

# Appendix

## “An Integrative Genomic and Proteomic Approach to Chemosensitivity

### Prediction”

Yan Ma, Zhenyu Ding, Yong Qian, Ying-wooi Wan, Kursad Tosun, Xianglin Shi, Vincent Castranova, E. James Harner, and Nancy L Guo

### Table of Contents

1	Data Sets Description .....	2
2	Missing Value Imputation .....	3
3	Defining Cutoffs of Drug Sensitivity and Resistance.....	4
4	Feature Selection and Model Construction.....	5
5	Assessing the Significance of Our Prediction Results.....	10
6	References.....	11

### Table of Figures

Supplementary Figure 1. Chemosensitivity profiles of different cancer types (using 0.5 SDs as cutoff)	4
Supplementary Figure 2. Drug response prediction by use of proteomic and genomic profiling by method I.....	7
Supplementary Figure 3. Prediction accuracy of the constructed optimal classifiers by method I.....	7
Supplementary Figure 4. Drug response prediction by use of proteomic and genomic profiling by method II.....	8
Supplementary Figure 5. Prediction accuracy of the constructed optimal classifiers by method II .....	9
Supplementary Figure 6. Final prediction accuracy of the constructed optimal classifiers .....	9
Supplementary Figure 7. Improvement of the integrated molecular chemosensitivity classifiers over protein expression-based classifiers. ....	10

# 1 Data Sets Description

In this study, we used three data files from the NCI DISCOVER database (<http://discover.nci.nih.gov/>). Specifically, we predicted drug response of 60 human cancer cell lines (NCI-60) to 118 anti-cancer drugs by using both proteomic and genomic profiling. Following are the sources and brief descriptions of each data.

## **Data file 1: Proteomic profiling of the NCI-60 cancer cell lines.**

Link: [http://discover.nci.nih.gov/host/2003\\_profilingtable7.xls](http://discover.nci.nih.gov/host/2003_profilingtable7.xls)

Description: 52 protein expression profiles were analyzed across 60 human cancer cell lines (the NCI-60).

Column A: HUGO name

Column B: Antibody name

Column C: Vendor

Columns D-BK: Protein expression values on the 60 human cancer cell lines

The data file was generated from a previous publication by Nishizuka et al. (1). They developed a protocol for making reverse-phase protein lysate microarrays with a larger number of spots than previously feasible. They analyzed the data points for 52 antibodies by using P-SCAN and a quantitative dose interpolation method on 60 human cancer cell lines (NCI-60). The clustered images of the protein expression profiles for the 60 cancer cell lines revealed biologically meaningful patterns. The protein expression patterns also conformed to the mRNA expression patterns for the same genes related to cell structure.

## **Data file 2: Drug activity data of 118 “Mechanism of Action” drugs.**

Link:

[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)

Description: It is a database of 118 anti-cancer drugs activities across the NCI-60 cancer cell lines, whose mechanisms of action are putatively understood. Some of these drugs are currently in routine clinical use for cancer treatment; others are either in clinical trials or in late stages of drug development.

Column A: Mechanism of action

Column B: Drug name

Column C: NSC number

Columns D-BK: Drug activities ( $-\log_{10} GI_{50}$ ) across 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 a compound consists of 60 such activity values, one for each cell line.

The data file was from a previous publication by Scherf et al. (2). They used cDNA microarrays to assess gene expression profiles in the NCI-60 lines, and correlated gene expression and drug activity patterns in the NCI-60. They concluded that clustering the cell lines based on gene expression entailed relationships very different from those obtained by clustering the cell lines based on their response to drugs. In addition, gene-drug relationships for the chemotherapeutic agents 5-fluororacil (5-FU) and *L*-asparaginase exemplified how variations in the transcriptional levels of particular genes relate to mechanisms of drug sensitivity and resistance.

### **Data file 3: A gene expression database for the molecular pharmacology of cancer.**

Link: <http://discover.nci.nih.gov/nature2000/natureintromain.jsp>

Description: Gene expression profiles were measured in the 60 cell lines. This data file contains 1,375 genes which were selected from the initial 9,703 gene spots after applying filtering rules. These genes showed strong patterns of variation among the cell lines and had at most four missing values across the 60 cell lines. The missing values resulted from insufficient resolution, image corruption, dust, or scratches on the slide, etc.

