

## Predicting Cancer Drug Response by Proteomic Profiling

Yan Ma,<sup>1</sup> Zhenyu Ding,<sup>2</sup> Yong Qian,<sup>4</sup> Xianglin Shi,<sup>4</sup> Vince Castranova,<sup>4</sup> E. James Harner,<sup>1</sup> and Lan Guo<sup>3</sup>

**Abstract Purpose:** Accurate prediction of an individual patient's drug response is an important prerequisite of personalized medicine. Recent pharmacogenomics research in chemosensitivity prediction has studied the gene-drug correlation based on transcriptional profiling. However, proteomic profiling will more directly solve the current functional and pharmacologic problems. We sought to determine whether proteomic signatures of untreated cells were sufficient for the prediction of drug response.

**Experimental Design:** In this study, a machine learning model system was developed to classify cell line chemosensitivity exclusively based on proteomic profiling. Using reverse-phase protein lysate microarrays, protein expression levels were measured by 52 antibodies in a panel of 60 human cancer cell (NCI-60) lines. The model system combined several well-known algorithms, including random forests, Relief, and the nearest neighbor methods, to construct the protein expression – based chemosensitivity classifiers. The classifiers were designed to be independent of the tissue origin of the cells.

**Results:** A total of 118 classifiers of the complete range of drug responses (sensitive, intermediate, and resistant) were generated for the evaluated anticancer drugs, one for each agent. The accuracy of chemosensitivity prediction of all the evaluated 118 agents was significantly higher ( $P < 0.02$ ) than that of random prediction. Furthermore, our study found that the proteomic determinants for chemosensitivity of 5-fluorouracil were also potential diagnostic markers of colon cancer.

**Conclusions:** The results showed that it was feasible to accurately predict chemosensitivity by proteomic approaches. This study provides a basis for the prediction of drug response based on protein markers in the untreated tumors.

Assessment of an individual patient's predisposition to drugs is essential to achieve the goal of personalized medicine in cancer therapy. Such an approach is needed for clinicians to decide which chemotherapeutic agents would be effective for a given patient, meanwhile, to avoid including those ineffective agents (and the entailed side effects) in treatment options. This decade has witnessed significant advances in pharmacogenomics research of predicting drug sensitivity by transcriptional

profiling (1–5). However, the pattern of transcriptional profiling does not necessarily correlate with the pattern of proteomic profiling. Transcriptional profiling can only reveal cancer information at the mRNA level. It is the protein that ultimately plays an essential role in cancer development and progression. It has been well established that the route from mRNA to protein involves several biological processes (i.e., translation and post-translational modification). Therefore, proteomic profiling will more directly address the current biological and pharmacologic issues (6). In this study, a predictive model system was presented to explore proteomic contributions to drug sensitivity. We predicted drug response of a panel of 60 human cancer cell (NCI-60) lines to 118 anticancer agents by proteomic profiling. The protein expression levels were measured on untreated cells. The focus here was on predicting response to therapy instead of analyzing molecular consequences of therapy. This study provides a basis for predicting drug response based on protein markers in the tumors of untreated patients.

It is especially challenging to predict chemosensitivity in the clinical context because drug responses reflect the properties intrinsic to both the target cells and the host metabolism (3). In this study, the analysis was limited to the intrinsic properties of cells exposed in culture by modeling the response of the panel of human cancer cell (NCI-60) lines. The NCI-60 set includes the cell lines derived from leukemias, melanomas, and carcinomas of ovarian, renal, breast, prostate, colon, lung, and central nervous system origin. These cell lines have been

**Authors' Affiliations:** Departments of <sup>1</sup>Statistics and <sup>2</sup>Computer Science and Electrical Engineering and <sup>3</sup>Mary Babb Randolph Cancer Center/Department of Community Medicine, West Virginia University; <sup>4</sup>The Pathology and Physiology Research Branch, Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, West Virginia  
Received 2/8/06; revised 4/4/06; accepted 5/11/06.

**Grant support:** NIH/National Center for Research Resources grant P20 RR16440-03 (L. Guo) and NIH/National Cancer Institute grant 1R01CA119028-01 (X. Shi).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

**Note:** Supplementary data for this article are available at Clinical Cancer Research Online (<http://clincancerres.aacrjournals.org/>).

**Requests for reprints:** Lan Guo, 1814 HSS, Mary Babb Randolph Cancer Center, P.O. Box 9300, Morgantown, WV 26506-9300. Phone: 304-293-6455; Fax: 304-293-4667; E-mail: lguo@hsc.wvu.edu and Yong Qian, The Pathology and Physiology Research Branch, Health Effects Laboratory Division, National Institute for Occupational Safety and Health, 1095 Willowdale Road, Morgantown, WV 26505-2888. Phone: 304-285-6286; Fax: 304-285-5938; E-mail: yaq2@cdc.gov.

