Data Mining the NCI60 to Predict Generalized Cytotoxicity

Adam C. Lee,[†] Kerby Shedden,[‡] Gustavo R. Rosania,[§] and Gordon M. Crippen*,[†]

Departments of Medicinal Chemistry and Pharmaceutical Sciences, College of Pharmacy, and Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109

Elimination of cytotoxic compounds in the early and later stages of drug discovery can help reduce the costs of research and development. Through the application of principal components analysis (PCA), we were able to data mine and prove that ~89% of the total log GI_{50} variance is due to the nonspecific cytotoxic nature of substances. Furthermore, PCA led to the identification of groups of structurally unrelated substances showing very specific toxicity profiles, such as a set of 45 substances toxic only to the Leukemia_SR cancer cell line. In an effort to predict nonspecific cytotoxicity on the basis of the mean log GI_{50}, we created a decision tree using MACCS keys that can correctly classify over 83% of the substances as cytotoxic/noncytotoxic *in silico*, on the basis of the cutoff of mean log $GI_{50} = -5.0$. Finally, we have established a linear model using least-squares in which nine of the 59 available NCI60 cancer cell lines can be used to predict the mean log GI_{50}. The model has $R^2 = 0.99$ and a root-mean-square deviation between the observed and calculated mean log GI_{50} (RMSE) = 0.09. Our predictive models can be applied to flag generally cytotoxic molecules in virtual and real chemical libraries, thus saving time and effort.

INTRODUCTION

With the advent of high-throughput screening (HTS), mountains of biological screening data have been produced and continue to accumulate. In fact, as of 2002, ~14% of research and development in the pharmaceutical industry is spent on biological screening.[1] In 2003, approximately one-third of the capital lost on all drug failures, $8 billion, was due to the inability to accurately predict toxicity during the early stages of drug development.[2] As it relates to the investigation of cytotoxicity and growth inhibition studies, numerous quantitative end points have been used, including protein analysis, enzyme release, exclusion or inclusion of dyes or radioactive markers, and metabolic alterations such as oxygen consumption and ATP levels. As pharmacokinetics and toxicity (ADMET) are now a consideration in the early stages of drug development, many recent efforts have been made by both academia and industry to address the prediction of specific and general cytotoxicity. Methods commonly utilized to assign a cytotoxicity score or to classify substances as being cytotoxic/noncytotoxic include neural networks, proteomic profiling, quantitative structure−property relationships (QSPRs), and quantitative structure−activity relationships (QSARs).[3−7] With a plethora of data sources available, it is possible to merge information from multiple HTS libraries in order to obtain a highly diverse set of druglike molecules which can be used to model both physiochemical and biological properties. Applying *in silico* screens to filter out molecules likely to fail ADMET, especially toxicity, is fiscally necessary considering that it can take over a decade and close to $1 billion to release a new and federally approved drug.[8]

PubChem, the database component of the National Institutes of Health (NIH) Molecular Libraries Initiative, serves as a public repository of chemical and biological activity data generated by the Molecular Libraries Screening Center Network and other screening centers.[9−11] PubChem is a user/depositor system, which accepts annotated chemical structures and related biological activity data. PubChem is broken down into three main database components: Compound, Substance, and BioAssay. PubChem Compound contains over 18 million unique chemical structures with their respective calculated physiochemical properties. PubChem Substance contains over 28 million records with structural data, descriptions of chemical samples from multiple sources, and links to 3D protein structures as well as PubMed citations. Finally, biological screening results for over 800 assays are stored in PubChem BioAssay, including the NCI60 human tumor cancer cell line HTS.[12]

When procuring data from multiple sources, quality and reliability are often issues. Although PubChem has taken some measures to ensure quality control (QC) in terms of referencing between compounds and substances, the quality of the structural data related to the bioassay data of PubChem was termed "user beware" because the structural content submitted by the depositor is accepted without review.[13] It should be mentioned that PubChem has neither the resources nor the assigned responsibility to curate the data. Notably, if errors are identified, they may be reported to and corrected by the depositor. Accepting screening data from multiple sources, one might expect inconsistent end points and instrument variations leading to standardization issues and precision errors.

In order to address the issues of QC and reliability, we have chosen to work with the NCI60 human tumor cell line anticancer drug screen, as it is one of the most recognized data sets assembled by a single organization with all assays

---

* Corresponding author e-mail: gcrippen@umich.edu.
† Department of Medicinal Chemistry, College of Pharmacy.
‡ Department of Statistics.
§ Department of Pharmaceutical Sciences, College of Pharmacy.
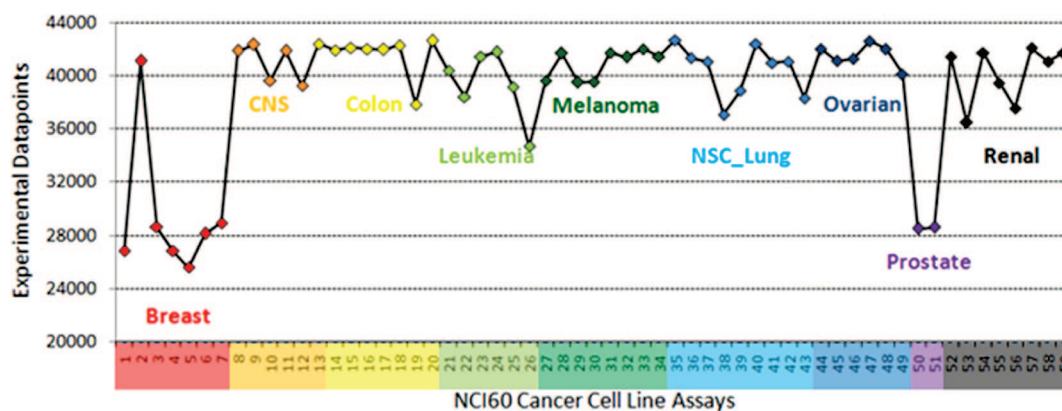
Figure 1. The number of available log GI_{50} values for the 43 474 substances with measurements in units of molarity, color-coded according to the class of cancer cell lines. The cell lines within each class have been alphabetized and numbered.

run at the same location. It provides a well-curated publicly available data set of toxicity profiles for 43 899 substances assayed *in vitro* against nine distinct organ-based classes of cancer: breast, colon, CNS, leukemia, lung, melanoma, ovarian, prostate, and renal. Furthermore, cytotoxic concentrations of substances determined *in vitro* have been shown to correlate well to lethal doses in laboratory animals and humans for a range of selected drugs and chemicals.[14,15] With a rich history spanning over 20 years, 59 of the 60 cancer cell lines are still currently available. The NCI60 contains *in vitro* screening data for up to three IC_{50} end points: GI_{50}, TGI, and LD_{50}, referring to the concentration of a substance, in units of molarity or micrograms per milliliter, required for 50% growth inhibition, total growth inhibition, and 50% lethal dose, respectively. The GI_{50} is our measurement of choice, as the lowest concentrations of substances are used for the observed effect. In this paper, we use only the log GI_{50} values where the concentration unit is molarity.

In a previous work, we contributed our novel method for automatically generating selective SMARTS strings which are able to classify cytotoxic molecules based on mean log GI_{50} cutoff of −5.0.[16] Each substance is associated with its respective mean log GI_{50}, which is the calculated mean of the available log GI_{50} data from the NCI60 for each respective substance. While the SMARTS produced work quite well as a filter, they fail to classify a significant portion of the NCI60. Here, we are more concerned with a robust model that accurately predicts mean log GI_{50}.