Column A: IMAGE Clone ID

Column B: Gene description

Column C: 5ACC: 5' genebank accession number

Column D: 3ACC: 3' genebank accession number

Columns E-BL: Gene expression levels expressed as  $\text{Log}_2(\text{ratio})$ , where ratio = the red/green fluorescence ration after computational balancing of the two channels.

This data file was also from Scherf et al. (2) as data file 2. Since only IMAGE Clone ID is provided for each gene, we used *MatchMiner* (<http://discover.nci.nih.gov/matchminer/MatchMinerLookup.jsp>) to search for the gene symbols. *MatchMiner* is a set of tools that enable the user to translate between disparate IDs for the same gene. It uses data from the UCSC, LocusLink, Unigene, OMIM, Affymetrix and Jackson data sources to determine how different IDs relate. Supported ID types include gene symbols and names, IMAGE and FISH clones, GenBank accession numbers, and UniGene cluster IDs.

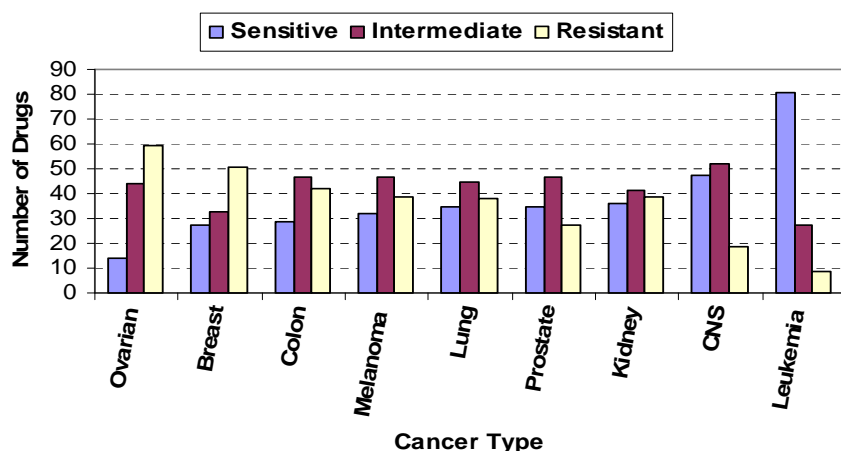
## **2 Missing Value Imputation**

Data file 3 of gene expression levels across the NIC-60 panel contains missing values. For instance, one gene (IMAGE Clone ID “48289”) has 40 out of 60 missing values. The remaining genes have either complete gene expression measurements across all the cell lines or up to 4 missing values. There are 1,374 genes left after removing the single gene with 40 missing measurements.

A nearest neighbor method (3) was used to provide accurate and robust estimate of missing values. Suppose gene  $g$  has missing values on array  $i$ , the weighted average of  $k$  nearest genes with values on array  $i$  is used to replace this missing value. Imputation results were found to be stable and accurate for  $k = 10 \sim 20$  neighbors (4). In these experiments, we tried  $k$  from 1 to 20 and found that the substituted values tended to converge starting from  $k = 11$ . By default, the number of neighbors is defined to be  $0.01 \times (\text{number of genes in the data})$ , so we chose  $k$  to be 13 which is also within the convergence range. Correlation was used as the similarity metric to search for the neighbors. *EMV* package in software *R* (<http://www.r-project.org>) was applied to replace missing values.

### 3 Defining Cutoffs of Drug Sensitivity and Resistance

The data file containing drug activity data of 118 anti-cancer agents was processed to define drug resistance and sensitivity of the NCI-60 lines, as described previously (5). Specifically, for each drug,  $\log_{10}(\text{GI}_{50})$  values were normalized across the 60 cell lines. Cell lines with  $\log_{10}(\text{GI}_{50})$  at least 0.5 SDs above the mean were defined as *resistant* to this drug; those with  $\log_{10}(\text{GI}_{50})$  at least 0.5 SDs below the mean were defined as *sensitive* to the drug; while the remaining cell lines with  $\log_{10}(\text{GI}_{50})$  within 0.5 SDs above or below the mean were defined as *intermediate* in the range of drug responses. We summarized the chemosensitivity profiles of each cancer type using this cutoff (i.e., 0.5 SDs) by averaging the profiles on the cell lines for each cancer type in the NCI-60 lines (Supplementary Figure 1).



