Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Selecting Clinically-Driven Biomarkers for Cancer Nanotechnology

John H. Phan[1], Andrew N. Young[2*], and May D. Wang[1*]
1 Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University.
313 Ferst Drive, Atlanta, GA, USA 30332
2 Department of Pathology & Laboratory Medicine, Emory University School of Medicine.

***Abstract*—The challenge of biomarker identification for bionanotechnology is that we need to find less than ten potential biomarkers from high throughput data so that quantum dot synthesis and imaging can be effective. Among all the extensive biomarker research, the novelty of our research is to reduce the number the biomarkers by studying the efficacy of several classifiers and error estimation methods. Specifically, we are using renal cancer expression data. The dataset consists of 31 microarray samples divided into four classes--clear cell, oncocytoma/chromophobe, papillary, and angiomyolipoma. Each class is compared to all other classes using error estimation methods for support vector machines (SVM), Fisher's discriminant (FD), and signed distance function (SDF). Prior knowledge of significant biomarker from a previous study is used to score the effectiveness of each classifier in correctly identifying these biomarkers. We have achieved intelligent model selection for biomarker identification so that the total number of nano-imaging targets is small.**

I. Introduction

Biomarker identification is generally an unsupervised process in which a dataset is blindly mined for significant biomarkers before any knowledge can be applied. In such a scenario, it is difficult to know if the most optimal algorithm is being used to find these biomarkers. For instance, if a set of biomarkers is identified, they can be validated with expert knowledge bases or literature to determine the accuracy of the algorithm in terms of true and false positives. If a different algorithm were to be used, it is highly likely that a different set (possibly with some overlap) of biomarkers would be discovered. This process of analysis followed by validation is problematic in terms of model selection for two reasons. First, it is very slow, since validation in traditional wet labs is not high-throughput compared to typical platforms such as microarrays and mass spectrometry. The number of candidate biomarkers may also be very large. Second, it is difficult to compare the performance of different algorithms if they identify different sets of significant biomarker candidates without knowledge of which biomarkers are true positives. Although high throughput technologies such as microarrays and mass spectrometry have enabled us to quickly process biological data, biomarker identification is limited by the slow validation process and frequent false positives due to a lack of samples.

This study proposes a method to improve the process of model selection for microarray data given a known set of significant biomarkers. Although number of known significant biomarkers is expected to be a small fraction of the total significant biomarkers, this small piece of knowledge provides a reference for determining the accuracy of a model for a given dataset. In almost all biological experiments, some knowledge of the biological processes involved may be enough to provide this vital information. In many cases of algorithm development, existing datasets are used for which a comprehensive biomarker identification and validation have been conducted. For this study, a renal cancer dataset consisting of several classes are used for which some genes are known to be differentially expressed [1].

Genetic biomarker identification is an important component in the development of predictive and personalized medicine. In this work, significant genetic biomarkers from a previous renal cancer study are used to select optimal classifier models. These classifier models are applied to gene expression microarray data originally used to identify the differentially expressed biomarkers through conventional methods. Renal cell carcinoma (RCC) is the most common malignant neoplasm of the adult kidney, comprising 3% of all human cancers [2]. Localized tumors can be detected by abdominal imaging and cured by surgery [3]. However, 25-40% of cases occur with extrarenal growth or metastases [4], and one-third of apparently localized lesions develop metastases during the postoperative course [5]. Advanced RCC responds poorly to systemic therapy and has a 5-year survival rate of less than 10% [6, 7]. Thus, biomarkers that improve diagnostic, prognostic or therapeutic classification would have significant clinical benefit. To address this need, microarrays have been used to discover candidate renal tumor expression markers [1, 5, 8, 9]. A limited number of selected markers have been validated by immunohistochemistry, resulting in novel bioassays with potential clinical utility for renal tumor classification. However, previous approaches to marker selection have not been standardized and thus may not have identified the optimal immunohistochemical targets. In addition, while immunohistochemistry is the current method of choice for tumor classification by pathologists, it is limited by difficulties in quantifying data and probing markers simultaneously in multiplex assays. Therefore, the aim of this study is to optimize statistical methods for selecting candidate biomarkers based on existing microarray data from renal tumors. Selected markers will be used to develop novel immunoassays based on nanoparticles such as fluorescent quantum dots, which provide the potential for quantitative measurements in multiplex analyses.

