molecular pharmaceutics

articles

Exploratory Chemoinformatic Analysis of Cell Type-Selective Anticancer Drug Targeting

Kerby Shedden[†] and Gus R. Rosania*[,‡]

Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, and Department of Pharmaceutics, University of Michigan College of Pharmacy, Ann Arbor, Michigan 48109

Received April 2, 2004

Abstract: In pharmaceutical development, structure−activity relationship studies aim to identify characteristics of chemical structures associated with well-defined activity end points. While this goal-driven approach is ideally suited for lead development purposes, a more exploration-driven approach is needed to discover cell type-selective drug targeting mechanisms in complex data sets. Growth inhibition profiles across different cancer cell lines are potentially informative with respect to molecular mechanisms targeting the activity of anticancer agents to specific tumor cells, yet only a small number of mechanistic associations between chemical structure and growth inhibition profiles have been discovered to date. Here, we have applied an exhaustive statistical analysis strategy to more than 10 000 compounds in the NCI's anticancer agent database to identify molecular substructures associated with specific cytotoxicity signatures against a panel of human tumor-derived cancer cell lines (the Developmental Therapeutics Program 60-cell line panel). Some of the most significant substructures conferring cell type-selective cytotoxic activity include a large family of delocalized lipophilic cations; chloropurines, chloropyrimidines, and thiazoles; organosulfur chelators and organometallic complexes; and an unexpectedly related family of alkyl-lysophospholipids and phosphate prodrugs. Information from cell-based assays and gene expression measurements have been related to substructures represented in the chemical space covered by the library, yielding several candidate targeting mechanisms.

Keywords: Chemoinformatics; structure−activity relationships; drug targeting; anticancer agents; drug discovery

Introduction

The central problem in the analysis of structure−activity relationships (SARs) is identifying characteristics of a compound's chemical structure that are associated with a specific biological property or functional activity. SAR studies in drug discovery focus on assays with well-defined end points, for example, inhibition of enzymatic activity. In these assays, potent compounds equate with promising leads. Thus, even if compounds are screened in multiple assays, the most promising drug leads can be ranked and prioritized on the basis of *a priori* performance thresholds set for each assay.

Growth inhibition (GI) profiles capture the specific cytotoxicity of a compound across a range of cell lines. Since the important information is contained in the entire profile rather than in responses of individual cell lines, GI profiles are not easy to study using standard SAR approaches. Nevertheless, GI profiles can be quite informative for basic research into the mechanisms of anticancer drug sensitivity and resistance, including the primary mode of action, secondary toxicities, metabolism, and transport.[1−8] While several studies looking at cytotoxicity profiles have focused on classification and prediction of compounds in different

* To whom correspondence should be addressed: University of Michigan College of Pharmacy, 428 Church St., Ann Arbor, MI 48109. E-mail: grosania@umich.edu.

[†] University of Michigan.

[‡] University of Michigan College of Pharmacy.

10.1021/mp049953k CCC: $27.50
Published on Web 06/15/2004

toxicity classes,[1,3,4,9−15] some of these toxicity classes may be exploited to enhance the selective activity of anticancer agents against specific types of tumor cells.[5,9,12,16−18]

Understanding the mechanism through which an anticancer agent selectively inhibits the growth of a particular cancer cell type can have practical applications in the development of targeted anticancer agents. Indeed, mechanistic SAR analyses of GI profiles not only point to mechanisms through which anticancer agents inhibit the growth of a particular cell type but also could be used as a starting point to develop "personalized" chemotherapeutic strategies.[9] To discover targeting mechanisms, one must start by identifying statistically significant associations between chemical structure and growth inhibition of particular cell types. Since GI profiles vary in a space having dimensions equal to the number of cell lines assayed (59 in our case), identifying statistically significant GI profiles is not a trivial amount of work. Identifying significant associations between GI profiles and chemical structure is even more challenging.

Here we describe the results of an exhaustive search for associations between GI profiles and chemical structure for more than 10 000 chemical structures in the NCI anticancer agent database. The basic idea is to partition the compounds into groups on the basis of whether small chemical structure fragments called "augmented atom codes" (AACs)[19] are present in the molecule. Then for each AAC, we make a statistical comparison of growth inhibition between compounds containing the AAC and compounds lacking it, for each cell line in the panel. The structure fragments yielding a significant result for this comparison are then expanded into larger fragments that determine a specific GI profile. Further checks are made to ensure that the fragment is not just linked to some other fragment that is the mechanistic actor. Finally, gene expression measurements performed on the 60 cell lines can help infer biological mechanisms underlying each chemical structure−growth inhibition association.[10,13−16,20] This method elucidates many chemical structure−growth inhibition associations with a moderate to high level of statistical significance. In this article, we present the 15 discovered associations that are most significant.

(1) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; 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.