**Supplementary Figure 1.** Chemosensitivity profiles of different cancer types (using 0.5 SDs as cutoff)

## 4 Feature Selection and Model Construction

In the NCI-60 gene expression data file, an expression profile  $\mathbf{x}_i = (x_{1i}, \dots, x_{Gi})'$  is associated with each cell line  $i$ . Gene expression data on  $G$  genes for  $n$  cell lines can be summarized by a

$G \times n$  matrix  $X = (x_{gi})$ , where  $x_{gi}$  denotes the expression measure of gene  $g$  in cell line  $i$ . In our study,  $X$  has dimension  $1,374 \times 60$  after data pre-processing (see Section 2).

Using similar notation, protein expression data on  $P$  proteins for  $n$  cell lines can be denoted as a

$P \times n$  matrix  $Y = (y_{pi})$ , where  $y_{pi}$  is the expression measure of protein  $p$  in sample  $i$ . In our experiment,  $Y$  has dimension  $52 \times 60$ , representing 52 proteins assayed in the 60 human cancer cell lines.

Each cell line has a class label with three values (i.e., three drug response categories: sensitive, intermediate, or resistant). Our previous project classified cell line chemosensitivity exclusively based on the protein expression profiles. In our current study, the goal is to predict the response of the cell lines to each anti-cancer drug using both proteomic and genomic profiles. Each cell line has an integrated gene-protein expression profile  $\mathbf{e}_i = (x_{1i}, \dots, x_{Gi}, y_{1i}, \dots, y_{Pi})'$  and a class label corresponding to a drug response. For each drug, chemosensitivity determinants might be a combination of protein expression profiles and gene expression profiles.

For each drug, the number of cell lines available for analysis is at most 60 with the missing values (some drugs were evaluated with fewer cell lines). Compared to the sample size, the number of features (i.e. proteins and genes) is very large. Some of these features might not be very informative in discriminating between different drug response classes. Including irrelevant features would compromise the performance of the classifiers. Thus, our objective was to obtain a small subset of relevant features (i.e., proteins and/or genes) and to achieve a good prediction accuracy of drug responses. Feature selection was performed with *varSelRF* package in software *R* (<http://www.r-project.org>). Chemosensitivity classification accuracy was estimated using the OOB error rates with the random forests algorithm in *R*. The OOB cases were not used in the feature selection.

Random forest is an ensemble learner (6). The general principle of ensemble methods is to construct a collection of diverse models from one single training data set, instead of using an individual fit of a method. Random forest uses un-pruned trees as its base models. The final classification decision is made by majority voting over all the trees. In order to build diverse tree classifiers, two sources of randomness are introduced in growing the trees: 1) Bootstrap samples are taken from the original learning data set, and different trees are built upon different bootstrap samples. 2) The variables evaluated for splitting the tree nodes are a random subset of the whole feature set. In growing each single tree, about one-third of the bootstrap sample is not used, which is called “out-of-bag” (OOB) cases. These out-of-bag cases can be used as a testing set in model performance assessment. The error estimate based on the OOB set is unbiased. Therefore, there is no need to perform cross-validation or