*Correspondence should be addressed to Dr. Andrew Young (andrew.n.young@emory.edu) , Dr. May D Wang, (maywang@bme.gatech.edu)

II. METHODOLOGY

A. Dataset

Microarray experiments were performed on frozen specimens from 13 clear cell RCC, 5 papillary RCC, 4 chromophobe RCC, 3 oncocytoma and 6 angiomyolipoma to obtain 8746 gene expression values per chip. The Emory University and Atlanta VA Medical Center Departments of Pathology & Laboratory Medicine diagnosed tumors using standard histopathologic criteria [4]. Carcinoma grading and staging were based on the standard Fuhrman nuclear grading system and Tumor-Node-Metastasis staging system respectively [4]. The microarray data was normalized using GCRMA procedure available at http://www.bioconductor.org/.

For each subclass, a set of biomarkers is known to be significant in that they are either up- or down-regulated in relation to samples in all other classes. These significant biomarkers are taken from a previous study on the same dataset [1]. In the clear cell (CC) class, 81 genes selected based on expression and gene ontology are differentially expressed between clear cell vs. other samples, 61 genes for chromophobe RCC/oncocytoma (CHR/ONC), 12 genes for papillary RCC (PAP), and 38 genes for angiomyolipoma (AML). These differentially expressed genes represent candidate biomarkers for diagnostic classification of renal tumor subtypes. Classification is clinically important because these tumor subtypes are associated with distinct clinical features, prognosis and response to therapy [4]. For example, advanced clear cell tumors respond variably to antiangiogenic and immunomodulatory regimens, while other subtypes are generally non-responsive [7, 10]. Of particular note, the CC markers include numerous genes related to immune response and angiogenesis, which may be relevant in predictive bioassays for therapeutic response.

B. Biomarker Identification

Biomarker identification is the process of reducing the often high dimension of ill-posed biological data (microarray or mass spec) and identifying features (biomarkers) which are able to differentiate biological samples into predefined classes. Statistically, differentiating features have well separated distributions such that the overlap is minimal. In large sample size studies, these distributions may be assumed to be Gaussian, thus simple methods which test the difference between distribution means, such as the t-test, may be used. In typical microarray experiments, however, sample sizes are often small and expression values are seldom normally distributed. For these cases, resampling methods have proven to be very effective error estimators and can be used in conjunction for any type of classifier [11, 12]. In this study, the support vector machine (SVM) [13], Fisher's discriminant (FD) [14], and signed distance function (SDF) [15] classifiers are used to rank each microarray feature by increasing error estimate. For each classifier several resampling error estimation methods are used: resubstitution (training error), resubstitution with bolstering (error smoothing) [16], leave-one-out cross validation, and 0.632 bootstrap [12, 17]. In addition, each error estimation method is tested with several different kernels: linear, radial basis gamma 1, radial basis gamma 10, and radial basis gamma 100 (bolstering is not used with radial basis). Therefore each of the three classifiers is applied to a dataset thirteen times.

Each of the four datasets (CC vs. All, CHR/ONC vs. All, PAP vs. All, and AML vs. ALL) is associated with a list of significant biomarkers discovered in a previous study [1]. Once error estimates have been computed for all features using one of the classification/error estimation methods, the difference between distributions of error estimates for significant and insignificant genes is computed using a simple metric:

$$S = \frac{EI - ES}{VI + VS}$$

where EI and ES are the expectation or mean of insignificant and significant gene error estimates, respectively. Likewise, VI and VS are the variance of insignificant and significant gene error estimates, respectively. A larger, positive, score implies that the differences in distribution of insignificant and significant genes are larger and correctly ordered (significant genes should have errors closer to zero). Since significant gene sets are invariant for a dataset, a comparison of these scores is essentially a comparison of gene ranking methods.

III. RESULTS

Thirty nine ranking methods were used (3 classifiers and 13 resampling methods) to compute error estimates for each of the four renal cancer datasets. The score for each method and dataset were computed using the available list of significant biomarkers. For all four datasets, the SVM classifier scored highest in distinguishing significant biomarkers from insignificant biomarkers compared to the FD and SDF classifiers. For the CC dataset, two methods – linear resubstitution with bolstering and linear bootstrap – scored very high with the bootstrap method slightly higher (fig 2). The linear resubstitution with bolstering method scored highest for the CHR/ONC dataset (fig 3). In contrast, both the PAP and AML datasets scored highest with the radial basis, gamma 1, resubstitution method (figs 4, 5). The SDF classifier was very competitive with the SVM, scoring higher for some methods, although not highest overall. The FD classifier was competitive with the SDF for only the CC dataset.

