An integrative genomic and proteomic approach to chemosensitivity prediction

YAN MA[1,2], ZHENYU DING[1,3], YONG QIAN[4], YING-WOOI WAN[1,3], KURSAD TOSUN[1,2], XIANGLIN SHI[4], VINCENT CASTRANOVA[4], E. JAMES HARNER[2] and NANCY L. GUO[1,3,5]

[1]Mary Babb Randolph Cancer Center, West Virginia University, Morgantown, WV 26506-9300; [2]Department of Statistics, West Virginia University, Morgantown, WV 26506-6330; [3]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506-6109; [4]The Pathology and Physiology Research Branch, Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505; [5]Department of Community Medicine, West Virginia University, Morgantown, WV 26506, USA

Received July 14, 2008; Accepted October 22, 2008

DOI: 10.3892/ijo_00000134

Abstract. New computational approaches are needed to integrate both protein expression and gene expression profiles, extending beyond the correlation analyses of gene and protein expression profiles in the current practices. Here, we developed an algorithm to classify cell line chemosensitivity based on integrated transcriptional and proteomic profiles. We sought to determine whether a combination of gene and protein expression profiles of untreated cells was able to enhance the performance of chemosensitivity prediction. An integrative feature selection scheme was employed to identify chemosensitivity determinants from genome-wide transcriptional profiles and 52 protein expression levels in 60 human cancer cell lines (the NCI-60). A set of 118 anti-cancer drugs whose mechanisms of action were putatively understood was evaluated. Classifiers of the complete range of drug response (sensitive, intermediate, or resistant) were generated for the evaluated anti-cancer drugs, one for each agent. The classifiers were designed to be independent of the cells' tissue origins. The classification accuracy of all the evaluated 118 agents was remarkably better ($P<0.001$) than that would be achieved by chance. Furthermore, 76 out of the 118 classifiers identified from integrated genomic and protein profiles significantly ($P<0.05$) improved the accuracy of protein expression-based classifiers identified previously. These results demonstrate that our integrated genomic and proteomic approach enhances the performance of chemosensitivity prediction. This study presents a new analytical framework to identify integrated gene and protein expression signatures for predicting cellular behavior and clinical outcome in general.

Introduction

Prediction of chemosensitivity in clinics is of great challenge, which involves both intrinsic properties of the cells and acquired resistance from treatment. There are multiple molecular mechanisms that affect chemosensitivity, which include alterations in drug influx and efflux, drug inactivation, expression or mutation of drug targets, DNA damage repair processes, cell cycle arrest and apoptosis (1). Most of these alterations are governed by changes in proteins and protein-genome interactions. There have been several studies integrating different levels of molecular profiles. A recent study integrated data on DNA copy number with gene expression levels and drug sensitivities in cancer cell lines (2). Another study integrated genomic and proteomic profiles to predict the tissue origin of cancer cell lines (3). Chemosensitivity prediction by integrated genomic and proteomic profiling is needed to reveal the mechanisms of chemosensitivity at multiple molecular levels. Such an approach could potentially be utilized to assess an individual patient's response to certain drugs in personalized cancer care. Previous attempts to integrate genomic and proteomic cancer profiling have been focused on the identification of correlated gene and protein expression patterns (3-6). Only the markers with correlated RNA and protein expression patterns were included in the prediction model and the markers without correlated RNA and protein expression were excluded from the prediction model (6). While the concordant gene and protein expression levels confirm the involvement of both genes and proteins in cancer progression and drug response, the discordant expression levels also provide insights into the critical

Correspondence to: Dr Nancy L. Guo, 1814 HSS, Mary Babb Randolph Cancer Center, P.O. Box 9300, Morgantown, WV 26506-9300, USA
E-mail: lguo@hsc.wvu.edu

Dr Yong Qian, Pathology and Physiology Research Branch, Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505, USA
E-mail: yaq2@cdc.gov

Key words: chemosensitivity, genomic profiling, proteomic profiling, bioinformatics, NCI-60 cell lines

biological processes. Both proteins and genes modulate their effects on cells through the changes in expression, activation, and post-translational modification, as well as the interactions between protein-protein, gene-gene and protein-gene. Notably, many DNA binding proteins, transcriptional factors and tumor suppressors regulate cancer drug sensitivity through protein-gene interactions. Therefore, a new perspective is needed to integrate both protein expression and gene expression profiles, extending beyond the correlation analyses of gene and protein expression profiles. In this study, we present a methodology to identify multi-leveled molecular signatures revealing complex mRNA and protein involvements in chemosensitivity. This methodology was tested by using an extensive database developed by the National Cancer Institutes' Developmental Therapeutics Program in a panel of 60 human cancer cell lines (the NCI-60). Previously, genome-wide transcriptional profiling was performed using cDNA microarrays (7). Later, proteomic profiling of the NCI-60 was carried out by using high-density reverse-phase lysate microarrays (5). Meanwhile, these cell lines were analyzed for their sensitivity to a broad range of chemical compounds, including anticancer drugs that were used either in clinics or in late development stages. These data are publicly available from the National Cancer Institute's Discover website (http://discover.nci.nih.gov/datasets.jsp).