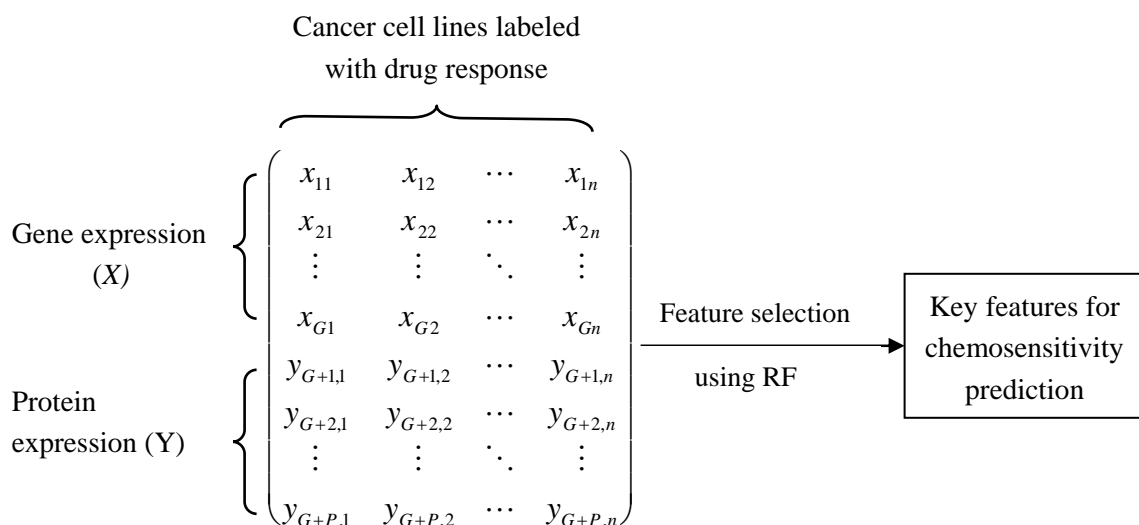
use a separate validation set (6). The random forests algorithm is well suited for processing large-scale microarray data and multi-classification problems (7).

The random forests algorithm provides variable importance evaluation. The importance of a variable is defined in terms of its contribution to prediction accuracy (6). If a feature plays a critical role in discriminating between distinct class labels, then randomly rearranging the values of this feature will decrease the classification accuracy significantly. The difference between the classification error based on the original data and the classification error based on randomly permuted data provides an importance measure of this feature. In this study, important features were selected with the *varSelRF* package in software *R*. The *varSelRF* package performs variable selection from random forests by eliminating the least important variables in a stepwise manner. The OOB error in each step is used as a feature selection criterion, instead of a model performance measure. The feature subset with the smallest OOB error was selected to build the chemosensitivity classifier.

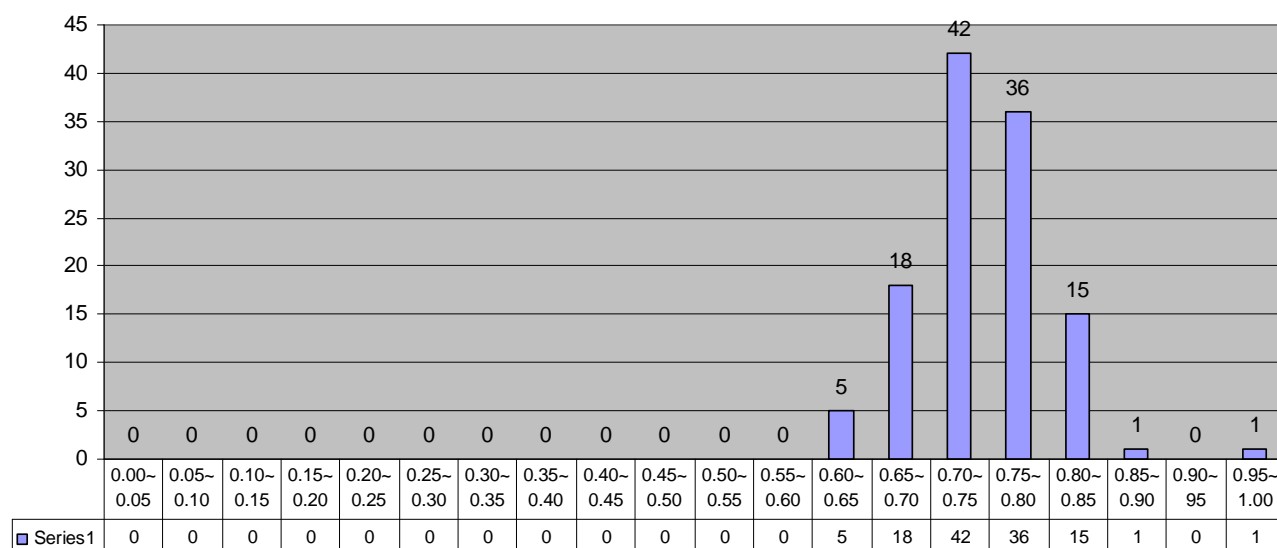
In our experiment, we performed the feature selection and then model construction in two ways. In Method I, for each drug, the gene expression profiles and protein expression profiles in the NCI-60 panel were integrated into a single file. The feature selection was performed on this integrated data file. Based on the identified features, a chemosensitivity classifier was constructed using the random forests algorithm. In Method II, for each drug, feature selection was performed separately on the gene expression profiles and the protein expression profiles. Then, the top gene features and top protein features were aggregated in a step-wise manner to build the classifier. Details are provided as following.