© 2006 American Association for Cancer Research.

doi:10.1158/1078-0432.CCR-06-0290

screened for drug activity of a broad range of chemical compounds. A sulforhodamine B assay was applied to examine the growth inhibition by measuring the total cellular protein changes on the stimulation with a particular chemical compound. The drug activities were assessed based on the pattern of growth inhibition within 48 hours. The data are available for the public (2). Here, the focus was on a 118-drug subset whose mechanisms of action are putatively known (2). Some of these drugs are currently in routine clinical use for cancer treatment, whereas others are either in clinical trials or in late stages of drug development.

We investigated the feasibility of drug response prediction by using protein expression levels. Both the proteomic profiles (6) and the drug activity database of the 118 agents (2) were generated by the National Cancer Institute and are available from the National Cancer Institute's Discover Web site.<sup>5</sup> The database of protein expression levels was generated by proteomic assays with 52-antibody reverse-phase protein lysate microarray in each individual cell line (6). The proteomic assays were done using reverse-phase protein lysate microarrays (6, 7). The protein samples were robotically planted on the chips followed by the measurements with antibodies. Each of the 52 antibodies is a specific antibody that recognizes a specific protein (6). The data and detailed information are available for the public (6). We sought to identify important protein markers to predict drug response of each individual cell line to the 118 anticancer agents. To construct the optimal classifiers, a computational model system was developed by integrating several state-of-the-art algorithms, including random forests (8), Relief (9, 10), and the nearest neighbor methods (10). To evaluate classifier accuracy, either a bootstrapped out-of-bag method (8) or 10-fold cross-validation (11) was used to assess the prediction performance. When compared with random prediction, all protein expression-based classifiers for the 118 drugs did accurately with statistical significance ( $P < 0.02$ ). Our results showed that it was feasible to predict drug response of cancer cell lines by proteomic profiling.

## Materials and Methods

**Proteomic profiling.** The protein expression data file was generated by Nishizuka et al. (6). A protocol was developed for making reverse-phase protein lysate microarrays with a larger number of spots than previously feasible. The data points for 52 antibodies were analyzed by using P-SCAN and a quantitative dose interpolation method on the 60 human cancer cell lines (NCI-60). The data file is available online.<sup>6</sup>

**Drug activity profiles.** The drug activity profiles of 118 anticancer agents were screened by Scherf et al. (2). Growth inhibition was assessed from the changes in total cellular protein after 48 hours of drug treatment using a sulforhodamine B assay. Drug activities ( $\log_{10} GI_{50}$ ) were recorded across the 60 human cancer cell lines.  $GI_{50}$  is the concentration required to inhibit cell growth by 50% compared with untreated controls. The activity profile of an agent consists of 60 such activity values, one for each cell line. The drug activity profiles of 118 agents are available online.<sup>7</sup>

**Defining drug sensitivity and resistance.** The data file containing drug activity data of 118 anticancer agents was processed to define drug resistance and sensitivity of the NCI-60 lines. Specifically, for each drug,  $\log_{10} (GI_{50})$  values were normalized across the 60 cell lines. Cell lines with  $\log_{10} (GI_{50})$  at least 0.5 SD above the mean were defined as resistant to this drug. Those with  $\log_{10} (GI_{50})$  at least 0.5 SD below the mean were defined as sensitive to the drug. The remaining cell lines with  $\log_{10} (GI_{50})$  within 0.5 SD were defined as intermediate in the range of drug responses.

**Classification methods.** For each drug, we formed a data set with 53 variables, including 52 protein variables and 1 drug response variable with the label of sensitive, intermediate, or resistant. The 52 protein expression variables were predictors, whereas the drug response was the predicted variable. Random forests (8) in software package *R*<sup>8</sup> was used as a classification technique. Random forests are a generalization of the classification tree algorithm. Instead of growing a single classification tree, the random forest algorithm constructs an ensemble of hundreds or thousands of trees. Each tree is built on a bootstrap sample from the original learning set. The variables used for splitting the tree nodes are a random subset of the whole variables set. The classification decision of a new instance is obtained by majority voting (unless the cutoff is user defined) over all trees. In random forests, about one third of the cases in the bootstrap sample are not used in growing the tree. These cases are called "out-of-bag" cases and are used to evaluate the algorithm performance. The out-of-bag method provides an unbiased evaluation of the prediction accuracy. Therefore, there is no need to use a separate test set or an additional cross-validation method for the evaluation (8). Several characteristics of random forests make it ideal for data sets that are high dimension, and most predictive variables are noisy (12).

The nearest neighbor methods (IB1 and NNge) implemented in software package WEKA 3.4<sup>9</sup> (10) were also used to construct the optimal classifiers for drug responses. IB1 is a basic instance-based learner. It uses normalized Euclidean distance to find the training instance closest to the given test instance and predicts the same class as this training instance. IB1 is a special case of IBk with  $k = 1$ . IBk implements the  $k$ -nearest neighbor algorithm. To classify a new instance  $x_0$ ,  $k$  training set instances closest in distance to  $x_0$  are obtained, and majority voting among these  $k$  neighbors determines the class of  $x_0$ . Some notable distance metrics are Euclidean distance, Mahalanobis distance, etc. (13). Despite its simplicity,  $k$ -nearest neighbor method has been successful in a large number of classification problems (14). NNge is a nearest neighbor method with generalization. It generates rules using nonnested generalized exemplars, which are rectangular regions of instance space used for calculating a distance function to classify new instances (10). Different from IB1 and IBk, NNge is a rule-based classifier. The "hypergeometric" model described above includes if-then rules (15). These two methods were applied to the drugs for which random forests were unable to achieve overall accuracy >50% in chemosensitivity prediction. The WEKA classifiers used 10-fold cross-validation to evaluate the prediction performance.

