J. Biomedical Science and Engineering, 2009, 2, 506-515 **JBiSE**
doi: 10.4236/jbise.2009.27073 Published Online November 2009 (http://www.SciRP.org/journal/jbise/).

Correlation of selected molecular markers in chemosensitivity prediction

David King, Thomas Keane, Wei Hu

Department of Computer Science, Houghton College, Houghton, NY 14744, USA.
Email: Wei.Hu@houghton.edu

Received 3 July 2009; revised 19 August 2009; accepted 20 August 2009.

ABSTRACT

Finding effective cancer treatment is a challenge, because the sensitivity of the cancer stems from both intrinsic cellular properties and acquired resistances from prior treatment. Previous research has revealed individual protein markers that are significant to chemosensitivity prediction. Our goal is to find correlated protein markers which are collectively significant to chemosensitivity prediction to complement the individual markers already reported. In order to do this, we used the D' correlation measurement to study the feature selection correlations for chemosensitivity prediction of 118 anti-cancer agents with putatively known mechanisms of action. Three datasets on the NCI-60 were utilized in this study: two protein datasets, one previously studied for chemosensitivity prediction and another novel to this topic, and one DNA copy number dataset. To validate our approach, we identified the protein markers that were strongly correlated by our analysis with the individual protein markers found in previous studies. Our feature analysis discovered highly correlated protein marker pairs, based on which we found individual protein markers with medical significance. While some of the markers uncovered were consistent with those previously reported, others were original to this work. Using these marker pairs we were able to further correlate the cellular functions associated with them. As an exploratory analysis, we discovered feature selection correlation patterns between and within different drug mechanisms of action for each of our datasets. In conclusion, the highly correlated protein marker pairs as well as their functions found by our feature analysis are validated by previous studies, and are shown to be medically significant, demonstrating D' as an effective measurement of correlation in the context of feature selection for the first time.

Keywords: Cancer; Chemosensitivity; Correlation; D'; Feature Selection; Genetic Algorithm; Markov Blanket; Memetic Algorithm; NCI-60

1. INTRODUCTION

The success of cancer treatment as well as the severity of the side effects of said treatment is heavily dependent on the sensitivity of the cancerous tissue to chemical treatment. Clinics face a great challenge in predicting treatment success, because chemosensitivity is determined by both intrinsic genomic and proteomic characteristics of the cancer as well as resistances induced through prior treatment. When trying to choose a therapy that will work best for a patient, it is important to evaluate their physical responses to different drugs. Because of this, many studies have been done to improve drug response prediction accuracy.

Data profiling of cancer cells at genomic, proteomic, chromosomal and functional levels has long been used in the analysis of pharmacological sensitivity of the cancer cells [1,2,3,4]. A primary source of cancer data in this field is a set of 60 human cancer cell lines provided by the National Cancer Institute (NCI-60) [5]. These cell lines have been in use since 1990 and over 100,000 chemical compounds have been tested on them [6]. The NCI-60 includes melanomas, leukemias and samples of ovarian, prostate, renal, breast, colon, lung and central nervous system cancers.

1.1. Related Works on the NCI-60

One study [7] used protein expression profiles to predict responses to a set of 118 anti-cancer agents with known or experimentally supported mechanisms of action [8]. Well known machine learning algorithms such as Random Forest, Nearest Neighbor and Relief were used to make chemosensitivity predictions. One Random Forest based classifier was built for each of the 118 drugs. To measure the significance of their predictions, this study compared the computed predictions against random pre-

dictions, which can be measured by a standard P-value. The P-value was the percentage of 1000 random predictions with higher accuracy than the calculated predictions. The study found chemosensitivity prediction accuracies ranging from 50 to 90%, with the vast majority being between 50 and 70%. Every prediction had *P*-values less than 0.019, and 97 of the predictions had *P*-values equal to 0.00.

A subsequent study by the same research group used a combination of the previously used proteomic data and new transcriptional data [9]. This integrative approach demonstrated its advantage, achieving higher accuracy and statistical significance, with *P*-values for all 118 drugs less than 0.001, calculated in the same manner as in [7].

A separate study [10] analyzed the correlation between DNA copy number variations, gene expression levels, and chemosensitivies to the same 118 drugs as in [7,9]. The analysis indicated that the correlations of gene expression and DNA copy number are particularly evident among leukemias and ovarian cancers.

An additional study [6] used four gene expression datasets, two of which were original to the paper, and one proteomic dataset. These data sets were used to observe the effectiveness of transcript profiling for the prediction of different protein expression levels. In addition, a consensus set selected from the four gene expression datasets was constructed. This consensus set was found to have a correlation to the protein dataset of 65%; a notable percentage that was higher than most reports done with mammalian cells. Further, this consensus dataset was used to predict tissue origin with a higher accuracy than any of its parent datasets.

1.2. Feature Selection and Motivation

New technologies in biomedical studies, such as microarrays, have made the analysis of large volumes of complex data a necessity [11]. Frequently, a majority of these data contain noise, i.e., features not relevant to a particular task at hand, such as classification of cancer types with gene expression data.

Both studies conducted by [7,9] used Random Forest as a feature selection technique to improve the accuracy of chemosensitivity predictions and to single out protein markers that were particularly important to this task.

In studying the effects of feature selection on chemosensitivity prediction, we observed disparity between expected and observed results. We ranked and ordered all features in the smaller protein dataset used in this study according to the Relief algorithm provided by Weka. We used Random Forest to make predictions based on incrementing feature subsets, using the top two ranked features, then three, four, etc. up to 40 features, as in **Figure 1**. We observed that contrary to our expectations, some higher ranked features decreased prediction accuracy, while some lower ranked features increased accuracy.

This led us to hypothesize that features contribute to the prediction accuracy collectively, rather than independently. To test this hypothesis, we developed a new technique using the D' measure [12] in order to study the correlations between feature pairs. As a demonstration of the utility of this technique, we apply it to those protein markers found to be significant in [9].

2. MATERIALS AND METHODS

2.1. Datasets

Three datasets derived from the NCI-60 were used in our study: two sets of protein data, and one of DNA data.

Protein expression data. The first protein expression dataset had 162 protein markers, hereafter referred to as Protein162, and was created by Shankavaram *et al* [6] and can be found at http://discover.nci.nih.Gov/datasets.jsp. The second dataset, which contains 52 protein markers (Protein52), available at http://discover.nci.nih.gov/host/2003_profilingtable7.xls., was generated by a study of the proteomic profiles of the NCI-60 [13], and was also used by two studies on chemosensitivity prediction [7,9].