### **Method I. Aggregate the proteomic and genomic data prior to feature selection**

For each drug, a dataset with 1,426 predictive variables (including 52 proteins, 1,374 mRNAs) and 1 drug response variable was generated. Feature selection was performed on the aggregated proteomic and genomic expression profiles. A classification model was then built upon the relevant features identified in the feature selection process. This method is graphically outlined in Figure 2. Initially, a large forest was constructed based on this dataset. Feature importance was assessed during a single run of random forest (Supplementary Figure 2). Then, a random forest was built on the dataset with 20% of the least important features removed. This procedure was repeated until two variables remained in the dataset. In each iteration, the random forest produced an error estimate. The feature subset with the smallest error rate was selected. Here, the OOB error rate was used as a feature selection criterion, instead of an accuracy measure of model performance. Using the selected feature subset, the chemosensitivity classification was estimated by using the OOB error rates with the random forests algorithm. The chemosensitivity prediction accuracy for all evaluated drugs is reported in Supplementary Figure 3.



**Supplementary Figure 2.** Drug response prediction by use of proteomic and genomic profiling by method I

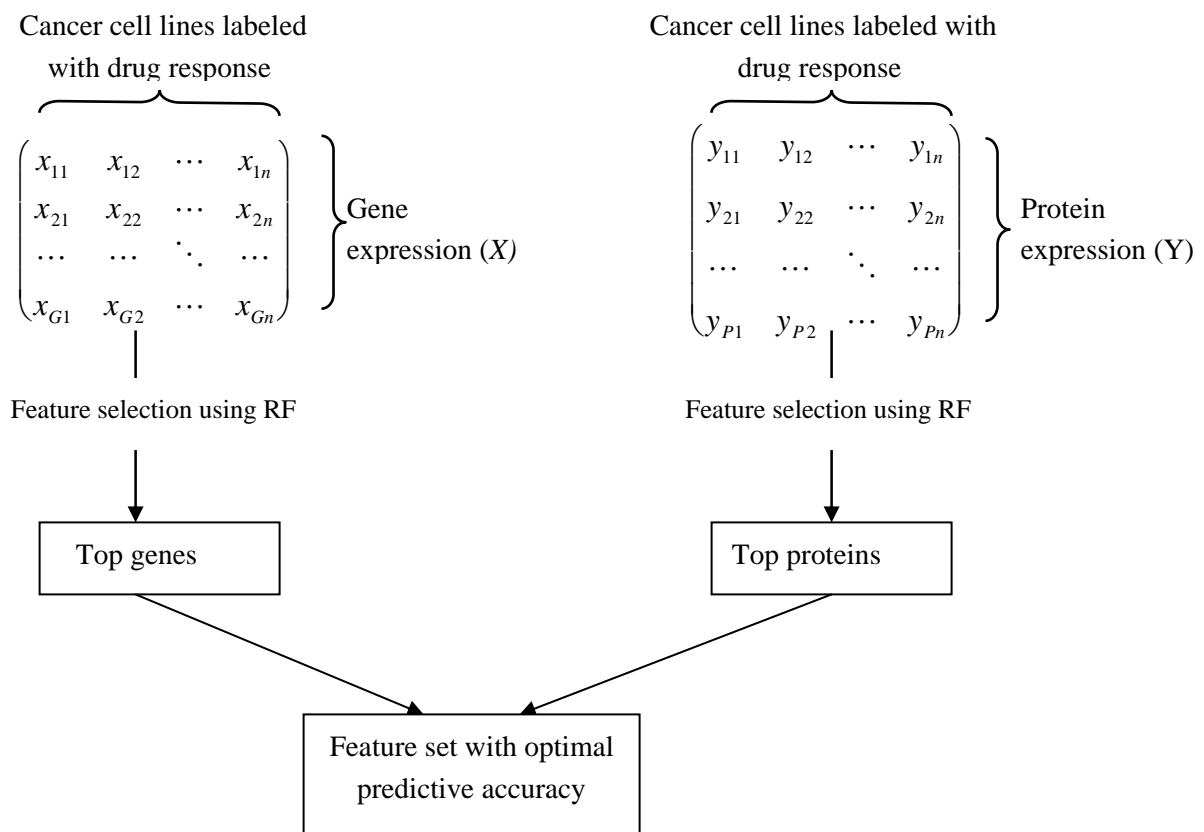


**Supplementary Figure 3.** Prediction accuracy of the constructed optimal classifiers by method I

### Method II. Aggregate the top-ranked proteins and genes from independent evaluation

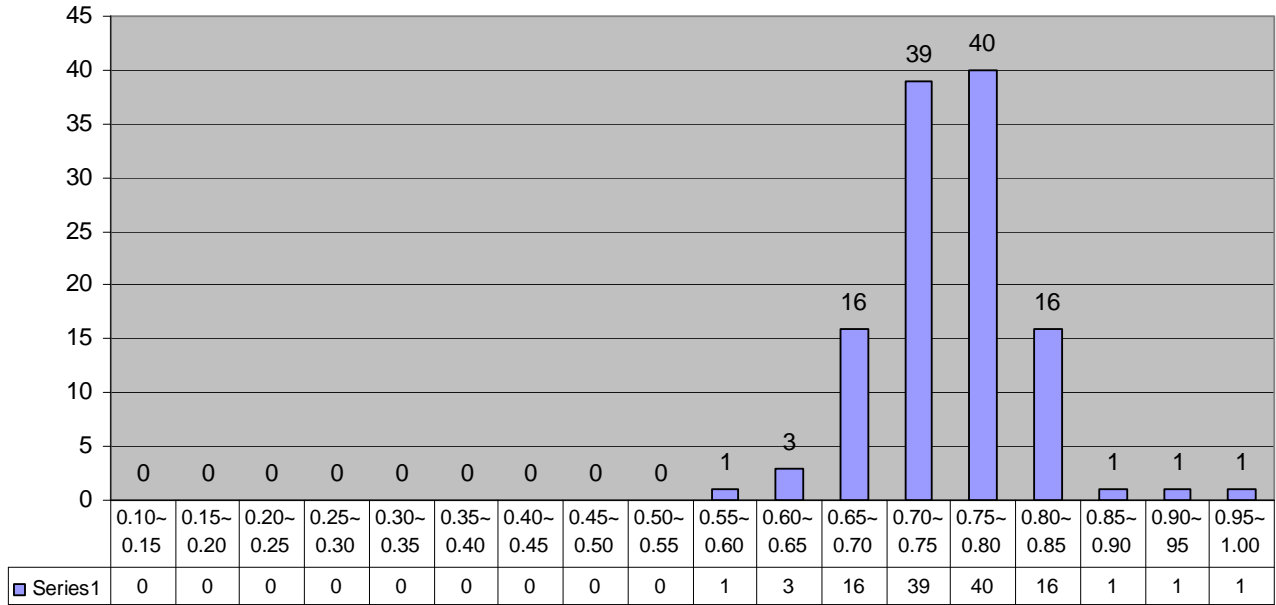
In Method II (Supplementary Figure 4), important gene and protein features were selected using the following three steps:

1. Starting from the proteomic profiling data, iteratively remove the least important protein variables. The subset of protein variables with the smallest error rate was selected.
2. Perform the same procedure as step 1 on the gene expression data to identify important gene features.
3. Add top-ranked protein variables identified in step1 to the top-ranked gene subset identified in step 2 one by one (starting from the most important protein variable). The feature set that generates the highest prediction accuracy with random forests is the optimal feature set. If protein variables do not improve the prediction accuracy, then the optimal feature set is the gene subset obtained in Step 2.