We investigated the feasibility of chemosensitivity prediction by integrating transcriptional and proteomic profiles. The goal was to identify multiple expression-leveled molecular signatures of drug response to reveal the interrelated and collaborative molecular mechanisms for chemosensitivity. We hypothesized that relevant gene and protein expression signatures would enhance the accuracy in drug sensitivity prediction. To test this hypothesis, the expression profiles of 9,706 genes and 52 proteins in each of the 60 cancer cell lines were analyzed. Two feature selection algorithms were explored to identify gene and protein chemosensitivity determinants. A panel of 118 anticancer drugs was selected to develop and test our methodology. Our methodology employed the random forests algorithm (8,9) implemented in open source software R (10) (http://www.r-project.org/). To obtain an unbiased evaluation, a bootstrapped out-of-bag method (8) was used to assess the prediction performance. Compared with random prediction, all chemosensitivity classifiers for the evaluated 118 drugs were significantly accurate (P<0.001). Seventy-six of the 118 integrative gene-protein expression-based classifiers were more accurate (P<0.05) than the classifiers exclusively based on protein expression levels (11). Our results showed that the multi-leveled gene and protein expression signatures remarkably increased the accuracy of drug response prediction. This study suggests that chemosensitivity mechanisms are more readily dissected by a systems perspective combining bioinformatics, genomics and proteomics.

Materials and methods

Database sources

Gene expression data. The gene expression data were generated by Scherf *et al* (7). Cell collection and mRNA purification were described previously (7). Gene expression levels were quantified using cDNA microarrays (Synteni, Inc.; now Incyte, Inc., Wilmington, DE) consisting of robotically spotted, PCR-amplified cDNAs on coated glass slides (7,12). Expression profiles of 9,706 genes were assayed and the data file is available on-line (http://discover.nci.nih.gov/datasets Nature2000.jsp).

Protein expression data. The protein expression data file was generated by Nishizuka *et al* (5). 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 (http://discover.nci.nih.gov/host/2003_profilingtable7.xls).

Drug activity data. The drug activity profiles of 118 anti-cancer agents were screened by Scherf *et al* (7). Growth inhibition was assessed from the changes in total cellular protein after 48 h of drug treatment using a sulphorhodamine 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 (http://discover.nci.nih.gov/nature2000/data/selected_data/dataviewer.jsp?baseFileName=a_matrix118&nsc=2&dataStart=3).

Data preprocessing

Gene screening and identification. The original gene expression data screened 9,706 genes in 60 human cancer cell lines. For data quality control, 1,374 genes were selected for analysis after removing genes that had more than four missing values in the 60 cancer cell lines. The missing values resulted from insufficient resolution, image corruption, dust, or scratches on the slides, etc. All these selected genes showed a strong pattern of variation across the 60 cancer cell lines. Since only IMAGE Clone ID was provided for each gene in the original data, we used MatchMiner (13) to search for the gene symbols.

Missing value replacement. A nearest neighbor method (14) was used to provide accurate and robust estimate of missing values. Suppose gene *g* had missing values on array *i*. The weighted average of *k* nearest genes with values on array *i* was used to replace this missing value. Imputation results were found to be stable and accurate for *k* = 10-20 neighbors (15). We experimented *k* from one to 20 and found that the substituted values tended to converge starting from *k* = 11. We chose *k* = 13 which is within the convergence range. Correlation was used as the similarity metric to search for the neighbors. The EMV package (14) in R was applied to replacing the missing values.

Defining 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. 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 SDs above the mean were defined as

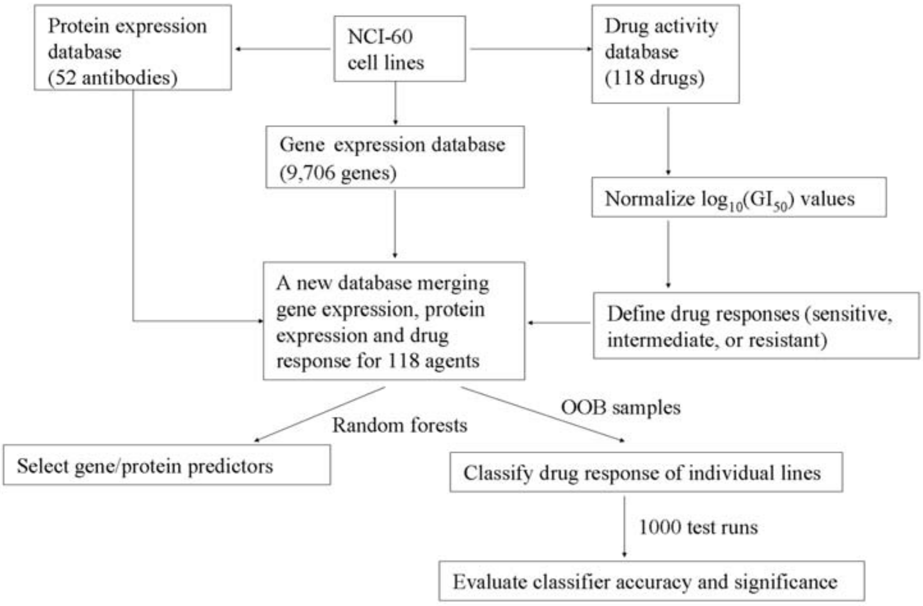
Protein expression database (52 antibodies)
NCI-60 cell lines
Drug activity database (118 drugs)
Gene expression database (9,706 genes)
Normalize $\log_{10}(GI_{50})$ values
A new database merging gene expression, protein expression and drug response for 118 agents
Define drug responses (sensitive, intermediate, or resistant)
Random forests
OOB samples
Select gene/protein predictors
Classify drug response of individual lines
1000 test runs
Evaluate classifier accuracy and significance