DNA copy number. The DNA copy number variation dataset was presented in a study of the correlation between mRNA and DNA copy number [10]. It is available at http://discover.nci.nih.gov/datasets.jsp.

Drug activity data. Our drug resistance information contained activity data from 118 anti-cancer agent activity profiles. They were screened by Scherf *et al* [8] and recorded using the NCI-60 cancer cell lines. The file containing this data can be found at http://discover.nci.nih.gov/nature2000/data/selected_data/dataviewer.jsp?baseFileName=a_matrix118&&nsc=2&dataStart=3.

Defining drug sensitivity and resistance. As in [7,9] we used a threshold to define sensitivity to a drug into

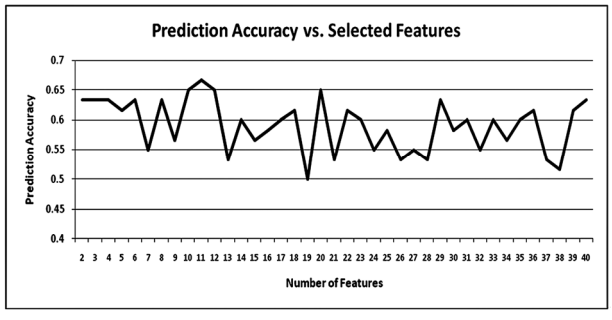

Figure 1. Random forest prediction accuracy. This plot shows the prediction accuracies of Random Forest using the same protein dataset used in [7,9]. The drug on which the prediction was performed was Bisentrene (NSC # 337766).

three categories. A $\log_{10}$ (GI_{50}) was taken for each cell line to determine sensitivity. Cell lines with sensitivities at least 0.5 standard deviations above the average were given the label 'resistant.' Those with sensitivities at least 0.5 standard deviations below the average were 'sensitive.' The remaining cell lines were defined as 'intermediate' [7,9].

2.2. D' Formula

A standard measurement for the correlation between pairs of events *i* and *j* in a set of sequences is *D'*, which can be defined by the following formulae:

$$D_{ij} = x_{ij} - p_i q_j \qquad D'_{ij} = \frac{D_{ij}}{D_{\max}}$$

where x_{ij} is the frequency at which both event *i* and event *j* occur in a single sequence, p_i is the frequency of event *i* and q_j is the frequency of event *j*. If $D_{ij} < 0$,

$$D_{max} = min\left[p_i q_j, (1-p_i)\left(1-q_j\right)\right], \text{ and if } D_{ij} > 0,$$

$$D_{max} = min\left[p_i \left(1-q_j\right), (1-p_i) q_j \right].$$

The *D'* formula was introduced by Richard Lewontin as a measurement of linkage disequilibrium of alleles at two or more loci on the same chromosome [12]. The *D'* formula has been shown to be a more reliable measurement than other measurements of correlation between pairs of events [14], but this study is the first to use it to correlate pairs of selected features.

2.3. Markov Blanked-Embedded Genetic Algorithm (MBEGA)

Genetic Algorithms have been used as a strategy for feature selection [15] due to their ability to generate better feature subsets than other feature selection algorithms. In some cases, these genetic algorithms are combined with memetic operations in order to fine tune results beyond what would be produced by classical genetic algorithms alone.

One particular implementation of these memetic algorithms is the Markov blanket-embedded genetic algorithm (MBEGA), which uses an approximation of a Markov blanket to reduce redundancy in selected features. Pseudocode for the MBEGA can be found in **Figure 2**.

In each generation of the algorithm, the MBEGA uses add and delete operations to add and delete features from some of the elite feature subsets in the population; the elite feature subsets are improved by adding important features and removing those that are less important. After the memetic operations, standard genetic algorithm techniques such as linear ranking, crossover and mutation methods occur to generate the next population [16].

The MBEGA was selected in our study for two reasons: First, the MBEGA generates a population of feature subsets in each generation, rather than generating a single final subset as in classical feature selection algothms. The feature subsets from each generation are represented as binary strings, with a 1 representing a present feature and 0 representing an absent feature, to calculate the D' values of our correlation analysis. Second, the MBEGA does not require a predefined number of features to be selected. Rather, the MBEGA gradually optimizes both the size of the feature subset as well as the accuracy of the classifier.

Markov Blanket Embedded Genetic Algorithm (MBEGA)

BEGIN
(1) **Initialize:** Randomly generate an initial population of feature subsets encoded as binary strings
(2) **For** the number of iterations to run
(3) Evaluate all feature subsets in the population based on prediction accuracy
(4) Select a number of elite feature subsets from the population to undergo the Markov blanket memetic operations
(5) **For** each feature subset create a set of all present features X and all absent features Y
Add operation **BEGIN**
1) Rank the features in Y according to their correlation to the class label.
2) Select a feature *Yi* in Y so that the larger the correlation of a feature in Y the more likely it will be picked.
3) Add *Yi* to X.
END
Delete operation **BEGIN**:
1) Order the features in X according to their correlation to the class label.
2) Select a feature *Xi* in X so that the larger the correlation of a feature in X the more likely it will be picked.
3) Eliminate all features in X which are less correlated than *Xi*. If no feature is eliminated, remove *Xi*.
END
(6) Replace the original elite feature subset with the improved feature subset.
(7) **End For**
(8) Perform crossover and mutation to create the next generation of feature subsets.
(9) **End For**
END

Figure 2. Pseudocode for MBEGA.

2.4. Correlation Analysis Using D' Formula

We used the D' formula to calculate correlation between pairs of features selected in each generation of the MBEGA. Because the MBEGA begins with a randomized feature subset and becomes more selective as the algorithm progresses, we decided to use only the last 20% of the feature subsets generated. We calculated the D' values for every pair of selected features, using the presence of one feature within the encoded binary sequence as event *i* and the presence of the other as event *j*.

2.5. Feature Selection Using Weka's Relief Algorithm

There are three primary types of feature selection algothms: filter, wrapper and embedded algorithms. Filter algorithms have advantages in their speed and scalability,

however they ignore feature dependencies. They also do not interact with classifiers, which is both an advantage, because they can select features independently, and a disadvantage, because they are unable to take the classifier into account when determining the feature subset. Wrapper algorithms, on the other hand, do interact with the classifier, and are therefore able to produce more informative feature subsets. They are also less prone to local optima. They are, however, computationally intense, and have a higher risk of over fitting. Embedded algorithms are built directly into a classifier. As such, they are able to interact with the classifier in the same manner as wrapper algorithms, but are far less computationally intense.

