Data Mining the NCI Cancer Cell Line Compound GI_{50} Values: Identifying Quinone Subtypes Effective Against Melanoma and Leukemia Cell Classes

Kenneth A. Marx,* Philip O'Neil, Patrick Hoffman, and M.L. Ujwal

AnVil, Inc., 25 Corporate Drive, Burlington, Massachusetts 01803

Received March 11, 2003

Using data mining techniques, we have studied a subset (1400) of compounds from the large public National Cancer Institute (NCI) compounds data repository. We first carried out a functional class identity assignment for the 60 NCI cancer testing cell lines via hierarchical clustering of gene expression data. Comprised of nine clinical tissue types, the 60 cell lines were placed into six classes—melanoma, leukemia, renal, lung, and colorectal, and the sixth class was comprised of mixed tissue cell lines not found in any of the other five classes. We then carried out supervised machine learning, using the GI_{50} values tested on a panel of 60 NCI cancer cell lines. For separate 3-class and 2-class problem clustering, we successfully carried out clear cell line class separation at high stringency, $p < 0.01$ (Bonferroni corrected t-statistic), using feature reduction clustering algorithms embedded in RadViz, an integrated high dimensional analytic and visualization tool. We started with the 1400 compound GI_{50} values as input and selected only those compounds, or features, significant in carrying out the classification. With this approach, we identified two small sets of compounds that were most effective in carrying out complete class separation of the melanoma, non-melanoma classes and leukemia, non-leukemia classes. To validate these results, we showed that these two compound sets' GI_{50} values were highly accurate classifiers using five standard analytical algorithms. One compound set was most effective against the melanoma class cell lines (14 compounds), and the other set was most effective against the leukemia class cell lines (30 compounds). The two compound classes were both significantly enriched in two different types of substituted p-quinones. The melanoma cell line class of 14 compounds was comprised of 11 compounds that were internal substituted p-quinones, and the leukemia cell line class of 30 compounds was comprised of 6 compounds that were external substituted p-quinones. Attempts to subclassify melanoma or leukemia cell lines based upon their clinical cancer subtype met with limited success. For example, using GI_{50} values for the 30 compounds we identified as effective against all leukemia cell lines, we could subclassify acute lymphoblastic leukemia (ALL) origin cell lines from non-ALL leukemia origin cell lines without significant overlap from non-leukemia cell lines. Based upon clustering using GI_{50} values for the 60 cancer cell lines laid out by the RadViz algorithm, these two compound subsets did not overlap with clusters containing any of the NCI's 92 compounds of known mechanism of action, a few of which are quinones. Given their structural patterns, the two p-quinone subtypes we identified would clearly be expected to possess different redox potentials/substrate specificities for enzymatic reduction in vivo. These two p-quinone subtypes represent valuable information that may be used in the elucidation of pharmacophores for the design of compounds to treat these two cancer tissue types in the clinic.

INTRODUCTION

The NCI Developmental Therapeutics Program's compound data set represents a unique public domain chemical compound resource. Started in 1955, it has been collected from a variety of sources including organic synthesis programs, natural products extracts, and more recently anti-AIDS assays. In a data mining study of eight large chemical structure databases, it was observed that the NCI data set had by far the largest number of compounds unique to it compared to all the others.[1] Currently, it is being used in an extensive drug evaluation effort at the NCI. For tens of thousands of tested compounds, the data set represents experimentally determined compound effectiveness against a panel of 60 cancer cell lines, representing nine clinical tissue types, expressed in terms of the $-\log[c_i]$, where c_i is the compound i concentration at which a growth inhibition of 50% is achieved (termed the GI_{50} value).

Numerous reports have appeared in which the NCI data set has been mined by the intramural NCI Informatics research group of Weinstein and collaborators. These reports include the application of statistics, supervised learning via cluster correlation, principle component analysis, and various neural network techniques.[2,3] Many literature citations have described compound class subsets, such as tubulin active compounds,[4] pyrimidine biosynthesis inhibitors,[5] and topoisomerase II inhibitors,[6] that possess similar mechanisms of action (MOA), share similar structures, or develop similar patterns of drug resistance. Also, compound structure classes such as the ellipticine derivatives have been studied and point to the validity of the concept that fingerprint patterns of activity in the NCI data set encode information concerning MOAs and other biological behavior of tested compounds.[7]

* Corresponding author phone: (978)934-3658; fax: (987)934-3013; e-mail: kenneth_marx@uml.edu. Corresponding author address: Department of Chemistry, University of Massachusetts Lowell, Lowell, MA 01854.

10.1021/ci034050+ CCC: $25.00
Published on Web 08/09/2003

More recently, gene expression analysis has been added to the data mining activity of the NCI compound data set. Microarray expression analysis of many thousands of genes has been performed in the 60 cancer cell lines, in the absence of any added drug. For the most part, these cells lines were shown to cluster on the basis of their gene expression patterns, in a manner similar to how they cluster based upon their histological tissue of origin.[8] Interestingly, the gene expression profiles of the 60 cancer cell lines have provided a way in which to predict chemosensitivity, using the GI_{50} value data, for a few hundred compound subset of the NCI data set.[9] And in a closely related study published after we completed our analysis in this report,[10] gene expression data on the 60 cancer cell lines were combined with NCI compound GI_{50} data and with a 27 000 feature database computed for the NCI compounds. Analyzing this combined data, two quinone core substructure types were identified that correlated highly with specific genes selectively overexpressed in melanoma and leukemia cell line classes.

In the present study, we use microarray based gene expression data to establish a number of "functional" classes of the 60 cancer cell lines. These functional classes are then used in a series of 2-class supervised learning problems, using a subset of 1400 of the NCI compounds' GI_{50} values as the input to a clustering algorithm in the RadViz program. RadViz combines embedded clustering algorithms based on feature reduction with the output displayed using a high dimensional visualization. The visual output displays the compounds' classification ability on the 60 cell lines in our particular examples. At $p < 0.01$ significance, RadViz identifies two small compound subsets that accurately classify the cancer cell line classes: melanoma from non-melanoma and leukemia from non-leukemia. We then demonstrate that analytic classifiers validate the two small compound subsets we selected. When we examined the small compound subsets, we found them to both be significantly enriched in quinone compounds. Moreover, the two sets of quinone compounds fell into two distinct subtypes. We then describe the relationship between these two quinone compound subtypes and those in the Blower et al.[10] study and relate these findings to the quinone enzymology literature.

METHODS

We chose to carry out classifications of cell line classes based upon a public domain data set comprised of 1400 NCI compounds' GI_{50} value data tested on all of the 60 cancer cell lines.[11] First, missing records needed to be replaced. For the ~4% missing values found in the data set, we tried and compared two approaches to missing value replacement: (1) record average replacement and (2) multiple imputation using Schafer's NORM software.[12] For the compounds identified as significant against leukemia and against melanoma at $p < 0.01$ in the 2-class problems presented later in this report, we compared the compound lists against one another (data not shown). Using either record average or multiple imputation as the missing value replacement method for the starting data set, there was close agreement (always > 90%) between the NCI compound lists selected in identical 2-class problem classifications. Therefore, in the present study, we used the record average replacement method for all the data presented.

Clustering of cell lines was done with R-Project software using the hierarchical clustering algorithm with "average" linkage method and a dissimilarity matrix computed as 1 − the Pearson correlations of the gene expression data.

AnVil Corporation's RadViz software was used for feature reduction and initial classification of the cell lines based on compound GI_{50} data.[11] RadViz is a visualization and classification tool that uses a spring analogy for placement of data points and incorporates machine learning feature reduction techniques as selectable algorithms.[13−15] The "force" that any feature exerts on a sample point is determined by Hooke's law: $f = kd$. The spring constant, k, ranging from 0.0 to1.0 is the value of the feature for that sample, and d is the distance between the sample point and the perimeter point on the RadViz circle assigned to that feature. The placement of a sample point is determined by the point where the total force determined vectorially from all features is 0. The default arrangement is to have all features equally spaced around the perimeter of the circle, but the feature reduction and class discrimination algorithms arrange the features unevenly in order to increase the separation of different classes of sample points. The feature reduction technique used in all figures in the present work is based on the t-statistic with Bonferroni correction for multiple tests. The circle is divided into n equal sectors or "pie slices," one for each class. Features assigned to each class are spaced evenly within the sector for that class, counterclockwise in order of significance (as determined by the t-statistic, comparing samples in the class with all other samples). As an example, for a 3-class problem, features are assigned to class 1 based on the sample's t-statistic, comparing class 1 samples with class 2 and 3 samples combined. Class 2 features are assigned based on the t-statistic comparing class 2 values with class 1 and 3 combined values, and class 3 features are assigned based on the t-statistic comparing class 3 values with class 1 and class 2 combined. Occasionally, when large portions of the perimeter of the circle have no features assigned to them, the data points would all cluster on one side of the circle, pulled by the unbalanced force of the features present in other sectors. In this case, a variation of the spring force calculation is used, where the features present are effectively divided into qualitatively different forces comprised of high and low k value classes. This is done via requiring k to range from −1.0 to 1.0. The net effect is to make some of the features pull (high or $+k$ values) and others "push" (low or $-k$ values) the points to spread them absolutely into the display space, but maintaining the relative point separations.

It should be stated that one can simply do feature reduction by choosing the top features by t-statistic significance and then apply those features to a standard classification algorithm. The t-statistic significance is a standard method for feature reduction in machine learning approaches, independently of RadViz. The top significance chemicals selected with the t-statistic are the same as those selected by RadViz. RadViz has this machine learning feature embedded in it and is responsible for the selections carried out here. The advantage of RadViz is that one immediately sees a "visual" clustering of the results of the t-statistic selection. Generally, the amount of visual class separation correlates to the accuracy of any classifier built from the reduced features. The additional advantage to this visualization is that subclusters, outliers, and misclassified points can quickly be seen in the graphical layout. One of the standard techniques to visualize clusters or class labels is to perform a Principle

