Published in final edited form as:
Prostate. 2009 July 1; 69(10): 1119–1127. doi:10.1002/pros.20961.

Optimizing Molecular Signatures for Predicting Prostate Cancer Recurrence

Yijun Sun[1] and **Steve Goodison**[2,*]

[1]Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, Florida

[2]Department of Surgery, University of Florida, Jacksonville, Florida

Abstract

BACKGROUND—The derivation of molecular signatures indicative of disease status and predictive of subsequent behavior could facilitate the optimal choice of treatment for prostate cancer patients.

METHODS—In this study, 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 advanced computational algorithms can derive more accurate prognostic signatures for prostate cancer.

RESULTS—At the 90% sensitivity level, a newly derived prognostic genetic signature achieved 85% specificity. This is the first reported genetic signature to outperform a clinically used postoperative nomogram. Furthermore, a hybrid prognostic signature derived by combination of the nomogram and gene expression data significantly outperformed both genetic and clinical signatures, and achieved a specificity of 95%.

CONCLUSIONS—Our study demonstrates the feasibility of utilizing gene expression information for highly accurate prostate cancer prognosis beyond the current clinical systems, and shows that more advanced computational modeling of tissue-derived microarray data is warranted before clinical application of molecular signatures is considered.

Keywords

microarray; nomogram; prostate cancer prognosis; predictive model

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 (data from the National Cancer Institute). 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,

[*]Correspondence to: Dr. Steve Goodison, Department of Surgery, Shands Health Science Center, University of Florida, Jacksonville, FL 32009. steve.goodison@jax.ufl.edu.

Yijun Sun and Steve Goodison contributed equally to this work.

Additional Supporting Information may be found in the online version of this article.

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 developed by Kattan et al. [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 accuracy of the nomogram does leave room for improvement, 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. Numerous exploratory studies have demonstrated the potential value of gene expression signatures in assessing the risk of postsurgical disease recurrence beyond the current clinical systems [6–9]. However, existing molecular predictive models were derived using relatively simple computational algorithms, and the critical issue of whether proposed gene signatures are ready for randomized, prospective clinical validation trials is still under debate in the oncology community [10–12]. Key to resolving this issue is the development of advanced algorithms that are capable of identifying relevant genes (features in bioinformatic terms) in a background of tens of thousands of genes, and on the basis of a limited number of patient tissue samples. This process is known as feature selection, and achieving this in high-dimensional data remains a major challenge in bioinformatics and machine learning [13]. Current limitations in feature selection performance seriously undermine the performance of currently used data analysis algorithms in terms of their speed and accuracy, and represent a major obstacle in the translation of molecular models to clinical applications. In order to overcome some of these restraints, we have previously derived a feature selection algorithm that addresses several major issues with prior work including computational efficiency and solution accuracy. We have experimentally demonstrated that our algorithm is capable of handling problems with extremely large input data dimensionality, to a point far beyond that needed for gene expression data analysis of genetically complex organisms [14–16]. The application of our approach to breast tissue microarray data sets has enabled us to derive highly accurate prognostic molecular signatures for breast cancer [17].

In this study, we conducted a computational analysis to investigate whether the application of our computational algorithm can lead to the derivation of more accurate prognostic molecular signatures for predicting prostate cancer recurrence. To this end, we analyzed a prostate tissue gene expression data set established at the Memorial Sloan Kettering Cancer Center (MSKCC) [5], and used a rigorous experimental protocol to compare the prognostic performance of newly identified genetic signatures with those previously derived. Receiver operator characteristic (ROC) curves and survival data analyses demonstrate the superior performance of the new gene signature over previous work. We further derived a hybrid prognostic signature, obtained by integrating gene expression data and clinical variables, that significantly outperformed both the gene signature and the predictive nomogram. Our results demonstrate that advanced computational modeling can significantly improve the accuracy of molecular prognostic signatures for prostate cancer.