Relief, a filter feature selection algorithm implemented in WEKA, was used to assess the features pairs found by our correlation analysis. Relief ranks features by assigning them weights according to their ability to

Table 1. Highly correlated protein marker pairs in protein52 based on significant chemosensitivity protein markers. The protein markers in column one and associated drugs, expressed as their NSC drug numbers, in column two were found in [9]. The remaining columns are the ten protein markers with the highest correlation to the protein marker in the first column. The markers notated with a * are those also selected by Weka's Relief algorithm.

Protein Marker	NSC Drug #	1	2	3	4	5	6	7	8	9	10
ISGF3g	56410	MAPK1	EP300*	MSN	NME1*	GSK3B	FADD	STAT1*	STAT3	STAT5A	MSH6
ISGF3g	354646	MGMT	VIL1	RIPK1*	EP300	EP300	MSN	GSK3B	FADD	STAT5A	MSH2
STAT3	56410	EP300*	EP300	MSN	GSK3B	FADD	ISGF3G*	STAT1*	STAT5A	STAT6*	MSH2
NME1	353451	FN1	MVP	RELA	MSN	CDH1*	MGMT	GSK3B	FADD	ISGF3G*	STAT5A
NME1	344007	KRT18	EP300	MAPK1	CDH1*	MGMT	GSK3B	FADD	ISGF3G	STAT3*	STAT5A
NME1	102816	TP53	EP300	MGMT*	CCNE	MAP2K1	CDH1*	MGMT*	GSK3B	FADD	STAT3*
NME1	107392	KRT8	MSN*	CDH1	MGMT*	GSK3B	FADD	ISGF3G	STAT3*	STAT5A	MSH2
MGMT	95466	KRT18*	CDH2	EP300	MSN	EP300	MSN*	CDH1	NME1*	GSK3B	ISGF3G*
CCNE	95441	KRT8*	CCNA2	CCNB1	VIL1*	CDH1	RELA	RIPK1	JAK1	MAP2K2	STAT5A
EP300	119875	KRT18	EP300	CDH2	KRT20	FN1	MSN	CCNB1*	JAK1	MAP2K1	MAP2K2
EP300	606497	EP300	CDH2	KRT20	FN1	KRT8	CCNB1	CCNE	RIPK1	STAT3*	MAP2K1
FN1	135758	KRT18	CDH2	KRT20	KRT8	CCNA2	CCNB1	CCNE	VIL1	MAP2K2	ISGF3G
MSN	301739	KRT20	MAPK1	MCP	MCM7	CDK6	G22P1	MVP	PGR	MAP2K2	FADD
MSN	755	KRT18	CDH2	MCM7	CDK6*	G22P1	MVP	PGR	MAP2K2	FADD	STAT1
MSN	376128	PCNA	MCP	CDK6	G22P1	MVP*	PGR	CCNE	VIL1	CDH1*	EP300
PGR	354646	MSN	MVP	CCNA2*	CCNB1*	CCNE	CDH1	CASP2	RIPK1*	EP300	FADD
STAT1	354646	KRT20	MAPK1	G22P1	MVP	MAP2K2*	MSN	NME1*	FADD	STAT3	STAT5A
STAT6	354646	EP300	PCNA	MAPK1	FADD	ISGF3G	STAT3	STAT5A	MSH2	MSH6	EP300
CASP2	264880	CCNA2*	CCNE	VIL1	CDH1	RELA*	RIPK1	JAK1*	STAT3	EP300	STAT1
CDH1	71261	MAPK1	CCNE	VIL1*	RELA*	CASP2	EP300	EP300	CDH1	NME1	MSH2*
MCP	740	TP53	EP300	EP300	KRT20	ACVR2*	MCM7*	CDK6	CCNB1	VIL1	EP300
KRT18	19893	TP53	EP300	CDH2	EP300	FN1*	KRT8	PGR	JAK1	MAP2K2	EP300
KRT18	757	TP53	EP300	RELA	STAT3	EP300	CDH1	GSK3B	STAT5A	MSH6	EP300
KRT18	33410	TP53	EP300	CDH2	EP300	KRT20*	RIPK1	MAP2K2	MSN	MSH6	EP300
KRT18	125973	TP53	EP300	CDH2	EP300*	KRT20	CDK6	CCNA2	CCNE	RELA	EP300
KRT18	658831	TP53	EP300	KRT20	FN1*	MAPK1	MSN	G22P1	JAK1	CDH1	EP300
KRT18	673188	TP53	EP300	CDH2	FN1	GSK3B*	FADD*	STAT5A	STAT6	MSH6	EP300
KRT18	671867	TP53	EP300	CDH2	EP300	CCNB1	CASP2	EP300	MAP2K1	MSH6	EP300
KRT18	664402	TP53	EP300	CDH2	EP300	MVP	PGR	JAK1	EP300	STAT6	EP300
KRT18	661746	TP53	EP300	CDH2	EP300	ACVR2	MCP	EP300	MSN	CDH1*	EP300
KRT18	673187	TP53	EP300	CDH2*	ACVR2	CDK6	VIL1	STAT6	MSH2*	MSH6*	EP300
KRT18	664404	TP53	EP300	CDH2	KRT20	RB1	MAPK1	EP300	EP300	STAT5A*	EP300
KRT18	671870	TP53	EP300	CDH2	KRT20*	FN1	PCNA	CDH1	CASP2	STAT6	EP300*
KRT18	666608	TP53	EP300	CDH2	EP300	KRT20	MAPK1	CCNB1	RIPK1	STAT3	MGMT
KRT18	600222	TP53	EP300*	CDH2*	RB1	G22P1	PGR	CCNA2	MSN	STAT3	STAT5A
KRT18	656178	EP300	EP300	EP300	MGMT	MAPK1	MSN	G22P1	MAP2K1	STAT5A*	STAT6
TP53	19893	KRT18	CDH2	RB1	MAPK1	EP300	CDK6	MSN	STAT6	MSH6	EP300
TP53	125973	KRT18*	KRT20	FN1*	MGMT	MAPK1	ERBB2	MCM7*	STAT6	MSH6	EP300
RELA	153353	CDH2*	MSN	G22P1	VIL1	CDH1	CASP2	RIPK1*	JAK1	MAP2K1	MAP2K2
G22P1	224131	EP300	MAPK1	MSN	MVP*	CCNA2	CCNB1	RIPK1	EP300*	MAP2K2	EP300

discriminate between neighboring patterns. In each iteration of the algorithm, an instance x containing features $(x_1,x_2,\ldots,x_n)$ is selected randomly, and one nearest neighbor from the same class (called NH) along with one nearest neighbor from a different class (called NM) are found. The weights of the features in x are updated such that they will be greater if x is similar to the NH and dissimilar to the NM, and less if the opposite is true.