In this study, we take a closer look at NCI60, as it refers to the existing 59 cancer cell lines. First, we address the issues of data set acquisition, analysis, and completeness. Next, we examine how principal component analysis (PCA) can be applied to large chemical data sets to extract hidden relations and attribute meaning to orthogonal toxicity profiles. We then apply binary decision trees for the *in silico* prediction of general cytotoxicity. Finally, we demonstrate how stepwise regression was used to develop a least-squares fit model allowing data from nine of the NCI60 cell lines to be used as an accurate predictor of generalized cytotoxicity across the 59 cell lines.

METHODS

All calculations were performed using the Molecular Operating Environment (MOE)[17] on a Dell Precision 380 workstation utilizing Red Hat Linux Enterprise, version 4.0.

Data Gathering. The PubChem FTP site[9] was our preferred data source, as structural data for each molecule could easily be obtained through the association of the PubChem Substance and BioAssay databases. The entire PubChem database is available in the following file formats: abstract syntax notation (ASN) and extensible markup language (XML). PubChem Substance and Compound downloads are also available in standard data file (SDF) format,[18] while BioAssay downloads are also available in comma separated value (CSV) format. MOE's built-in functions were used to import and merge the required structural data and log GI_{50} profiles into a flat table for computational analysis.

Data Analysis. Understanding the landscape of a data set is necessary in order to avoid "garbage in garbage out", especially in cases where the data set is an incomplete matrix, as is the case with the NCI60 data set. Preliminary analysis of the toxicity profiles showed that only 88.2% of the assay data was complete; that is, data were available for 88.2% of the $59 \times 43\,889$ possible experimental end points. Approximately half of the experimental data consists of upper- or lower-threshold concentrations signifying minimal, log $GI_{50} = -4.0$ or −5.0, or maximal, log $GI_{50} = -8.0$, activity. The remaining portion of the experimental data shows some quantitative level of activity which is not threshold. Furthermore, only 4824 substances have been screened against the entire NCI60 data set. The NCI60 data set provides log GI_{50} data based on the measurements taken in one of three concentration units: molarity (M; 43 474 substances), micrograms per milliliter (μg/mL; 369 substances), and volumetric (48 substances). There is no log GI_{50} data for 108 substances. In order to determine the units used for the volumetric formats, one must contact the contributor, according to the NCBI help desk. See Figure 1 for the distribution of available experimental log GI_{50} values in molarity across the different cell lines.

Figure 2 describes the mean log GI_{50} for 43 474 compounds of NCI60 using 1000 bins with the following statistics: mean = −4.518, standard deviation = 0.7447, mode = −4.0, minimum = −11.74, and maximum = 4.0. Interesting features to note are the significant skew of the data toward −4.0, a shoulder at −5.0, and the maximum mean log $GI_{50} = 4$. The maximum mean log $GI_{50} = 4.0$ is an anomaly, which we assume to be a data entry problem where the submitter intended to instead enter −4.0. In order

DATA MINING NCI60

J. Chem. Inf. Model., Vol. 48, No. 7, 2008 **1381**
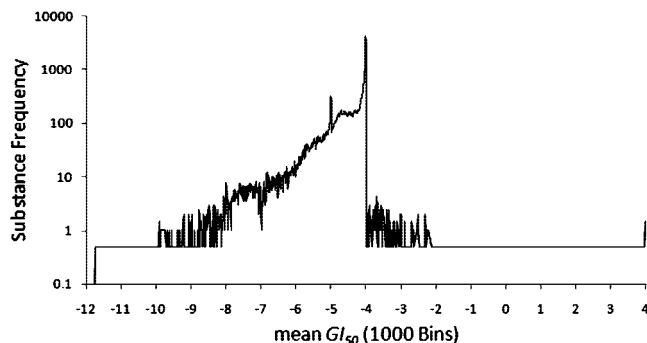
Figure 2. The mean log GI_{50} of the NCI60 substances shown with a granularity of 1000 bins.

(a)

(b)
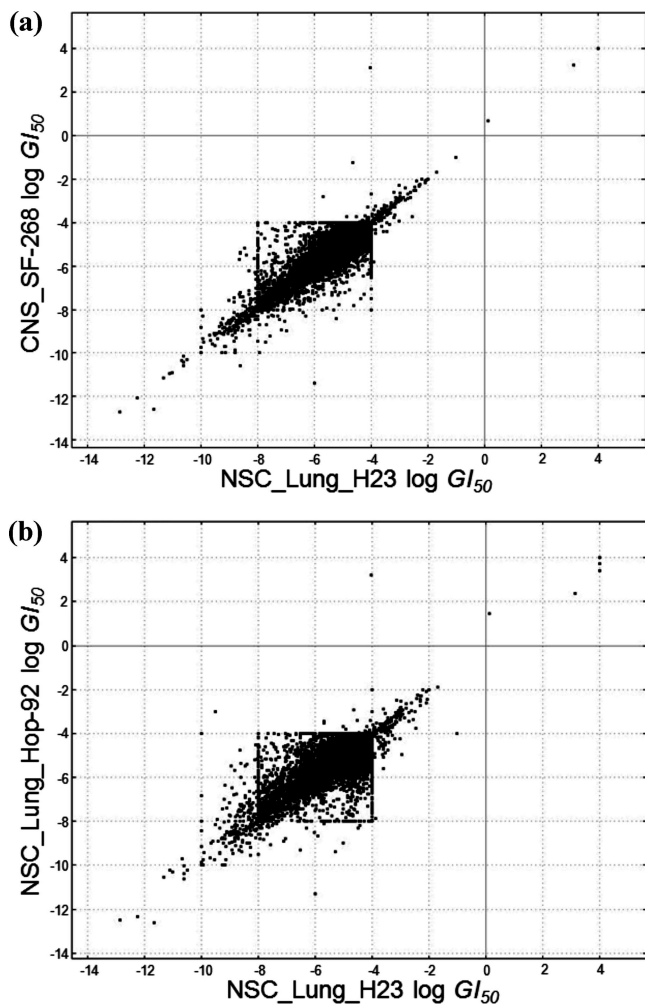
Figure 3. (a) Substance log GI_{50} correlation for 40 131 data points between a non-small-cell lung and CNS cancer cell lines having $R^2 = 0.86$. (b) Substance log GI_{50} correlation for 35 680 data points between two non-small-cell lung cancer cell lines having $R^2 = 0.73$.

to conform to the most common standard upper threshold, the mean log GI_{50} of all substances greater than -4.0 was adjusted to -4.0. Approximately 80% of all substances have a mean log GI_{50} greater than -5.0, explaining the skew toward inactivity. Finally, 5782 and 413 substances, respectively, have mean log $GI_{50} = -4$ and -5 (indicating no activity) for all assays against which those substances were screened. These values correspond to the maximal allowable concentration of a substance used for assays over a specific time frame. The National Cancer Institutes Developmental

Therapeutics Program (NCI/DTP) homepage designates a link to important changes to the NCI60 cell screen, specifying the addition of a one-dose 59 cell assay at a 10^{-5} M concentration (corresponding to log $GI_{50} = -5$) in an attempt "to increase substance throughput and reduce data turnaround time to suppliers while maintaining efficient identification of active compounds." This is followed by the regular five-dose assay used to determine GI_{50}, TGI, and LD_{50}.[12] For some period in the past, only the five-dose assay was used with a maximum recorded concentration of 10^{-4} M, corresponding to log $GI_{50} = -4$ (inactive).

Correlation analysis between the activity data from different NCI60 screens of the same substances, that is, on different cell lines, can also be applied to detect anomalies. In this case, two of the threshold values are visualized. In Figure 3a and b, the upper log GI_{50} threshold is apparent at -4, and a lower threshold is seen at -8. The lower threshold was used for only 178 substances. In this case, the threshold values are the maximum and minimum cutoffs in the NCI60 dose response experiments. Note, another apparent threshold of log $GI_{50} = -10$ can also be detected for some experiments.