MATERIALS AND METHODS

Data Set

We analyzed the gene expression and clinical data set used in the study published by Stephenson et al. [5] Senior author Dr. William Gerald of MSKCC kindly provided updated clinical information for this study. 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 three consecutive increases in the serum level of PSA after radical prostatectomy, and 40 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 Stephenson et al. [5] 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 manufacturer's instructions. Image processing was performed using Affymetrix Microarray Suite 5.0 to produce cel.files, which were used directly in our analyses. In line with the majority of microarray analyses, for genes to be incorporated into the published MSKCC predictive models [5], data were filtered using several criteria that included a significant differential expression between the two classes (*P*-value <0.001), a fold change >1.3, and a "present" call in greater than 80% of the samples in either class. If feasible, it is preferable to allow a learning algorithm to decide without bias which genes are useful for prediction, without the use of any arbitrary preprocessing filters. In our study, except for a simple re-scaling of the expression values of each gene to be between 0 and 1 (see Supplementary Data), no other preprocessing was performed.

Feature Selection Algorithm

We have previously derived a feature selection algorithm that addresses several major issues with prior work, including their problems with computational complexity, solution accuracy, algorithmic implementation, and capability to handle problems with large data dimensionality [14–18]. The key idea is to decompose an arbitrary complex model into a set of locally linear ones through local learning, and then estimate feature relevance globally within a large margin framework. The algorithm is a generic feature selection method that performs without making any assumptions about the underlying data distribution. It avoids any combinatorial search, and thus allows one to process many thousands of features within 1 min on a personal computer (Figure S4 in Supplemental Data). 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-optimum solutions in the presence of thousands of irrelevant features. For details of the computational algorithm see our previous publications [14–18] and Supplementary Data Section 2. The Matlab implementation of the algorithm is available upon request for validating the reported results and academic research.

Experimental Procedure

To avoid possible overfitting of a computational model to training data, we used a rigorous experimental protocol with the leave-one-out cross validation (LOOCV) method to estimate classifier parameters and prediction performance [17,19], as depicted in Figure 1. 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 (see Supplementary Data), 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 the simulation study presented in the Supplemental Data, we demonstrated 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 (Refs. [13] and [17], and Supplemental Data).

Statistical Analysis

Kaplan–Meier survival plots and log-rank tests [20] 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 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 was considered statistically significant.

RESULTS

Using the iterative analytical approach depicted in Figure 1, we developed two computational models to predict the biochemical recurrence of prostate cancer in a cohort of 79 patients who had clinically localized prostate cancer treated by radical prostatectomy. Biochemical recurrence of disease was defined as three consecutive increases in the serum level of PSA. The first model was based exclusively on gene expression data obtained from tissue samples, and the second combined 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 clinically used postoperative nomogram [3].

ROC curve analysis was performed to compare the prediction performance of the two novel prognosis models and the nomogram (Fig. 2). The nomogram performed reasonably well, consistent with multiple studies reported in the literature [3,4], but our 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*-value <0.0001) and the gene-expression model (*P*-value <0.05). The odds ratio (OR) of the hybrid and genetic models, reported in Table I, shows 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.

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 Figure 3, shows a significant difference in the probability of remaining free of disease recurrence in patients with a good or bad prognosis (*P*-value <0.001). The Mantel–Cox estimate of hazard ratio for biochemical recurrence of prostate cancer within 5 years for the hybrid model was 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) depicted in Figure 3. At the 5-year end point, all three approaches had similar low relapse rates in patients with good prognosis, but the patients assigned to the bad-prognosis group by the hybrid model had 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).

To avoid possible overfitting of the computational model, we used the LOOCV method to estimate classifier parameters and prediction performance [17,19]. As the name suggests, LOOCV involves using microarray data from a single sample as validation data, and the data from the remaining samples as the training data. The experiment is repeated until each sample has been tested. Over the 79 iterations of LOOCV, a total of 11 genes were identified in the optimal genetic prognostic signature (Table II). The mean expression of each gene in the 79 tumor samples obtained from patients with, and without, disease recurrence was visualized by creating individual scatter plots (Figure S2 in Supplemental Data). The observed pattern (under- or over-expressed) in the recurrent cases for each gene, and the frequency of occurrence of each gene over 79 model iterations, are listed in Table II. A high occurrence rate is an indication of the relative importance of the corresponding gene for predicting disease recurrence. In the hybrid modeling approach, the nomogram output was selected in all 79 computational iterations, and 4, 5, and 6 genes were identified in 69, 9, and 1 iteration(s), respectively. A total of five different genes were included in the optimal hybrid models (Table II). 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).