3. RESULTS AND DISCUSSION

To generate sequences for D' analysis, we ran a ten-fold cross validation of the MBEGA on all three datasets. Each fold of the MBEGA ran for 100 generations, with a population of 51. Each fold generated 5100 sequences, of which we used the last 20% generated, or 1020 from each fold. The final number of sequences used for the D' analysis was 10200 for each dataset.

3.1. Correlation Analysis of Individual Protein Markers from Previous Study

A previous study [9] discovered 18 individual protein markers from Protein52 along with their functions, including transcriptional factoring, tumor suppressing, DNA repair, cell adhesion, and apoptosis, among others, that are significant to the prediction of chemosensitivity to 33 of the 118 anti-cancer agents. These drugs represent 12 out of the 15 total mechanisms of action present in the 118 anti-cancer agents, with a large number of them being tubulin active antimitotic agents. In order to investigate the protein markers highly correlated with those found in [9], for each protein marker/drug combination identified there, we found the ten protein markers with the highest D' value. We also sought to validate these pairs using a ten-fold cross validation of the Relief feature selection algorithm provided by Weka, which measures feature significance individually. As seen in **Table 1**, the highly correlated protein marker pairs our analysis discovered are validated not only by the protein markers reported in [9], but also by Relief.

In order to discover any patterns in the correlation of the protein markers selected in **Table 1**, we took a frequency count of them, as illustrated in **Figure 3**. While most protein markers had frequencies within the same range, mostly between 4 and 10, there were some which clearly stood out. In particular, the protein marker CDH2, a cell adhesion protein, is highly correlated with 19 out of the 40 protein markers in **Table 1**. CDH2 was not selected in the previous study, but is very similar in both function and family to CDH1, which was selected. Another protein with a high frequency, 16 out of 40, was TP53 whose function is tumor suppression and apoptosis. We found that in most occurrences TP53 was paired with protein marker KRT18. Both of these protein markers are involved in protein death, and both were found to be strong chemosensitivity predictors for the drug Taxol in the previous study [9]. Lastly, we noticed STAT5A is both from the same family as and is highly correlated with the protein markers STAT1 and STAT6, both of which were highlighted in previous study [9].

We were also interested in observing how the functions of the individual protein markers from [9] correlated with the functions of the highly correlated protein markers found in **Table 1**. We grouped the previously reported protein markers according to their function, and then selected the protein markers that were most frequently correlated with them. We only included those protein functions which had three or more protein markers associated with them, as in **Table 2**.

Because we used two protein datasets in this study, we wanted to conduct the same analysis on the Protein162 dataset in order to explore the possibility of discovering new protein markers highly correlated with those found in [9]. All but 2 of the previously reported protein markers from Protein52, G22P1 and CCNE, were also present in Protein162, so we used the same protein marker/drug combinations as in **Table 1** when generating **Table 3** for Protein162.

We also created a selection frequency histogram for

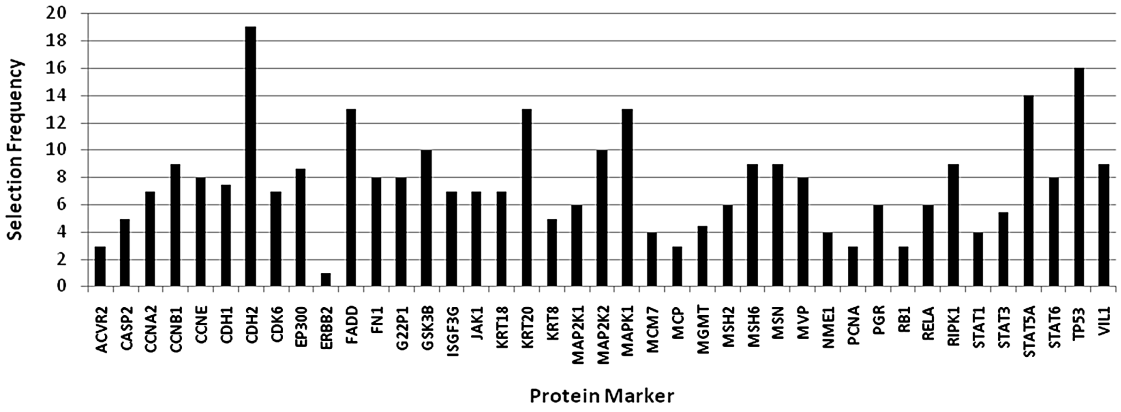

Figure 3. Frequency of protein52 protein markers present in **Table 1**.

Table 2. Correlation of the functions of protein markers in **Table 1**.

Protein Function	Reported Protein Markers	Correlated Protein Markers	Functions of Correlated Protein Markers
Transcriptional Factor	ISGF3G	STAT5A	Transcriptional Factor
	STAT3	FADD	Apoptosis
	EP300		
	STAT1		
	STAT6		
	RELA		
Integrin Signaling	NME1	STAT6	Transcriptional Factor, Inteferon Signaling
	EP300	MSH6	DNA Repair
	TP53	KRT18	Structural Protein; Biomarker of cell death
		FADD	Apoptosis
Tumor Suppressors	CCNE	FADD	Apoptosis
	TP53	FN1	Cell Adhesion; Integrin Signaling
	RELA	GSK3B	Hormonal Control
		KRT18	Structural Protein; Biomarker of cell death
Cell Apoptosis	CASP2	CDH2	Cell Adhesion
	KRT18	KRT20	Structural Protein
	TP53	MSH6	DNA Repair
	RELA	TP53	Tumor Suppressor; Cell Cycle and Apoptosis

Table 3. Highly correlated protein marker pairs in protein162 based on significant chemosensitivity protein markers. The protein markers in column one and associated drugs, expressed as their NSC drug numbers, in column two were found in [9]. The remaining columns are the ten protein markers with the highest correlation to the protein marker in the first column. The protein markers G22P1 and CCNE present in **Table 1** are excluded here because neither is present in Protein162. The markers notated with a * are those also selected by Weka's Relief algorithm.