**Feature selection algorithms.** The mean decrease in accuracy measure implemented in the random forest algorithm was used to rank the importance of the features in prediction. This measure determines the variable importance in terms of the contribution to prediction accuracy. Mean decrease in accuracy is defined as follows: for each tree, the algorithm randomly rearranges the values of the  $m^{\text{th}}$  variable for the out-of-bag set, puts this permuted set down the tree, and gets new classifications for the forest. The importance of the  $m^{\text{th}}$  variable can be defined in "mean decrease in accuracy" as the difference between the out-of-bag error rate for randomly permuted  $m^{\text{th}}$  variable and the original out-of-bag error rate. This method was used with the random forest package implemented in *R* to construct the optimal classifiers.

<sup>5</sup> <http://discover.nci.nih.gov/datasets.jsp>.

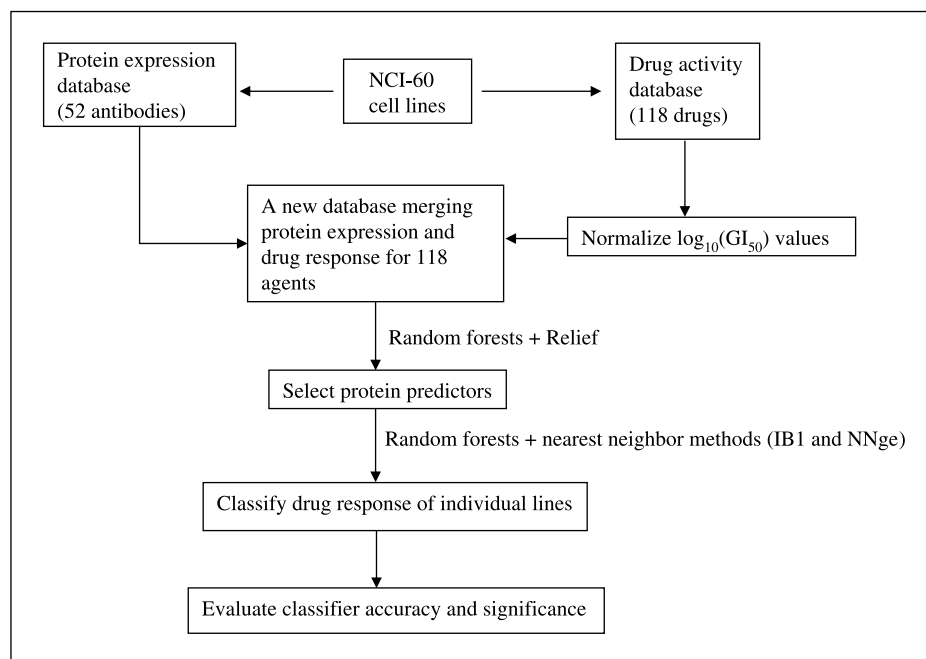
<sup>6</sup> [http://discover.nci.nih.gov/host/2003\\_profilingtable7.xls](http://discover.nci.nih.gov/host/2003_profilingtable7.xls).

<sup>7</sup> [http://discover.nci.nih.gov/nature2000/data/selected\\_data/dataviewer.jsp?baseFileName=a\\_matrix118&nsc=2&dataStart=3](http://discover.nci.nih.gov/nature2000/data/selected_data/dataviewer.jsp?baseFileName=a_matrix118&nsc=2&dataStart=3).

<sup>8</sup> <http://www.r-project.org>.

<sup>9</sup> <http://www.cs.waikato.ac.nz/ml/weka/>.

Fig. 1. Architecture of the model system for drug response prediction.



When the random forest package failed to achieve accuracy >50% in drug response prediction, the Relief method implemented in WEKA 3.4 was used as a filter to rank the proteins. Relief evaluates the importance of a variable by repeatedly sampling an instance and checking the value of the given variable for the nearest instance from the same and different classes. The values of the attributes of the nearest neighbors are compared with the sampled instance and used to update the relevance scores for each attribute. As approximated in Eq. A, Relief computes the weight of attribute  $A$  as follows:

$$W[A] = \frac{P(\text{different value of } A \mid \text{near miss}) - P(\text{different value of } A \mid \text{near hit})}{2} \quad (A)$$

Relief assigns more weight to those attributes that have the same value for instances from the same class and differentiate between instances from different classes (9, 10).