DISCUSSION

A number of studies have been conducted describing the use of microarray technologies for prostate cancer diagnosis and prognosis [8,9,21–23], and the notion that molecular models can provide prediction performance close to those achieved by current clinical systems has been established [5,8,9]. However, to date, these predictive models have been derived using simple computational algorithms, and whether these approaches achieved optimal performance when using genetic information is rarely addressed in the literature. We have recently developed a new feature selection algorithm [14–17] to address limitations inherent to current microarray data analysis strategies.

While high-throughput microarray technologies greatly facilitate the search for molecular disease biomarkers through multivariate data analyses, they also pose serious challenges to existing learning algorithms. With a limited number of patient samples and high-dimensional data per sample, a learning algorithm can easily overfit training data, resulting in models with over-optimistic error rates, but with a very poor generalization performance on unseen test data—a phenomenon called the curse of dimensionality in machine learning [24,25]. As described above, one needs to perform computational feature selection to identify the small fraction of genes that potentially drive disease, in this case tumor progression. Existing feature selection algorithms rely on combinatorial searches that have no guarantee of optimality in the presence of tens of thousands of irrelevant genes, and are seriously limited by computational complexity. For this reason, many gene identification algorithms resort to filter methods that evaluate genes individually based on statistical measures such as a Fisher score and/or a *P*-value of *t*-tests [7,8]. Although filter methods can provide a working solution for exploratory purposes, the obtained gene signatures are far from optimal for clinical applications.

The application of our feature selection algorithm to the MSKCC data set enabled us to derive a genetic signature that predicts biochemical 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 [5] that the nomogram and gene expression models can provide complementary information for predicting the biochemical recurrence of prostate cancer. Though the nomogram performs very well when the estimated 7-year disease PFP is larger than 90%, it assigns a significant number of non-recurrence patients to the bad-prognosis group. It is evident in Figure 4 that microarray data provide additional information to stratify these patients. If a threshold for the probability of recurrence was applied using the nomogram data only, for example, at 0.7 on the *x*-axis of Figure 4, then several non-recurrent cases below the threshold would be wrongly classified. However, if we have plot both nomogram and microarray data and add a threshold to the decision based on microarray data, say at 0.475 on the *y*-axis of Figure 4, only a couple non-recurrent cases would be wrongly classified (below nomogram threshold and to the right of the microarray threshold). While it is clear that the hybrid signature performs extremely well thus far, we should emphasize that in many cases clinical data are not available, or are not consistent across institutions, and thus it is important that 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 (RP) family that acts to stabilize rRNA structure, regulate catalytic function, and integrate translation with other cellular processes, but recent studies have shown that many 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 RPs not only as targets of tumor suppressors or proto-oncogenes, but also as more direct mediators of aspects of tumor progression [26]. RPL23 has been shown by Dai et al. [27] to be part of a multiprotein 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 [28]. 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 [29]. 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 [30]. EI24 is a direct target of p53 transcriptional activation and is thought to be involved in the formation of reactive oxygen species [31]. 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 [32]. 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 [33]. There is growing evidence for a pivotal role of GTPases in tumor progression [34], 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 [35]. 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 [10,17,36]. 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 data set. The ROC curves of our analyses depicted in Figure 2 show that, in this cohort, there is now very little room for improvement. Although cross-validation analyses between microarray platforms, and even between institutes presents another set of problems, the findings described here suggest that the application of this computational approach to larger scale cohort studies may lead to the derivation of prognostic prostate cancer signatures that are worthy of clinical validation trials.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Dr. William L. Gerald, MD, PhD, of the Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY, for data sharing and for valued discussion. This work was supported in part by the National Cancer Institute under grant RO1 CA108597 (S.G.), and the Susan Komen Breast Cancer Foundation under grant No. BCTR0707587 (Y.S.).