Protein Marker	NSC Drug #	1	2	3	4	5	6	7	8	9	10
ISGF3g	56410	ANXA1	CASP7	CRK	EP300	EP300	EP300	JAK1	RELA	RIPK1	TP53
ISGF3g	354646	AKAP5	AP2M1	CDC2*	MSN	PRKCI	RIPK1*	SMARCB1	FASLG	TP53	VIL2
STAT3	56410	CCNA2	CDH1	CDKN2A	HRAS	KRT18	MAPK1*	MVP	STAT1*	STAT3	VASP
NME1	353451	ANXA4	EP300	EP300	FN1	GTF2B*	IRS1	MGMT	NCAM1	PCNA	TP53
NME1	344007	CASP7	CASP7	CDH2	ENAH	EP300	HSPA4	JAK1	MCC	PRKCI	TP53
NME1	102816	CASP2	CASP7	EP300	FADD	ISGF3G	MAP2K1	MGMT	MSN	NCAM1*	TUBB2A
NME1	107392	ANXA1	CASP7	EP300	EP300	MAP2K2	NCAM1	PCNA	PRKCA	RB1	RELA
MGMT	95466	ACVR2A	BCAR1	EP300	FADD	MGMT*	RB1	STAT1	TP53	TP53	YWHAG
EP300	119875	PARP1	CTNNB1*	EP300	GRB2*	JAK1	MCC	MSN	RIPK1	STAT6*	TP53
EP300	606497	ADNP	ATXN2*	EP300	EP300	GRB2	JAK1	MCP*	MSH6	STAT3*	TP53
FN1	135758	AKAP8	CDK4	CTTN	EP300	EP300	EP300	GSTP1	PCNA	STAT6	FASLG
MSN	301739	ADNP	CDH1	EP300	JAK1	MAP2K2	MSN*	MVP	RIPK1	SMARCB1	STAT3
MSN	755	CDK5	CTNNB1	EP300	ISGF3G	JAK1	KRT18	MAPK1	MCC	RB1	RIPK1
MSN	376128	PARP1	CDC2	CDH2	EP300	EP300	MGMT	MVP*	PTPN6	RB1*	TP53
PGR	354646	CDH2	ENAH	EP300	EP300	EP300	EP300	MSN	RELA	EXOC4	TP53
STAT1	354646	CASP2	EP300	EP300	EP300	FADD	MCM7	MSH6*	PRKCB1	RELA	TYR
STAT6	354646	PARP1	CDK4*	CDK7	CRK	ENAH	EP300	MSH6*	RB1	RELA	STAT3
CASP2	264880	CASP7	CTNNB1*	DSG1	EP300	EP300	EP300	FADD*	ISGF3G	KRT7*	PCNA
CDH1	71261	FN1	KRT19	MGMT	PTPN11	RELA*	RELA	RELA	STAT6	TP53	TRADD
MCP	740	CASP7	DSG1	EP300	ESR1	FADD	KRT19	MAP2K2	MCM7	RB1*	SMARCB1
KRT18	19893	PARP1	PARP1	CDKN2A	EP300	MCM7*	MSN	PRKCA	RB1	RELA	STAT5A
KRT18	757	CDK4	ENAH	EP300	EP300	EP300	KRT19*	MCP	SMARCB1	STAT5A	STAT6
KRT18	33410	AKAP8	EP300	EP300	EP300	EP300	JAK1	KLK3	KRT19	STAT1	VASP
KRT18	125973	CDC2	CDK5	GSTP1	IRS1	KRT19*	TP53	TP53	TP53	TP53	VIL2
KRT18	658831	ATXN2	CCNA2	IRS1	MCM7*	MSH2*	EXOC4	SMARCB1*	TP53	TP53	TRADD
KRT18	673188	AKAP5	CDK5	EP300	JAK1	KLK3	KRT19*	MAP2K2	RELA	TGFB1I1	VASP
KRT18	671867	ADNP	BCAR1	CCNA2	CDC2	EP300	KLK3	MAP2K1	MAPK1	MVP	STAT3
KRT18	664402	ADNP	CCNA2	EP300	EP300	KRT19*	KRT7	PTPN11	RELA	RELA	STAT6
KRT18	661746	AP2M1	CCNB1	CDH2	DSG1	EP300	KRT19	PRSS8	RELA	STAT5A	VASP
KRT18	673187	CCNA2	EP300	EP300	ERBB2	FADD	XRCC6	MGMT	MVP*	STAT3	TP53
KRT18	664404	CDK6	EP300	ISGF3G	KLK3	MCC	MSH6	PRKCB1	STAT1	STAT6	FASLG
KRT18	671870	CASP2	CCNB1	CDH2	EP300	MAP2K1	MCP	MLH1	PCNA	RELA	EXOC4
KRT18	666608	CDH1*	CDK7	ENAH	EP300	FADD	GSTP1	KLK3	KRT20	PRSS8*	VIL1
KRT18	600222	ADNP	AKAP5	AKAP8	CDH2	EP300*	GSTP1*	KLK3	MAP2K1	PTPN11	TYR
KRT18	656178	CDH2	EP300	GTF2B	KRT19	KRT8*	EXOC4	TP53	TP53	TP53	TRADD
TP53	19893	CASP7	EP300	EP300	ESR1	GTF2B	KRT8	MAP2K2	STAT1	STAT1	STAT6
TP53	125973	CASP7	CCNA2	CDH1	EP300	JAK1	KLK3	KRT19*	PRKCA	STAT1	TP53
RELA	153353	ADNP	CASP7	CDK7	EP300	EP300	EP300	PRKCI	PRSS8	PTPN11	STAT3

Table 3, illustrated in **Figure 4**. We observed that the frequencies were lower by roughly a factor of 2 for this dataset when compared to Protein52. We believe this is because the number of unique protein markers in Protein162 was roughly twice that of the unique protein markers in Protein52; however we chose the top ten most correlated pairs in both instances.

Many of the most selected protein markers from **Table 1**, including CDH2, TP53, and STAT5A, had only an average or even low frequency in **Table 3**. We selected 8 protein markers from **Table 3** whose average frequencies were above 4. These were KRT18, KLK3, CCNA2, ADNP, MVP, RIPK1, SMARCB1, and ENAH. Because the Protein162 dataset contains protein markers not present in Protein52, we found 5 protein markers which were not reported in the previous study [9]. These proteins, as well as their cellular functions and associated drugs can be seen in **Table 3**. The most frequently selected protein marker from **Table 3** was KRT19, a structural protein from the same family as KRT18, a protein marker found to be significant in [9], and KRT20, a protein marker frequently selected in **Table 1**. KLK3 had the second highest frequency of selection. Serum levels of the KLK3 protein, called PSA, are used to diagnose and monitor prostatic carcinoma. Members of the KLK family are also thought to be biomarkers for cancers and diseases. CCNA2 has a functional relationship with CDC2, another protein marker with an above-average selection frequency in **Table 3**. ADNP affects both normal cell growth and cancer proliferation. In addition, ADNP is a transcription factor, a trait held in common with six of the eighteen significant protein markers in [9]. MVP is a protein which is over-expressed in multi-drug resistant cancer cells, and is potentially useful as a signal for drug resistance. MVP also bears a functional relation with STAT1, one of the important protein markers reported in [9]. RIPK1 is an apoptosis protein related to cell death, much like the KRT18, TP53 and CASP2 found in both the previous study [9] and in **Table 1**. SMARCB1 functions as a tumor suppressor, but mutations within the protein are associated with rhabdoid tumors. ENAH is a cell adhesion protein which is present in some breast cancers, and may be used as a marker for such.