MOE's correlation matrix tool allowed us to select n database fields (in this case, cancer cell lines containing log GI_{50} data) for which it calculates an $n \times n$ matrix of pairwise log GI_{50} correlations (R), where $n \leq 25$. By selecting a specific pairwise correlation from the matrix, we can visualize the scatter plot between any two of the selected cancer cell lines. When examining the scatter plots of pairs of NCI60 log GI_{50} data, very strong correlations were observed. The R^2 ranged from 0.66 to 0.88. It is interesting to note that some cell lines of similar tissue type were less correlated than those of differing organ types, as shown in Figure 3a and b.

The initial analysis is performed to ensure the quality and mining potential for the data set. In this case, the vast majority of the log GI_{50} data points occur over the range $[-8, -4]$, corresponding to 4 orders of magnitude in concentrations. Just over 10% of the substances have complete toxicity profiles, 43% of the data points are at threshold values, and 12% of the possible data points are missing. It is encouraging to note that there is a 4 orders of magnitude difference between the high and low threshold values, representing the difference between 0.1 mM and low 10 nM concentrations. This range is good for the differentiation of profiles based on activity, as required concentrations for drug leads are typically in the low micromolar range. It is expected that the vast majority of molecules in an HTS data set will be inactive against multiple or all targets of interest; however, it is desirable that the distribution of activity data span several orders of magnitude for a large set of substances.

Imputing Missing Data. Missing values are a problem for all data analysts. If only a few substances had missing values, we could simply omit them from the data set, but in this case, many substances have missing log GI_{50} observations for multiple assays. Removing these substances drastically reduces the data set from 43 474 to 4824 substances, which would diminish the predictive power of the resulting model. We implemented a common strategy used in linear models in order to preserve the size of the predictor subset by imputing the missing data, null values, with the mean
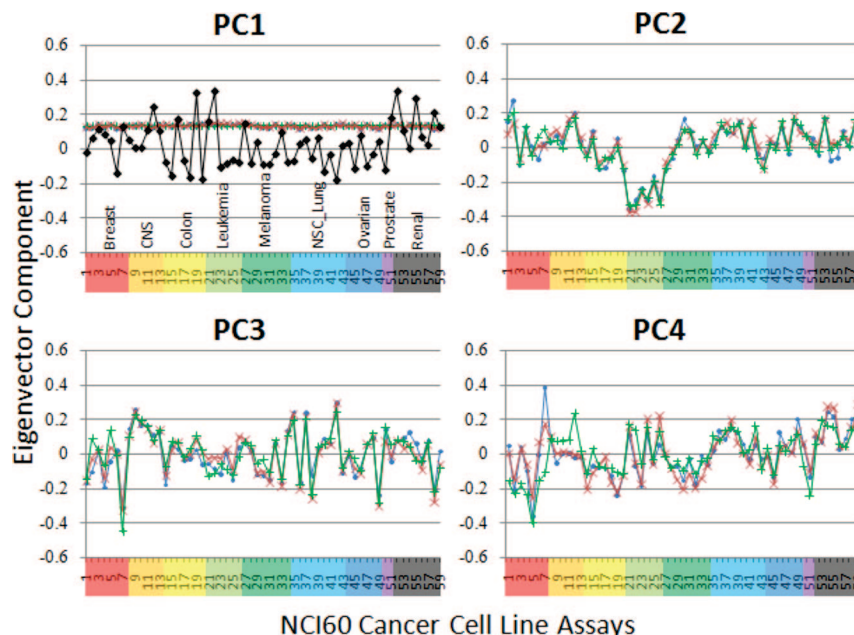
Figure 4. Plots of eigenvector componenets versus NCI60 cancer cell line assays. The *x* axis represents the 59 cancer cell line assays and has been color-coded and organized as described in Figure 1. The *y* axis represents the components of the eigenvectors for each respective principal component. (−•−, black) Random toxicity profiles for 4824 entries, each cell within the range [−4, −9], plotted only for PC1. (−•−, blue) From the 4824 substances having complete profiles. (−×−, red) From the 43 474 substances using imputed data for missing values. (−+−, green) From the 43 474 substances ignoring missing values.
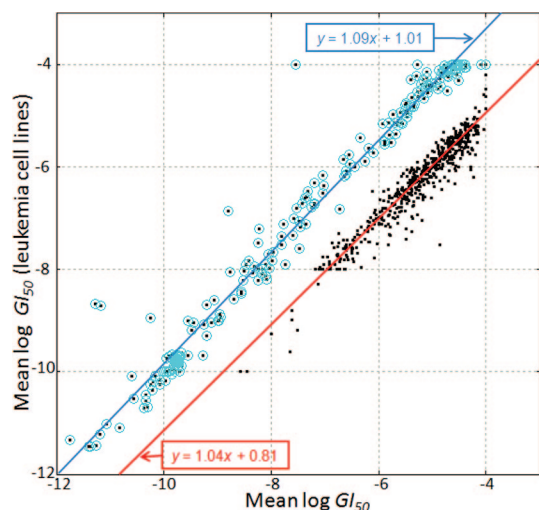

Figure 5. (1) The fit line through 543 substances determined by PC2 (•), which are most responsible for increased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines. (2) The fit line for 239 substances determined by PC2 (•, in a blue circle), which are responsible for decreased cytotoxicity of the leukemia cell lines compared to the nonleukemia cancer cell lines.

log GI_{50} of all the screening data for each respective substance.[19]

Data Mining with PCA. Principal component analysis was used for two purposes. First, PCA was used for validation purposes to ensure that we did not inadvertently skew our matrix of toxicity profiles by imputing 12% of the missing assay data with the mean log GI_{50} for each respective substance. PCA was performed on a matrix of random values from [−9.0, −4.0] to show that the first principal components of the data sets containing experimental and imputed data were not due to mean centering. Also, PCA of the nonimputed data set was performed using the covariance matrix derived from eq 1, such that all existing pairs of log GI_{50} values between assays were considered. The covariance
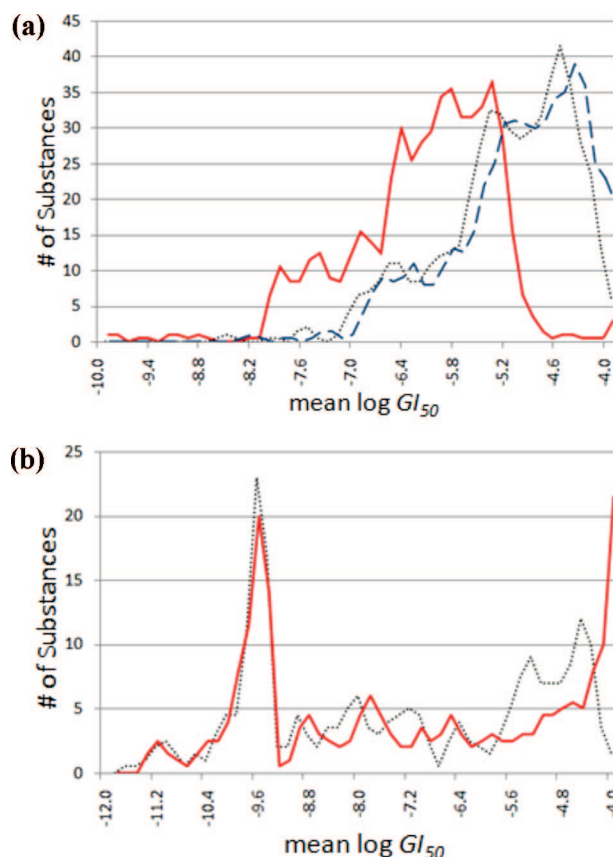
Figure 6. (a) The 543 substances determined by PC2 which are most responsible for increased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines. The mean log GI_{50}'s of the leukemia cell lines (−, red), all cell lines (•••, black), and the nonleukemia cancer cell lines (--, blue) are depicted. (b) The 239 substances determined by PC2 which are most responsible for decreased leukemia cell line toxicity relative to the remaining 53 NCI60 cell lines. The mean log GI_{50}'s of the leukemia cell lines (−, red) and all cell lines (•••, black) are depicted.