References

1. Han M, Partin AW, Pound CR, Epstein JI, Walsh PC. Long-term biochemical disease-free and cancer-specific survival following anatomic radical retropubic prostatectomy. The 15-year Johns Hopkins experience. Urol Clin North Am. 2001; 28:555–565. [PubMed: 11590814]

2. Blute ML, Bergstralh EJ, Iocca A, Scherer B, Zincke H. Use of Gleason score, prostate specific antigen, seminal vesicle and margin status to predict biochemical failure after radical prostatectomy. J Urol. 2001; 165:119–125. [PubMed: 11125379]

3. Kattan MW, Wheeler TM, Scardino PT. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. J Clin Oncol. 1999; 17:1499–1507. [PubMed: 10334537]

4. Stephenson AJ, Scardino PT, Eastham JA, Bianco FJ Jr, Dotan ZA, DiBlasio CJ, Reuther A, Klein EA, Kattan MW. Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. J Clin Oncol. 2005; 23:7005–7012. [PubMed: 16192588]

5. Stephenson AJ, Smith A, Kattan MW, Satagopan J, Reuter VE, Scardino PT, Gerald WL. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. Cancer. 2005; 104:290–298. [PubMed: 15948174]

6. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science. 1999; 286:531–537. [PubMed: 10521349]

7. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–536. [PubMed: 11823860]

8. LaTulippe E, Satagopan J, Smith A, Scher H, Scardino P, Reuter V, Gerald WL. Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. Cancer Res. 2002; 62:4499–4506. [PubMed: 12154061]

9. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002; 1:203–209. [PubMed: 12086878]

10. Sawyers CL. The cancer biomarker problem. Nature. 2008; 452:548–552. [PubMed: 18385728]

11. Loi S, Sotiriou C, Buyse M, Rutgers E, Van't Veer L, Piccart M, Cardoso F. Molecular forecasting of breast cancer: Time to move forward with clinical testing. J Clin Oncol. 2006; 24:721–722. author reply 2–3. [PubMed: 16446348]

12. Brenton JD, Carey LA, Ahmed AA, Caldas C. Molecular classification and molecular forecasting of breast cancer: Ready for clinical application? J Clin Oncol. 2005; 23:7350–7360. [PubMed: 16145060]

13. Lafferty L, Wasserman J. Challenges in statistical machine learning. Stat Sinica. 2006; 16:307–322.

14. Sun, Y.; Todorovic, S.; Goodison, S. A feature selection algorithm capable of handling extremely large data dimensionality. Proceedings of the SIAM International Conference on Data Mining; 2008. p. 530-540.

15. Sun Y, Wu D. Feature extraction through local learning. Stat Anal Data Min. 2009 in press.

16. Sun Y. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. IEEE Trans Pattern Anal Mach Intell. 2007; 29:1035–1051. [PubMed: 17431301]

17. Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics. 2007; 23:30–37. [PubMed: 17130137]

18. Sun, Y.; Todorovic, S.; Goodison, S. Toward optimal feature selection through local learning. Technical Report. Available online: http://plaza.ufl.edu/sunyijun/Paper/FeatureSelection.pdf

19. Wessels LF, Reinders MJ, Hart AA, Veenman CJ, Dai H, He YD, van't Veer LJ. A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics. 2005; 21:3755–3762. [PubMed: 15817694]

20. Kirkwood, B.; Sterne, J. Essential medical statistics. Oxford: Blackwell Publishing; 2003.

21. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, Frierson HF Jr, Hampton GM. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res. 2001; 61:5974–5978. [PubMed: 11507037]

22. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM. Delineation of prognostic biomarkers in prostate cancer. Nature. 2001; 412:822–826. [PubMed: 11518967]

23. Luo J, Duggan DJ, Chen Y, Sauvageot J, Ewing CM, Bittner ML, Trent JM, Isaacs WB. Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling. Cancer Res. 2001; 61:4683–4688. [PubMed: 11406537]

24. Trunk GV. A problem of dimensionality: A simple example. IEEE Trans Pattern Anal Mach Intell. 1979; 1:306–307. [PubMed: 21868861]

25. Garey, MR.; Johnson, DS. Computers and intractability: A guide to the theory of NP-completeness. New York: W.H. Freeman; 2006.