Figure 1. Scheme of the model system for drug response prediction.

resistant to this drug. Those with $\log_{10}$ (GI_{50}) at least 0.5 SDs below the mean were defined as sensitive to the drug. The remaining cell lines with $\log_{10}$ (GI_{50}) within 0.5 SDs were defined as intermediate in the range of drug responses.

Constructing chemosensitivity classifiers

Biomarker identification and classifier construction. The random forest algorithm is an ensemble of tree classifiers. The basic step of random forests is to form diverse tree classifiers from a single training set. Each tree is built using a different bootstrap sample from the original data. About one-third of the cases are not used in the construction of a tree. These cases are called out-of-bag (OOB) cases. Random forests introduce two sources of randomness into the algorithm: i) each tree is grown out of a random sample of the original data and ii) each node in the tree is split by the best variable from a randomly selected subset of variables. In this study, two functions of random forests implemented in R were used to identify chemosensitivity determinants and to construct drug response classifiers. The feature selection experiments were performed using the varSelRF package of R (9). The feature subset with the smallest OOB error was chosen as the optimal feature subset. Using the identified feature subset, the random forests algorithm was used to obtain a classification error using the OOB error rates. The OOB samples were not used in the feature selection.

Classification accuracy evaluation. In order to assess the significance of our prediction results, it is necessary to demonstrate that our prediction results are significantly better than those of random predictions. For each drug (each data file), the class labels of the 60 cell lines were randomly permuted while keeping the number of instances in each group fixed. The matches between the rearranged class labels and the original ones were recorded. The percentage of the matches was calculated as the overall accuracy for the random prediction. This procedure was repeated 1,000 times. The p-value was calculated as the number of random predictions that exceeded our prediction accuracy in 1,000 test runs. The experimental details and prediction results are provided in Appendix (http://www.hsc.wvu.edu/mbrcc/fs/GuoLab/publications.asp).

Results

In this study, we sought to predict drug response by integrated analyses of transcriptional profiles and proteomic profiles. In clinics, treatment response to chemotherapy is categorized into complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD) according to the RECIST guidelines (16). A patient's response to treatment is defined based on the change in the tumor size before and after the chemotherapy. Responders include patients with complete response (CR) and partial response (PR). In this study, we used cancer cell lines to investigate chemosensitivity. Chemosensitivity was defined based on the drug activity profiles in 60 human cancer cell lines (the NCI-60). For each evaluated agent, the complete range of drug responses across these 60 cancer cell lines was partitioned into three classes (sensitive, intermediate, or resistant) based on the normalized growth-inhibitory activities (GI_{50} values) (11,17) (Materials and methods). Chemosensitivity prediction was approached as a supervised multi-classification problem.

The NCI-60 set includes the cell lines derived from leukemias, melanomas, as well as carcinomas of ovarian, renal, breast, prostate, colon, lung and central nervous system origin. The transcriptional profiling of 9,706 genes was performed using cDNA microarrays in the NCI-60 (7). The proteomic profiles were generated using 52 antibody reverse-phase protein assays in the NCI-60 (5). In this study, a panel of 118 anticancer drugs was selected to develop and test our methodology. The mechanisms of action of these 118 agents were putatively understood (7). The overall scheme of our model system is depicted in Fig. 1. To construct the supervised expression-based classifiers for chemosensitivity prediction, we created a new database by merging the transcriptional profiles, proteomic profiles and the responses of the NCI-60 cell lines to the 118 drugs (Fig. 1). For each drug, the data file contained the expression levels of 1,374 genes (after data preprocessing), the protein expression levels measured by 52 antibodies and the response of each cell line to this drug. For each cell line, the gene and protein expression levels 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. We sought to identify collaborative gene and protein expression patterns to increase the accuracy in chemosensitivity prediction.

The feature selection and classification evaluation was performed in following steps: i) leave six samples out and refer to them as the out-of-bag (OOB) samples; ii) perform feature selection to identify the optimal subset using the remaining samples; iii) create a random forest with the identified feature subset using the remaining samples; iv) classify the OOB samples and note the error rates; v) repeat for another set of OOB samples (the procedure was repeated for 10 times); and vi) compute the average error.