DATA MINING NCI60

J. Chem. Inf. Model., Vol. 48, No. 7, 2008 **1383**
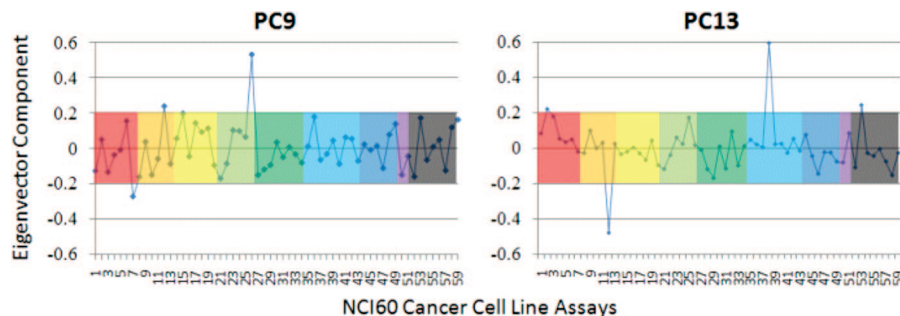
Figure 7. PC9 and PC13 eigenvector componenents versus NCI60 cancer cell line asssays. PC9 specifically identifies the Leukemia_SR cancer cell line related eigenvector component. PC13 identifies NSC_Lung_Hop_92 and CSN_SNB_75. The axes and color coding scheme are described in Figure 1.
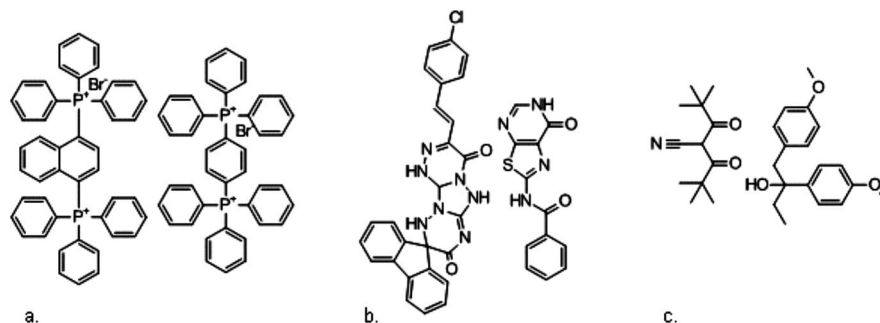

Figure 8. Comparison of molecules from the 45 substances having specific toxicity for the Leukemia_SR cancer cell line. The left molecule of each pair has the greatest similarity to the right one. (a) $S_{tan} = 0.92$, (b) $S_{tan} = 0.57$, (c) $S_{tan} = 0.24$.

matrix is a square 59×59 matrix, where each row i and column j correspond to their respective NCI60 cancer cell line. Let the matrix $\mathbf{X}$ be defined by $x_{k,j}$, where $x_{k,j}$ is the log GI_{50} for substance k and cell line j. Then, $\bar{x}_j$ is the mean log GI_{50} of the substances for cell line j. Now, let k refer to the substances having experimental log GI_{50} data for assays i and j, such that $\bar{x}_i$ and $\bar{x}_j$ refer to the mean log GI_{50} values for the substances having experimental log GI_{50} data for both assays i and j. Note that $i = j$ is allowed in eq 1. Validation for using an imputed data set can be established by showing a high correlation between the components of the eigenvectors responsible for the majority of log GI_{50} variance from the imputed and nonimputed data sets.

$$C_{i,j} = n^{-1} \sum_{k=1}^{n} x_{k,i} x_{k,j} - \bar{x}_i \bar{x}_j \qquad (1)$$

PCA was then used to extract interesting features of the NCI60 for further investigations. MOE's PCA tool can output the importance of substance contribution to the principal components based on the respective log GI_{50} toxicity profile of $\mathbf{X}$. This is done by first calculating the sample average vector $\mathbf{x} = [\bar{x}_1,..., \bar{x}_{59}]$ and covariance matrix $\mathbf{C}$ based on the matrix of toxicity profiles for all n substances. $\mathbf{C}$ is diagonalized such that $\mathbf{C} = Q^T D D Q$, where Q, the PCA transform, is orthogonal and D is diagonal-sorted from top left to bottom right. There are p nonzero diagonal values in D, the square roots of the eigenvalues of S, corresponding to the principal components. If we take the $p \times n$ matrix $\mathbf{Z} = Q(\mathbf{X} - \mathbf{x})$, such that $\mathbf{Z}$ has identity covariance and zero mean, there exists a p vector of the form $z_i = Q(x_i - \mathbf{x})$, where the p components of each z_i can be taken as the relative weights for the respective principal components corresponding to each substance. Hence, the z_{ip} values are weights for each substance (i) and principal component (p). The sub-
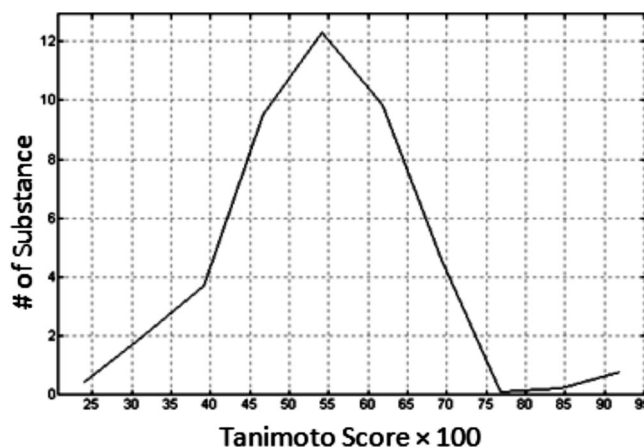
Figure 9. Histogram of the highest S_{tan} derived by comparing pairs of substances using the MACCS fingerprint on the set of 45 substances specifically targeting the Leukemia_SR cancer cell line. Mean = 0.55, $\sigma = 0.13$, min = 0.24, max = 0.92.

stances most responsible for a particular principal component's variance have the largest magnitudes of z_{ip} values.

Predictive Binary Decision Tree. We applied two concepts in the construction of a binary decision tree. First, consider the "ideal" fingerprint where the different descriptors' occurrence are statistically independent and each descriptor evenly divides a data set of n molecules with some property value, such as mean log GI_{50}. The "ideal" fingerprint has length n_d, where n_d is the minimum number of descriptors to uniquely identify all n molecules in the training set.

$$n_d = \text{ceil}(\log_2 n) \qquad (2)$$

Using this concept, we have applied the MACCS keys as our base fingerprint. The MACCS key which most evenly divides the substances at any given node in our tree is given weight when making our branching decision. As our goal

Table 1. Comparative Summary of Recent Cytotoxicity Study Results

method	data set[a]	training set size	test set size	R^2	log GI$_{50}$ cutoff	correctly classified (%)[b]	improvement[c]
neural network[3]	19 libraries	8298	2000	N/A	N/A	73[d]	N/A
QSPR[4]	NCI60	27 000	N/A	0.67	−6	N/A	× 10
proteomic profiling[5]	NCI60[e]	2/3	1/3	N/A	N/A	60	N/A
QSAR[6]	37 cpds[f]	4/5	1/5	0.73	N/A	N/A	N/A
random forest[7]	NCI60	≤ 42 723	≤ 42 723	N/A	−5	76[g]	N/A
decision tree	NCI60	21 763	2015	0.54	−5	83	× 4
					−6	93	× 10
least squares fit	NCI60	2412	2412	0.99	−5	96.7	× 2
					−6	99.1	× 5
			38 650	0.99	−5	97.9	× 6
					−6	99.8	× 31