(2) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.

(3) Shi, L. M.; Fan, Y.; Lee, J. K.; Waltham, M.; Andrews, D. T.; Scherf, U.; Paull, K. D.; Weinstein, J. N. Mining and visualizing large anticancer drug discovery databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 367−379.

(4) Koutsoukos, A. D.; Rubinstein, L. V.; Faraggi, D.; Simon, R. M.; Kalyandrug, S.; Weinstein, J. N.; Kohn, K. W.; Paull, K. D. Discrimination techniques applied to the NCI in vitro anti-tumour drug screen: predicting biochemical mechanism of action. *Stat. Med.* **1994**, *13*, 719−730.

(5) 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.

(6) Alvarez, M.; Paull, K.; Monks, A.; Hose, C.; Lee, J. S.; Weinstein, J.; Grever, M.; Bates, S.; Fojo, T. Generation of a drug resistance profile by quantitation of mdr-1/P-glycoprotein in the cell lines of the National Cancer Institute Anticancer Drug Screen. *J. Clin. Invest.* **1995**, *95*, 2205−2214.

(7) Wallqvist, A.; Monks, A.; Rabow, A. A.; Thanki, N.; Shoemaker, R. H.; Covell, D. G. Mining the NCI screening database: explorations of agents involved in cell cycle regulation. *Prog. Cell Cycle Res.* **2003**, *5*, 173−179.

(8) Alvarez, M.; Robey, R.; Sandor, V.; Nishiyama, K.; Matsumoto, Y.; Paull, K.; Bates, S.; Fojo, T. Using the National Cancer Institute anticancer drug screen to assess the effect of MRP expression on drug sensitivity profiles. *Mol. Pharmacol.* **1998**, *54*, 802−814.

(9) Shedden, K.; Townsend, L. B.; Drach, J. C.; Rosania, G. R. A rational approach to personalized anticancer therapy: chemoinformatic analysis reveals mechanistic gene-drug associations. *Pharm. Res.* **2003**, *20*, 843−847.

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

(11) Monga, M.; Sausville, E. A. Developmental therapeutics program at the NCI: molecular target and drug discovery process. *Leukemia* **2002**, *16*, 520−526.

(12) Keskin, O.; Bahar, I.; Jernigan, R. L.; Beutler, J. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Characterization of anticancer agents by their growth inhibitory activity and relationships to mechanism of action and structure. *Anticancer Drug Des.* **2000**, *15*, 79−98.

(13) Amundson, S. A.; Myers, T. G.; Scudiero, D.; Kitada, S.; Reed, J. C.; Fornace, A. J., Jr. An informatics approach identifying markers of chemosensitivity in human cancer cell lines. *Cancer Res.* **2000**, *60*, 6101−6110.

(14) Yamori, T. Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics. *Cancer Chemother. Pharmacol.* **2003**, *52* (Suppl. 1), S74−S79.

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

(16) Wallqvist, A.; Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Establishing connections between microarray expression data and chemotherapeutic cancer pharmacology. *Mol. Cancer Ther.* **2002**, *1*, 311−320.

(17) Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. *J. Med. Chem.* **2002**, *45*, 818−840.

(18) Johnson, D. E.; Blower, P. E., Jr.; Myatt, G. J.; Wolfgang, G. H. Chem-tox informatics: data mining using a medicinal chemistry building block approach. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 92−101.

(19) Todeschini, R. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, Germany, 2000; Vol. xxi, p 667.

Materials and Methods

Data Source. The Developmental Therapeutics Program (DTP) at the National Cancer Institute (NCI) has compiled a database of roughly 41 000 small molecules that have also been screened for chemosensitivity on a reference set of human tumor-derived cell lines, commonly known as the "60-cell line panel".[11,14,21] For each compound, the chemical structure is available as a connection table specifying all chemical bonds in the molecule.

Growth inhibition (GI) measurements are available as GI_{50} values, the concentration of a compound slowing growth at 48 h by 50% relative to the growth of untreated cells. We analyzed GI_{50} values on the $\log_2$ scale, and only considered GI_{50} values for 59 cell lines for which gene expression data are available. These 59 cell lines are the usual NCI 60 cell lines except that breast cell line MDA-N has been omitted. Compounds that were missing GI_{50} values for more than 10 of these 59 cell lines were dropped from the analysis. Compounds for which the standard deviation of the observed GI_{50} values was less than 0.3 were also dropped, leaving 10 589 compounds for further analysis. A standard deviation of 0.3 corresponds roughly to requiring at least one-third of the cells in a cell line to differ by more than 25% from the mean GI_{50} for the compound.