Two feature selection schemes were employed to identify the chemosensitivity determinants for each of the evaluated drugs (Fig. 2). Two functions of the random forests algorithm were used in the feature selection and classifier construction. Specifically, in the NCI-60 transcriptional data file, an expression profile $\mathrm{x_i} = x_{1i},...x_{Gi})$ is associate with each cell line *i*. Gene expression data on *G* genes for *n* cell lines can be represented by a *G* x *n* matrix, $\mathrm{X} = (x_{gi})$, where x_{gi} denotes the expression level of gene *g* in cell line *i*. In this study, *X* has dimension 1,374 x 60 after data preprocessing. Using similar notation, protein expression data on *P* proteins for *n* cell lines can be denoted as a *P* x *n* matrix, $\mathrm{Y} = (y_{pi})$, where y_{pi} is the expression level of protein *p* in sample *i*. *Y* has dimension 52 x 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). Each cell line has an integrated gene-protein expression profile $e_i = (x_{1i}...x_{Gi},y_{1i},...y_{Pi})'$ and a class label corresponding to a drug response. These expression profiles were handled by matrix transpose functions in random forests.

In feature selection Method I, for each drug, the gene expression and protein expression profiles in the NCI-60 panel were first combined into a single file. Then, feature selection was performed on this combined data file using a backward elimination to identify the optimal feature subset (9). This optimal feature subset contained top-ranked gene and/or protein attributes that generated the highest prediction accuracy in the integrative feature selection (Fig. 2A). In the backward elimination, a random forest was first generated with 2,000 trees to obtain an importance rank of the gene and protein variables. Based on this importance rank, we repeatedly eliminated 20% of the least important variables from the remaining dataset until two variables left. In each iterative step, a new forest of 1,000 trees was constructed. The classification of drug response for a cancer cell line was determined by the majority vote of these 1,000 trees in the forest. The subset with the smallest OOB error rate was identified as the optimal feature subset. In Method II, for each drug, feature selection was performed separately on the gene expression profiles and the protein expression profiles to find the top gene subset and top protein subset, respectively, using the backward elimination as described above. Then, the identified top gene features and top protein features were integrated in a stepwise manner to build the classifier (Fig. 2B). Specifically, the top-ranked protein attributes were added to the top gene subset one by one (starting from the most important protein attribute) until the optimal accuracy was achieved. The feature subset that generated the highest prediction accuracy is the optimal subset. If the addition of protein attributes did not increase the prediction accuracy, then the optimal subset was the top gene subset. The second method was designed to account for unbalanced data. In this study, the available gene variables outnumber the protein variables. After selecting the top gene set and top protein set, different feature entities are reduced to comparable dimensions. In both approaches, the optimal feature subset was identified as the one with the smallest OOB error rate in the feature selection process using the varSelRF function in R (9) (Fig. 2C). The reported prediction accuracy was evaluated with the OOB error rates using the random forests algorithm (9,18). The OOB samples used in the evaluation were not included in the feature selection. The reported chemosensitivity classification performances were not the OOB error rates in the feature selection.

During the feature selection process, we did not follow the '1-Standard Error (1-SE) rule' as suggested by Diaz-Uriarte *et al* (9). This rule chooses the smallest gene subset, whose error rate is within one standard error of the minimum error rate of all forests. Instead, we used '0-Standard Error (0-SE) rule', which identifies the gene subset with the smallest OOB error rate. As the random forests algorithm does not guarantee the same feature subsets in each run, we chose the feature subset with the lowest prediction error in three runs with the varSelRF function.

Both feature selection methods have certain advantages in constructing the chemosensitivity classifiers for the evaluated drugs. Therefore, the optimal classifiers from both methods were chosen as the results. The optimal classifiers used between two and 40 predictors, with an average of 10 predictors in each classifier. Eighty-nine optimal classifiers were constructed exclusively based on gene expression profiles, while 29 optimal classifiers used a combination of gene expression and protein expression profiles. The overall accuracy of the optimal classifiers is summarized in Fig. 3A. We evaluated the prediction results by comparing them with the random prediction in 1,000 test runs (see Materials and methods for details). The results showed that, for all the evaluated 118 drugs, none of the random predictions in 1,000 iterations achieved our accuracy ($P<0.001$). Compared with the previous study on protein expression-based chemosensitivity prediction (11), 76 of the classifiers identified using this integrated approach performed better ($P<0.05$) than the classifiers exclusively based on protein expression levels (Fig. 3B). These results demonstrate that the integrated analyses of transcriptional profiles and proteomic profiles can enhance the performance of chemosensitivity prediction.