[a] The NCI60 is evolving; that is, newer versions are larger and have more complete log GI$_{50}$ toxicity profiles. [b] Substances with log GI$_{50}$ greater than or equal to the cutoff value are considered nontoxic, while those with log GI$_{50}$ values less than the cutoff value are considered toxic. All percentages were rounded down. [c] Improvement over random selection is based on the respective cutoff value. [d] A total of 20% of all substances were unclassified and assumed to be incorrect. [e] The data set consisted of 118 drugs from the NCI60. [f] The data set consisted of 37 naphthoquinone ester derivatives. [g] The value 76% is the average of correctly classified substances for the individual cancer cell lines, not the mean log GI$_{50}$.

was to not only reduce the number of substances at each child node, but also to reduce the range of mean log GI$_{50}$, a weighted accuracy factor was also included in our branching decision. Branching is allowed to continue as long as a MACCS key exists that can divide a node into children nodes, each containing no fewer than two substances. If branching cannot occur, the node is taken to be a leaf. Thus, the trained binary tree has a MACCS key at each nonterminal node, and a prediction value at each leaf equal to the average of the mean log GI$_{50}$'s of the training set substances at each respective leaf node. Equations 3–8 describe how decisions are made at each node.

$\mathbf{h}_k$ and $\hat{\mathbf{h}}_k$ are binary vectors with lengths equal to the number of substances m at a node with $\mathbf{h}_k$ representing the hit profile and $\hat{\mathbf{h}}_k$ the inverse hit profile with respect to MACCS key k. $|\mathbf{h}_k|$ and $|\hat{\mathbf{h}}_k|$ are the sum totals of the hits and misses for the respective profiles, see eqs 3–5.

$$|\mathbf{h}_k| = \mathbf{h}_k \bullet \mathbf{h}_k = \sum_{j=1}^{m} h_{kj} \tag{3}$$

$$|\hat{\mathbf{h}}_k| = \hat{\mathbf{h}}_k \bullet \hat{\mathbf{h}}_k = \sum_{j=1}^{m} \hat{h}_{kj} \tag{4}$$

$$m = |\mathbf{h}_k| + |\hat{\mathbf{h}}_k| \tag{5}$$

Equation 6 defines the idealness score, $S_{1,k}$, for MACCS key k where $0.5 \leq S_{1,k} \leq 1.0$. A value of 0.5 indicates that

the substances at a particular node are evenly divided into children nodes, whereas a score of 1.0 indicates that one child node contains all of the substances and the other contains none.

$$S_{1,k} = \frac{\max[|\mathbf{h}_k|, |\hat{\mathbf{h}}_k|]}{m} \quad \text{for } k = 1,..., 166 \tag{6}$$

Equation 7 defines the accuracy score $S_{2,k}$, for MACCS key k where $0 < S_{2,k} \leq 1$. r_k and $\hat{r}_k$ are the log GI$_{50}$ ranges at the two child nodes for MACCS key k. Values less than 1.0 represent child nodes with smaller log GI$_{50}$ ranges than the parent node. Values closer to zero reflect child nodes with more narrow log GI$_{50}$ ranges.

$$S_{2,k} = \frac{r_k|\mathbf{h}_k| + \hat{r}_k|\hat{\mathbf{h}}_k|}{\max_k[r_k, \hat{r}_k] \times m} \quad \text{for } k = 1,..., 166 \tag{7}$$

Equations 6 and 7 have been normalized such that the lower values of the scores are desirable. Equation 8 describes how the final decision is made at each node and considers the ability of a particular MACCS key to evenly split the substances and at the same time minimize the log GI$_{50}$ range of the children nodes. Weighting factors for $S_{1,k}$ and $S_{2,k}$ were systematically determined by spawning several decision trees and varying the weights. The final weighting scheme was the one resulting in the highest fit R^2 based on correlating
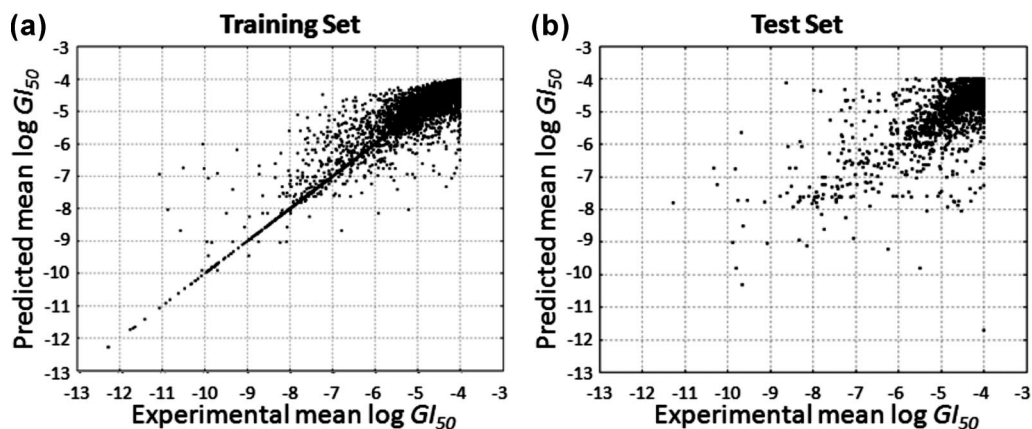
Figure 10. Prediction results using a decision tree. (a) The training set had a fit $R^2 = 0.9284$ with RMSE = 0.2285. (b) The test set exhibited predictive $R^2 = 0.5432$ with RMSE = 0.7119.

DATA MINING NCI60

J. Chem. Inf. Model., Vol. 48, No. 7, 2008 **1385**

Table 2. Decision Tree Results for Cutoff log $GI_{50} = -5.0$

confusion matrix	cutoff log $GI_{50} = -5.0$			cutoff log $GI_{50} = -6.0$		
	pred. nontoxic	pred. toxic	% correct	pred. nontoxic	pred. toxic	% correct
exptl. nontoxic	1248	252	83.2	1744	97	94.7
exptl. toxic	90	425	82.5	49	125	71.8

the mean of all log GI_{50} values of the leaf and experimental log GI_{50} values from our training set.

$$S = \min_{k}[S_{1,k} + 5.0 \times S_{2,k}] \qquad (8)$$

Substances were randomly divided into 10 subsets of nearly equal size. The predictive binary decision tree was trained on 90% of the substances and validated on the remaining 10% in an attempt to maximize two relevant metrics: R^2 and the binning of cytotoxic (mean log GI_{50} < −5.0) and noncytotoxic (mean log GI_{50} > −5.0) substances. Cross-validation was performed on different training and test sets based on the random classification of the substances. A final validation was performed to ensure that the predictive nature of this procedure was not an artifact by randomly assigning the mean log GI_{50} values to different substances within the training set and then rebuilding the decision tree. There was essentially no log GI_{50} correlation (R^2 < 0.02) between the predicted values for the training set and these randomly assigned values.

Prediction Using Least Squares Fit. A more accurate prediction model was devised by randomly dividing the 4284 substances having complete toxicity profiles into equal-sized training and test sets. The experimental log GI_{50} values for each cell line were then correlated to the mean log GI_{50}. Starting with the cell line whose log GI_{50} values have the highest R^2 with the mean log GI_{50} values, we applied forward stepwise linear regression with the constraint that no more than one cell line from each class can be used. The final model was validated on the test set. A second validation was performed using the remaining 38 650 substances with incomplete toxicity profiles.