26. Kobayashi T, Sasaki Y, Oshima Y, Yamamoto H, Mita H, Suzuki H, Toyota M, Tokino T, Itoh F, Imai K, Shinomura Y. Activation of the ribosomal protein L13 gene in human gastrointestinal cancer. Int J Mol Med. 2006; 18:161–170. [PubMed: 16786168]

27. Dai MS, Zeng SX, Jin Y, Sun XX, David L, Lu H. Ribosomal protein L23 activates p53 by inhibiting MDM2 function in response to ribosomal perturbation but not to translation inhibition. Mol Cell Biol. 2004; 24:7654–7668. [PubMed: 15314173]

28. Onel K, Cordon-Cardo C. MDM2 and prognosis. Mol Cancer Res. 2004; 2:1–8. [PubMed: 14757840]

29. Dahiya R, McCarville J, Lee C, Hu W, Kaur G, Carroll P, Deng G. Deletion of chromosome 11p15, p12, q22, q 23–24 loci in human prostate cancer. Int J Cancer. 1997; 72:283–288. [PubMed: 9219834]

30. Zhao X, Ayer RE, Davis SL, Ames SJ, Florence B, Torchinsky C, Liou JS, Shen L, Spanjaard RA. 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. 2005; 65:2125–2129. [PubMed: 15781622]

31. Gu Z, Flemington C, Chittenden T, Zambetti GP. ei24, a p53 response gene involved in growth suppression and apoptosis. Mol Cell Biol. 2000; 20:233–241. [PubMed: 10594026]

32. Vadlamudi RK, Kumar R. P21-activated kinases in human cancer. Cancer Metastasis Rev. 2003; 22:385–393. [PubMed: 12884913]

33. Schrantz N, da Silva Correia J, Fowler B, Ge Q, Sun Z, Bokoch GM. Mechanism of p21-activated kinase 6-mediated inhibition of androgen receptor signaling. J Biol Chem. 2004; 279:1922–1931. [PubMed: 14573606]

34. Goodison S, Yuan J, Sloan D, Kim R, Li C, Popescu NC, Urquidi V. The RhoGAP protein DLC-1 functions as a metastasis suppressor in breast cancer cells. Cancer Res. 2005; 65:6042–6053. [PubMed: 16024604]

35. Okabe T, Nakamura T, Nishimura YN, Kohu K, Ohwada S, Morishita Y, Akiyama T. 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. 2003; 278:9920–9927. [PubMed: 12531901]

36. Sun, Y.; Goodison, S. Predicting breast cancer metastasis by integrating both clinical and genetic markers. Proceedings of the International Conference on Bioinformatics and Computational Biology; 2007. p. 229-235.

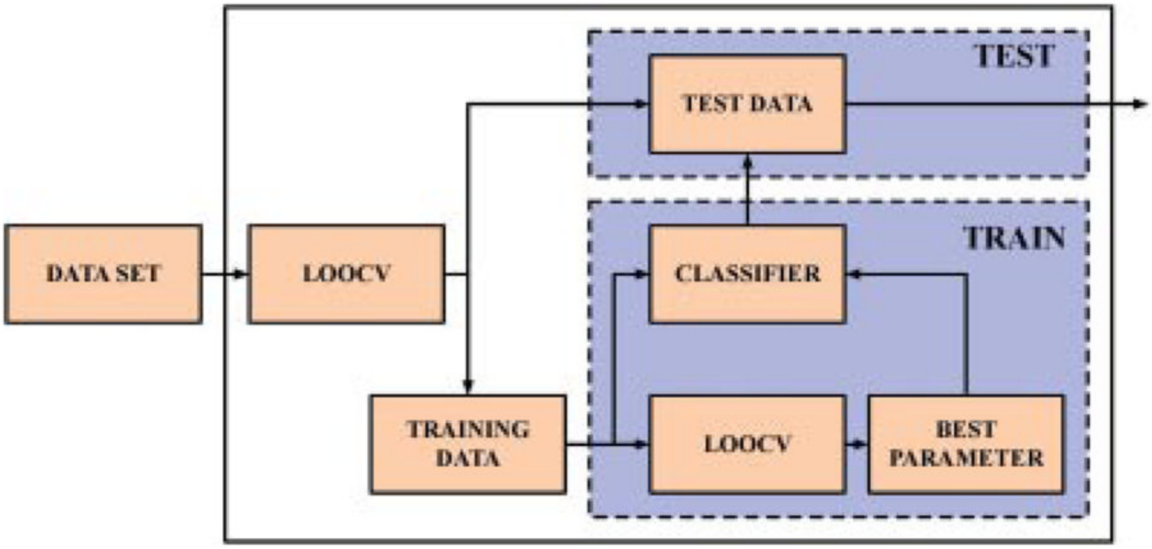

Fig. 1.