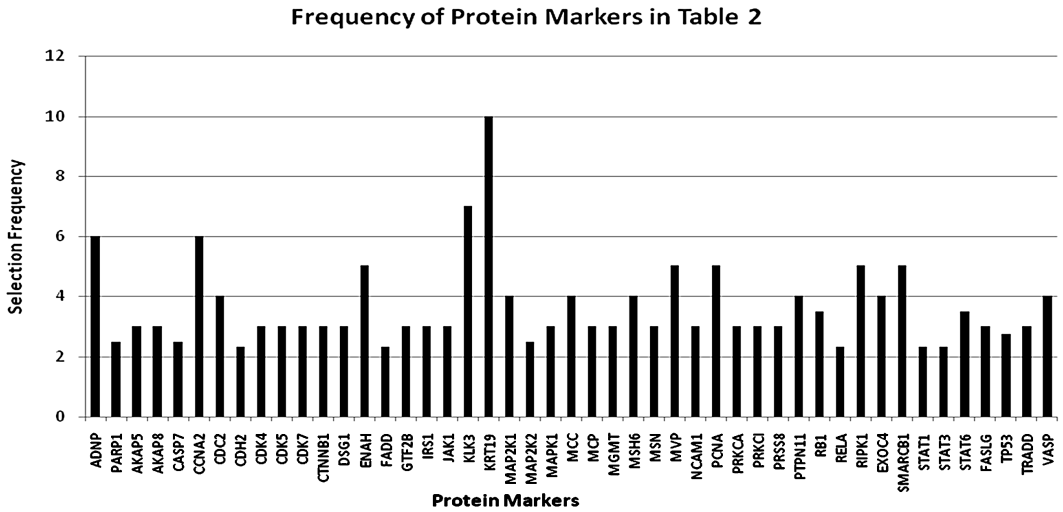

Figure 4. Frequency of protein162 protein markers present in **Table 3**. This plot shows the frequency with which protein markers from the Protein162 dataset were selected in **Table 3**. Only those protein markers with an average frequency above 2 are shown due to limited space.

Table 4. Highly correlated protein markers for the evaluated anticancer drugs**.** Protein markers denoted with a * were unique to the Protein162 dataset, and as such not reported in previous study [9].

Protein Marker	Function	NSC Drug Numbers
KRT19*	Structural Protein	71261, 740, 757, 33410, 125973, 673188, 664402, 661746, 656178
KLK3*	Biomarker of Prostatic Carcinoma	33410, 673188, 671867, 664404, 666608, 600222, 125973
ADNP*	Cell Growth, Cancer Proliferation, Transcription Factor	125973, 301739, 671867, 664402, 600222, 153353
CCNA2	Binding & Activating Agent	56410, 658831, 671867, 664402, 673187, 125973
MVP	Mediating Drug Resistance, Over-expressed in multi-drug resistance cancer cells	56410, 301739, 376128, 671867, 673187
RIPK1	Apoptosis Protein	56410, 354646, 119875, 301739, 755
SMARCB1*	Tumor Suppressor	354646, 301739, 740, 757, 658831
ENAH*	Cell Adhesion Protein	344007, 354646, 354646, 757, 666608