RESULTS AND DISCUSSION

We chose to adopt this strategy of imputing missing data with mean values to maintain the largest possible set of substances for mining. In order to validate that imputing the mean log GI_{50} did not inadvertently skew the overall data set, PCA was performed on four different data sets: the complete set of 4824 substances having complete toxicity profiles, the imputed set consisting of all 43 474 substances with the imputed mean log GI_{50} for the missing values respective to each substance, the nonimputed data set consisting of all 43 474 substances with missing values, and a random data set modeled after the complete set having 4824 entries each with 59 random data points ranging from [−9, −4]. Figure 4 compares the first four principal components, accounting for over 92% of the log GI_{50} variance for all data sets. There is extremely good correlation between the components of the eigenvectors for these principal components relating to all data sets except for the one with randomly assigned data. As expected, the correlation of principal components between the data sets continues to degrade when examining the lower-order components.

Since the components of the eigenvectors are strongly correlated, we can assume that the imputed mean values did not severely impair the quality of the data set. A similar analysis, comparing the principal components of the set of substances having complete toxicity profiles and those having greater than or equal to 90% complete toxicity profiles, showed an even closer fit to the nonimputed data set depicted in Figure 4 than any of the other data sets.

It was interesting to note that PC1 accounts for over 89% of the log GI_{50} variance for all of the data sets with real experimental values, while PC1 for the data set having randomly generated log GI_{50} values only accounted for 2% of the variance. Aside from the random data set, all of the components of the eigenvector for PC1 were found to be approximately equal in magnitude (even more so in the case of the nonimputed data set). Given the high explanatory power of this component, we see that many substances tend to be uniformly toxic across all of the NCI60 assays. Even when the 176 and 6203 substances having only threshold log GI_{50} values are eliminated from the respective data sets of complete toxicity profiles and the one including imputed values for all substances, PC1 still accounts for over 88% of the total log GI_{50} variance. Since this pattern dominates in both the imputed and complete data sets, while the random data set deviates and explains very little, we can conclude that this is not an artifact of our procedure. Since PC1 corresponds to uniform log GI_{50} across all cell lines, we can artificially remove this component by mean-centering each substance's toxicity profile and performing PCA once again. Indeed, PC2 in Figure 4 from the first analysis becomes the first principal component in the new analysis. In the original PCA, PC2 explained approximately 1.2% of the log GI_{50} variance. As the first principal component in the new analysis, it explained 12% of the log GI_{50} variance and maintained the eigenvector components from the first analysis. In any case, the result that most compounds show uniform toxicity (high or low) across all cell lines is hardly surprising, but it leads to our least-squares model that greatly reduces the effort required for screening compounds.

PC2 corresponds to a rather uniform level of toxicity across most cancer cell lines, with the majority of its components between −0.17 and +0.17. It is very interesting to note that all six leukemia cell lines (RPMI_8226, SR, CCRF_CEM, K_562, MOLT_4, and HL_60(TB)) have eigenvector components less than −0.2. For the complete subset, CNS_SNB_75 and Breast_HS578T have component values greater than 0.2. The increased absolute value of the eigenvector components corresponds to certain cancer cell lines, indicating that there are groups of substances for which these cell lines are either more sensitive or more resistant. PC2 is responsible for just over 1% of the total variance over all cancer cell lines. We attempted to relate the values of the eigenvector components to the doubling times of these outstanding cancer cell lines and noted that the leukemia cell lines' doubling times (19.6−33.5 hours) are approximately half that of CNS_SNB_75 (62.8 hours) and Breast-HS578T (53.8 hours).[12] Unfortunately, this trend does not follow for the remainder of the cell lines with eigenvector values close to −0.2 and +0.2. The outstanding feature of PC2 is that it identified all of the leukemia cell lines.

We speculate that PC2 may have identified the leukemia cell lines due to the fact that a greater portion of their surface

areas is available to exogenous substances. The leukemia cell lines are grown in vials suspended in solution, whereas the other cancer cell lines are grown on plates and require attachment to the plate wall, reducing the exposed surface area. Therefore, it is plausible that the leukemia cancer cells are more susceptible to toxic substances due to increased surface area exposure.

In order to substantiate this claim, we first identified the substances most responsible for PC2 using MOE's PCA tool. We found over 4000 substances with $|z_{i2}| \geq 1.0$ and removed them from the complete data set in order to examine the eigenvectors produced by PCA in their absence. By removing the substances with $|z_{i2}| \geq 1.0$, we were able to reduce the predominance of this feature 10-fold. Instead of being the second principal component and accounting for over 1% of the total log GI_{50} variance, the leukemia cancer cell lines were identified in the ninth principal component along with several other cancer cell lines with absolute values of eigenvector components greater than 0.2 and accounting for less than 0.1% of the total log GI_{50} variance. Furthermore, there are 782 substances with $|z_{i2}| \geq 2.0$. Notably, less than 5% of the data was imputed with 313 substances having no imputed data whatsoever. A total of 543 of these substances, 228 with no imputed values, have fairly uniform increased cytotoxicity against the leukemia cell lines compared to the other 53 cancer cell lines. The remaining 239 substances, 75 with no imputed values, were shown to be slightly less cytotoxic to the leukemia cell lines on average. The more sensitive nature of the leukemia cell lines is illustrated by the scatter plot in Figure 5 between the mean log GI_{50} for the leukemia cell lines and the mean log GI_{50} for the nonleukemia cell lines for the 543 substances of interest. The plot shows that there is an approximately 1.8 orders of magnitude difference between the respective mean log GI_{50} of the leukemia and other cell lines. Also, the mean log GI_{50} distributions of the leukemia cell lines are shifted to significantly higher levels of toxicity than the distributions of the nonleukemia cell lines, as seen in Figure 6a and b. As a side note, only 33 of the 782 substances most responsible for PC2, $|z_{i2}| \geq 2.0$, were uniformly more toxic to the leukemia cell lines than all others of the NCI60. Even though the substances responsible for PC2 do not exhibit the highest levels of cytotoxicity for all leukemia cell lines, they do show a tendency to be uniformly more toxic for the majority of the leukemia cell lines.

We examined the layout of the eigenvector components for the latter principal components looking for outstanding features similar to that of PC2. While no other classes were uniformly identified as having all cancer cell lines with the absolute value of the eigenvector components greater than 0.2, we did find several principal components that identified a few cancer cell lines with the absolute value of the eigenvector components significantly greater than 0.2. The two principal components with few outstanding eigenvector components accounting for the largest log GI_{50} variance are depicted in Figure 7.

Of 600 substances found to be responsible for PC9 with $|z_{i9}| \geq 1.0$, only 45 showed specific cytotoxicity against the Leukemia_SR cell line; that is, the concentrations of substances required to inhibit cell growth for the Leukemia_SR cell line were 2−4 orders of magnitude lower than the concentration necessary for the remainder of the NCI60

cell lines. We found that most of these substances are structurally dissimilar. Only one pair, the phosphonium molecules, show significant structural similarity with a Tanimoto similarity coefficient $S_{tan} = 0.92$ based on the MACCS keys, as depicted in Figure 8a. Figure 8b and c illustrate two other molecules and their respective highest-scoring (most similar) match within this set of molecules. In the game of fingerprint-based similarity searching, one typically does not consider molecules with an $S_{tan} < 0.70$ as structurally similar. Most of the 45 substances exhibit multicyclic ring systems with aromatic components, and we have identified a few purine and pyrimidine derivatives. Figure 9 depicts a histogram of the highest S_{tan} for each of the 45 substances when compared to the other respective substances in the data set. The mean S_{tan} for each molecule and its most structurally similar pair of the set of 45 substances was 0.55. Having established that these substances are structurally dissimilar yet share similar toxicity profiles, we conjecture that these substances do not share a common mechanism of action against the Leukemia_SR cell line, which is contrary to the belief that substances sharing the same toxicity profiles also follow the same mechanism of activity.[20] We are not implying that toxicity profiles cannot be used for predicting the mechanism of activity, but rather we believe that there may be several viable Leukemia_SR cell specific targets which when activated lead to cell death on the basis of different mechanisms of action.