Computational experimental procedure.[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

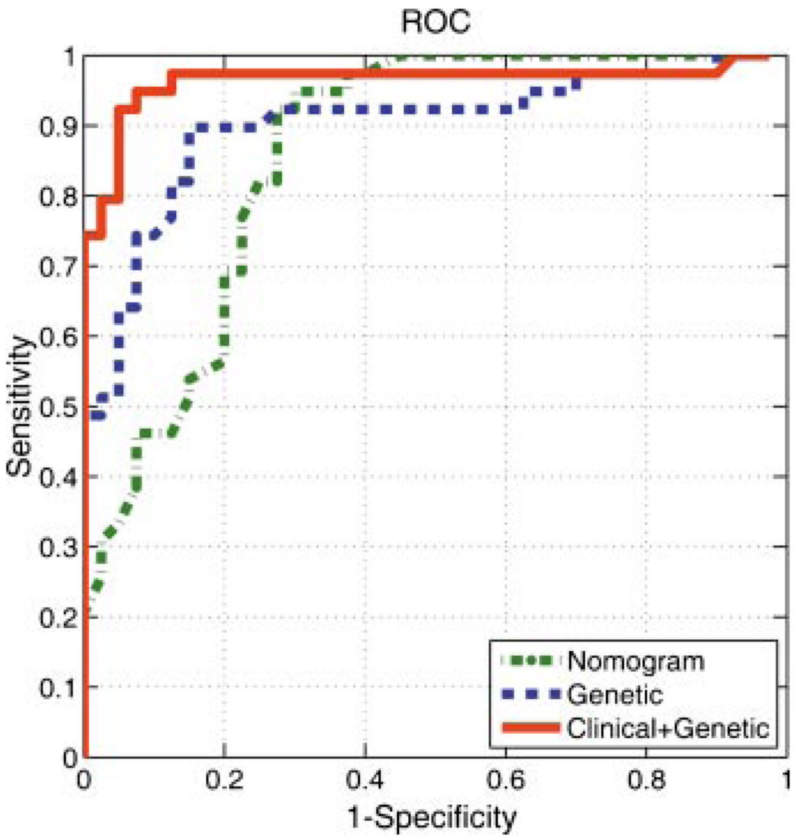

Fig. 2.

Receiver operating characteristic (ROC) plot comparing the prediction performance of the clinical predictive nomogram, genetic prognostic signature, and hybrid model (combination of nomogram and genetic). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