In addition to the proposed classifier scoring method, standard receiver-operator characteristic (ROC) curves were used to compare the best and worst classifier models for each of the four subtypes (fig 6). A larger area under the ROC curve indicates a better classifier. The ROC curves clearly validate the selection of best and worst classifier models, since the area under the curve of the worst classifier model in some cases is close to 50%, indicating a nearly random classification.

These results suggest that, for specific datasets, the choice of classifier and error estimation method plays a significant role in identifying differentiating biomarkers. Although the score metric for each method is very simple,

this metric produces very different results when random significant biomarkers are selected (not shown). This suggests that the score may be used to ascertain the probability, or p-value, that a set of biomarkers is significant.

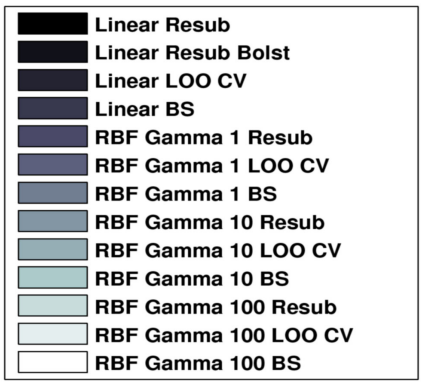

Fig. 1. Thirteen different error estimation methods are used for each of the three classifiers.

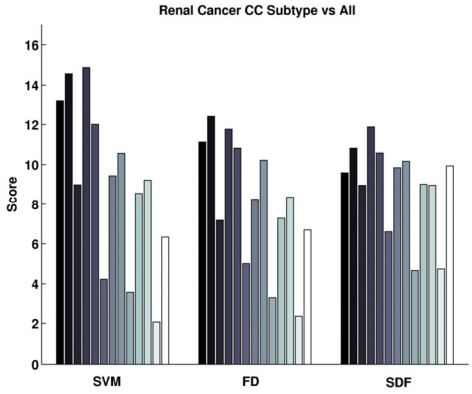

Fig. 2. Clear cell renal cell carcinoma vs. remaining samples analyzed using all classifiers and error estimation methods. The SVM linear bootstrap method is the most accurate differentiator of significant and insignificant biomarkers.

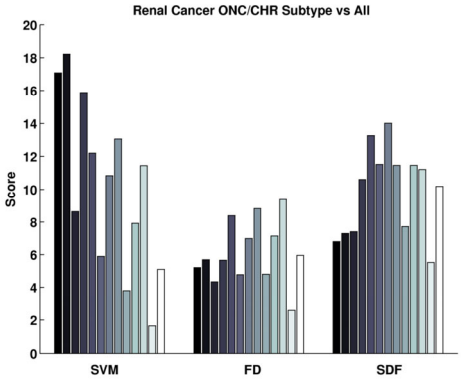

Fig. 3. Chromophobe renal cell carcinoma/oncocytoma vs. remaining samples analyzed using all classifiers and error estimation methods. The SVM linear resubstitution with bolstering method is the most accurate differentiator of significant and insignificant biomarkers.

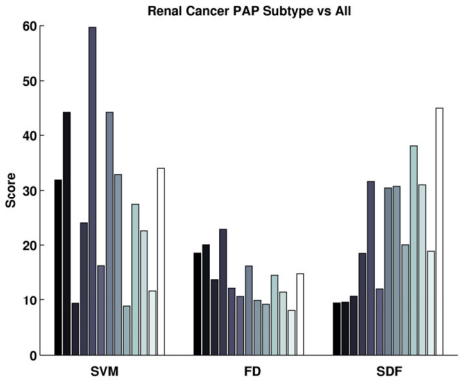

Fig. 4. Papillary renal cell carcinoma vs. remaining samples analyzed using all classifiers and error estimation methods. The radial basis resubstitution with a gamma parameter of 1 is the most accurate differentiator.