PC13 identifies the Lung Hop_92 and CNS_SNB_75 cell lines. Even fewer substances are responsible for this principal component, as it is responsible for less than 0.3% of the total system variance. The analysis of the substances responsible for toxicity profiles matching the landscape of PC13, while similar to that of PC9, can neither confirm nor reject our hypothesis that multiple mechanisms of actions may be in play when examining cytotoxic substances with similar toxicity profiles.

Several groups have recently published their ability to predict cytotoxicity.[3−7] Table 1 contains a summary of the methods used and results obtained, including the results in our study. Our method used the MACCS keys, paying attention to narrowing the range of log GI_{50} in all children nodes of our decision tree. While we were unable to achieve reliable results in a leave-some-out cross-validation study using only the substances in the data set with complete toxicity profiles, we did derive a predictive model that achieved $R^2 = 0.53$ and RMSE = 0.71, using the data set having imputed values for less than 10% of the cancer cell lines, see Figure 10. While this number is not outstanding, it allows for improved discrimination between toxic versus nontoxic substances. Furthermore, we achieve similar results when using different 90:10 training/test splits of the data set. When taking log $GI_{50} = -5$ as the cutoff, we were able to correctly classify 83% of all substances and 82% of the cytotoxic substances in our test set. With a log GI_{50} cutoff $= -6$, we correctly classified 93% of all substances in our test set, but only 72% of those taken to be cytotoxic. See Table 2 for the results presented as a confusion matrix.

Finally, with least-squares fitting, we achieved excellent prediction using equal-sized randomly selected training and test sets using the 4824 substances with complete toxicity profiles; that is, no data were imputed. While still skewed toward log $GI_{50} = -4.0$, this data set was more uniformly

DATA MINING NCI60

J. Chem. Inf. Model., Vol. 48, No. 7, 2008 **1387**
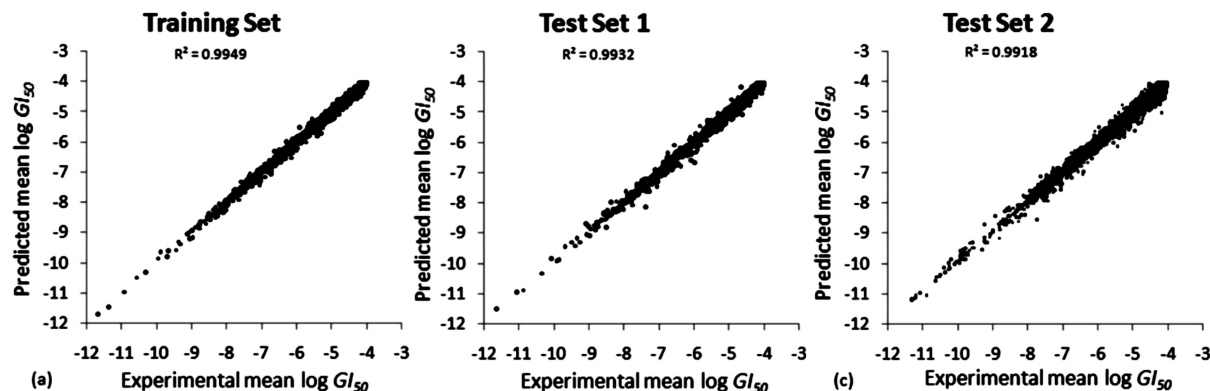
Figure 11. Prediction results using the least-squares fit and forward stepwise regression. (a) The training set consisted of 2412 complete toxicity profiles and had a fit $R^2 = 0.9949$ and RMSE = 0.07855. (b) Test set 1 consisted of 2412 complete toxicity profiles and had $R^2 = 0.9932$ and RMSE = 0.09000. (c) Test set 2 consisted of 38 650 incomplete toxicity profiles and had $R^2 = 0.9918$ and RMSE = 0.06195. The toxicity profiles used in a and b contained no imputed data, whereas the toxicity profiles used to determine c had imputed data.

distributed than the data set used when training and testing our decision tree. If one considers log $GI_{50} = -5.0$ as the cutoff value between toxic and nontoxic substances, then the data set is evenly distributed. Random selection was performed by first sorting the substances by ascending mean log GI_{50} and then dealing them one-by-one into the training and test sets. This ensured that there was an even distribution between the two sets. We identified the Ovarian_OVCAR-8 cancer cell line as having the best fit to the mean log GI_{50}. Applying forward stepwise linear regression and allowing only one cancer cell line from each class to be included in the model yielded eq 9.

$$
\begin{aligned}
\text{mean log } GI_{50} = \ &-0.08234 \\
&+0.06395 \times \text{Breast\_HS-578T} \\
&+0.13046 \times \text{CNS\_U251} \\
&+0.08476 \times \text{Colon\_HCC-2998} \\
&+0.09574 \times \text{Leukemia\_K-562} \\
&+0.13151 \times \text{Melanoma\_M14} \\
&+0.14644 \times \text{NSCLung\_NCI-H23} \\
&+0.09082 \times \text{Ovarian\_SK-OV-4} \\
&+0.10767 \times \text{Prostate\_PC-3} \\
&+0.12995 \times \text{Renal\_CAKI-1}
\end{aligned}
\tag{9}
$$

When applied to the test set, 96.7% and 99.1% assignment accuracy was obtained. The predictive R^2 was found to be greater than 0.99 and RMSE less than or equal to 0.09, see Figure 11a and b. While this method is superbly predictive for mean log GI_{50}, it requires that assay data be available for nine of the 59 available NCI60 cancer cell lines. One surprising observation shown in Table 1 based on the test set of 2412 substances with complete toxicity profiles was the relative lack of improvement over random selection when compared to our decision tree model. This may be attributed to the distributions of the test sets. In the test set described for the decision tree, there was an approximately 25% and 9% chance that a randomly selected substance would be toxic on the basis of the respective log GI_{50} cutoffs of -5.0 and -6.0. With the least-squares fit derived through stepwise linear regression, the respective chances are 50% and 18%. On the basis of the data set distribution, the second method can only improve half as much as the first. Thus, the quality of prediction should never be based on improvement over random selection alone. To verify the robustness of this method, we also

used the least-squares fit model to predict the mean log GI_{50} for the remaining 38 650 substances with toxicity profiles containing imputed data. We found that the correlation coefficient was little changed and the RMSE improved to 0.06. This was expected due to the incidences of imputed mean log GI_{50} values for all missing data, see Figure 11c. Here, significant improvement over random selection was seen. This was due to the change in ratio of toxic:nontoxic substances. The ratio for the test set having 2412 substances with the more complete toxicity profiles was 1200:1212 and 428:1984 for the log GI_{50} cutoff values of -5 and -6, respectively, whereas the ratios for the test set containing 38 650 substances with imputed data were 5268:33382 and 1215:37435 for the same respective log GI_{50} cutoff values. The ratio of toxic:nontoxic is inversely proportional to the improvement over random. When considering the improvement over random selection, one must also consider the classification ratio, as the maximum improvement over an evenly divided set 1:1 (50% chance to correctly classify any substance) is 2-fold, as was the case with our unimputed test set. In our case with 1215:37435 toxic:nontoxic substances it is possible to improve the identification of toxic substances by ~32 fold (38650/1215).