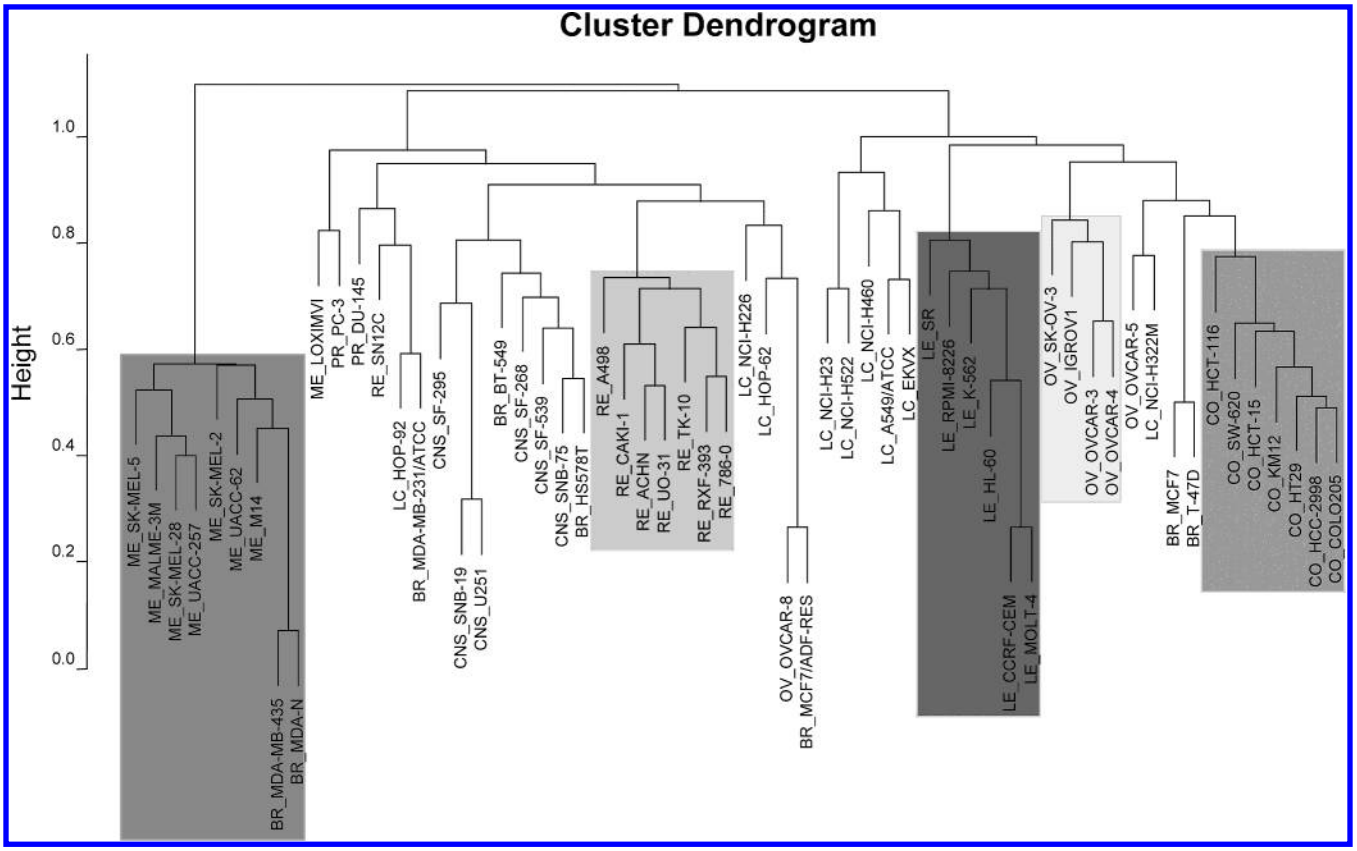

Figure 1. Cancer cell line functional class definition using a hierarchical clustering (1-Pearson coefficient) dendrogram for 60 cancer cell lines based upon gene expression data. Five well defined clusters are shown highlighted. We treat the highlighted cell line clusters as the truth for the purpose of carrying out studies to identify which chemical compounds are highly significant in their classifying ability.

Component Analysis and show the points in a 2d or 3d scatter plot using the first few Principle Components as axes. Often this display shows clear class separation, but the most important features contributing to the PCA are not easily seen. RadViz is a "visual" classifier that can help one understand important features and how many features are related.

The selected features were validated using several classifiers from Weka 3.1.9 (Waikato Environment for Knowledge Analysis, University of Waikato, New Zealand). The classifiers used were IB1 (nearest neighbor), IB3 (three nearest neighbor), logistic regression, Naïve Bayes Classifier, support vector machine, and neural network with back-propagation.

Both ChemOffice 7.0 (CambridgeSoft Corp.) and the NCI website were used to identify compound structures via their NSC numbers and substructure searching to identify quinone compounds in the larger data set were carried out using ChemFinder Ultra 7.0 (CambridgeSoft).

RESULTS AND DISCUSSION

Identifying Functional Cancer Cell Line Classes using Gene Expression Data. Initially, we decided to identify cancer cell line classes that we could use in a subsequent supervised learning approach to stringently select compound subsets capable of classifying the individual cancer cell classes. In implementing this approach, we used the 1376 gene expression values determined for the 60 NCI cancer cell lines by the Whitehead group.[11] These data are found in the T-matrix, comprised of 60 cell lines (representing nine cancer tissue types) × 1376 genes, with the gene expression value determined for each gene from each of the 60 cell lines. For the purpose of identifying cancer cell classes, we chose to examine gene expression based clustering. In Figure 1, we present a hierarchical clustering dendrogram using the 1-Pearson distances calculated from the T-Matrix. There are five well defined clusters, shown highlighted, that were observed. Four of the clusters in Figure 1 (renal, leukemia, ovarian, and colorectal from second left to right) are comprised entirely of cell line members originating from that one tissue tumor type. These represent pure cell line classes. In only the melanoma class instance does the class contain two members of another clinical tumor type, two breast cancer cell lines—MDA-MB-435 and MDA-N. The two breast cancer cell lines seem to be related to melanoma cell lines via a neuroendocrine origin as has already been observed and remarked upon.[11] Since these two breast cancer cell lines possess gene expression profiles that closely resembled the melanoma cell cluster, they behave "functionally" as melanoma cells and are treated as members of that class in the present report. In the supervised learning studies that follow, we treat these functional clusters as the truth—for the purpose of carrying out studies to identify which chemical compounds' assayed activities are highly significant in their ability to carry out cell classification. The remaining cell lines in the Figure 1 dendrogram, those not found in any of the five functional classes, are defined as being in a sixth class—the non-melanoma, leukemia, renal, ovarian, colorectal class.

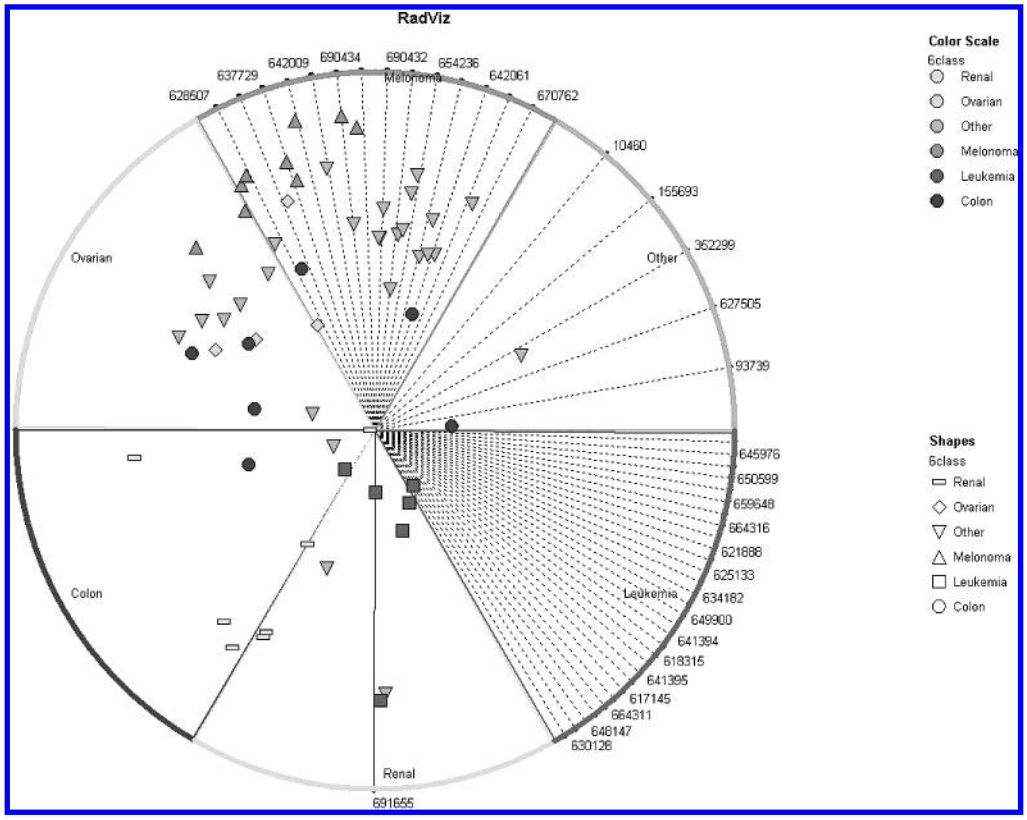

Figure 2. RadViz display of the 6-class classification of cancer cell lines— melanoma, leukemia, ovarian, renal, colorectal, non-(melanoma, leukemia, ovarian, renal, colorectal) where $p < 0.01$ is the criterion for selecting those compounds capable of acting as classifiers. Only four groups of compounds (for four classes—the other two classes produced no compounds significant at this level) were identified to be significant, and the class discrimination algorithm produced the layout shown. The four compound groups are seen arrayed within their labeled class sectors (their NSC numbers are given) along the circumference of the circle. The cell lines are coded by the symbols representing their tumor class type.

6-Class Cancer Cell Classification. High class number classification problems are difficult to implement with high accuracies where the data are not clearly separatable into distinct classes. This fact is well illustrated in Figure 2. Here a 6-class [melanoma, leukemia, ovarian, renal, colorectal, non-(melanoma/leukemia/ovarian/renal/colorectal] classification was carried out using RadViz, which combines an analytic class discrimination layout algorithm, employing feature reduction, with a high dimensional visualization resulting from the algorithm's output.[13−15] Starting with the 1400 compounds' GI_{50} data, those compounds were selected that were effective in carrying out the classification at the $p < 0.01$ (Bonferroni corrected t-statistic) significance level. Clearly, the selected NCI compounds were not able to separate the 60 cell lines into the six classes we identified in Figure 1. The six shape coded cell line classes are not clustered into their appropriate sector shaped class regions of the RadViz space. In addition, there is significant cell line class/cluster overlap. Only four cell line classes had compounds that satisfied the significance criterion for their classification. And within the four classes, the renal cell line class had only one compound that met the criterion for its classification.

3-Class Cancer Cell Classification and Validation of Results. These results clearly illustrate the difficulty of classification when there are more than a few classes. For this reason, we have chosen to implement a series of classifications that contain fewer classes, namely the 2-class and 3-class problems. Next, we carried out a 3-class problem, comprised of the classes melanoma, leukemia, non-melanoma/leukemia and two separate 2-class problems—melanoma, non-melanoma and leukemia, non-leukemia. Then, we examined the compounds selected at the $p < 0.01$ criterion. At this significance level, we used the class discrimination layout algorithm features of RadViz to select compounds and array them to maximize the cluster separation in the figures that follow. In Figure 3, we present the results of the 3-class problem for melanoma, leukemia, and non-melanoma/leukemia classes. These results show that the leukemia class is clearly separated from the non-melanoma/leukemia class (other), while there is only slight overlap of the melanoma class with the non-melanoma/leukemia class. Only one melanoma cell line resides at the border of the non-melanoma/leukemia class. In this respect, the 3-class problem is clearly superior to the 6-class problem in terms of producing clear class clustering. There are, respectively, 14 and 30 most effective compounds for distinguishing melanoma and leukemia from each other and from the non-melanoma/leukemia class in the 3-class problem.