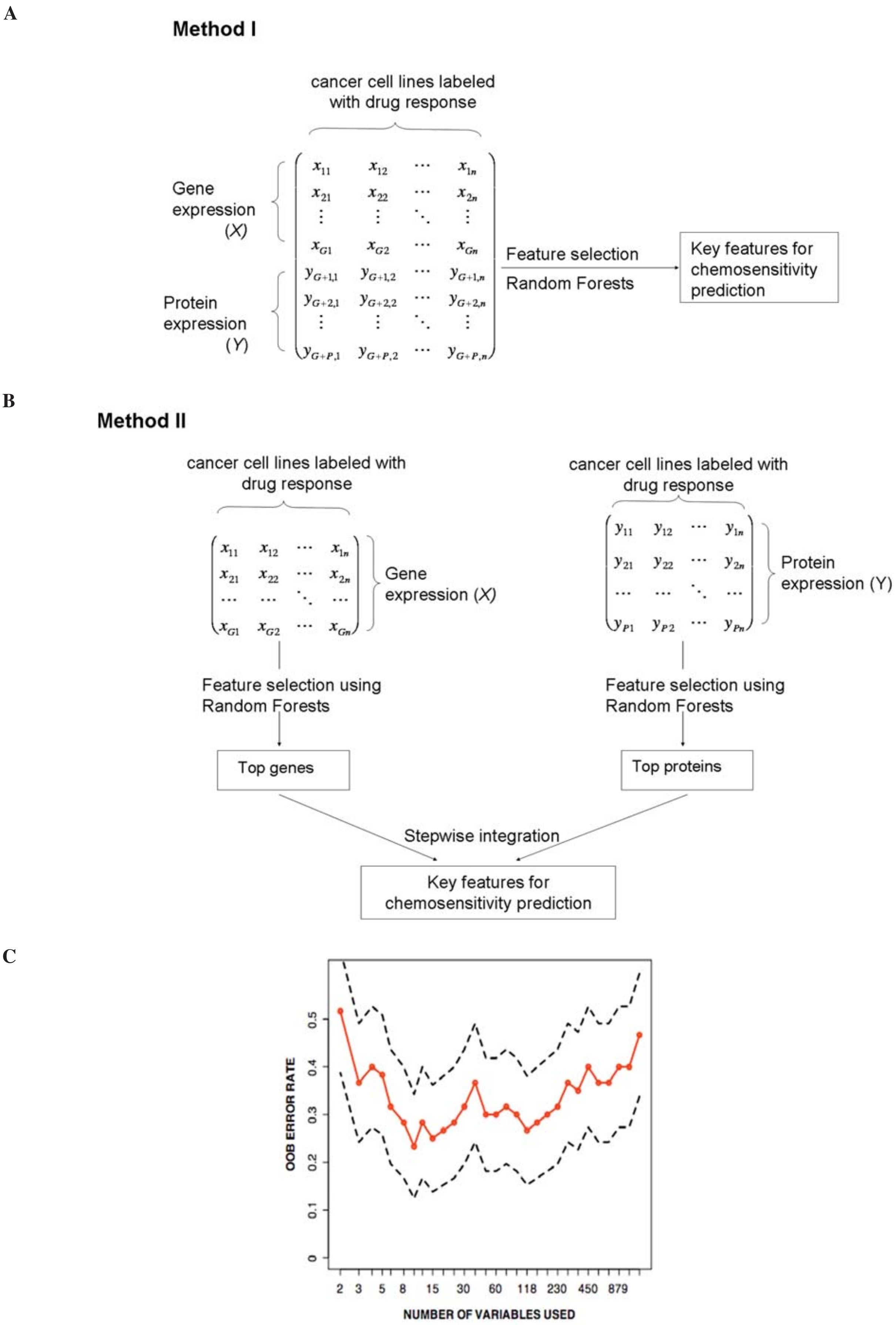
A
Method I
cancer cell lines labeled with drug response
Gene expression (X)
Protein expression (Y)
Feature selection
Random Forests
Key features for chemosensitivity prediction
B
Method II
cancer cell lines labeled with drug response
Gene expression (X)
cancer cell lines labeled with drug response
Protein expression (Y)
Feature selection using Random Forests
Feature selection using Random Forests
Top genes
Top proteins
Stepwise integration
Key features for chemosensitivity prediction
C
OOB ERROR RATE
0
0.1
0.2
0.3
0.4
0.5
2
3
5
8
15
30
60
118
230
450
879
NUMBER OF VARIABLES USED

Figure 2. Integrative feature selection scheme for identifying gene expression and protein expression signatures for chemosensitivity prediction. (A) Feature Selection Method I: the genomic and proteomic profiles were combined first, and the feature selection was performed on this combined profile. (B) Feature Selection Method II: feature selection was first performed separately on the genomic profile and proteomic profile. Then, the identified top gene subset and top protein subset were integrated in a stepwise manner, until the optimal accuracy was achieved. (C) The OOB error rate during feature selection. The optimal feature subset was the one with the smallest OOB error rate. In this example, the subset with top 10 features was the optimal feature subset.

Table I. Identified chemosensitivity protein markers for the evaluated anticancer drugs.

