, no. 2, pp. 290–298, 2005.

[5] A. J. Stephenson, P. T. Scardino, J. A. Eastham, *et. al*, “Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy,” *J. Clin. Oncol.*, vol. 23, no. 28, pp. 7005–7012, 2005.

[6] L. J. van’t Veer, H. Dai, M. J. van de Vijver, *et. al*, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[7] Y. Wang, J. G. Klijn, Y. Zhang, *et. al*, “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,” *Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[8] E. LaTulippe, J. Satagopan, A. Smith, *et. al*, “Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease,” *Cancer Res.*, vol. 62, no. 15, pp. 4499–4506, 2002.

[9] D. Singh, P.G. Febbo, K. Ross, *et. al*, “Gene expression correlates of clinical prostate cancer behavior,” *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

TABLE II

Genes identified in genetic and hybrid (marked by ¶) predictive models. The value inside and outside of the brackets in the last column is the number of models in which a gene was selected in the hybrid and genetic models, respectively.

Gene symbol	Gene title	Mean expression in recurrent tumors	Occurrence frequencies
PAK3¶	P21 (CDKN1A)-activated kinase 3	Underexpressed	78 (79)
RPL23¶	ribosomal protein L23	Overexpressed	79 (79)
E124¶	etoposide induced 2.4 mRNA	Overexpressed	79 (79)
TGFB3¶	transforming growth factor, beta 3	Underexpressed	79 (3)
RBM34¶	RNA binding motif protein 34	Overexpressed	62 (8)
PCOLN3	procollagen (type III) N-endopeptidase	Underexpressed	78
FUT7	fucosyltransferase 7	Underexpressed	30
RICS	Rho GTPase-activating protein	Overexpressed	8
MAP4K4	mitogen-activated protein	Overexpressed	5
CUTL1	CCAAT displacement protein	Overexpressed	2
ZNF324B	zinc finger protein 324B	Underexpressed	1

[10] J. B. Welsh, L. M. Sapinoso, A. I. Su, *et. al*, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," *Cancer Res.*, vol. 61, no. 16, pp. 5974–5978, 2001.

[11] C. Sawyers, "The cancer biomarker problem," *Nature*, vol. 452, pp. 548–552, 2008.

[12] S. Loi, C. Sotiriou, M. Buyse, *et. al*, "Molecular forecasting of breast cancer: time to move forward with clinical testing," *J. Clin. Oncol.*, vol. 24, no. 4, pp. 721–722, 2006.

[13] J. Lafferty and L. Wasserman. Challenges in statistical machine learning. *Statist. Sinica*, vol. 16, pp. 307–322, 2006.

[14] Y. Sun, S. Todorovic, and S. Goodison, "A feature selection algorithm capable of handling extremely large data dimensionality," in *Proc. 8th SIAM Intl. Conf. Data Mining*, Atlanta, GA, 2008, pp. 530–540.

[15] Y. Sun, "Iterative RELIEF for feature weighting: algorithms, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1035–1051, 2007.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[17] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. 9th Intl. Conf. Mach. Learn.*, Aberdeen, Scotland, United Kingdom, 1992, pp. 249–256.

[18] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[19] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, 2002.

[21] A. Y. Ng, "Feature selection, L1 vs. L2 regularization, and rotational invariance," in *Proc. 21st Intl. Conf. Mach. Learn.*, Banff, Alberta, Canada, 2004, pp. 78–86.

[22] S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, *et. al*, "Delineation of prognostic biomarkers in prostate cancer," *Nature*, vol. 412, no. 6849, pp. 822–826, 2001.

[23] J. Luo, D. J. Duggan, Y. Chen, *et. al*, "Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling," *Cancer Res.*, vol. 61, no. 12, pp. 4683–4888, 2001.

[24] T. Kobayashi, Y. Sasaki, Y. Oshima, *et. al*, "Activation of the ribosomal protein L13 gene in human gastrointestinal cancer," *Int. J. Mol. Med.*, vol. 18, no. 1, pp. 161–170, 2006.

[25] M. S. Dai, S. X. Zeng, Y. Jin, X. X. Sun, L. David, and H. Lu, "Ribosomal protein L23 activates p53 by inhibiting MDM2 function in response to ribosomal perturbation but not to rranslation inhibition," *Mol. Cell. Biol.*, vol. 24, no. 17, pp. 7654–7668, 2004.

[26] K. Onel and C. Cordon-Cardo, "MDM2 and prognosis," *Mol. Cancer Res.*, vol. 2, no. 1, pp. 1–8, 2004.

[27] R. Dahiya, J. McCarville, C. Lee, *et. al*, "Deletion of chromosome 11p15, p12, q22, q23-24 loci in human prostate cancer," *Int. J. Cancer*, vol. 72, no. 2, pp. 283–288, 1997.

[28] X. Zhao, R. E. Ayer, S. L. Davis, *et. al*, "Apoptosis factor EI24/PIG8 is a novel endoplasmic reticulum-localized Bcl-2-binding protein which is associated with suppression of breast cancer invasiveness," *Cancer Res*, vol. 65, no. 6, pp. 2125–2129, 2005.

[29] Z. Gu, C. Flemington, T. Chittenden, and G. P. Zambetti, "ei24, a p53 response gene involved in growth suppression and apoptosis," *Mol. Cell. Biol*, vol. 20, no. 1, pp. 233–241, 2000.

[30] R. K. Vadlamudi and R. Kumar, "P21-activated kinases in human cancer," *Cancer Metastasis Rev.*, vol. 22, no. 4, pp. 385–393, 2003.

[31] N. Schrantz, J. da Silva Correia, B. Fowler, Q. Ge, Z. Sun, and G. M. Bokoch, "Mechanism of p21-activated kinase 6-mediated inhibition of androgen receptor signaling," *J. Biol. Chem.*, vol. 279, no. 3, pp. 1922–1931, 2004.

[32] E. A. Clark, T. R. Golub, E. S. Lander, and R. O. Hynes, "Genomic analysis of metastasis reveals an essential role for RhoC," *Nature*, vol. 406, no. 6795, pp. 532–535, 2000.

[33] S. Goodison, J. Yuan, D. Sloan, *et. al*, "The RhoGAP protein DLC-1 functions as a metastasis suppressor in breast cancer cells," *Cancer Res.*, vol. 65, no. 14, pp. 6042–6053, 2005.

[34] T. Okabe, T. Nakamura, Y. N. Nishimura, K. Kohu, S. Ohwada, and Y. Morishita, "RICS, a novel GTPase-activating protein for Cdc42 and Rac1, is involved in the beta-catenin-N-cadherin and N-methyl-D-aspartate receptor signaling," *J. Biol. Chem.*, vol. 278, no. 11, pp. 9920–9927, 2003.