To assess the accuracy of the 3-class problem approach in selecting compounds significant in their classification ability, we undertook the study presented in Table 1 below. In this study, the 3-class problem selected compound output from Figure 3 was tested for its prediction accuracy using four different analytic classification algorithms. Using the "hold-one-out" method, a total of 60 training-test cycles were carried out. In each cycle, training was carried out on 59 cell lines and then testing on the one cell line held out, for each of the 60 cancer cell lines. For each cycle, a different sample was held out. Overall, the four classifiers correctly

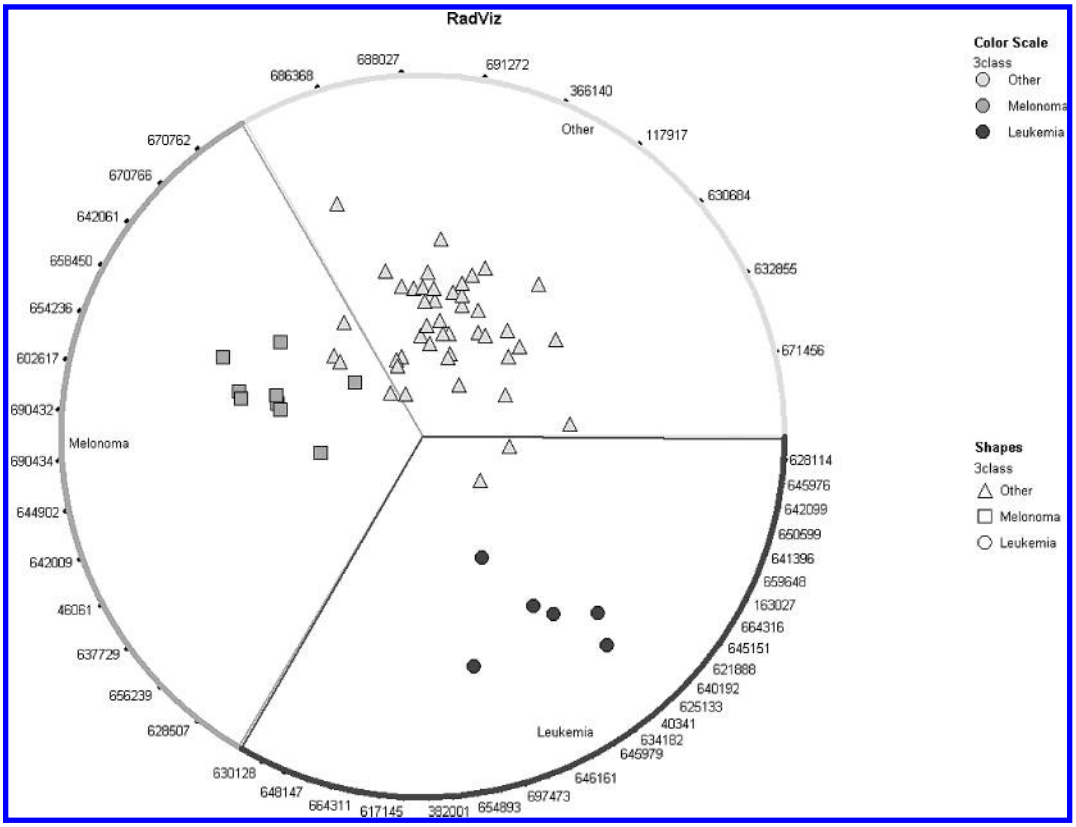

Figure 3. RadViz result for 3-class problem classification of melanoma, leukemia, and non(melanoma, leukemia) cancer cell types at the $p < 0.01$ criterion. The cell lines are symbol coded as described in the figure. There were 14 compounds effective against melanoma, and they are laid out (NSC numbers shown) on the melanoma sector of the circle. Similarly, for leukemia, 30 compounds were identified (not all NSC numbers are shown) and are laid out in that sector, and 8 compounds least effective against both melanoma and leukemia are laid out in the third sector.

Table 1. Validation of RadViz Results by Comparison to Analytical Algorithm Classification in the 3-Class Problem

classifier	total correct (60)	non (45)	Mel (9)	Lk (6)
IB1	59	44	9	6
IB3	60	45	9	6
Naïve Bayes	55	44	7	4
neural network	59	45	8	6

classified 99% (176 correct out of 178) of the non-melanoma/leukemia cases, 92% of the melanoma cases (33 correct out of 36), and 92% of the leukemia cases (30 correct out of 32), with a total rate of 97% correct. These results provide good validation for the accuracy of the compounds selected by the RadViz algorithm on the 3-class problem we presented in Figure 3.

2-Class Cancer Cell Classifications and Validation of Results. We next compared the 3-class problem result to implementing a series of 2-class problems also at $p < 0.01$ significance, for the melanoma, non-melanoma and leukemia, non-leukemia classifications. We present these respective 2-class problem results in Figures 4 and 5. In contrast to the 3-class problem result, the 2-class problem result in Figure 4 at the same significance, $p < 0.01$, for the melanoma, non-melanoma classes produced clear and accurate class separations of the 60 cancer cell lines. There was no overlap of the classes as occurred for the 3-class problem in Figure 3 between the melanoma and non-melanoma/leukemia classes. The same 14 compounds were selected in both the 3-class and 2-class problems as being most effective (lowest GI_{50} values) against melanoma class cells. In the 2-class leukemia, non-leukemia problem in Figure 5, it is clear that there is as great or greater a separation of classes than occurred in the 3-class problem in Figure 3. Again, the same compounds (in this case 30 compounds) were identified as most effective against the leukemia class cells in both the 2-class and 3-class problems. Conversely, only eight compounds were selected that were least effective against both melanoma and leukemia, but most effective against the non-melanoma/leukemia class. This was a very different result than the 3 and 50 least effective compounds we identified for the separate 2-class melanoma, non-melanoma and leukemia, non-leukemia problems in Figures 4 and 5.

Next, we wanted to validate the results we obtained from our RadViz methodology for compound selection in both 2-class problems. Therefore, we used the compounds selected in the 2-class melanoma, non-melanoma and leukemia, non-leukemia classification problems to determine classification accuracies from six analytic classification techniques using the same selected compounds' GI_{50} values as a classifier set. These comparison results are presented in Table 2 for the compounds selected at both $p < 0.01$ and $p < 0.001$. The table entries display the frequency of correct classification based upon a 60-fold repetition of the training-test process using the hold-out-one method. What is clear from the results is that the accuracies achieved by all of these analytic techniques was always above 90% using the compound subsets selected in Figures 4 and 5. As an example of the high accuracies obtained, using the 80 compounds selected for the leukemia, non-leukemia classification that satisfied the $p < 0.01$ criterion, the average accuracy achieved by the six analytic algorithms was 99.3%. This also corresponds

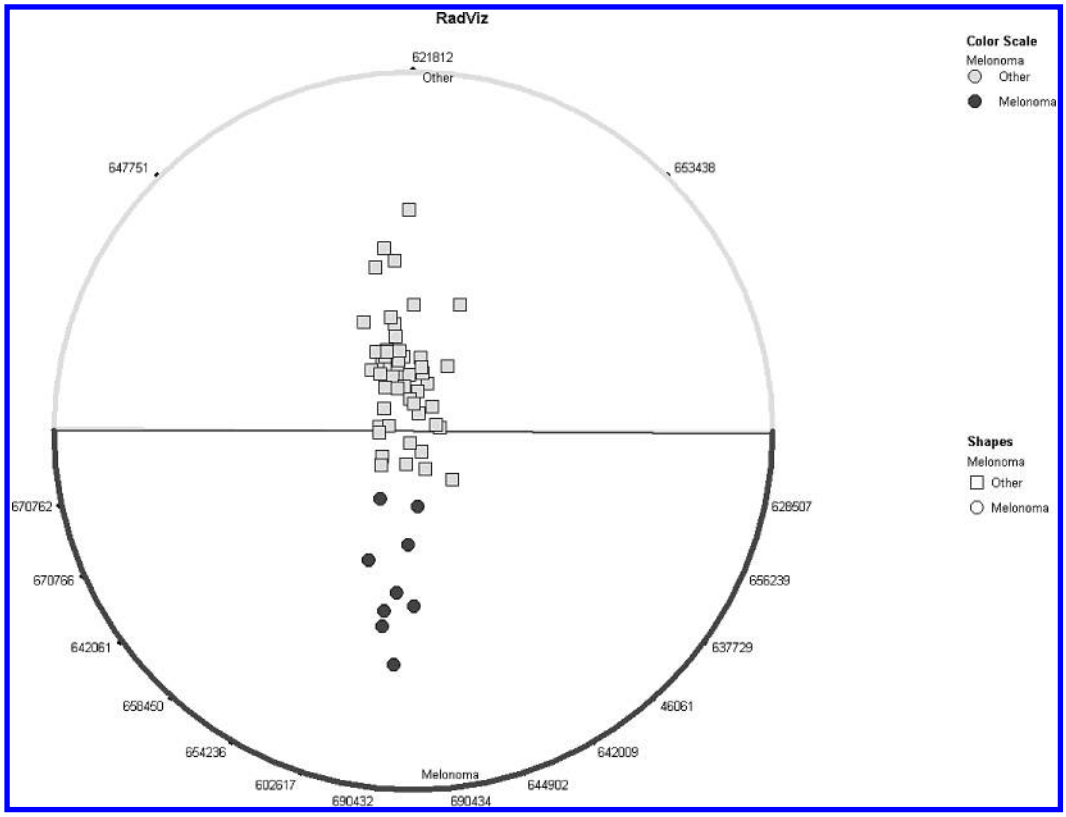

Figure 4. RadViz result for the 2-class melanoma, non-melanoma problem where $p < 0.01$ is the selection criterion. Cell lines are symbol coded as described in the figure. A total of 17 compounds were identified to be significant, and the class discrimination algorithm produced the sector layout shown. A total of 14 compounds most effective against melanoma are laid out in that class sector, while 3 compounds most effective against non-melanoma cells are laid out in that class sector.

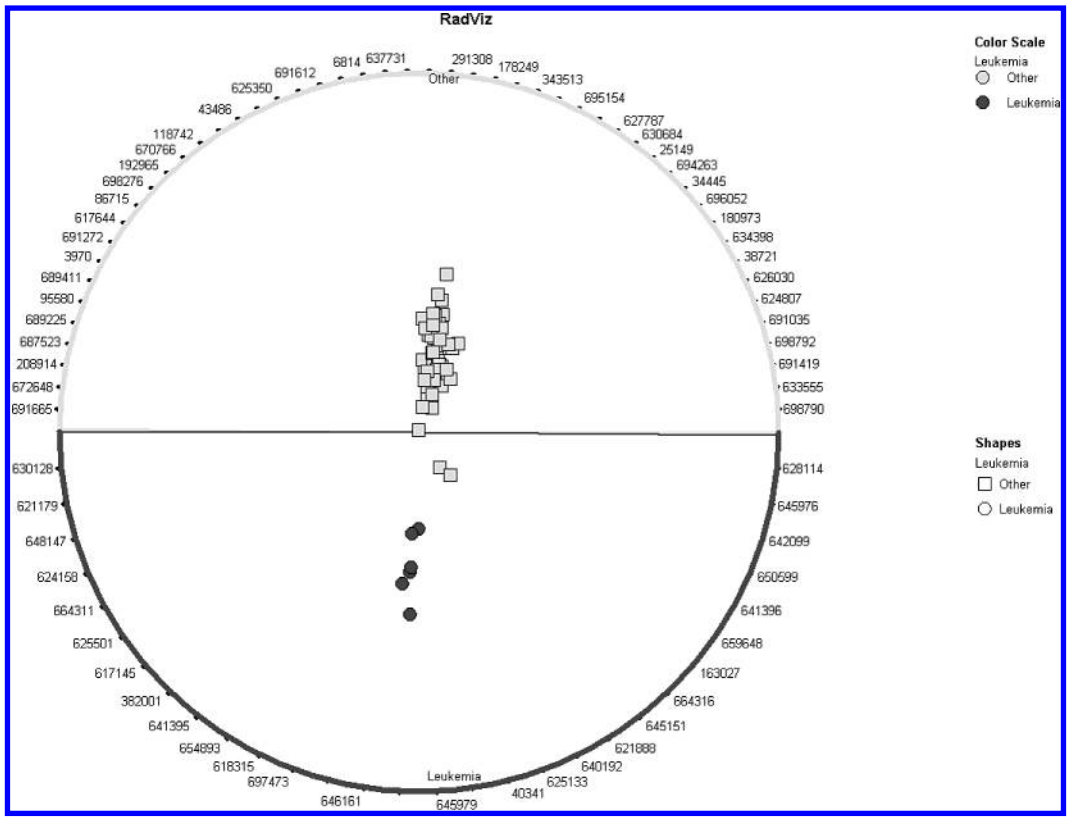

Figure 5. RadViz result for the 2-class leukemia, non-leukemia classification problem for the $p < 0.01$ criterion. The cells are symbol coded as described in the figure. Under this stringency condition, 80 compounds were selected as significant—the 30 most effective against leukemia were laid out in that class sector, and the 50 least effective were laid out in the non-leukemia class sector.

to only a 0.7% error rate. Based upon repetitively selecting 80 compounds randomly and using these to carry out the same classification, the average level of accuracy was calculated to be 95.7%, corresponding to a 4.3% error rate

Table 2. Comparison of Accuracies of Test Series of Classification Algorithms Using Compounds Selected with the RadViz Method

60-fold	number of chemicals	IB1	IB3	logistic regression	Naïve Bayes	support vector	neural network
			$p < 0.01$ Significance Compounds				
melanoma	17	59/60	60/60	54/60	59/60	59/60	59/60
leukemia	80	60/60	60/60	59/60	59/60	60/60	60/60
			$p < 0.001$ Significance Compounds				
melanoma	10	57/60	59/60	60/60	58/60	58/60	57/60
leukemia	50	60/60	60/60	59/60	58/60	60/60	60/60

(data not shown). Therefore, using the RadViz selected compounds represented a greater than 6-fold lowered level of error compared to the randomly selected compounds, thus validating our selection methodology.