Data Normalization. GI_{50} values were processed to remove effects due to either the average toxicity of each compound or the average sensitivity of each cell line. We fit a linear model of the form $GI_{50ij} = A_i + B_j + R_{ij}$, where GI_{50ij} is the experimental GI_{50} value for compound i in cell line j. The A_i values correspond to differing average toxicities across the compounds. For example, metal-containing compounds tend to have lower GI_{50} values than compounds containing no metal. The B_j values correspond to the overall sensitivity of each cell line across all compounds. For example, leukemic cell lines tend to be more sensitive to growth inhibitory agents overall, while renal cell lines tend to be more resistant. All subsequent analysis is carried out on the R_{ij} values, which we will term GI_{50} values henceforth. A GI_{50} value of zero indicates that compound i has average toxicity in cell line j, after correcting for overall compound toxicity and overall cell line sensitivity. Values of GI_{50} much greater than zero indicate lower than expected toxicity, and values of GI_{50} much less than zero indicate higher than expected toxicity. Missing values were filled in with zero.

Chemical Structure Descriptors. For analysis of chemical structure−activity relationships, the compounds were computationally fragmented into a simpler set of substructures using the connection tables in the DTP database. Specifically, we identified all possible chemical structure fragments consisting of a central atom together with information about the peripheral atoms to which the central atom is directly bonded (i.e., the element type of each atom and the order of each bond, without distinguishing aromatic, nonaromatic, or partial bonding characteristics). These fragments are called augmented atom codes (AACs)[19] and have been used for many years in QSAR studies. For example, a central nitrogen atom with single bonds to two carbon atoms and double bonds to one carbon atom is an AAC. This AAC is denoted N:C1;C1;C2, where the letter before the colon is the chemical symbol for the central atom and the letters following the colon are the chemical symbols of the atoms to which the central atom is bonded (each followed by a number indicating the number of bonds to the central atom).

Initial Screening of AACs. Candidate AACs with specific effects on GI_{50} are identified by comparing the average GI_{50} values for compounds containing the AAC to the average GI_{50} of compounds lacking it for each cell line. If this ratio exceeds 2 or is less than $1/2$ in at least one cell line, then the AAC is selected for further analysis. Only AACs present in at least 30 compounds were considered for selection. In the calculation of false positive rates (*vide infra*), if at least 30 compounds are used to form the average, the probability of obtaining a 2-fold change by chance in at least one of the 59 cell lines is less than 1 in 10^4. Since 130 AACs are present in at least 30 compounds, it follows that the expected number of false positives is less than 0.01.

Estimation of False Positive Rates. The false positive rate is the probability of finding an AAC for which at least one cell line's mean GI_{50} value for compounds containing the AAC differs by 2-fold from the mean GI_{50} value for compounds lacking the AAC, by random chance. To calculate this false positive rate, we carried out a statistical analysis of the $|R_{ij}|$ values (the absolute values of the adjusted GI_{50} values defined above) for compounds deemed unaffected by biological variation in the cell lines. To identify such compounds, we returned to the raw (unadjusted) GI_{50} data and selected the compounds for which no GI_{50} values were missing, and no GI_{50} values were at the ceiling level of 10^{-4} M. These 2427 compounds were sorted in ascending order by range (greatest value minus least value across the cell lines), and compounds with zero range were omitted. Next, we selected 250 compounds with a range in the second quartile, for which variation is likely due to experimental noise in the assay. After all 59×250 data points for these compounds have been pooled together, the corresponding adjusted $|R_{ij}|^{1.14}$ values closely follow an exponential distribution with mean of 0.2 (data not shown). Thus, we can simulate from the error distribution of the GI_{50} assay by simulating e from a standard exponential distribution, taking E to be equal to $(0.2e)^{1/1.14}$, and then multiplying E by -1 with a probability of 0.5. Using 10^5 simulated data sets of 59 "cell lines" and 30 "compounds" from this distribution, we found that fewer than 1 in 10^4 of the simulated data sets had at least one cell line with a $\geq$2-fold change, and the

(20) Wallqvist, A.; Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Linking the growth inhibition response from the National Cancer Institute's anticancer screen to gene expression levels and other molecular target data. *Bioinformatics* **2003**, *19*, 2212−2224.

(21) Vekris, A.; Meynard, D.; Haaz, M. C.; Bayssas, M.; Bonnet, J.; Robert, J. Molecular determinants of the cytotoxicity of platinum compounds: the contribution of in silico research. *Cancer Res.* **2004**, *64*, 356−362.

expected number of false positives is therefore less than 130×10^{-4} ($\approx$0.01).

Discrimination of Extreme and Average Subsets of Compounds. For SAR analysis, a comparison was made between the distribution of GI_{50} values for compounds containing a specific chemical substructure and the distribution of GI_{50} values for compounds lacking it. The comparison was based on kernel density estimates (Gaussian kernel with a bandwidth of $1.06\sigma/n^{1/5}$) prepared for each of the two sets of compounds. Individual compounds whose GI_{50} falls at a point in the GI_{50} distribution where the density for compounds containing the AAC is at least 3 times greater than the density for compounds lacking the AAC are deemed to be "extreme", while the remaining compounds are deemed to be "average". Substructure expansion analysis, described below, was then performed to determine if any structural features beyond the AAC could be used to discriminate the extreme from the average compounds.