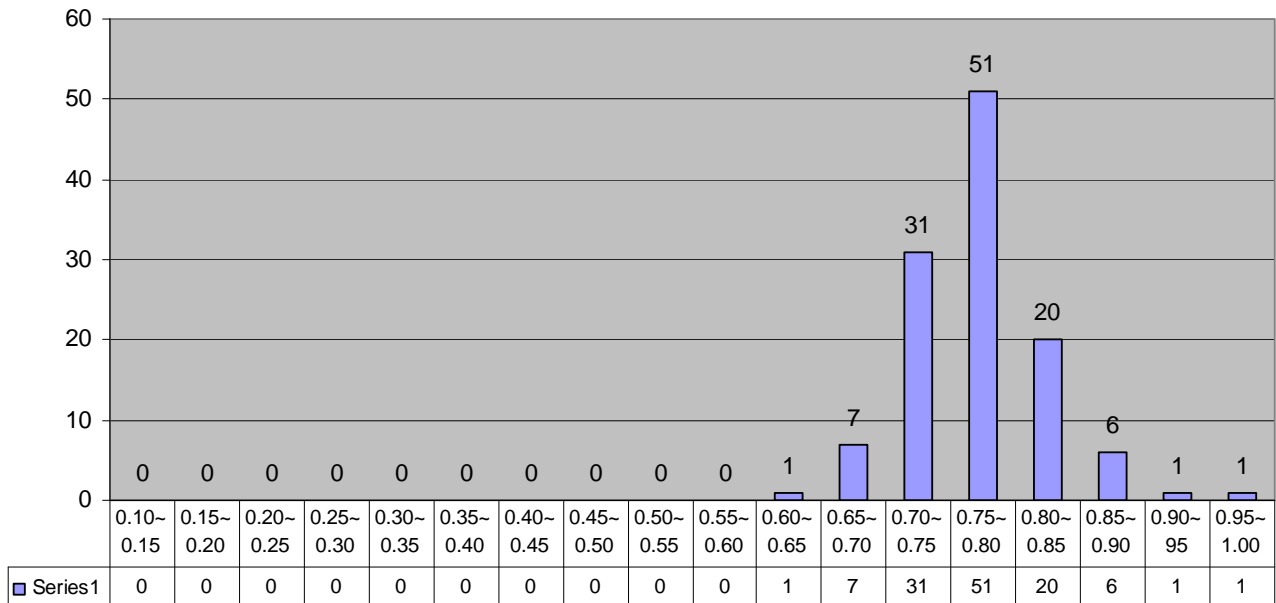
**Supplementary Figure 4.** Drug response prediction by use of proteomic and genomic profiling by method II

The distribution of the prediction accuracies from all drugs are shown in Supplementary Figure 5.



**Supplementary Figure 5.** Prediction accuracy of the constructed optimal classifiers by method II

Both methods have certain advantages in constructing the chemosensitivity classifiers for the evaluated drugs. Therefore, we chose the better results from both methods as our final results (Supplementary Figure 6).