The apparently high level of accuracy, 95.7%, is counterintuitive and is due to the fact that any 80 compounds randomly selected from the 1400 will always include a small number of the significant compounds. These significant compounds are sufficient to cause the 95.7% accuracy level to be achieved. Thought of in another way, it should be noted that in the case of classification of melanoma vs non-melanoma, if all samples are simply classified as non-melanoma, good guessing, the accuracy is 85% since 51 of the 60 cell lines will be correctly classified. Similarly, classifying all samples in the leukemia classification problem as non-leukemia results in 90% accuracy (54 correct out of the 60). Thus, only results with higher than 85% or 90% accuracy for the two class problems show reasonable discrimination of the classes. In cases such as this involving highly unbalanced numbers of class members, the error rate is a better measure of performance. For the leukemia, non-leukemia classification, using the 80 compounds that satisfied the $p < 0.01$ criterion, there were a total of only two errors by the six analytic algorithms. Based upon repetitively selecting 80 compounds randomly and using these to carry out the same classification, there was an average of 12.9 errors per random set across the same six algorithms (with a range of 7−19 errors). In addition, each of the random sets contained some of the compounds from the classification set (from 2 to 6 compounds).

The number of errors in a random 80 compound set is negatively correlated with the number of compounds in common with the 80 compound classification set as well as the rank of the compounds by t-statistic significance. To illustrate this, we assigned each compound in the 80 compound classification set a score based on significance, with the most significant compound given a score of 80, and the 80th compound given a score of 1. For each of the random sets we computed the sum of the scores of the compounds in common with the classification set. These ranged from 95 to 299, with higher values corresponding to more compounds in common and more highly significant compounds in common. The correlation between the number of errors for a random set and its sum of scores was highly negative, −0.79. Thus the accuracy was highly predictable based on the compounds in common with the 80 compound classification set.

We also carried out 2-class classifications for the melanoma, non-melanoma and leukemia, non-leukemia problems at the more stringent $p < 0.001$ selection criterion (data not shown). From these analyses, we identified two compound subsets, comprised of 8 and 20 compounds, respectively, that were most effective against melanoma and leukemia. These represent smaller compound subsets than those significant at the less stringent $p < 0.01$ level, where 14 and 30 compounds were identified, respectively. We also applied this stringent criterion, $p < 0.001$, to 2-class problems from a number of the other cancer cell line classes identified in Figure 1. In most of these 2-class problems, we found that few to no compounds met this criterion.

To provide insight into the RadViz displays of class separations we observed here, we have taken the output from the 2-class leukemia, non-leukemia problem for $p < 0.001$ selection stringency and displayed it in what we call a "multiline-viz" plot in Figure 6. The corresponding RadViz figure was not displayed here, but two clearly separated class clusters were observed similar to the Figure 5, $p < 0.01$, result. The RadViz result yielded 50 compounds significant in the 2-class classification. In the "multiline-viz" plot in Figure 6, the GI_{50} values ± SD of these two compound classes are clearly seen as distinct on the *Y* axis. The 50 compounds' average GI_{50} values (for all 60 cell lines) and ± 1 SD limits are displayed for the non-leukemia class in the left panel and for the leukemia class in the right-hand panel. There is a clear symmetry in these two compound classes relative to the mean values for all compounds (horizontal reference line). There were 20 compounds most effective against leukemia and least effective against non-leukemia cells. In a RadViz plot the result is to pull the points (cell lines) in this class toward the most effective compounds' axes in that class sector. The other 30 compounds were most effective against the non-leukemia cells—pulling this class toward these compounds' axes in that class sector—and least effective against leukemia cells. Those that are high on the left and low on the right in Figure 6 will pull more strongly on the non-leukemia cell lines in RadViz, while those that are low on the left and high on the right will pull more strongly on the leukemia cell lines.

Compounds Most Effective against Melanoma and Leukemia Do not Fall into Any of the 6 NCI MOA Classes. The NCI data set lists 92 compounds of known Mechanism Of Action (MOA): alkylating agents, antimitotic agents, topoisomerase I inhibitors, topoisomerase II inhibitors, RNA/DNA antimetabolites, DNA antimetabolites.[2] We decided to determine whether the most effective compounds against melanoma and leukemia identified in the 2-class problems we presented in Figures 4 and 5 fall into any one of these six MOA compound classes. Therefore, in Figure 7, we carried out a 2-class classification in a somewhat different way, using the 14 melanoma and the 30 leukemia compounds, effective at $p < 0.01$, from Figures 4 and 5 as the two classes. The RadViz class discrimination algorithm was used to lay out the 60 cell lines based upon the 2-class compounds' GI_{50} values. As was expected, this produced the two well separated vertical rectangular symbol classes—melanoma at lower left and leukemia at upper right. Then the 92 known MOA compounds were simply placed upon this optimized RadViz display using a six symbol code to represent the six MOA classes. Since none of the MOA compound classes overlap with any of the compounds in either the melanoma or the leukemia class, there is no clear indication of molecular mechanism of action for the melanoma or leukemia compound classes we have selected. Of

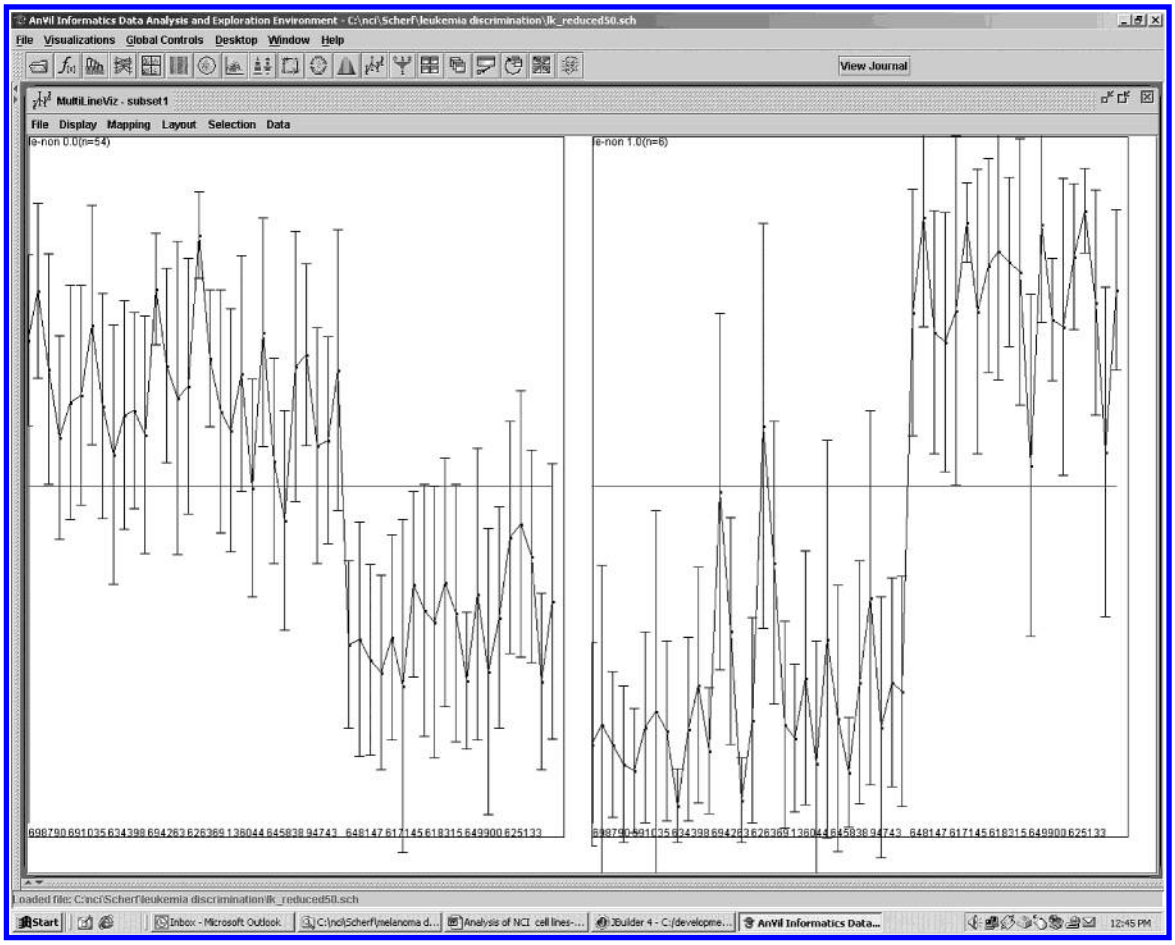

Figure 6. MultiLineViz plot of the average GI_{50} values $\pm$ 1 SD (for all 60 cell lines) for the top 50 compounds identified in the 2-class leukemia, non-leukemia problem in the previous figure. The left panel displays values for the non-leukemia class, while the right panel displays those for the leukemia class.

the six MOA classes, five appeared to cluster evenly about the central region of the display. This indicates that they are equally effective against all cell lines. Only the tubulin MOA class (horizontal rectangles) had all compound members clustering away from the center toward the melanoma compound class on the lower left. This suggests that overall this class is more effective against the melanoma cell class than against the leukemia cell class.

Quinone Compound Subtype Preferentially Effective against Melanoma. Next, we decided to examine the chemical identity of the compounds selected as effective at the $p < 0.01$ level in the separate melanoma, non-melanoma and leukemia, non-leukemia 2-class problems we presented in Figures 4 and 5. To summarize, for the 14 compounds selected as most effective against melanoma, 11 are p-quinones. Only three compounds do not contain p-quinone structures. They are NSC 654236, 46061, and 637729. Although the latter compound is an o-quinone, we do not include it in our listing. Of the 11 p-quinones, all 11 are internal ring quinone structures. We display these structures in Figure 8. These internal ring quinones possess either two neighbor aromatic five- or six-member fused rings, some of which are heteroatom containing, on either side of the quinone ring or an aromatic fused ring neighbor on one side and covalent non-H substitutions off the other side of the quinone. These substitutions all have electronegative atoms covalently bonded to either or both the o and m positions of the quinone ring, except for one compound which has an −OH substituent off the adjacent ring. In eight of the cases, the internal ring quinones are directly bonded to two electronegative atoms, either heteroatoms contained within the aromatic fused ring or as small covalent substituents. And in two more cases, the internal quinone ring is bonded to one electronegative atom within a neighboring fused ring.