Protein markers	Antibody	Function	Drugs
ISGF3g	ISGF3g	Transcriptional factor Interferon signaling	Porfiromycin Morpholino-adriamycin
STAT3	Stat3	Transcriptional factor Interferon signaling	Porfiromycin
NME1	Nm23	Tumor suppressor Integrin signaling	Mitozolamide Piperazine mustard Azacytidine 5-Hydroxypicolinaldehyde thiosemicarbazone
MGMT	MGMT Ab-1	DNA metyl/alkyl-transferase DNA repair	PCNU
CCNE	Cyclin E (G1- and S-phase Cyclin) Ab-2	Cell cycle protein	Semustine (MeCCNU) Mechlorethamine
EP300	p300/CBP Ab-1	Transcriptional co-factor Tumor suppressor	Cisplatin Camptothecin,20-ester (S)
FN1	Fibronectin Ab-1	Cell adhesion protein Integrin signaling	Piperazinedione
MSN	Moesin Ab-1	Cytoskeleton protein Integrin signaling	Mitoxantrone Thiopurine (6MP) Dolastatin-10
PGR	Progesterone receptor (PgR) Ab-2	Steroid receptor Transcriptional receptor	Morpholino-adriamycin
STAT1	Stat1 C-terminus)	Transcriptional factor Inteferon signaling	Morpholino-adriamycin
STAT6	Stat6	Transcriptional factor Inteferon signaling	Morpholino-adriamycin
CASP2	Caspase-2/ICH-1L	Apoptosis protein	5-6-Dihydro-5-azacytidine
CDH1	E-Cadherin	Cell adhesion protein A parameter of cell metastasis	ß-2'-Deoxythioguanosine
MCP	CD46 (Mambrane Cofactor Protein) Ab-2	Immune response Inhibition of complement activation	Methotrexate
KRT18	Keratin 18 Ab-1	Structural protein A biomarker of cell death	Fluorouracil (5FU) Colchicine Colcichine-derivative Taxol (Paclitaxel) Taxol analog
TP53	p53 tumor suppressor protein Ab-8	Tumor suppressor Cell cycle and apoptosis	Fluorouracil (5FU) Taxol (Paclitaxel)
RELA	NF-κB p65	Transcriptional factor Cell cycle and apoptosis	L-Alanosine
G22P1	Ku (p70) Ab-4	DNA binding protein DNA repair DNA repair	N-phosphonoacetyl-L-aspartic-acid

The identified protein markers for chemosensitivity classification are listed in Table I. Literature review found that the changes in protein expression of all these identified protein markers either are correlated to or are directly involved in the sensitivity of various cancers to chemotherapy. Anticancer chemotherapy mainly targets a variety of signaling

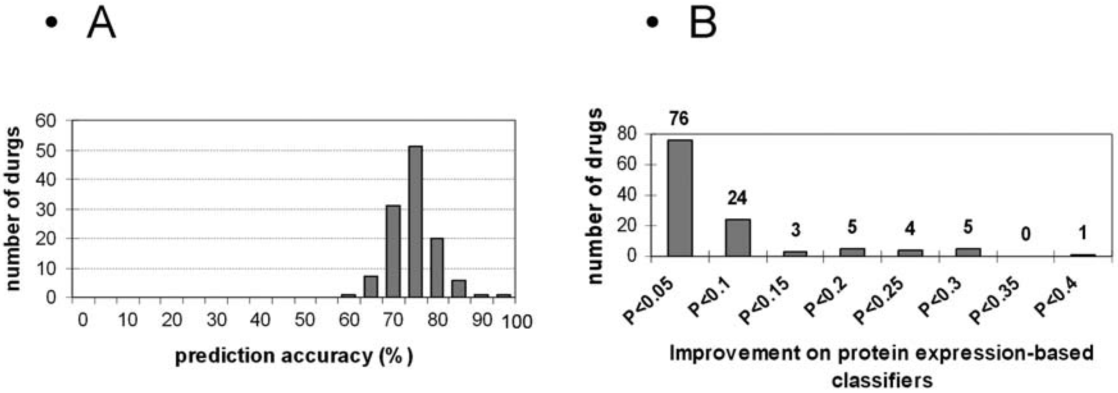

Figure 3. Overall accuracy for the 118 chemosensitivity classifiers. (A) Distribution of classification accuracy for the 118 drugs. The prediction accuracy is the percentage of correctly classified instances. (B) Performance improvement of the integrative chemosensitivity classifiers over protein expression-based classifiers identified in a previous study (11).

Table II. Identified gene and protein chemosensitivity markers for Taxol (Paclitaxel) (NSC 125973).

Chemosensitivity markers	Protein name	Function
Gene markers		
MMP2	Matrix metalloproteinases 2	Breakdown extracellular matrix
TXNDC5	Thioredoxin domain containing 5	Protein disulfide reductases, isomerases or oxidases
LEPROT	Leptin receptor overlapping transcript	Unknown
OSBPL1A	Oxysterol binding protein-like 1a	Sterol sensors
TWSG1	Twisted gastrulation homolog 1	Submandibular salivary gland ontogenesis
GALNT2	Udp-n-acetyl-alpha-d-galactosamine: polypeptiden-acetylgalactosaminyl-transferase 2	Glycosylation
PYCR1	Pyrroline-5-carboxylate reductase 1	Metabolism
Protein markers		
KRT18	Keratin 18	Structural protein A biomarker of cell death
TP53	p53 tumor suppressor protein	Tumor suppressor Cell cycle and apoptosis

transduction pathways to induce genome-wide response and ultimately leads to tumor clearance. The alteration of cell signaling transduction, either intrinsically inherent or acquired during treatment, plays a major role in modulating a particular cancer's sensitivity to chemotherapy (1). Our identified protein markers include eight transcription factors, three tumor suppressor proteins, two DNA repairing proteins, three cell cycle/apoptosis proteins, five cell cytoskeleton/adhesion proteins and one immune regulator protein (Table I).