**Evaluating classifier accuracy.** To assess the significance of our prediction results, it is necessary to show that our prediction results are significantly better than those of random prediction. For each drug, the original class distributions were maintained and the class labels of the 60 cell lines were randomly permuted. The random permutation produced 60 class labels while keeping the class distribution fixed. The matches between the rearranged class labels and the original ones were recorded. The percentage of the matches was calculated as the accuracy measure for the random prediction. This procedure was repeated for 1,000 times. Based on the generated 1,000 accuracy measures, the  $P$  was calculated as the upper percentile of our prediction accuracy in the profile of 1,000 random prediction results. If the prediction accuracy produced by our classifier exceeds the 95th percentile of those 1,000 random prediction accuracies, it is concluded that our prediction is significantly better than random prediction ( $P < 0.05$ ). The experimental details and prediction results are provided in Supplementary Materials.

**Unsupervised hierarchical clustering.** Unsupervised hierarchical clustering was done using the online tool CIMminer<sup>10</sup> developed by the National Cancer Institute (16). The distance was computed based on

correlation, and the clustering method was complete linkage for both the samples and the proteins. A heat map was generated by using CIMminer.

## Results

We sought to predict the complete range of drug responses by proteomic profiling. The architecture of our prediction model system was delineated in Fig. 1. We approached the chemosensitivity prediction as a supervised multiclassification problem. For each chemotherapeutic agent, the complete range of drug responses across the 60 cancer cell lines was partitioned into three classes (sensitive, intermediate, or resistant) based on the normalized growth-inhibitory activities ( $GI_{50}$  values). The partition scheme generated a relatively balanced data set for the chemosensitivity profiling (see Supplementary Materials for details).

To compare chemosensitivity of different tissue types, we averaged the number of drugs in each class on the cell lines with the same tissue origin in the NCI-60 panel (Fig. 2). For the 118 agents, the ovarian cancer lines were the least sensitive and the most resistant to these drugs, whereas the leukemias were the most sensitive and the least resistant. There were six cell lines from ovarian tumors in the NCI-60 panel (i.e., OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, IGROV1, and SK-OV-3). On the average, they were sensitive to 14 drugs and resistant to 60 drugs. On the other end of the spectrum, the six cell lines from leukemias, CCRF-CEM, K-562, MOLT-4, HL-60 (TB), SR, and RPMI-8226, were sensitive to 81 drugs and resistant to 9 drugs on the average.

Based on the drug activity database, we compared the chemosensitivity profiles of the 118 anticancer agents. For each agent, we calculated the percentage of cell lines in the NCI-60 panel that were sensitive to this compound. Our analysis showed that these 118 drugs had varied sensitivity across the 60 cancer cell lines. More than 50% of the NCI-60 lines were sensitive to each of the five drugs, including Camptothecin, 7-Cl

<sup>10</sup> <http://discover.nci.nih.gov/cimminer/>.

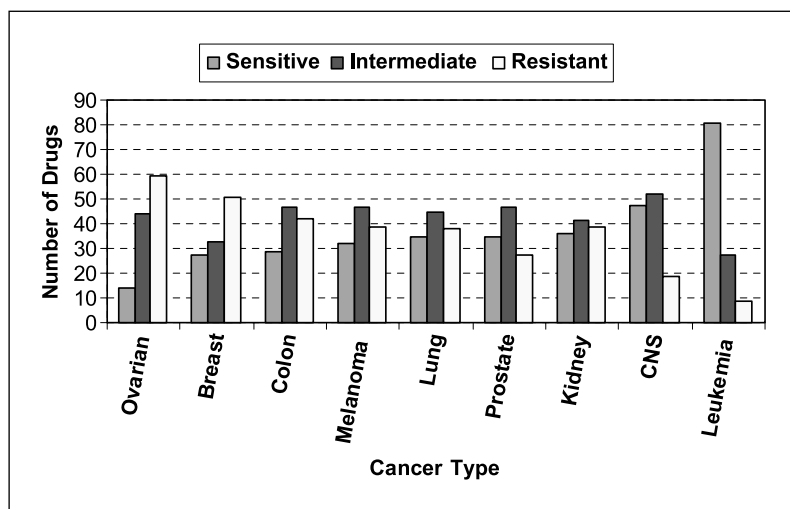


Fig. 2. Chemosensitivity profiles of different cancer types. For each tissue type, the number of drugs in each class was averaged on the number of cell lines with the same tissue origin. CNS, central nervous system.

(NSC 249910), Camptothecin,9-NH2 (S) (NSC 603071), Aminopterin-derivative (NSC 184692), an-antifol (NSC 623017), and DUP785 (brequiar) (NSC 368390; Fig. 3). None of the 60 cell lines was sensitive to Colchicine-derivative (NSC 33410) or Vincristine-sulfate (NSC 67574; Fig. 3).

The mechanisms of action of these 118 agents are putatively understood (2). To compare the chemosensitivity profiles of different mechanisms of action, we did the following analysis. For each drug mechanism category, the number of cell lines in each class (sensitive, intermediate, or resistant) was averaged for the drugs with the same mechanisms (Fig. 4). The results indicated that, on the average, drugs with different mechanisms of action had relatively uniform chemosensitivity profiles in the 60 cancer cell lines.