After our analysis, but before this manuscript was prepared for publication, the work of Blower et al.[10] appeared. These authors carried out an analysis simultaneously correlating gene expression data for the 60 cancer cell lines with GI_{50} values for a set of 4463 compounds for which a 27 000 feature set had been calculated. They identified a subclass of compounds containing a benzothiophenedione core structure that were most highly correlated with the expression patterns of Rab7 and other melanoma specific genes. Carrying out query searches using the benzothiophenedione substructure, these authors identified 23 members of this substructure class. There is clearly some overlap between the internal quinone subtype we have defined in the present study and the benzothiophenedione core structure members identified by Blower et al..[10] Out of the 11 internal quinone compounds we identified, three are of the benzothiophenedione core structure class. However, these three are not among the most effective compounds against melanoma, being ranked 6th, 7th, and 10th in effectiveness as shown in Figure 8. Our top five most effective compounds are distinct from the core substructures that Blower et al.[10] identified as

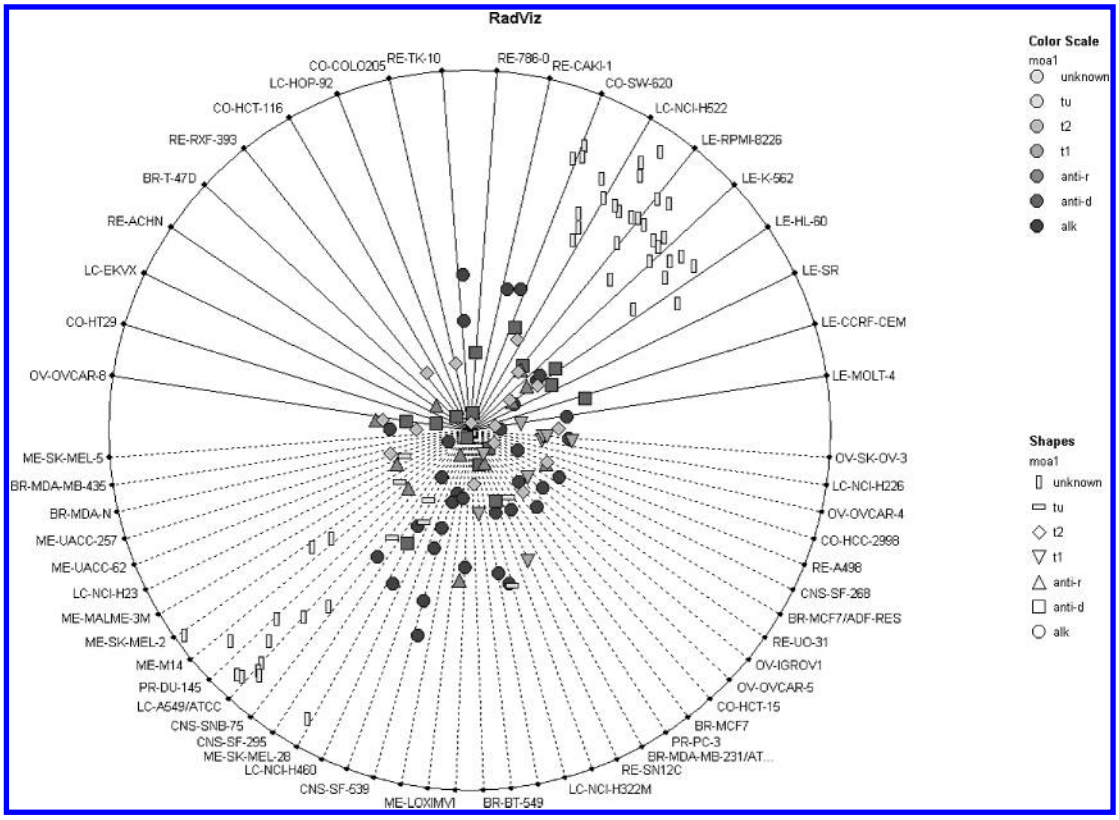

Figure 7. RadViz class discrimination layout of the 60 cell lines, based upon their GI_{50} values, used to classify in the 2-class problem, where the classes are the 14 melanoma compounds and the 30 leukemia compounds, both selected at $p < 0.01$ in Figures 4 and 5. There are 20 cell lines in the top leukemia class sector and 30 cell lines in the bottom melanoma class sector. The two vertical rectangle symbol clusters represent melanoma compounds (at lower left) and leukemia compounds (at upper right). Then the 92 compounds of six known MOA types were passively added into the display using the six symbol code shown.

most effective against either melanoma (benzothiophenedione) or leukemia (indolonaphthoquinone).

Quinone Compound Subtype Preferentially Effective against Leukemia. There were 30 compounds selected as most effective against leukemia in the leukemia, non-leukemia 2-class problem, of which eight are structures containing p-quinones. The 22 other structures did not contain p-quinones. These are NSC 630128, 621179, 624158, 664311, 625501, 617145, 382001, 654893, 697473, 646161, 649900, 645979, 634183, 625133, 621888, 645151, 664316, 163027, 650599, 642099, 645976, and 628114. In contrast to the internal ring quinones comprising the melanoma class, 6 out of the 8 leukemia p-quinones were external ring quinones. We display just these structures in Figure 9 A. In contrast to the internal ring quinones, these external ring quinones had only one aromatic fused ring neighbor which had no ring heteroatoms in all cases. Also different, the quinone was itself at the periphery of the molecule and had no non-H substituents off the exterior side of the ring at either o or m positions. Thus, the "external" and "internal" quinone rings should possess different electron densities and redox potentials for the quinoid oxygens. Besides redox potentials, other possible subtype differences may exist such as solubility, steric differences relative to metabolic enzyme active sites, differential cellular adsorption, etc.

Again from the work of Blower et al.[10] these investigators identified a subclass of compounds, comprised of an indolonaphthoquinone core structure. These compounds were most highly correlated with the expression patterns of LCP1, lymphocyte cytosolic protein 1 (L-plastin located on chromosome 13), HS1, a hematopoietic lineage specific gene, and other leukemia specific genes. Carrying out query searches using this substructure, these authors identified 20 members of this core substructure class. In contrast to the melanoma quinone subtype case we described above, in this situation there is more overlap between the external quinone subtype we have defined in the present study and the indolonaphthoquinone core structure members identified by Blower et al..[10] Out of the six external quinone compounds we identified, four are of the indolonaphthoquinone core structure class. They ranked 1st, 2nd, 4th, and 6th in effectiveness in our study as shown in Figure 9A. Therefore, in this case, there is significant overlap between the indolonaphthoquinone core structure class identified by Blower et al.[10] as most effective against leukemia and our external quinone subtype class. This overlap between Blower et al.[10] and our results is somewhat remarkable since we included no gene expression data in our analysis of the GI_{50} values, as did the Blower study. This suggests two things. The first is that there is sufficient information inherent in the GI_{50} values to carry out the basic core discovery presented here, without the need to include gene expression data in the analysis. The second is that the class discrimination layout algorithm of RadViz, used here to select and array compound's axes to maximize the cluster separation, is a highly effective data mining tool.

Uniqueness of the Two Quinone Subtypes to the Respective Cancer Cell Line Class. No compounds were found effective at the $p < 0.01$ criterion for cell classes other than melanoma and leukemia, except for the single com-

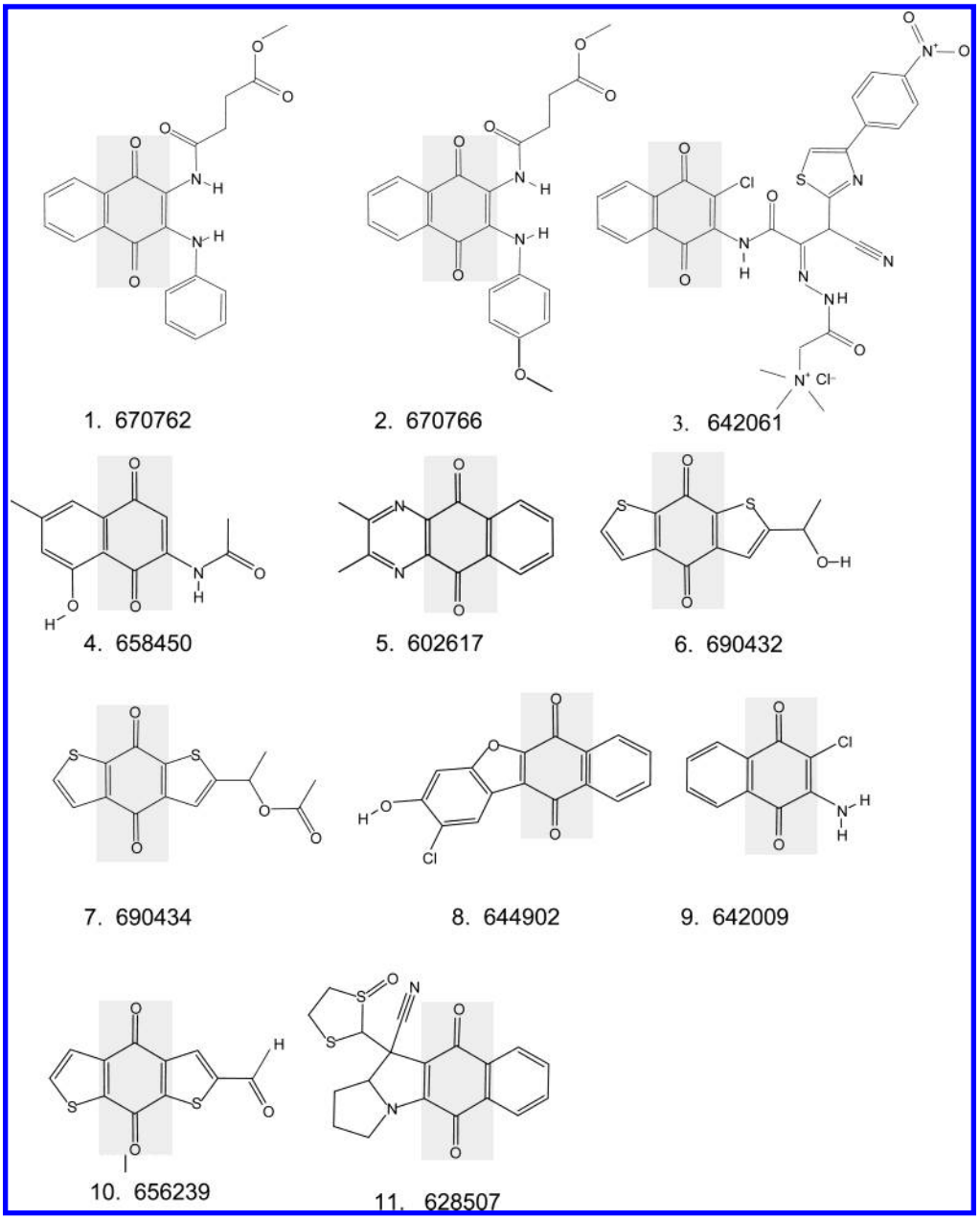

Figure 8. The 11 internal quinone subtype compounds most effective against melanoma in their rank order of effectiveness.

pound NSC 691655 effective against the renal cancer cell class (see Figure 2). Therefore, we decided to restrict our view of quinone subtype uniqueness entirely to the melanoma and leukemia cell line classes. To ascertain the uniqueness of the two quinone subsets found effective against melanoma and leukemia, we first determined the extent of occurrence of p-quinones of all types in our starting data set. To do this, we examined the entire data set via substructure searching using the ChemFinder 7.0 software. We identified 131 quinones in our starting data set of nearly 1400 compounds, representing 9.4% of all compounds. This percentage is larger than that of the ~0.8% of quinones (2051) identified within the 250 251 NCI compounds available in ChemFinder via substructure searching. Within the 131 quinones we identified, a large number are significantly different from both the internal and external quinone subtypes we defined above. As examples, 28 were in a class of quinones with a p-quinone fused ring neighbor in the reduced dihydroxy state. Another five were a class of quinones, where the quinone ring was part of a large covalent circular structure 15−20 bonded atoms in extent. The former quinone class would be expected to have very different redox properties from the internal and external p-quinone subtypes we have defined here. And the large size of the latter quinone class may prevent these from serving as good substrates for reducing enzymes in the cell.

Within the group of 131 quinones are the 10 internal quinones effective against melanoma at $p < 0.01$ and the six external quinones effective against leukemia at $p < 0.01$. However, also contained within the 131 quinones are nine quinones which fit the external quinones' definition but were not found in the list effective against leukemia at $p < 0.01$. Likewise, there were 31 quinones among the 131 which fit the internal quinone definition but were not found in the list effective against melanoma at $p < 0.01$. Therefore, the internal and external quinone subtypes we identified as effective against melanoma and leukemia, respectively, represent a significant fraction, 25% (10/41) of all the internal quinones and 40% (6/15) of all the external quinones in the data set.