In this study, 29 classifiers used a combination of gene expression and protein expression profiles. For instance, Taxol (Paclitaxel; NSC 125973) is used in the treatment of ovarian, breast and non-small cell lung cancer. The chemosensitivity classifier for Taxol contained seven gene markers and two protein markers (Table II). The overall accuracy of chemosensitivity prediction was 0.817, which was significantly ($P<0.02$) higher than the protein expression-based classifier identified previously (11). One of the selected protein markers is p53 protein, which affects cell functions through regulating many downstream gene expression. It has been well-established that p53 plays major roles in DNA repair, cell cycle arrest, cell apoptosis induction and cell pro-signaling alteration. Interestingly, it was found that p53 has a regulatory role in our selected gene markers MMP2 (19,20) and PYCR1 (21). MMP2 gene encodes MMP2 protein, a member of peptidase enzyme families responsible for the degradation of extracellular matrix components. The change in MMP2 expression is directly related to angiogenesis, tumor growth and metastasis (22). PYCR1 encodes pyrroline-5-carboxylate reductase (P5CR), which catalyzes the reduction of $\Delta(1)$-pyrroline-5-carboxylate (P5C) to praline using NAD(P)H as the cofactor (21). This enzyme may be involved in oxidation of the anti-tumor drug thioproline (21). These results demonstrated that the complementary classifiers with both proteomics and genomics information more accurately reflect the biological nature of cells, which leads to generating a highly accurate prediction system for chemosensitivity.

Discussion

Chemosensitivity mechanisms involve collaborative biological processes at the transcriptional level, translational regulation, posttranslational modification, proteasome function and protein-protein interactions (17). Current DNA microarrays are an extremely powerful technology for measuring mRNA expression of each particular gene. Technologies for globally and quantitatively measuring protein expression are also becoming available (5,23,24). Genome-wide transcriptional profiles of the NCI-60 using Affymetrix U133 and U95 chips were recently released, together with new proteomic profiles of the NCI-60 with 162 antibodies for 94 proteins (3). Although such large-scale data are proven invaluable in distinguishing cancer types and drug responses, new computational approaches are needed to integrate these diverse data types and assimilate them into biological models to predict chemosensitivity. Previous integrative analyses were focused on the correlation between different levels of expression patterns. In those studies, only the markers with concordant RNA and protein expression were included in the prediction models, while the markers with discordant RNA and protein expression were excluded from the prediction models. Those approaches might potentially miss some important biological information regarding chemosensitivity mechanisms, such as protein-protein interactions and protein-gene interactions. Furthermore, it was not clear whether the combination of protein and gene expression data could enhance the prediction accuracy. Here, we developed an integrated approach, extending beyond the correlation analysis, to identifying the gene and protein expression signatures. Two computational feature selection schemes were investigated in this study. An extensive database including genome-wide transcriptional profiles and 52-antibody protein assays on the NCI-60 cell lines was employed to develop and test our methodology. The results demonstrated that this integrated approach enhanced the performance of drug response prediction. The identified multi-level signatures provided insight into gene-protein collaborations in chemosensitivity.

A particular challenge of integrative chemosensitivity prediction is the small amount of available protein expression data due to the technical difficulties in proteomics (5). By the time of this analysis was finished, we had only found one proteomic data set generated from the NCI-60 panel. This data set contained protein expression levels measured by 52 antibodies (5). The number of the features in the proteomic data is 4% of that in the transcriptional data set used in this study. These unbalanced data make it difficult to construct integrated gene-protein expression-based chemosensitivity classifiers. We developed two stepwise feature selection schemes to account for the unbalanced gene expression and protein expression profiles. The optimal classifiers built from both approaches were selected as the results. As the variable importance measures in the randomForest package were not reliable in situations where variables vary in the scale of measurement (25), a more robust function, valSelRF by Diaz-Uriarte *et al* (9), was used in integrative feature selection on proteomics and genomics data in this study. After the feature subsets were identified, the OOB error rates using random forests were reported as classification performances. The constructed chemosensitivity classifiers were remarkably accurate ($P<0.001$) using the proposed methodologies. In the evaluation, 76 (64%) classifiers identified from both genomic and proteomic profiles outperformed the ones exclusively based on protein expression levels and 29 (25%) integrated classifiers outperformed the ones exclusively based on gene expression levels. These results demonstrated that our analytical approach to integrating protein expression and gene expression profiles is successful. The majority (75%) of optimal classifiers were exclusively based on gene expression profiles, which might result from the unbalanced number of genes and proteins in the data.

This study presented a new perspective for integrating different types of microarray data for predicting drug response. The random forests algorithm was used in this study for two reasons: i) random forests are well suited for processing large-scale microarray data and multi-classification problems (9) and ii) as the majority of previous protein expression-based chemosensitivity classifiers were constructed with random forests (11), utilizing random forests for integrative chemosensitivity classification can minimize the performance discrepancy due to the use of different analytical methods. It should be noted that other algorithms, such as support vector machine, *k*-nearest neighbor, Relief, and wrapper (26), may also be used in this general feature selection scheme. In the analyses, several established packages in R were used, including the *k*-NN imputation algorithm (the EMV package) (14), and the random forests feature selection function (the valSelRF package) (9). In addition, we developed software scripts to integrate different types of microarray data for efficient large-scale computation. After we conducted this analysis, additional microarray data for the NCI-60 (3), including the CGH (2) and microRNA profiles (27), as well as new cancer cell line profiles (28) have become available. This integrative methodology will be tested with new data in future research.