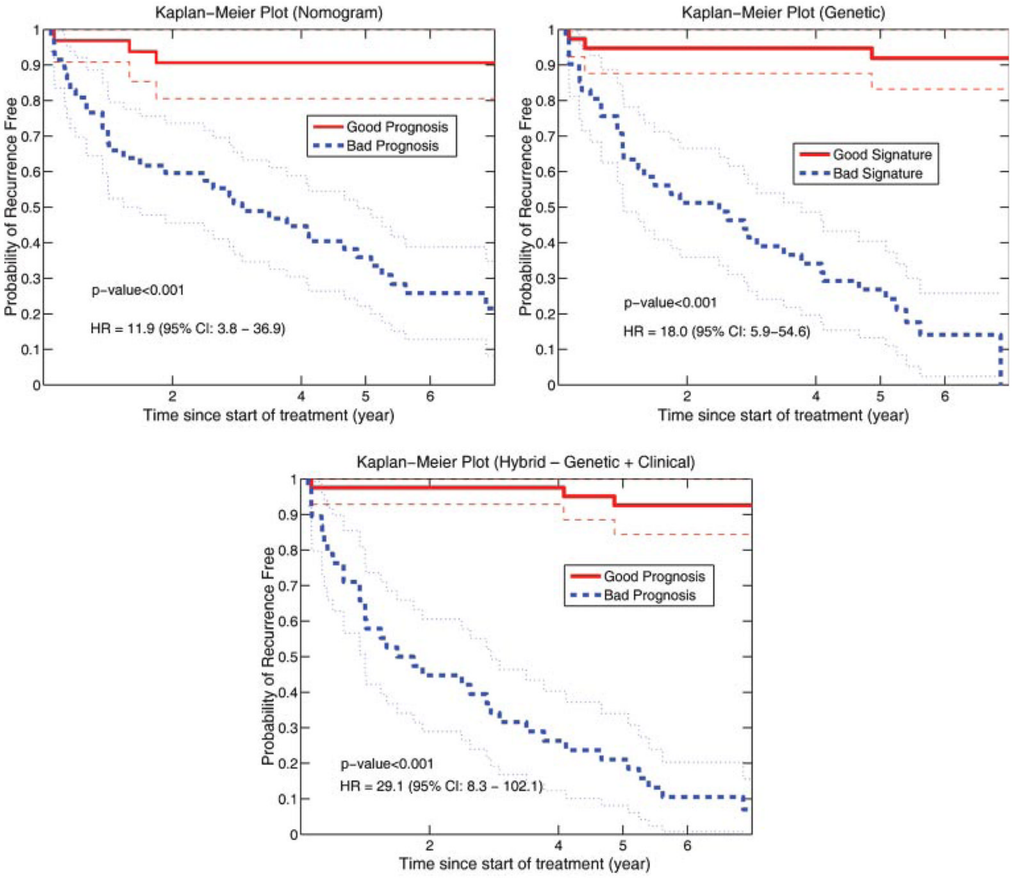

Fig. 3.

Kaplan–Meier survival curve probabilities of remaining free of disease recurrence for patients with a good or a bad prognosis as defined using the clinical nomogram (Nomogram), molecular prognostic signature derived from gene expression microarray data (Genetic), or a hybrid prognostic signature derived by the combination of the genetic and nomogram data (Hybrid Genetic and Clinical). The *P*-values were computed by log-rank test.[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