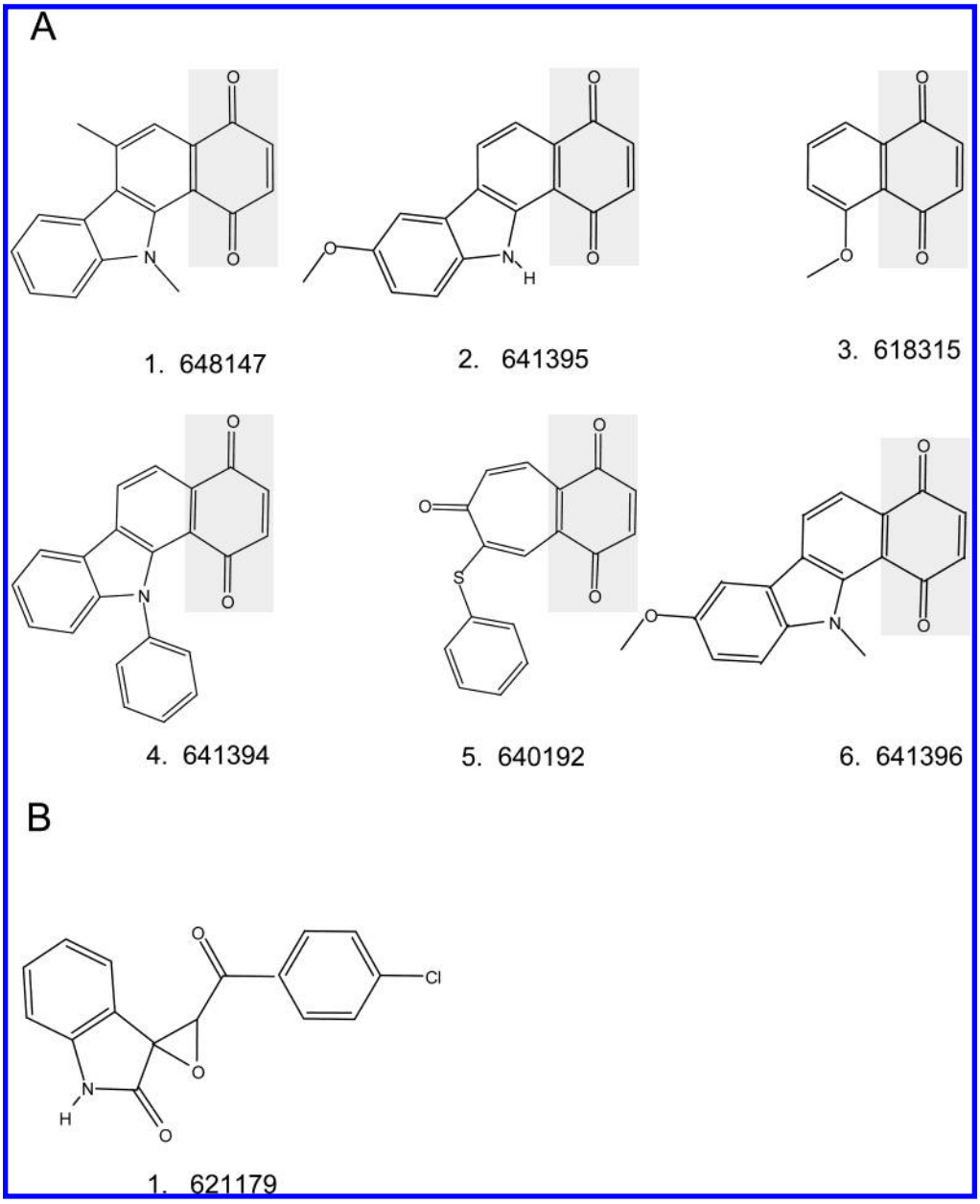

Figure 9. (A) The six external quinone subtype compounds most effective against leukemia in their rank order of effectiveness. (B) The single compound found effective against both melanoma and leukemia at $p < 0.05$ in Figure 10. It is not a quinone but an epoxide.

We next carried out a calculation, as a type of control, to exclude the possibility that individual compounds of the internal (melanoma) or external (leukemia) quinone subtype classes could cross over and be effective against the other cancer cell line class at a selection criterion slightly less stringent than the $p < 0.01$ condition that was used to select them. Therefore, the following 2-class problem classification was carried out at the $p < 0.01$ significance criterion. The two cell line classes used were as follows: combined melanoma and leukemia as the first and non-melanoma and non-leukemia as the second. No compounds were found effective against both melanoma and leukemia cell lines in the combined class. Therefore, we relaxed the 2-class significance criterion to $p < 0.05$ and obtained the result shown in Figure 10. Clearly, it was far easier to separate the leukemia cell lines than the melanoma cell lines in this combined class from the non-melanoma, non-leukemia class. Even at this relaxed criterion, only one compound was found significant against the combined melanoma and leukemia class. This compound, NSC 621179—presented in Figure 9B, was not a quinone but an epoxide and was one of the 30 compounds selected as effective against leukemia at $p < 0.01$ (second out of 30 in rank of effectiveness). Therefore, the quinone subtype specificity we have identified in this study maintains its tumor cell class selectivity, and no members of either class cross over to become effective against the other subtype's tumor cell line class, even under the lower significance criterion we tested here.

It is interesting to note in Figure 7 that a small number of the 92 MOA compounds in the topoisomerase II class are dihydroxyanthraquinones. Examples of these structures are the compounds—mitoxantrone, daunorubicin, doxorubicin, and deoxydoxorubucin. While the dihydroxyanthraquinones represent a structural class that contains the central quinone ring, the adjacent dihydroxy fused ring structure was not a feature of the 11 member internal quinone structure class that we defined as specifically effective against melanoma cells. In only one case out of the 11, NSC 658450, was a

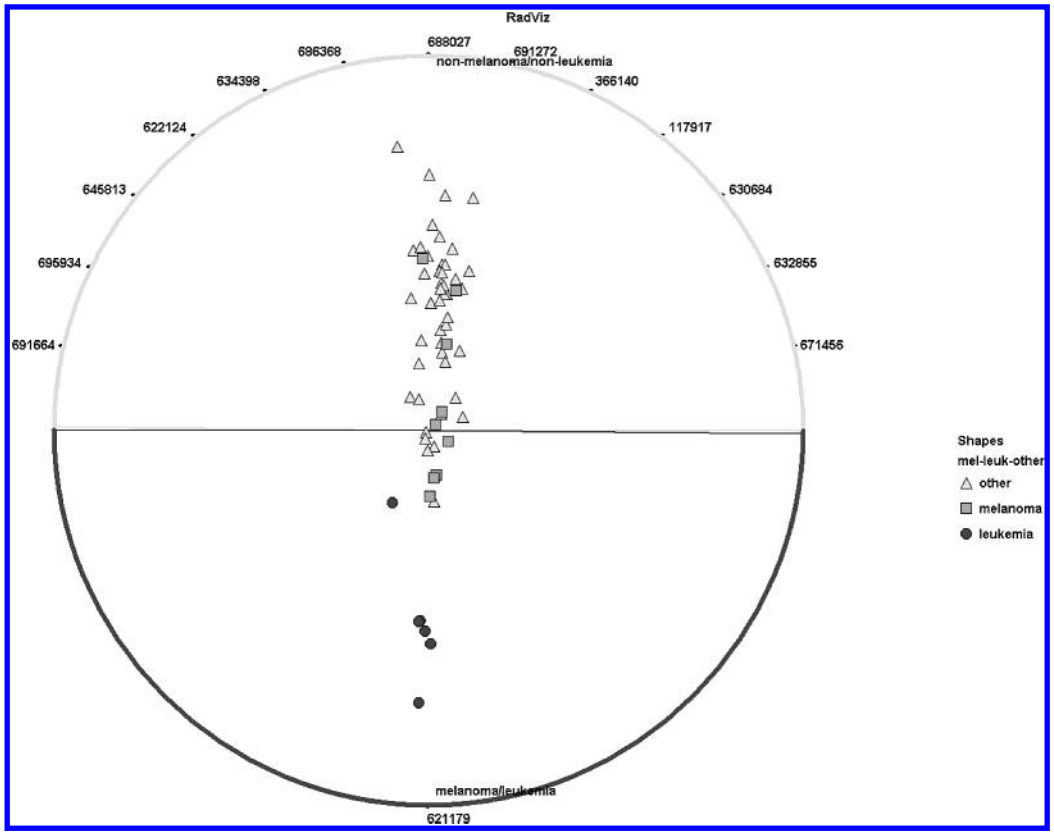

Figure 10. RadViz result for the 2-class problem where the first class is melanoma and leukemia and the second class is non-melanoma and non-leukemia and for which the selection criterion is $p < 0.05$. Cell lines are symbol coded as described in the figure.

single hydroxy group located on the adjacent fused ring. Within the 30 compound subset found most effective against leukemia, there was only one compound (NSC 659648—the 24th most effective out of 30) that contained a dihydroxy substituent ring fused to the quinone ring structure. Given this fact, it is interesting to note that in Figure 7, the topoisomerase II class of compounds overall resides closer to the leukemia cell line dimensions than the melanoma. When we examined the location of the four dihydroxyanthraquinone compounds—mitoxantrone, daunorubicin, doxorubicin, and deoxydoxorubicin, we found that doxorubicin and mitoxantrone were located near the center of the RadViz display indicating equal effectiveness against the leukemia and melanoma enriched cell lines. In contrast, deoxydoxorubicin and daunorubicin were located toward the leukemia cell line dimensions in the display. This indicates that within the topoisomerase II class compounds, deoxydoxorubicin and daunorubicin were among the most effective against leukemia as compared to melanoma cell lines. However, even deoxydoxorubicin and daunorubicin were located nowhere near the 30 compound cluster most effective against leukemia that contained the external quinone subtype members. Therefore, the two quinone subtypes we have defined represent compounds relatively distinct in their effect on cells from the 92 known MOA compounds.

Subclassification of Leukemia Cell Lines. The 60 cancer cell lines, representing nine major cancer types chosen by the NCI for the GI_{50} determination of different compounds, were, in many cases, derived from different cancer clinical subtypes. Thus, we asked the question whether we could subclassify either the melanoma or the leukemia cell lines into distinct clinical subclasses based upon using our two respective compound classes—the 14 compounds most effective against melanoma and the 30 most effective against leukemia. Therefore, we carried out a 3-class RadViz based classification for the melanotic melanoma, other melanoma, and non-melanoma cell classes at $p < 0.05$, using the most effective 14 compounds identified for the $p < 0.01$ selection criterion in Figure 4 as most effective against all melanomas. There was a small separation observed between the two members of the other melanoma class and the seven melanotic melanomas brought about by 2 and 12 of the compounds, respectively. However, there was also significant interspersion of the non-melanoma cell lines with the two melanoma classes. Therefore, we judged this subclassification to be a failure and have chosen not to present the figure.