CONCLUSIONS

Refining chemical data sets can facilitate the process of drug development by helping to minimize the high attrition due to poor ADMET during the clinical phases.[21] The largest problem with current public domain chemical and biological activity data is the lack of a curation procedure and QC. We have shown that, even in the cases where curated data sets are available, one must carefully evaluate the data in order to ensure the greatest accuracy for data mining purposes.

In this work, we have preprocessed and validated the quality of a large reliable subset of NCI60 for data mining toxicity profiles. Here, we have used PCA to validate the use of larger training sets by demonstrating that PC1 was not an artifact due to imputing the mean log GI_{50} and further establishing a correlation between the components of the eigenvectors for the principal components responsible for 92% of the total log GI_{50} variance. The same steps can be used to refine screening data from multiple assays, whether they include a subset of the NCI60 assays, a combination of

the NCI60 and other biological screens, or any selection of HTS drawing their activity data from a common set of substances. Further examination of the PCA results led to interesting deductions regarding the general landscape of a data set. In this case, over 89% of the total system variance relates to the generally cytotoxic nature of substances, and the leukemia cell lines behaved differently. PC2, responsible for ~1.2% of the log GI_{50} variance, identifies the leukemia cell lines. Finally, it is possible to extract the substances responsible for the principal components in order to mine for similarities or differences. Here, PC9 and PC13 indicate that multiple mechanisms may share similar toxicity profiles based on the NCI60.

On the basis of our analysis of the substances responsible for PC9 and the latter principal components, we have found the chemical structures to be of such diversity that it would be impossible to derive the underlying QSARs based on the limited size of the data set and the likelihood that the substances' cytotoxic natures are due to different mechanisms of action. QSPR investigations may provide valuable insights regarding the compounds responsible for these principal components. Identifying QSARs within the larger groups of substances responsible for PC1 and PC2 on the basis of structurally similar subsets of these substances may also be possible. However, as it stands, both QSAR and QSPR investigations are beyond the scope of this work.

We derived two predictive methods: one using a binary decision tree, the other using forward stepwise linear regression and least-squares fit. In our study, greater than 10-fold enrichment over random selection can be expected for substances with mean log $GI_{50} < -6$, using both of our methods based on our subset of NCI60. The least-squares model further offers a very accurate method for determining the level of general cytoxicity for substances that need not be limited to the NCI data set. While it has shown improved results over past cytotoxicity prediction methods, there is one caveat, that future predictions on substances outside of NCI60 cannot be performed completely *in silico*. This pitfall is also our major discovery; that is, only 9 of the 59 available NCI60 cancer cell lines are required to implement our model. In other words, for the purposes of an initial screen to identify generally cytotoxic substances, the 59 cancer cell lines are highly redundant. While selectivity is one of the main goals with antitumor agents, our method can flag nonselective substances with significantly low mean log GI_{50} values for early elimination. Selective substances will not be flagged, as they have higher mean log GI_{50} values for the majority of cancer cell line assays, deemphasizing the increased toxicity of selective substance for only a few cancer cell lines. The goal is to flag compounds as early as possible for elimination in the drug discovery pipeline in order to save time and money through streamlining the toxicity detection system while decreasing the overall demand on chemical and biological resources.

Being able to more accurately identify nonspecific cytotoxins brings to light two new questions relevant to the Food and Drug Administration's approval of current and new drug substances. (1) Are noncytotoxic molecules more "druglike" for drugs that are not meant to be anticancer agents? (2) Are drug molecules less cytotoxic than nondrugs? Answering these questions might help to further access the overall generalizabilty of our methods and minimize the attrition rates of new potential drugs in the later stages of the drug development pipeline.

REFERENCES AND NOTES

(1) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.

(2) Acton, G. Toxicogenomics and Predictive Toxicology: Market and Business Outlook, 2004. http://www.vivogroup.com/reports.html (accessed Feb 29, 2008); http://www.the-infoshop.com/study/cd25153_toxicogenomics.html (accessed Feb 29, 2008).

(3) Molnár, L.; Keserű, G. M.; Papp, Á.; Lőrincz, Z.; Ambrus, G.; Darvas, F. A neural network based classification scheme for cytotoxicity predictions: Validation on 30,000 compounds. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1037–1039.

(4) Huang, R.; Wallqvist, A.; Covell, D. G. Assessment of in Vitro Activities in the National Cancer Institute's Anticancer Screen with Respect to Chemical Structure, Target Specificity, and Mechanism of Action. *J. Med. Chem.* **2006**, *49*, 1964–1979.

(5) Ma, Y.; Ding, Z.; Qian, Y.; Shi, X.; Castranova, V.; Harner, E. J.; Guo, L. Predicting Cancer Drug Response by Proteomic Profiling. *Clin. Cancer Res.* **2006**, *12* (15), 4583–4589.

(6) Saíz-Urra, L.; Maykel, P. G.; Tiejeira, M. 2D-autocorrelation descriptors for predicting cytotoxicity of napthoquinone ester derivatives against oral human epidermoid carcinoma. *Bioorg. Med. Chem.* **2007**, *15*, 3565–3571.

(7) Guha, R. Flexible Web Service Infrastructure for the Development and Deployment of Predictive Models. *J. Chem. Inf. Model.* **2008**, *48*, 456–464.

(8) Berkowitz, B. A.; Sachs, G. Life Cycle of a Blockbuster Drug. *Mol. Interventions* **2002**, *2*, 6–11.

(9) NCBI: PubChem Project: National Center for Biotechnology Information, Bethesda, MD 2008. http://pubchem.ncbi.nlm.nih.gov/ (accessed Jan 8, 2008).

(10) NIH Roadmap: Molecular Libraries and Initative: National Institutes of Health, Bethesda, MD 2005. http://nihroadmap.nih.gov/molecularlibraries/ (accessed Jan 8, 2008).

(11) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138–1139.

(12) DTP: Developmental Therapeutics Program NCI/NIH. http://dtp.nci.nih.gov/ (accessed Jan 8, 2008).

(13) Richard, A.; Gold, L. S.; Nicklaus, M. Chemical Structure Indexing of Toxicity Data on the Internet: Moving Toward a Flat World. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 314–325.

(14) Ekwall, B. Screening of toxic compounds in mammalian cell cultures. *Ann. N.Y. Acad. Sci.* **1983**, *407*, 64–77.

(15) Rowan, A. N.; Berlin, A.; Becking, G. C.; Ekwall, B.; Fernicola, N.; Friedrich, J.; Gournar, M. I.; Kaloyanova, F.; Krishna Murti, C. R.; Ordonez, B.; Sanockij, I. V.; Stammati, A. L. In *Short-term Toxicity Tests for Non-genotoxic Effects*; Bourdeau, P., Somers, E., Richardson, G. M., Hickman, J. R., Eds.; John Wiley & Sons Ltd: New York, 1990; Chapter 2, pp 7−9.

(16) Wang, H.; Klinginsmith, J.; Dong, X.; Lee, A. C.; Guha, R.; Wu, Y.; Crippen, G.; Wild, D. Chemical Data Mining of the NCI Human Tumor Cell Line Database. *J. Chem. Inf. Model.* **2007**, *6*, 2063–2076.

(17) *MOE: Molecular Operating Environment*, ver. 2007.0902; Chemical Computing Group: Montreal, Quebec, Canada, 2007.

(18) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255.

(19) Hastie, T. J.; Tibshirani, R. J. *Generalized Additive Models*; Chapman and Hall/CRC: New York, 1990; Vol. 43, Chapter 6, p 166.

(20) Zaharevitz, D.; Holbeck, S.; Bowerman, C.; Svetlik, P. COMPARE: a web accessible tool for investigating mechanisms of cell growth inhibition. *J. Mol. Graphics Modell.* **2002**, *20*, 297–303.

(21) Van de Waterbeemd, H.; Gifford, E. ADMET in *Silico* Modeling: Towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.

CI800097K