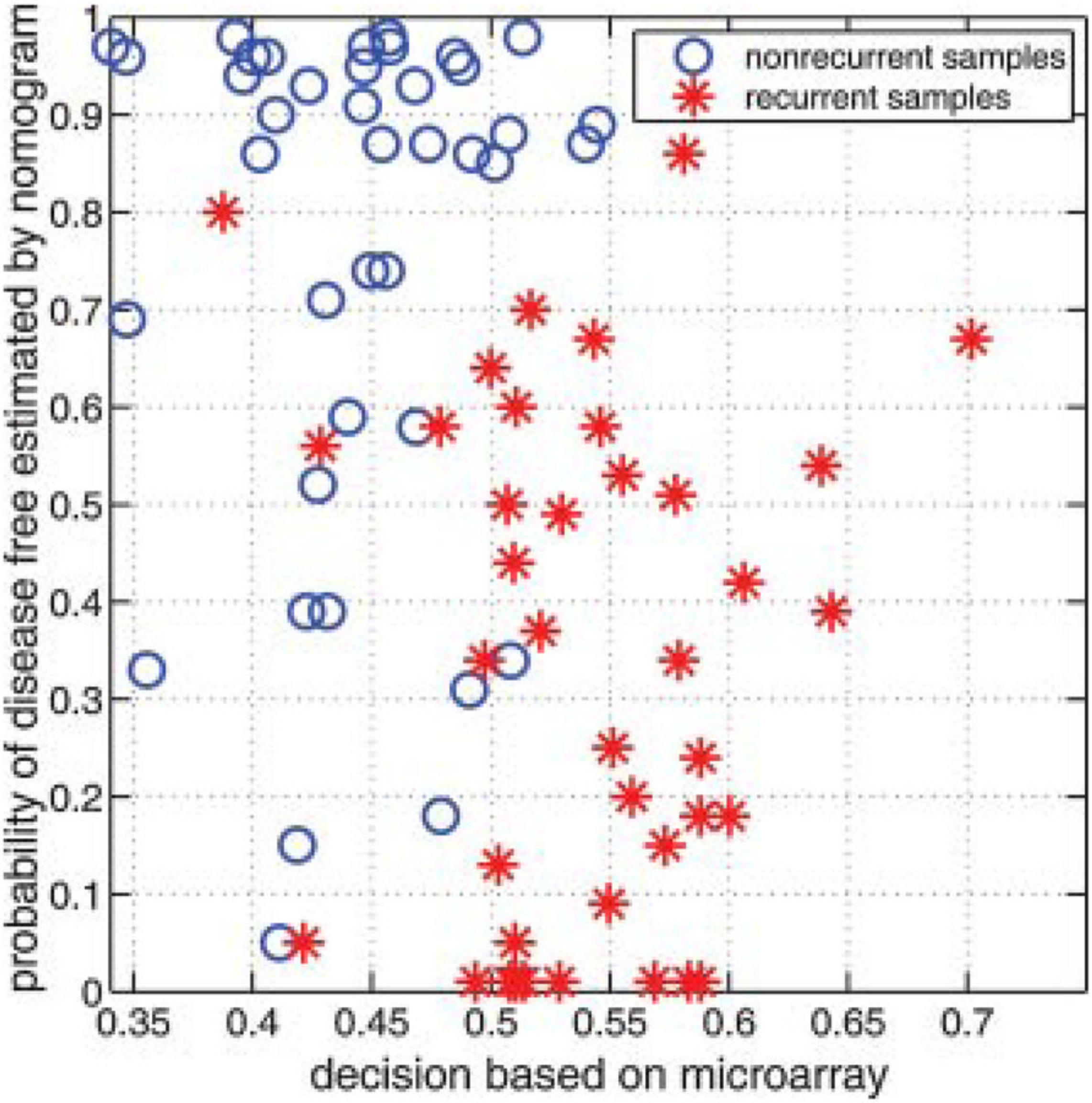

Fig. 4.

Scatter plot showing the distribution of prostate cancer patients with (asterisks) and without (circles) documented biochemical recurrence when plotted using both nomogram and microarray data. The distribution demonstrates that the genetic and clinical markers contain complementary information in assessing the risk of a patient developing biochemical disease recurrence. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

TABLE I

Prediction Results of the Clinically Used Predictive Nomogram, Our Genetic Prognostic Signature Derived From Microarray Gene Expression Data, and the Hybrid Predictive Model (Combination of Genetic and Nomogram Data)

				Hazard ratio (HR)	
Methods	**AUC (95% CI)**	**Specificity (%)**	**Odds ratio (95% CI)**	**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

The specificity was computed at the 90% sensitivity level.

TABLE II

Genes Identified in the Genetic Prognostic Signature and the Hybrid Genetic and Clinical (Marked by ¶) Predictive Model

Gene symbol	Gene title	Mean expression in recurrent tumors	*P*-value	Occurrence frequencies
PAK3¶	P21 (CDKN1A)-activated kinase 3	Under-expressed	<9.0e −6	78 (79)
RPL23¶	Ribosomal protein L23	Over-expressed	<5.0e −5	79 (79)
E124¶	Etoposide-induced 2.4 mRNA	Over-expressed	<3.0e −7	79 (79)
TGFB3¶	Transforming growth factor, beta 3	Under-expressed	<1.0e −5	79 (3)
RBM34¶	RNA-binding motif protein 34	Over-expressed	<3.0e −4	62 (8)
PCOLN3	Procollagen (type III) N-endopeptidase	Under-expressed	<3.0e −5	78
FUT7	Fucosyl transferase 7 (alpha (1,3) fucosyl transferase)	Under-expressed	<3.0e −3	30
RICS Rho	GTPase-activating protein	Over-expressed	<3.0e −6	8
MAP4K4	Mitogen-activated protein kinase 4	Over-expressed	<3.0e −5	5
CUTL1	Cut-like 1, CCAAT displacement protein (*Drosophila*)	Over-expressed	<3.0e −5	2
ZNF324B	Zinc finger protein 324B	Under-expressed	<5.0e −4	1

The *P*-values, computed using a *t*-test, quantify the up- or down-regulation of a gene between patients with, and without recurrence. The value inside and outside of the brackets in the last column is the number of iterative models in which a gene was selected in the hybrid and genetic models, respectively.