We next carried out a 3-class RadViz based leukemia cell subclassification for the acute lymphoblastic leukemia (ALL), non-ALL leukemia (other), and non-leukemia cell classes at $p < 0.05$. To carry out the subclassification, we used the most effective 30 compounds identified for the $p < 0.01$ selection criterion in Figure 5 as most effective against all leukemias. This result, presented in Figure 11, was more successful than the previously discussed unsuccessful attempt to carry out a 3-class melanoma subclassification. Six of the 30 compounds were most effective against the ALL class; while 12 of the 30 compounds were most effective against the non-ALL leukemia. None of the 30 compounds were effective against non-leukemia cells at $p < 0.05$. The large group of non-leukemia cells are pulled to the left in this figure, in the absence of any selected compounds effective against non-leukemia cells, due to the use of zero-centered normalization in the RadViz display. This display feature helps to emphasize the intrinsic computed separation of the

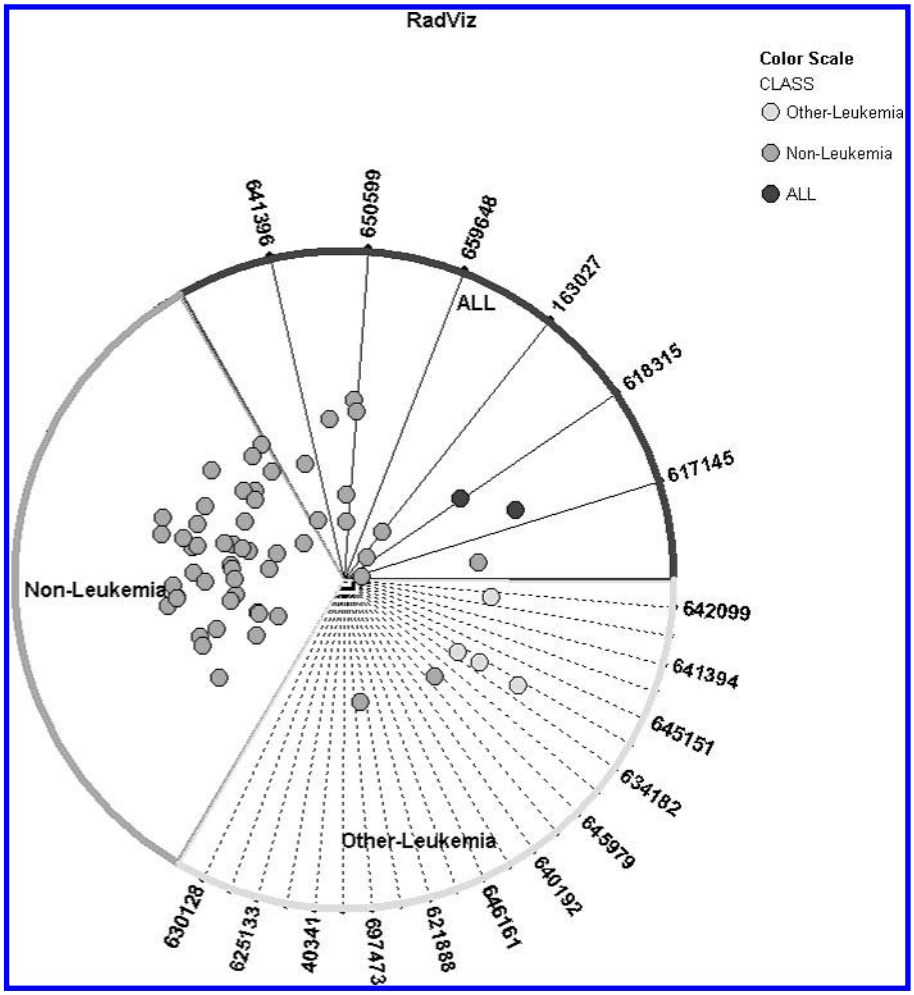

Figure 11. RadViz result for the 3-class problem classifying the following three classes: acute lymphoblastic leukemia (ALL), non-ALL leukemia (other-leukemia), and non-leukemia cell classes at $p < 0.05$. We used as input the 30 compounds identified in Figure 5 as most effective against all leukemias at the $p < 0.01$ selection criterion. Cell lines are symbol coded as described in the figure. The NSC numbers of the compounds selected to classify the classes are presented in the order of their ranking from most effective to least effective moving counterclockwise within each class sector.

three cell line clusters by using more of the available display area to present their relative separations. In Figure 11, it is clear that there is a separation of the two ALL cell lines (CCRF-CEM and MOLT-4) from the non-ALL leukemia subclass. These two ALL cell lines were also the most closely clustered leukemia cells in the Figure 1 gene expression based clustering dendrogram. This suggests the interesting possibility that the chemical identity of the compounds most effective against the two ALL cell lines are linked to the gene functions most responsible for closely clustering these two ALL cell lines in Figure 1. With the exception of two non-leukemia cell lines out of the 54 total, there is not any significant overlap of this non-leukemia class with the two leukemia subclasses. These two cell lines are as follows: CO-SW-620, a colon cancer cell line that lies between the ALL and non-ALL leukemia cell line clusters, and LC-NCF-H522, a lung cancer cell line that is immediately adjacent to the non-ALL leukemia (other) cell line cluster.

Next, we asked what relationship the six compounds effective against the ALL cell subclass had with the 12 compounds effective against the non-ALL leukemia cell subclass. To begin this comparison, we first noticed that there was nearly an inverse correlation between the rankings of the 30 compounds identified as most effective against leukemia in the leukemia vs non-leukemia classification (Figure 5 result) and the compound rankings observed as most effective against the two leukemia subclasses in Figure 11. To display this correlation in rankings, we present Figure 12. In this figure, the correlation of effectiveness rankings for the 18 compounds selected in the Figure 11 result for the ALL and non-ALL leukemia cell subclassifications (*Y*-axis) are plotted, respectively, in panels A and B against the 30 compounds selected in the Figure 5 result for all leukemia cell vs non-leukemia cell classification (*X*-axis). The 30 compounds are symbol coded according to whether they occur in the ALL leukemia, non-ALL leukemia, or neither classes. Filled symbols correspond to compounds in a given class list that belong to the external quinone subtype class we have identified as effective against leukemia. In panel C, we present symbols on the $Y = 0$ line that all belong to the third class (neither) of compounds that were not effective against either ALL leukemia or non-ALL leukemia. We have presented the data in this way in order to emphasize two features. The first is that a significant majority of the most effective of the 30 compounds identified in the Figure 5 leukemia vs non-leukemia cell line classification (*X*-axis - #2−6, 8−10) were not effective in discriminating ALL from non-ALL leukemia cell lines. Another way of viewing this fact is that predominantly the lowest ranking of the 30 compounds (least effective) used to classify all leukemias

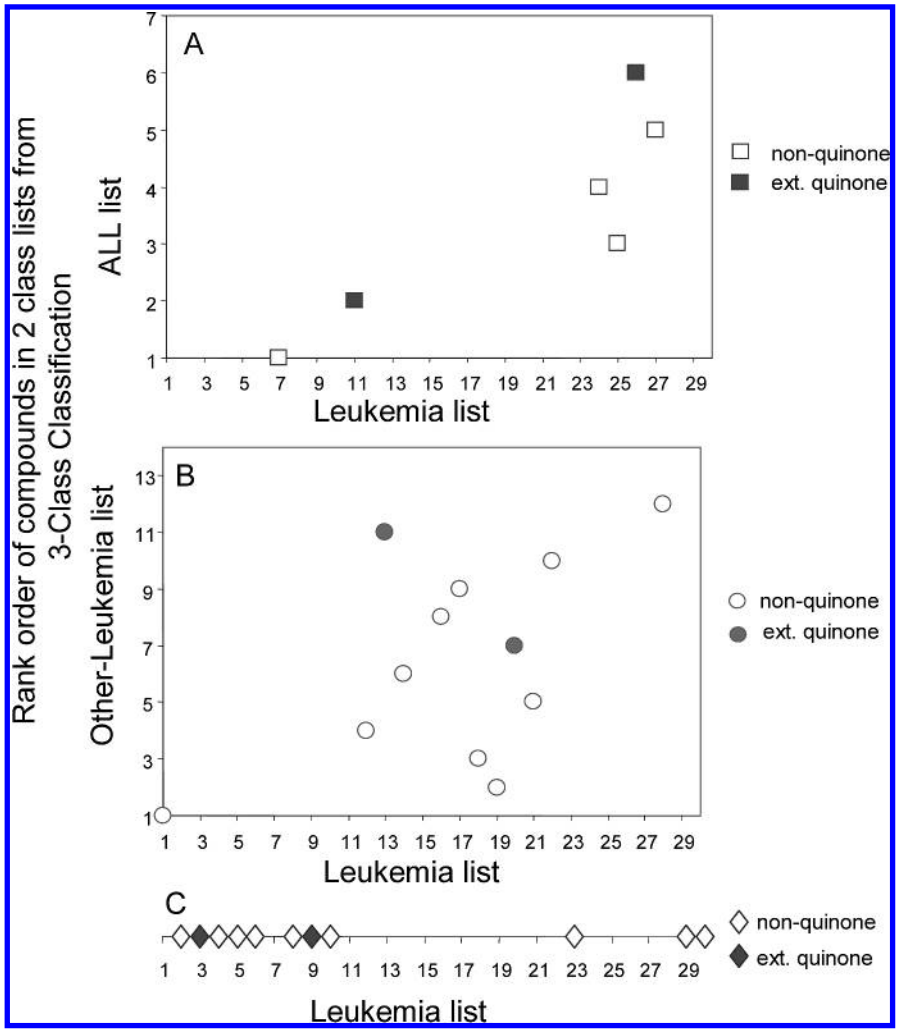

Figure 12. Correlation of effectiveness rankings for the 18 compounds selected in Figure 11 result for the ALL and non-ALL leukemia cell subclassifications (*Y*-axis) with the 30 compounds selected in Figure 5 result for all leukemia cell vs non-leukemia cell classification (*X*-axis). The 30 compounds are symbol coded according to whether they occur in the ALL leukemia subclass (panel A), non-ALL leukemia subclass (panel B) neither subclass ($Y = 0$ line C) that were not effective against either ALL leukemia or non-ALL leukemia. Filled symbols correspond to compounds that belong to the external quinone subtype class we have identified as effective against leukemia.

from non-leukemias in Figure 5 are the most effective at subclassifying ALL from non-ALL leukemias. This observation is perhaps not surprising, since compounds most effective at classifying all leukemia cell lines together with the tightest possible clustering vs non-leukemia cell lines (Figure 5) would not be expected to be good at maximizing the separation between ALL and non-ALL leukemia subclasses. The second point obvious in Figure 12 is that the distribution of the six external quinone subtype compounds we identified in the 30 compounds are evenly distributed, with two to each of the three categories/panels in Figure 12. Therefore, we can conclude that the external quinone subtype we have defined is not significantly biased in its effectiveness against either ALL or non-ALL leukemias.

NAD(P)H: Quinone Oxidoreductase 1-Quinone Substrates and Leukemias. Different redox potentials and enzymatic reactivities are likely to be the key to how these quinone subtypes differentially affect melanoma and leukemia cells. A strong candidate enzyme for this differential reactivity is NAD(P)H:quinone oxidoreductase 1 (QRI, NQO1, also DT-diaphorase; EC 1.6.99.2). This enzyme is a homodimeric flavoprotein, mainly cytosolic in location, that is expressed in normal cells and at high levels in many types of tumors.[16] Two electron reduction of a variety of substrates is catalyzed by this enzyme with the requirement for a pyridine nucleotide cofactor. The most efficient substrates are quinones.[17] The enzyme has been crystallized, and the X-ray structures of the apoenzyme at 1.7-A resolution and its complex with substrate duroquinone (2,5A) are known.[18,19] NAD(P)H:quinone oxidoreductase 1 is a chemoprotective enzyme that protects cells from oxidative challenge and has the ability to carry out two electron reduction of quinones without generation of reactive oxygen radicals. Antitumor quinones, of the type we have identified above in the NCI data set, may be bioactivated by this enzyme to forms that are cytotoxic.[17] This catalytic property makes this enzyme an excellent target for enzyme-directed drug development.[19] Reductive activation is particularly well-suited for treatment of hypoxic tumors, where the bioreduction of the chemical agent to hydroquinone cannot be reversed by endogenous oxygen.[20]

Interestingly, there are a number of reports that correlate this enzyme with the following increased risks. Adult leukemia was correlated with a specific NAD(P)H:quinone oxidoreductase 1 allele that possesses diminished activity.[21] Lack of a functional NAD(P)H:quinone oxidoreductase 1 allele was correlated with pediatric leukemias that have MLL

fusions.[22] And NAD(P)H:quinone oxidoreductase 1 genetic polymorphisms were associated with therapy-related leukemia/myelodysplastic syndrome and de novo acute myeloid leukemia.[23] These reports, associating leukemias with particular aspects of NAD(P)H:quinone oxidoreductase 1, suggest a correlation pointing to the enzyme as likely being a significant factor in why the external quinone subtypes, acting as particularly potent and effective substrates, exhibit their differential selectivity toward leukemias. It may be that the internal quinone subtypes, effective against the melanoma cell class but not leukemia, are less effective substrates for the NAD(P)H:quinone oxidoreductase 1 expressed in the leukemia cells but are highly effective as substrates for the melanoma enzyme, which may be a different isotype. We believe that only through experiments or calculations to determination the redox potentials of the different quinone compounds, influenced by the type and distribution of substituent groups, will the exact nature of the compound selectivity exhibited by the subtypes in our study be known.