The proteomic profiles of the 60 cancer cell lines were generated by the National Cancer Institute to screen compounds for anticancer activities (6). To construct the supervised protein expression-based classifiers for the prediction of drug response, we created a new database by merging the proteomic profiles and the responses of the NCI-60 lines to the 118 agents (Fig. 1). For each agent, the data file contained both the protein expression levels measured by the 52 antibodies in each cell line (6) and the response of each line to this drug. For each cell line, the protein expression levels measured by the 52

antibodies were predictors, whereas the drug response was the predicted variable. A classifier was constructed for each drug, independent of the tissue origin of the cells. In this study, we sought to construct the optimal classifiers of chemosensitivity prediction for the 118 anticancer drugs.

By exclusively using the protein expression data, we investigated the feasibility of predicting drug response of each line. The goal was to identify the optimal classifiers that achieve the highest prediction accuracy of drug response with the minimum number of proteins. The random forest algorithm (8) implemented in software package *R* was first used to construct the classifiers. The random forest package was used as both a classifier and a feature selection method to rank the importance of each protein in chemosensitivity prediction (Fig. 1). Based on the ranking, the protein variables were filtered from the prediction model in a stepwise manner. The optimal classifier contained the minimum number of proteins that generated the highest overall prediction accuracy (defined as the percentage of correctly predicted instances). Specifically, for each drug, the lowest ranking proteins were sequentially removed. The bottom 2 proteins were removed first, and a subset of top 50 proteins was included in the prediction model. Then, the bottom 5 proteins were removed from the prediction model for each iteration. When the subset contained 10

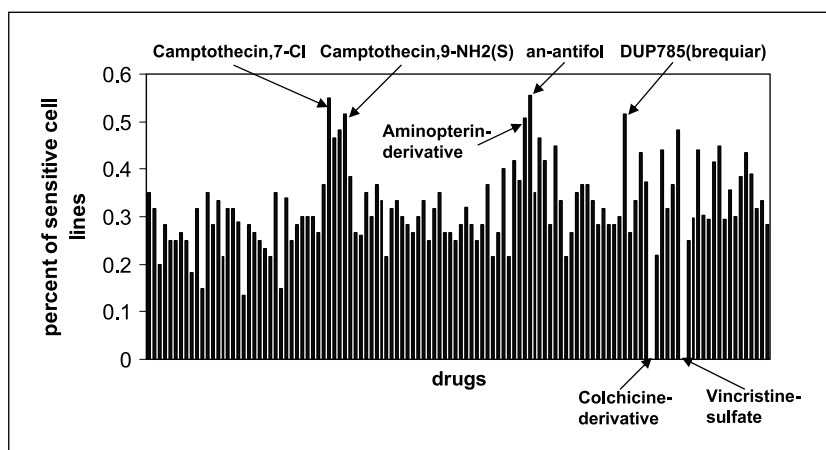
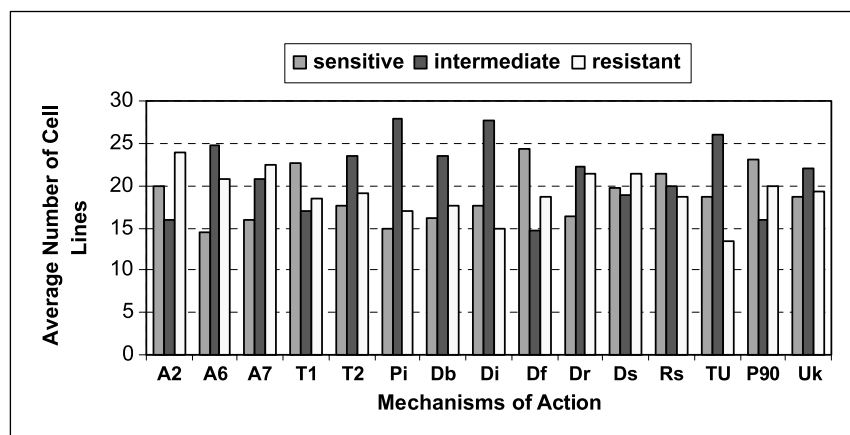


Fig. 3. Percentage of sensitive cell lines to the 118 drugs. Left to right, the drugs were listed according to the order in the original data file. For each drug, the percentage of the cell lines in the NCI-60 panel that were sensitive to this drug was calculated.

**Fig. 4.** Drug responses of the NCI-60 lines averaged for each mechanism of action. For each category of mechanisms of action, the number of cell lines in each class was averaged for the number of drugs belonging to the corresponding category.