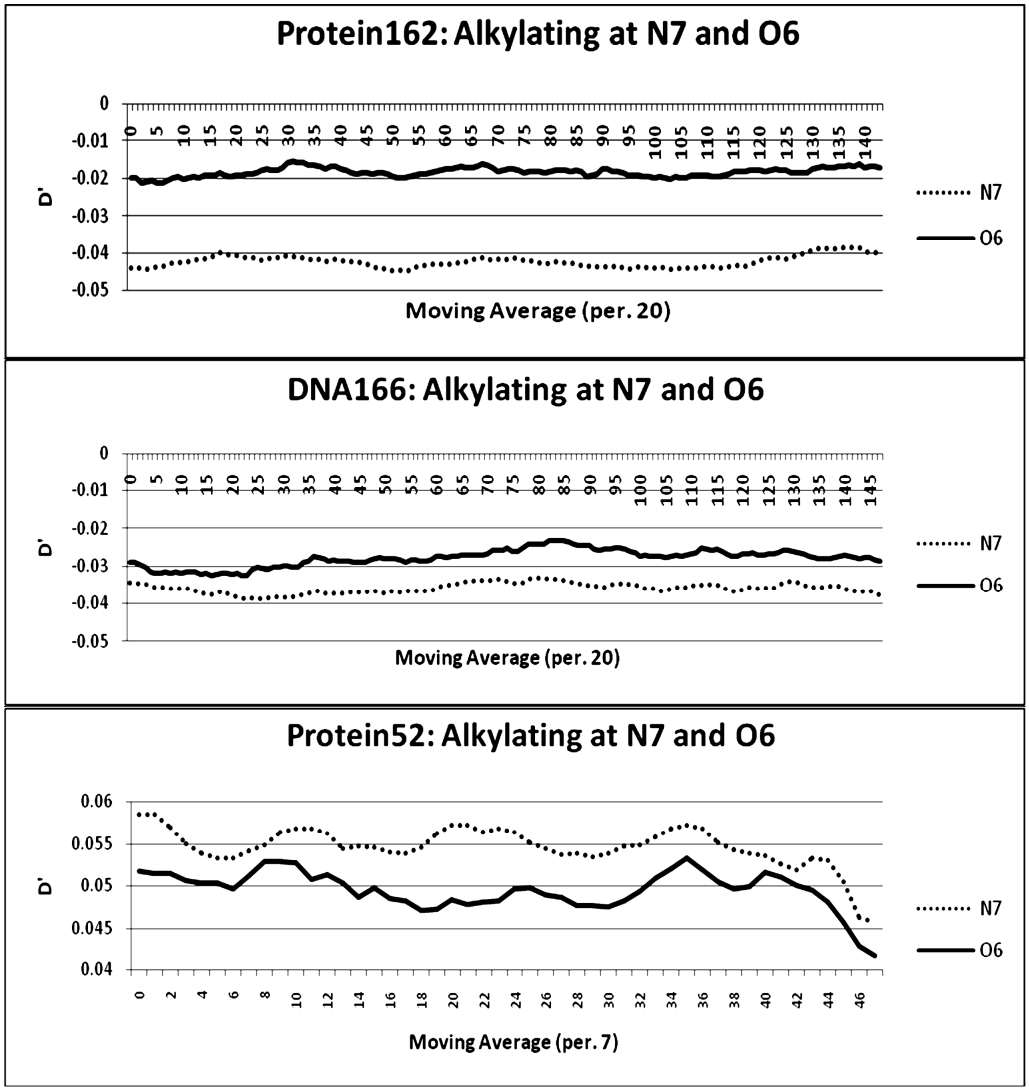

Figure 5. D' patterns categorized according to mechanism of action. The drugs which alkylate at N7 and O6 produce similar patterns of D' values in each of the three datasets. For readability, the curves displayed are moving average curves with a period of 20 for Protein162 and DNA166 and a period of 7 for Protein52.

3.2. Feature Correlation Analysis of All 118 Drugs

In addition to simply calculating the D' values of the feature pairs, we also calculated the average D' values for each feature based on the D' values of all pairs associated with that feature. We performed this analysis for both protein datasets as well as the DNA166 dataset, using the protein markers as features for the protein datasets, and the DNA copy number variations as features for the DNA166 dataset. As an exploratory analysis, we attempted to use the average D' values to find the trends of feature correlation within and between the mechanisms of action of the 118 anti-cancer agents.

Each dataset generated unique patterns of feature correlation in each of the 118 drugs. We did observe, however, similar patterns of feature correlation in drugs with related mechanisms. In particular, we noticed that topoisomerase I inhibitors and topoisomerase II inhibitors have very similar trends of feature correlation to one another in all three datasets. Drugs which alkylate at positions N7 (24 drugs) and O6 (7 drugs) of guanine also have very similar trends of feature correlation, as shown in **Figure 5**. This implies that related drug mechanisms tend to produce similar patterns of correlation between feature pairs. Our analysis indicated that this is not necessarily true of drugs with similar chemical structure.

We also grouped drugs with similar mechanisms into three larger categories: drugs which alkylate at specific positions of guanine (Alkylating), drugs which inhibit topoisomerase (TIM Inhibitors), and all other drugs (Other). The D' values of these larger categories were generated by averaging the D' values of the individual drugs within that larger category. We found the correlation

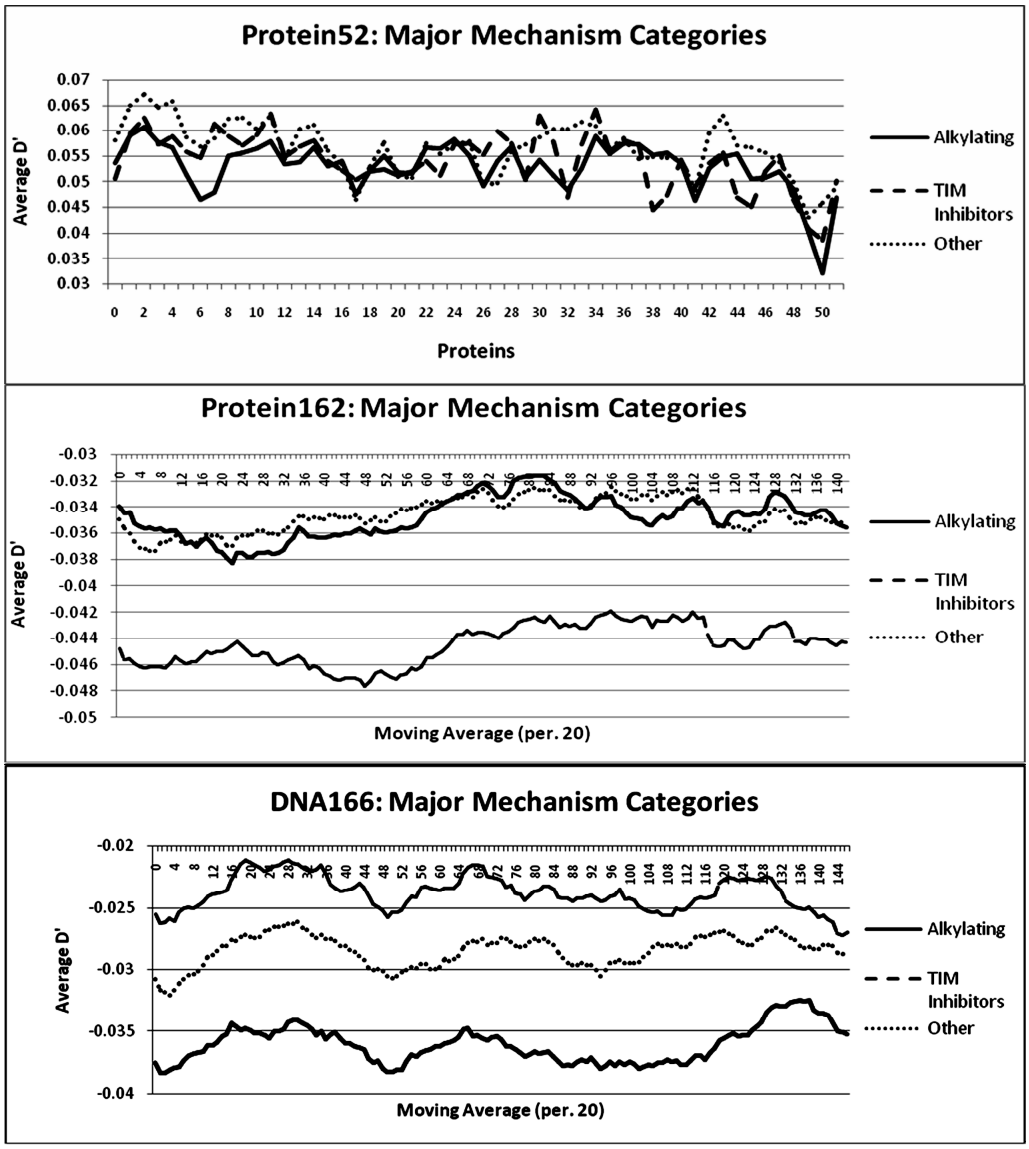

Figure 6. D' patterns according to major mechanism of action categories. Each dataset yields different levels of D' values between the three categories. The values produced by Protein52 are very similar in all three mechanism categories, whereas TIM Inhibitor values for Protein162 are distinct from the others. All three categories produce distinct values in DNA166. For readability, the plots of Protein162 and DNA166 are moving averages with a period of 20, whereas the plot of Protein52 shows the curve without a moving average.

trends of these three categories to be different for all three datasets.