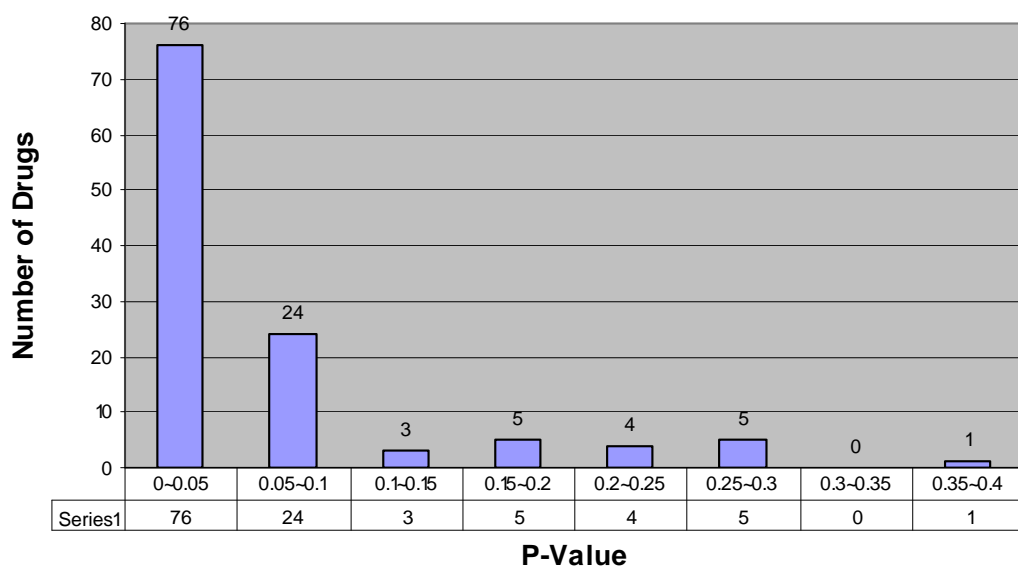


**Supplementary Figure 6.** Final prediction accuracy of the constructed optimal classifiers

## 5 Assessing the Significance of Our Prediction Results

In order to assess the significance of our prediction results, it is necessary to demonstrate that our prediction results are significantly better than random prediction. Two methods were used for the purpose. In the first method, for each drug we maintained the original class distributions and randomly permuted the class labels. For instance, a drug is examined on 60 cell lines, and the first 18 are labeled as “intermediate”, the next 23 as “resistant”, and the last 19 as “sensitive”. Random permutation produces 60 class labels, while keeping the class distribution fixed (18 intermediate, 23 resistant, and 19 sensitive). Using this method, the matches between the rearranged class labels and the original ones were recorded. The percentage of matches was calculated as the accuracy measure for random prediction. Repeating this procedure 1000 times generated 1000 accuracies. The  $P$  value was calculated as the upper percentile of our prediction accuracy in the profile of 1000 random prediction results. If the prediction accuracy produced by our classifier exceeds the 95<sup>th</sup> percentile of those 1000 random prediction accuracies, it is concluded that our prediction is significantly better than random prediction ( $P < 0.05$ ). For all 118 drugs, we obtained  $P$ -value of zero. Therefore, our chemosensitivity classifiers are significantly more accurate than random prediction ( $P < 0.001$ ).

We also evaluated the improvement of these prediction results compared to our previous study, in which the classifiers were exclusively based on protein expression profiles (5). Supplementary Figure 7 shows the distribution of  $P$ -values resulting from the significance tests. Seventy-six out of 118 classifiers are more accurate using the integration of both genomic and proteomic classifiers ( $P < 0.05$ ) than the protein expression-based classifiers identified previously (5).



**Supplementary Figure 7.** Improvement of the integrated molecular chemosensitivity classifiers over protein expression-based classifiers.

## 6 References

1. Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M 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.
2. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L et al. A gene expression database for the molecular pharmacology of cancer. *Nat.Genet.* 2000;24:236-44.
3. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R et al. Missing value estimation methods for DNA microarrays. *Bioinformatics.* 2001;17:520-5.
4. Dudoit S, Fridlyand J. Classification in microarray experiments. In: Speed T. *Statistical analysis of gene expression microarray data*. Chapman & Hall/CRC, 2003:93-159.
5. Ma Y, Ding Z, Qian Y, Shi X, Castranova V, Harner EJ et al. Predicting Cancer Drug Response by Proteomic Profiling. *Clinical Cancer Research* 2006;12:4583-9.
6. Breiman L. Random Forests. *Machine Learning* 2001;45:5-32 .
7. Diaz-Uriarte R, Alvarez dA. Gene selection and classification of microarray data using random forest. *BMC.Bioinformatics.* 2006;7:3.