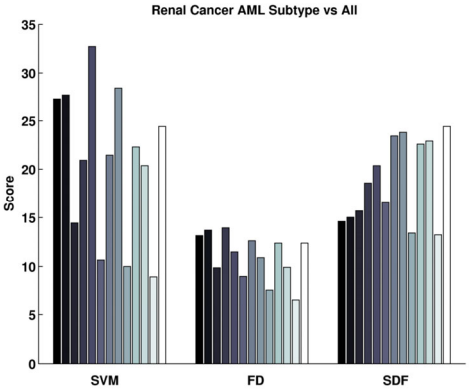

Fig. 5. Angiomyolipoma vs. remaining samples analyzed using all classifiers and error estimation methods. The radial basis resubstitution with a gamma parameter of 1 is the most accurate differentiator.

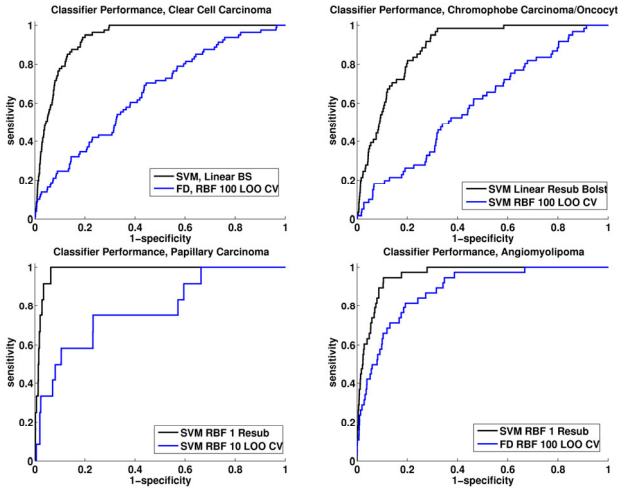

Fig. 6. Receiver-operator characteristic (ROC) curves comparing the best classifier/kernel/error estimation method (black line) with the worst (blue line) for each of the four subtypes.

IV. Discussion

In order to apply quantum dot technology to cancer diagnostic problems, quantum dots should be conjugated to highly specific and sensitive biomarkers which accurately represent disease states. Ideally, only a small number of these biomarkers, should be used in a single assay to ensure

that fast and accurate interpretation is possible in the clinical setting. Existing algorithms, however, are inefficient and the biomarker selection process should be reexamined to increase the confidence of biomarkers and to reduce the overall life-cycle of the validation process.

In a typical biomarker selection process, the initial data analysis is more or less blind. In other words, machine learning parameters and statistical assumptions are arbitrary and often too simplistic for the problem at hand. The resulting biomarkers discovered in the process are then validated through a lengthy and possibly expensive study, often leading to the realization that the analytical parameters and assumptions need to be slightly modified. But because no stringent method for selecting initial parameters and assumptions for a particular dataset does not mean that initial parameter selection needs to be completely arbitrary, especially when involving the full validation process, be it literary or experimental.

Enough biological knowledge has been gathered for almost every major human disease that the biomarker selection process can be modified with the potential for increased efficiency. The process can be modified by assuming that some biomarkers are known in advance. Then initial machine learning parameters can be selected to optimize the number of correct biomarkers discovered. Essentially, this method simplifies and improves the overall process of biomarker selection by incorporating prior knowledge into not only the classification of samples, but also into the feature selection process. Once optimal machine learning parameters have been selected for a dataset, the algorithm can be used to discover new biomarkers. Because these new biomarkers are related to existing, validated biomarkers, they may inherently be more relevant than biomarkers discovered with arbitrary parameters.

The results of this study suggest that classifiers and error estimation methods for selecting significant biomarkers differ significantly in terms of the resulting rank order of biomarkers. Optimal machine learning parameters were selected for a multiple class renal cancer microarray dataset based on significant genes identified in previous studies. The results of this study will be used to continually improve the panel of significant biomarkers for use with quantum dot nanotechnology in clinical applications.

ACKNOWLEDGMENTS

This research has been supported by grants from National Institutes of Health (Bioengineering Research Partnership R01CA108468, Center of Cancer Nanotechnology Excellence U54CA119338), and Georgia Cancer Coalition (Distinguished Cancer Scholar Award to Professor Wang).