In Protein52, we observed that while each of the larger categories carries unique drugs and drug mechanisms, the averaged D' values of all three of these categories were very similar, with the averages being 0.053 for the Alkylating category, 0.054 for the TIM Inhibitor category, and 0.06 for the Other category. All averaged D' values were positive.

The same analysis of the Protein162 dataset revealed that while the averaged D' values of the Alkylating and the Other categories where very similar, with average D' values –0.0348 and –0.0345 respectively, the TIM Inhibitors category was quite distinct with an average D' value of –0.044. All averaged D' values were negative.

For DNA166, all three curves have distinct averaged D' values, with the Alkylating category having an average D' of –0.0382, the Other category an average of –0.0286 and the TIM Inhibitors category an average of –0.0240. Again, all D' values were negative.

These plots, available in **Figure 6**, illustrate the benefit of using multiple datasets in this type of study. While all of our datasets are based on the NCI-60, they each provide unique insight into physical responses of the cell lines to the 118 anti-cancer drugs. If we were only using the DNA data, we might be tempted to claim

that these three mechanism categories produce distinctly different D' values, whereas if we were only using the data from Protein52, we might claim the opposite. It is only when a wide range of data are used in study that a holistic understanding of the effects of these 118 drugs becomes possible.

4. CONCLUSIONS

We found that each of our datasets provides a unique insight into the analysis of feature correlations in the study of the chemosensitivity of the NCI-60 cancer lines. Two of the three (Protein52, DNA166) have been used in previous studies for the prediction of chemosensitivity of cancerous cells, and though Protein162 was novel to this topic, we have shown that both Protein162 and Protein52 contain a number of protein markers that are both medically significant and highly correlated to individual protein markers highlighted in previous study [9].

In addition, we have shown D' to be an accurate measure of correlation in the context of feature selection for the first time.

5. ACKNOWLEDGEMENTS

We would like to thank Houghton College for its financial support.

REFERENCES

[1] Staunton, J. E., Slonim, D. K., Coller, H. A., Tamayo, P., Angelo, M. J., Park, J., Scherf, U., Lee, J. K., Reinhold, W. O., Weinstein, J. N., Mesirov, J. P., Lander, E. S., and Golub, T. R., (2001) Chemosensitivity prediction by transcriptional profiling, PNAS, **98**, 10787−10792.

[2] Potti, A., Dressman, H. K., Bild, A., Riedel, R. F., Chan, G., Sayer, R., Cragun, J., Cottrill, H., Kelley, M. J., Petersen, R., Harpole, D., Marks, J., Berchuck, A., Ginsburg, G. S., Febbo, P., Lancaster, J., and Nevins, J. R., (2006) Genomic signatures to guide the use of chemotherapeutics. Nature Medicine, **12**, 1294−1300.

[3] Paweletz, C. P., Charboneau, L., Bichsel, V. E., Simone, N. L., Chen, T., Gillespie, J. W., Emmert-Buck, M. R., Roth, M. J., Petricoin, E. F., and Liott, L. A., (2001) Reverse phase protein microarrays which capture disease progression show activation of prosurvival pathways at the cancer invasion front, Oncogene, **20**, 1981−1989.

[4] Lee, J. K., Bussey, K. J., Gwadry, F. G., Reinhold, W., Riddick, S. L. Pelletier, S. Nishizuka, G. Szakacs, J. Annereau, G., Shankavaram, U., Lababidi, S., Smith, L. H., Gottesman, M. M., and Weinstein, J. N., (2003) Comparing cDNA and oligonucleotide array data: Concordance of gene expression across platforms for the NCI-60 cancer cells, Genome, Biol., **4**, R82.

[5] Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O., (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet., **24**, 227−235.

[6] Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Scherf, U., Kahn, A., Dolginow, D., Cossman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K., and Weinstein, J. N., (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study, Mol. Cancer Ther., **6**, 820−832.

[7] Ma Y., Ding Z., Qian Y., Shi X., Castranova V., Harner, E. J., and Guo L., (2006) Predicting cancer drug response by proteomic profiling, Clin. Cancer Res., **12**, 4583−4589.

[8] Scherf U., Ross D. T., Waltham M., Smith L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E. A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N., (2000) A gene expression database for the molecular pharmacology of cancer, Nat. Genet., **24**, 236−244.

[9] Ma, Y., Ding, Z., Qian, Y., Wan, Y., Tosun, K., Shi, X., Castranova, V., Harner, E. J., and Guo, N. L., (2009) An integrative genomic and proteomic approach to chemosensitivity prediction, International Journal of Oncology, **34**, 107−115.

[10] Bussey, K. J., Chin, K., Lababidi, S., Reimers, M., Reinhold, W. C., Kuo, W., Gwadry, F., Ajay, Kouros-Mehr, H., Fridlyand, J., Jain, A., Collins, C., Nishizuka, S., Tonon, G., Roschke, A., Gehlhaus, K., Kirsch, I., Scudi ero, D. A., Gray, J. W., and Weinstein, J. N., (2006) 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.

[11] Saeys, Y., Inza, I., and Larranaga, P., (2007) A review of feature selection techniques in bioinformatics. Bioinformatics, 23: 2507−2517.

[12] Lewontin, R. C., (1964) The interaction of selection and linkage, I. General considerations, Heterotic Models, Genetics, **49**, 49−67.

[13] Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W. C., Waltham, M., Kouros-Mehr, H., Bussey, K. J., Lee, J. K., Espina, V., Munson, P. J., Petricoin, E., Liotta, L. A., and Weinstein, J. N., (2003) 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.

[14] Hedrick, P. W., (1987) Gametic disequilibrium measures: proceed with caution, Genetics, **117**, 331−341.

[15] Leardic R., Boggia, R., and Terrile, M., (2005) Genetic algorithms as a strategy for feature selection, Journal of Chemometrics, **6**, 267−281.

[16] Zhu, Z., Ong, Y., and Dash, M., (2007) Markov blanket-embedded genetic algorithm for gene selection, Pattern Recognition, **40**, 3236−3248.