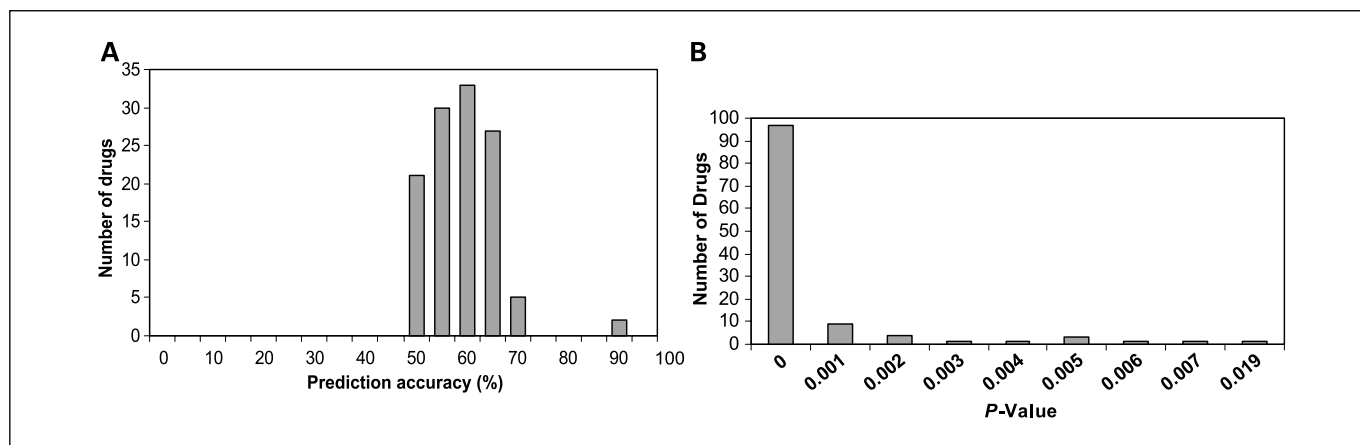
proteins, the bottom 1 protein was removed at a time. For each drug, the optimal classifier was the one achieving the highest prediction accuracy with the minimum number of proteins. In our study, the smallest feature set of the constructed optimal classifiers consisted of three proteins. For the 118 drugs, 115 had overall prediction accuracy >50% by using random forests. The random forest algorithm uses an out-of-bag error based on the bootstrapped samples to evaluate the classification results. The reported prediction accuracy evaluated by the out-of-bag error was proven to be unbiased (8). Therefore, there is no need for any additional cross-validation or an independent validation set to evaluate the results (8).

Three drugs had relatively low overall prediction accuracy (<50%) by using random forests. To identify the optimal classifiers, we used several methods implemented in software package WEKA 3.4 (10). Specifically, the Relief algorithm was used as a filter to identify the protein markers and the nearest neighbor methods (IB1 and NNge) were deployed as the classifiers. For these three drugs, the lower ranked proteins were filtered from the prediction models based on the order of importance computed by Relief. The optimal protein subset generated the highest prediction accuracy by using the nearest neighbor method (IB1 or NNge). The prediction results using the WEKA techniques were evaluated by 10-fold cross-

validation. The estimated accuracy by this validation method has been proven to have the lowest bias and variance among all cross-validation methods, including the leave-one-out method (11). It, thus, provides an objective evaluation of the performance of our prediction models in general.

Overall, the constructed optimal classifiers used between 3 and 26 protein predictors, with an average of 8 predictors in each classifier. The overall accuracy of the optimal classifiers for the 118 drugs was summarized in Fig. 5A. We evaluated the prediction results by comparing them with the random prediction in 1,000 test runs (see Materials and Methods and Supplementary Materials for details). The results showed that, for 97 drugs, none of the random predictions in 1,000 iterations achieved our accuracy ( $P = 0.00$ ). Our prediction accuracy is significantly better than random prediction at  $P < 0.007$  level for 117 drugs and at  $P < 0.019$  level for the remaining 1 drug (Fig. 5B).

This study identified the protein markers for predicting chemosensitivity of the 118 agents in 60 cancer cell lines. The markers can, in principle, provide a basis to devise the optimal combination of therapies directed specifically to eliminate the cancer cells while minimizing toxicity to the normal cells (17). In addition, the markers can theoretically portrait a unique molecular signature for detection and diagnosis of a cancer



**Fig. 5.** Overall accuracy for the 118 chemosensitivity classifiers. *A*, distribution of classification accuracy for the 118 drugs. The prediction accuracy is the percentage of correctly classified instances. *B*, evaluation of the classification accuracy measured by  $P$ s. It represents the possibility that our prediction accuracy was achieved by random prediction in 1,000 test runs.