In the report by Blower et al.,[10] these investigators examined the melanoma and leukemia cell classes and calculated their highest gene expression change correlations with their compound set. From this analysis they produced the two respective quinone core structures, that we discussed previously, as most highly correlated with the growth inhibiting properties of the respective cell classes. From these data, they made interesting observations concerning the potential cellular consequences of compounds from the melanoma core structure members. They were able to do this because they had the broad gene functions with which to postulate mechanisms. Since our data contains no gene correlation information, we decided to briefly recapitulate their discussion here, because it has relevance to some members of our melanoma quinone subtype class that are shared in common in the two studies.

In the case of the melanoma cell class, they identified the Rab7 gene as most highly correlated with compounds that were query selected as members of the benzodithiophenedione core structure class. The Rab7 gene is a member of the GTP binding protein family that helps provide specificity in the docking of cellular transport vesicles. It also appears that the Rab7 protein is a key regulator of aggregation and fusion of late endocytic lysosomes.[24] A number of other genes whose expression levels highly correlate with the same compounds express proteins involved in other lysosomal functions, particularly vacuolar proton pump activity. These investigators expressed the opinion that a link exists between the quinone oxidation potential, the proton pump, and the electron transport chain. They suggested the possibility that benzodithiophenedione compounds may act directly as surrogate oxidizing agents, effectively competing with ubiquinone in the electron transport chain and thereby disrupting an essential cellular redox process. They believe that effectiveness in this type of mechanism would be, as we have already argued above, based upon the redox potential of the individual quinone core containing compounds, as influenced by individual substituent type and distribution. Such a mechanism may indeed be the primary cause of cytotoxicity for these benzodithiophenediones as well as certain of the members of our melanoma specific quinone subtype. If so, then this provides a different way, in addition to being a potential substrate for NAD(P)H:quinone oxidoreductase 1, by which these quinones may produce cytotoxic cellular perturbations.

CONCLUSIONS

Functional clusters of the 60 NCI cancer cell lines were established based upon hierarchical clustering dendrograms using the 1-Pearson metric calculated from gene expression data for the 60 NCI cancer cell lines. We identified five nearly pure compound classes where the cell lines were comprised of the pure clinical tumor types. A sixth class represented all cell lines not in the five cell line classes. Using these six cell line classes, we unsuccessfully carried out a 6-class RadViz classification of the cell lines starting with the 1400 compounds GI_{50} values. Classification of data with more than a few classes is not usually successful. Therefore, we next successfully carried out a series of simpler problems using RadViz: a 3-class problem and a series of 2-class problems at high stringencies, $p < 0.01$ and $p < 0.001$, identifying small subsets of compounds that could carry out these classifications.

The results obtained from the 3-class and 2-class problem approach using the RadViz methodology was validated using the compound subsets selected. The compound subsets were used to test classification accuracies obtained with the following analytic algorithms: IB1, IB3, logistic regression, Naïve Bayes, Support Vector, and Neural Network. Using the average of all algorithms' results, far greater accuracy was achieved using the compound subsets selected in the 2-class problems than were achieved from randomly selected compound subsets of the same size.

The selected compound subsets most effective against melanoma and leukemia had no cluster overlap with the 92 compounds of known MOA available at the NCI. Moreover, there was no overlap with the topoisomerase II MOA class containing four known quinone compounds. By next examining the selected compound subset structures directly, we determined that a large fraction of each of the significant ($p < 0.01$) compounds against melanoma (14) and leukemia (30) were quinones (11 and 8, respectively). However, for each tumor class type, the quinones formed two structurally distinct subtypes. For melanoma, the quinone rings were internal, while for leukemia the quinone rings were external. The external quinone class was found to not be preferentially effective against either ALL or non-ALL leukemias, in the 18 of 30 compounds that were found to carry out a modest clustering of ALL vs non-ALL leukemia cell lines. These differences suggest important biological variation in the different tumor classes' ability to adsorb, metabolize, and respond to these different quinone subtypes. The differences can be due to a number of possible mechanisms including the following: enzymatic selectivity or rate differences for the quinone subtypes acting as substrates, where NAD(P)H:quinone oxidoreductase 1 is a strong candidate enzyme, or to differences in the tumors' response to the enzymatic product species resulting from the two quinone subtypes, or to competition with ubiquinone of the electron transport chain, thereby disrupting cellular redox processes.

ACKNOWLEDGMENT

AnVil and the authors gratefully acknowledge support from an SBIR Phase I Grant R43 CA94429-01 from the National Cancer Institute.

REFERENCES AND NOTES

(1) Voigt, K.; Bruggeman, R. Toxicology Databases in the Metadatabank of Online Databases. *Toxicology* **1995**, *100*, 225−240.

(2) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**, *275*, 343−349.

(3) Shi, L. M.; Fan, Y.; Lee, J. K.; Waltham, M.; Andrews, D. T.; Scherf, U.; Paul, K. D.; Weinstein, J. N. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 367−379

(4) Bai, R. L.; Paul, K. D.; Herald, C. L.; Malspeis, L.; Pettit, G. R.; Hamel, E. Halichondrin B and homahalichondrin B, marine natural products binding in the vinca domain of tubulin-based mechanism of action by analysis of fifferential cytotoxicity data. *J. Biol. Chem*. **1991**, *266*, 15882−15889.

(5) Cleveland, E. S.; Monks, A.; Vaigro-Wolff, A.; Zaharevitz, D. W.; Paul, K.; Ardalan, K.; Cooney, D. A.; Ford, H., Jr. Site of action of two novel pyramidine biosynthesis inhibitors accurately predicted by COMPARE program. *Biochem. Pharmacol*. **1995**, *49*, 947−954.

(6) Gupta, M.; Abdel-Megeed, M.; Hoki, Y.; Kohlhagen, G.; Paul, K.; Pommier, Y. Eukaryotic DNA topoisomerases mediated DNA cleavage induced by new inhibitor: NSC 665517. *Mol. Pharmacol*. **1995**, *48*, 658−665.

(7) Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connors, P. M.; Paul, K. D.; Friend, S. H.; Weinstein, J. N. Mining the National Cancer Institute Anticancer Drug Discovery Database: cluster analysis of ellipticine analogues with p53-inverse and central nervous system-selective patterns of activity. *Mol. Pharmacol.* **1998**, *53*, 241−251.

(8) 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.; Pergamenshcikov, A.; Lee, J. C. F.; Lashkari, D.; Shalon, D.; Myers, T. G.; Weinstein, J. N.; Botstein, D.; Brown, P. O. Systematic variation of gene expression patterns in human cancer cell lines. *Nat. Genet*. **2000**, *24*, 227−235.

(9) Staunton, J. E.; Slonim, D. K.; Coller, H. A.; Tamayo, P.; Angelo, M. P.; Park, J.; Sherf, U.; Lee, J. K.; Reinhold: W. O.; Weinstein, J. N.; Mesirov, J. P.; Landers, E. S.; Golub, T. R. Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci*. **2001**, *98*, 10787−10792.

(10) Blower, P. E.; Yang, C.; Fligner, M. A.; Verducci, J. S.; Yu, L.; Richman, S.; Weinstein, J. N. Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J.* **2002**, *2*, 259−271.

(11) Scherf, W.; 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.; Weinstein, J. N. A gene expression database for the molecular pharmacology of cancer. *Nature* **2000**, *24*, 236−247.

(12) Schafer, J. L. *Analysis of Incomplete Multivariate Data, Monographs on Statistics and Applied Probability*; Chapman & Hall/CRC; 1997; p 72.

(13) RadViz, URL: www.anvilinfo.com.

(14) Hoffman, P.; Grinstein, G.; Marx, K.; Grosse, I.; Stanley, E. DNA visual and analytical data mining. *IEEE Visualization* 1997 Proceedings, Phoenix, pp 437−441.

(15) Hoffman, P.; Grinstein, G. Multidimensional information visualization for data mining with application for machine learning classifiers: *Information Visualization in Data Mining and Knowledge Discovery*; Morgan-Kaufmann: San Francisco, 2000.

(16) Ross, D. NAD(P)H: quinone oxidoreductases. *Encyclopedia of Molecular Medicine;* 2001; pp 2208−2212.

(17) Ross, D.; Beall, H.; Traver, R. D.; Siegel, D.; Phillips, R. M.; Gibson, N. W. Bioactivation of quinones by DT-Diaphorase. Molecular, biochemical and chemical studies. *Oncol. Res.* **1994**, *6*, 493−500.

(18) Faig, M.; Bianchet, M. A.; Talalay, P.; Chen, S.; Winski, S.; Ross, D.; Amzel, L. M. Structure of recombinant human and mouse NAD(P)H: quinone oxidoreductase: Species comparison and structural changes with substrate binding and release. *Proc. Natl. Acad. Sci.* **2000**, *97*, 3177−3182.

(19) Faig, M.; Bianchet, M. A.; Winsky, S.; Moody, C. J.; Hudnott, A. H.; Ross, D.; Amzel, L. M. Structure-based development of anticancer drugs: complexes of NAD(P)H: quinone oxidoreductase 1 with chemotherapeutic quinones. *Structure (Cambridge)* **2001**, *9*, 659−667.

(20) Wolkenberg, S. E. In situ activation of antitumor agents. *Tetrahedron Lett*. **2001**, 1−5.

(21) Smith, M. T.; Wang, Y.; Kane, E.; Rollinson, S.; Wiemels, J. L.; Roman, E.; Roddam, P.; Cartwright, R.; Morgan, G. Low NAD(P)H: quinone oxidoreductase I activity is associated with increased risk of acute leukemia in adults. *Blood* **2001**, *97*, 1422−1426.

(22) Wiemels, J. L.; Pagnamenta, A.; Taylor, G. M.; Eden, O. B.; Alexander, F. E.; Greaves, M. F. A lack of a functional NAD(P)H: quinone oxidoreductase allele is selectively associated with pediatric leukemias that have MLL fusions. United Kingdom Childhood Cancer Study Investigators. *Cancer Res*. **1999**, *59*, 4095−4099.

(23) Naoe, T.; Takeyama, K.; Yokozawa, T.; Kiyoi, H.; Seto, M.; Uike, N.; Ino, T.; Utsunomiya, A.; Maruta, A.; Jin-nai, I.; Kamada, N.; Kubota, Y.; Nakamura, H.; Shimazaki, C.; Horiike, S.; Kodera, Y.; Saito, H.; Ueda, R.; Wiemels, J.; Ohno, R. Analysis of the genetic polymorphism in NQO1, GST-M1, GST-T1 and CYP3A4 in 469 Japanese patients with therapy related leukemia/myelodysplastic syndrome and de novo acute myeloid leukemia. *Clin. Cancer Res*. **2000**, *6*, 4091−4095.

(24) Bucci, C.; Thompsen, P.; Nicoziani, P.; McCarthy, J.; van Deurs, B. Rab7: a key to lysosome biogenesis. *Mol. Biol. Cell* **2000**, *11*, 467−480.

CI034050+