REFERENCES

[1] A. N. Schuetz, Yin-Goen, Q., Amin, M.B., Moreno, C.S., Cohen, C., Hornsby, C.D., Yang, W.L., Petros, J.A., Issa, M.M., Pattaras, J.G., Ogan, K., Marshall, F.F, Young, A.N., "Molecular classification of renal tumors by gene expression profiling," *J Mol Diagn*, 2004.

[2] A. Jemal, Tiwari, R.C., Murray, T., Ghafoor, A., Samuels, A., Ward, E., Feuer, E.J., Thun, M.J., "Cancer Statistics, 2004," *CA Cancer J Clin*, vol. 54, pp. 8-29, 2004.

[3] Y. Homma, Kawabe, K., Kitamura, T., Nishimura, Y., Shinohara, M., Kondo, Y., Saito, I., Minowada, S., Asakage, Y., "Increased incidental detection and reduced mortality in renal cancer--recent retrospective analysis at eight institutions," *Int J Urol*, vol. 2, pp. 77-80, 1995.

[4] M. B. Amin, Tamboli, P., Javidan, J., Stricker, H., de-Peralta Venturina, M., Deshpande, A., Menon, M., "Prognostic impact of histologic subtyping of adult renal epithelial neoplasms: an experience of 405 cases," *Am J Surg Pathol*, vol. 26, pp. 281-291, 2002.

[5] M. A. Gieseg, Cody, T., Man, M.Z., Madore, S.J., Rubin, M.A., "Expression profiling of human renal carcinomas with functional taxonomic analysis," *BMC Bioinformatics*, vol. 3, 2002.

[6] A. Zisman, Pantuck, A.J., Dorey, F., Said, J.W., Shvarts, O., Quintana, D., Gitlitz, B.J., deKernion, J.B., Figlin, R.A., Belldegrun, A.S., "Improved prognostication of renal cell carcinoma using an integrated staging system," *J Clin Oncol*, vol. 19, pp. 1649-1657, 2001.

[7] A. J. Pantuck, Zeng, G., Belldegrun, A.S., Figlin, R.A., "Pathobiology prognosis, and targeted therapy for renal cell carcinoma: exploiting the hypoxia-induced pathway," *Clin Cancer Res*, vol. 9, pp. 4641-4652, 2003.

[8] A. N. Young, Amin, M.B., Moreno, C.S., Lim, S.D., Cohen, C., Petros, J.A., Marshall, F.F., Neish, A.S., "Expression profiling of renal epithelial neoplasms: a method for tumor classification and discovery of diagnostic molecular markers," *Am J Pathol*, vol. 158, pp. 1639-1651, 2001.

[9] M. Takahashi, Yang, X.J., Sugimura, J., Backdahl, J., Tretiakova, M, Qian, C.N., Gray, S.G., Knapp, R., Anema, J., Kahnoski, R., Nicol, D., Vogelzang, N.J., Furge, K.A., Kanayama, H., Kagawa, S., Teh, B.T., "Molecular subclassification of kidney tumors and discovery of new diagnostic markers," *Oncogene*, vol. 22, pp. 6810-6818, 2003.

[10] R. J. Motzer, Bacik, J., Mariani, T., Russo, P., Mazmudar, M., Reuter, V., "Treatment outcome and survival associated with metastatic renal cell carcinoma of non-clear cell histology," *J Clin Oncol*, vol. 20, pp. 2376-2381, 2002.

[11] U. Braga-Neto, Dougherty, E., "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, pp. 374-380, 2004.

[12] W. J. Fu, Carroll, R.J., Wang, S., "Estimating misclassification error with small samples via bootstrap cross-validation," *Bioinformatics*, vol. 21, pp. 1979-1986, 2005.

[13] N. Cristianini, Shawe-Taylor, J., *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000.

[14] S. Mika, Ratsh, G., Weston, J., Scholkopf, B., Muller, K., "Fisher discriminant analysis with kernels," 1999.

[15] E. M. Boczko, Young, T., "The signed distance function: a new tool for binary classification," *arXiv:cs.LG/0511105v1*, 2005.

[16] C. Sima, Braga-Neto, U., Dougherty, E.R., "Superior feature-set ranking for small samples using bolstered error estimation," *Bioinformatics*, vol. 21, pp. 1046-1054, 2005.

[17] B. Efron, "Estimating the error rate of a prediction rule: some improvements on cross-validation," *J of the Am Stat Assoc*, vol. 78, pp. 316-331, 1983.