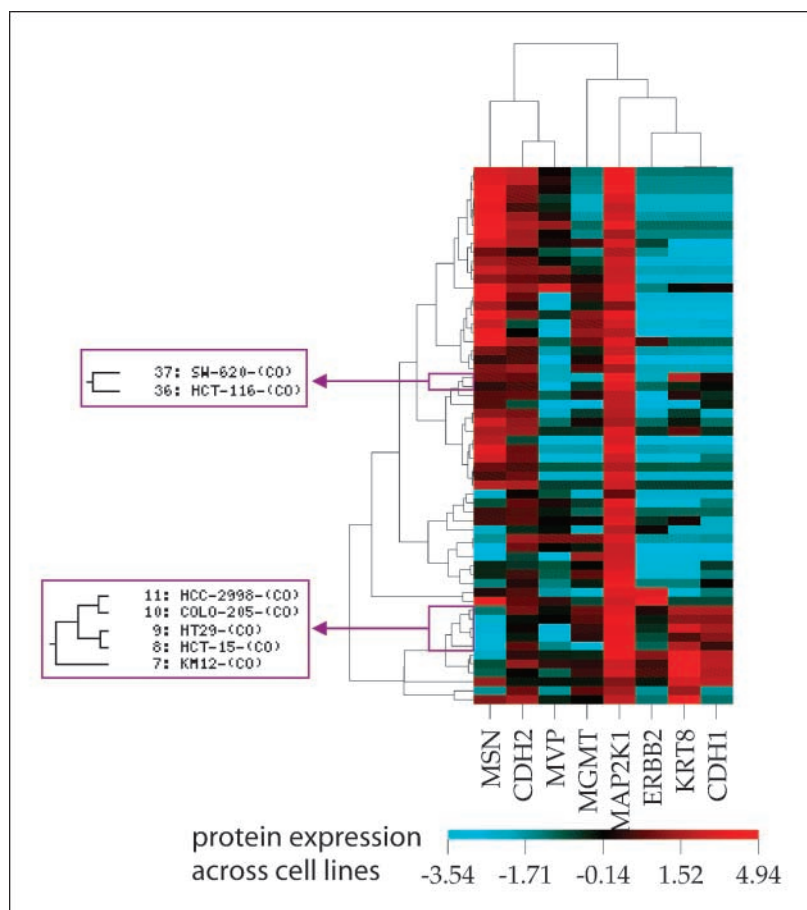
(6, 17, 18). Among the studied drugs, 5-fluorouracil (NSC 19893) has been included in the treatment combinations for patients with stage III colon cancer (19). In this analysis, eight protein markers were identified for the prediction of drug response to 5-fluorouracil, including CDH1, CDH2, KRT8, ERBB2, MSN, MVP, MAP2K1, and MGMT. All of these proteins, except for KRT8, are involved in the pathogenesis of colon cancer. To investigate the feasibility of using these markers to diagnose colon cancer, we did unsupervised hierarchical clustering on the 60 cancer cell lines by using the expression levels of these eight proteins. There were a total of seven colon cancer cell lines in the NCI-60 panel. Five of them, KM-12, HCT-15, HT29, COLO-205, and HCC-2998, were aggregated together (Fig. 6). The remaining two cell lines, HCT-116 and SW-620, were clustered together but separately from these five cell lines (Fig. 6). The results showed that the identified protein markers provided a basis not only for detection and diagnosis of colon cancer but also for devising the optimal therapeutic combination targeted specifically to eliminate the cancer cells.

### Discussion

An implicit prerequisite of personalized medicine is that an individual patient's response to drugs can be predicted (3). Previous studies in drug sensitivity investigated the gene-drug correlation by transcriptional profiling (1–5). However, transcriptional profiling only reveals the information of mRNA,

and the majority of current diagnostic markers and therapeutic targets are proteins, not mRNA (6). Proteomic profiling quantifies protein expression levels, and thus, it will yield more direct answers to functional and pharmacologic questions (6). Here, we reported a machine learning methodology for protein expression-based prediction of chemosensitivity. In this study, the model system was applied to predicting cytotoxicity of 118 anticancer agents by using protein expression profiles in the 60 untreated cell lines (NCI-60). The feasibility of predicting drug response was investigated exclusively based on protein expression data.

A particular limitation of protein expression-based chemosensitivity prediction is the small amount of available protein expression data due to the technical difficulties in proteomics (6). Thus far, we have only found one proteomic data set done on the NCI-60 panel. The data set contains protein expression levels measured by 52 antibodies (6). The available features in the studied data set are much less than those in a data set generated by a gene chip that can quantify the level of thousands of genes simultaneously. The limited data resource made it even more difficult to construct protein expression-based classifiers for the prediction of chemosensitivity. Another limitation is the small size of the samples. The NCI-60 panel contains a total of 60 cell lines, with 2 to 9 lines representing each histologic origin. In this study, tissue origin or cancer type was not used as a predictor. All the cell lines were treated equally, and the tissue types were not revealed in the classification. To evaluate the prediction performance, we



**Fig. 6.** Unsupervised hierarchical clustering of the NCI-60 panel based on the eight protein markers. The colon cancer lines were aggregated into two groups. One group contained five cell lines, whereas the other contained two lines. The eight protein markers were the chemosensitivity predictors for 5-fluorouracil, which is used for the treatment of colon cancer.

used either a bootstrapped out-of-bag error (8) or 10-fold cross-validation method. The bootstrapped out-of-bag method uses two thirds of the samples as the training set and the remaining samples as the validation set. In the 10-fold cross-validation method, the data are partitioned into 10-fold. Each time, 9-fold is used as the training set and the remaining 1-fold as the validation set. This process is repeated 10 times until every sample is validated once. Compared with the leave-one-out method, the disadvantage of both evaluation methods is that they further reduce the size of the samples used to generate the model. Consequently, the prediction accuracy can be potentially compromised. However, both methods provide an unbiased evaluation for the prediction performance (8, 11). In addition, we approached the prediction of drug sensitivity as a multiclassification problem. The complete range of drug responses was partitioned into three categories: sensitive, resistant, and intermediate. As shown in a computational analysis of classification schemes (20), a multiclassification algorithm is inherently more difficult than a binary one and generally yields compromised prediction accuracy.