In this study, a novel scheme was developed to identify integrative gene and protein expression signatures to predict chemosensitivity. This is a general approach to systematically evaluate genome-wide DNA, RNA, and protein contributions in cancer progression and drug sensitivity. This methodology was tested by using an extensive database developed by the National Cancer Institute. In these large-scale case studies, cell line chemosensitivity classifiers were constructed for a broad range of anticancer drugs. The results demonstrated that our identified integrative gene and protein signatures were able to enhance the chemosensitivity prediction accuracy. This study indicated that cancer mechanisms are more readily revealed by a systems approach integrating genomics, proteomics, and bioinformatics. This study provides a new computational model to integrate genomic and proteomic profiles to predict cellular behavior and cancer outcomes in general.

Acknowledgements

Dr Nancy L. Guo was supported by the NIH/NCRR P20 RR16440-03 grant and West Virginia University Research Development Grant (RDG NT10017W). Dr Xianglin Shi is supported by the NIH/NCI 1R01CA119028-01grant.

References

1. Longley DB and Johnston PG: Molecular mechanisms of drug resistance. J Pathol 205: 275-292, 2005.
2. Bussey KJ, Chin K, Lababidi S, *et al*: Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. Mol Cancer Ther 5: 853-867, 2006.
3. Shankavaram UT, Reinhold WC, Nishizuka S, *et al*: Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol Cancer Ther 6: 820-832, 2007.
4. Chen G, Gharib TG, Huang CC, *et al*: Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 1: 304-313, 2002.
5. 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 USA 100: 14229-14234, 2003.
6. Varambally S, Yu J, Laxman B, *et al*: Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell 8: 393-406, 2005.
7. Scherf U, Ross DT, Waltham M, *et al*: A gene expression database for the molecular pharmacology of cancer. Nat Genet 24: 236-244, 2000.
8. Breiman L: Random Forests. Machine Learning 45: 5-32, 2001.
9. Diaz-Uriarte R and Alvarez dA: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7: 3, 2006.
10. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. 1st edition. Springer, New York, 2005.
11. Ma Y, Ding Z, Qian Y, *et al*: Predicting cancer drug response by proteomic profiling. Clinical Cancer Res 12: 4583-4589, 2006.
12. Shalon D, Smith SJ and Brown PO: A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 6: 639-645, 1996.
13. Bussey KJ, Kane D, Sunshine M, *et al*: MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biol 4: R27, 2003.
14. Troyanskaya O, Cantor M, Sherlock G, *et al*: Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520-525, 2001.
15. Speed T: Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC, 2003.
16. Therasse P, Arbuck SG, Eisenhauer EA, *et al*: New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 92: 205-216, 2000.
17. Staunton JE, Slonim DK, Coller HA, *et al*: Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci USA 98: 10787-10792, 2001.
18. Ambroise C and McLachlan GJ: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA 99: 6562-6566, 2002.
19. Lee JG, Dahi S, Mahimkar R, *et al*: Intronic regulation of matrix metalloproteinase-2 revealed by in vivo transcriptional analysis in ischemia. Proc Natl Acad Sci USA 102: 16345-16350, 2005.
20. Mertens PR, Steinmann K, Fonso-Jaume MA, En-Nia A, Sun Y and Lovett DH: Combinatorial interactions of p53, activating protein-2, and YB-1 with a single enhancer element regulate gelatinase A expression in neoplastic cells. J Biol Chem 277: 24875-24882, 2002.
21. Meng Z, Lou Z, Liu Z, *et al*: Crystal structure of human pyrroline-5-carboxylate reductase. J Mol Biol 359: 1364-1377, 2006.
22. Klein G, Vellenga E, Fraaije MW, Kamps WA and de Bont ES: The possible role of matrix metalloproteinase (MMP)-2 and MMP-9 in cancer, e.g. acute leukemia. Crit Rev Oncol Hematol 50: 87-100, 2004.
23. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH and Aebersold R: Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17: 994-999, 1999.
24. Ideker T, Thorsson V, Ranish JA, *et al*: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292: 929-934, 2001.
25. Strobl C, Boulesteix AL, Zeileis A and Hothorn T: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8: 25, 2007.
26. Witten IH and Frank E: Data Mining: In: Practical Machine Learning Tools and Techniques. 2nd edition. Morgan Kaufmann, 2005.
27. Blower PE, Verducci JS, Lin S, *et al*: MicroRNA expression profiles for the NCI-60 cancer cell panel. Mol Cancer Ther 6: 1483-1491, 2007.
28. Lee JK, Havaleshko DM, Cho H, *et al*: A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. Proc Natl Acad Sci USA 104: 13086-13091, 2007.