Given the above limitations and difficulties, the observed accuracies of the constructed classifiers are notable. Our classification accuracy was much higher than that of random prediction, with all the 118 evaluated agents being predictable with statistical significance ( $P < 0.02$ ). Specifically, 117 agents reached the significance level at  $P < 0.007$  and the remaining one at  $P < 0.019$ . The results showed that it was feasible to use a data set of only 60 diverse cell lines and 52 protein expression features to generate accurate and statistically significant chemosensitivity classifiers. Furthermore, we have also identified a proteomic signature for detection and diagnosis of colon cancer (Fig. 6).

In current study, we constructed a total of 118 protein expression-based classifiers, one for each anticancer drug. To identify the protein markers, random forests and Relief were used as protein filters. To achieve the optimal prediction results, random forests and the nearest neighbor methods were used as classifiers. The majority of the optimal classifiers were

built on random forests. The remaining ones were developed using the WEKA techniques (Relief and the nearest neighbor methods). This model system combined several sound algorithms and identified accurate classifiers achieving statistical significance. This framework provided a unique platform for integrating state-of-the-art machine learning methods and enabled the efficient and reliable performance in solving large-scale biomedical applications.

To the best of our knowledge, this is the first study to accurately ( $P < 0.02$ ) predict cell line chemosensitivity exclusively based on proteomic profiling. Furthermore, we improved on the previous work (3) by including the intermediate level in the prediction of drug response. Staunton et al. (3) built gene expression-based binary chemosensitivity classifiers by excluding the cell lines with intermediate response levels. They pointed out that their prediction models should be extended by including the intermediate levels for future clinical applications (3). In our analysis, the percentage of intermediate responses is considerable (Figs. 2 and 4) and should not be ignored. To achieve the goal of individualized therapy, drug sensitivity prediction must be extended beyond the cell line models and include primary patient material in the analysis (3). The NCI-60 panel was originally from clinical cancers. Generally speaking, they represent the biological properties of the corresponding cancer types. Using these cell lines to do various analyses allows for reproducible and stable experimental results. About the clinical samples and clinical testing, the same methodologies in molecular biology and bioinformatics can be applied. The present study showed the feasibility of screening samples for proteomic determinants of chemosensitivity to progress toward the goal of personalized medicine of cancer treatment.

## Acknowledgments

We thank Drs. David Lalka and Tim Vincent (West Virginia University, Morgantown, WV) for their valuable help and Dr. Shen Xiao (Food and Drug Administration, Bethesda, MD) for the thoughtful discussions.

## References

- Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006;439:353–7.
- Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000;24:236–44.
- Staunton JE, Slonim DK, Collier HA, et al. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A* 2001;98:10787–92.
- Szakacs G, Annereau JP, Lababidi S, et al. Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. *Cancer Cell* 2004;6:129–37.
- Solit DB, Garraway LA, Pratilas CA, et al. BRAF mutation predicts sensitivity to MEK inhibition. *Nature* 2006;439:358–62.
- Nishizuka S, Charboneau L, Young L, et al. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc Natl Acad Sci U S A* 2003;100:14229–34.
- Paweletz CP, Charboneau L, Bichsel VE, et al. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* 2001;20:1981–9.
- Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
- Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans Knowl Data Eng* 2003;15:1437–47.
- Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. New York: Morgan Kaufmann; 2005.
- Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int Jt Conf Artif Intell* 1995;1137–43.
- Diaz-Uriarte R, Alvarez dA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 2006;7:3.
- Aha D, Kibler D. Instance-based learning algorithms. *Mach Learn* 1996;6:37–66.
- Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer-Verlag; 2001.
- Basili R, Serafini A, Stellato A. Classification of musical genre: a machine learning approach. Proceedings of ISMIR 2004, 5th International Conference on Music Information Retrieval; 2004.
- Weinstein JN, Myers TG, O'Connor PM, et al. An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997;275:343–9.
- Munagala K, Tibshirani R, Brown PO. Cancer characterization and feature set extraction by discriminative margin clustering. *BMC Bioinformatics* 2004;5:21.
- Nishizuka S, Chen ST, Gwady FG, et al. Diagnostic markers that distinguish colon and ovarian adenocarcinomas: identification by genomic, proteomic, and tissue array profiling. *Cancer Res* 2003;63:5243–50.
- Benson AB III. Adjuvant chemotherapy of stage III colon cancer. *Semin Oncol* 2005;32:74–7.
- Rifkin R, Mukherjee S, Tamayo P, et al. An analytical method for multi-class molecular cancer classification. *SIAM Rev* 2003;45:706–23.