SAR Analysis by Fragment Expansion. The fragment expansion algorithm proceeds in an iterative fashion, with a given AAC as the initial fragment. At each step, all possible extensions of the fragment that can be obtained by adding a single "floating atom" are considered. To determine which atom is added, each possible extension is evaluated in terms of its representation in the average versus extreme subsets of compounds. Specifically, a χ^2 statistic is calculated comparing the frequency with which the expanded fragment occurs in the two compound sets. If the χ^2 statistic exceeds 10, if the proportion in the extreme compounds exceeds 0.6, and if this proportion is greater than the corresponding proportion in the average compounds, then the floating atom is considered to be a "discriminating candidate" for the extension. Among all discriminating candidates, the one with the highest χ^2 statistic is added to the fragment. If no discriminating candidate is found, the floating atom with the greatest frequency in the extreme compounds is added, as long as that frequency exceeds 0.8. This process is repeated, joining additional floating atoms to the fragment until a termination criterion is met. The process terminates if no atom can be added to the fragment, or if the average agreement between the fragment and the extreme compounds drops below 75% of the fragment size. In this manner, each AAC with a significant effect on GI_{50} is extended to a larger fragment that is more highly specific to the distinctive pattern of growth inhibition associated with the AAC (see Table 2 of the Supporting Information).

Gene Expression Data. Microarray measurements of gene expression in the 59-cell line panel are used to aid in inferring the biological and chemical mechanism underlying each significant structure−GI association.[10,13−16,20] We used the triplicate array measurements obtained using Affymetrix U95A chips from Novartis. Arrays were scale-normalized to have equal medians. For each AAC's GI_{50} profile (i.e., the 59 log ratios between the average GI_{50} for compounds containing an AAC and the average GI_{50} for compounds lacking it), we calculated Pearson correlation coefficients between the values in the GI_{50} profile and log-transformed gene expression measurements for each of the three sets of arrays. For some analyses, the six leukemic cell lines were excluded when correlations were formed. This reduces the number of significant correlations by 75%, since there are many compounds in the database exhibiting leukemia-specific toxicity, and many genes are specifically expressed or repressed in leukemic cell lines. Genes arising from this relationship are unlikely to be related to the biological response to a drug. A list of genes was prepared for each association, where the genes are ranked according to the least (in magnitude) of the correlation coefficients for each of the triplicate experiments (Table 1 of the Supporting Information).

Results

Identification of Relationships between AACs and Chemosensitivity Profiles. The DTP database contains a diverse set of natural products and synthetic compounds that also have GI_{50} data for our reference set of 59 cell lines. We selected 10 589 compounds with low levels of missing data and moderate to high variation in the assay readout. In these compounds, we identified 747 distinct AACs to use as starting points for studying the relationship between chemical substructures of compounds and their growth inhibitory activity. Figure 1A shows the number of compounds matching each AAC, ranked according to the number of matches for each AAC. From the initial set of 747 AACs, we selected for further analysis 130 AACs (bold in Figure 1A) that were present in at least 30 compounds.

For each cell line−AAC pair, we then looked at the log ratios of the average GI_{50} for compounds containing the AAC to the average GI_{50} for compounds lacking it. Figure 1B summarizes these fold changes over the entire database; for each AAC, the largest fold change (in magnitude) across the cell lines is selected, and these fold changes are plotted as a function of rank. As discussed in Materials and Methods, less than 0.01 AAC is expected to have a 2-fold change in at least one cell line by chance. We found 15 AACs with fold change exceeding 2 in at least one cell line (bold in Figure 1B). This gives a false discovery rate of less than 0.01/15 (=0.001), so overall, the 15 selected AACs are highly statistically significant; it is unlikely that even one of them is a false positive.

These 15 selected AACs are represented in 1701 distinct compounds of 10 589 that were analyzed, so they are fairly common in the agents screened by the NCI. Several of the 15 AACs satisfy the 2-fold change condition for more than one cell line, so there are a total of 45 AAC−cell line pairs in which AAC-specific effects on GI_{50} are seen. As a result of meeting the fold change condition, these are the most likely AACs to be functionally significant in determining cell type-specific growth inhibitory activity.

Applying a two-way clustering algorithm to the GI_{50} fold changes in the 15 AACs and the 59 cell lines reveals that certain sets of AACs lead to similar chemosensitivity profiles (Figure 2). Specifically, there are several distinct clusters that are formed by more than one agent that share similar

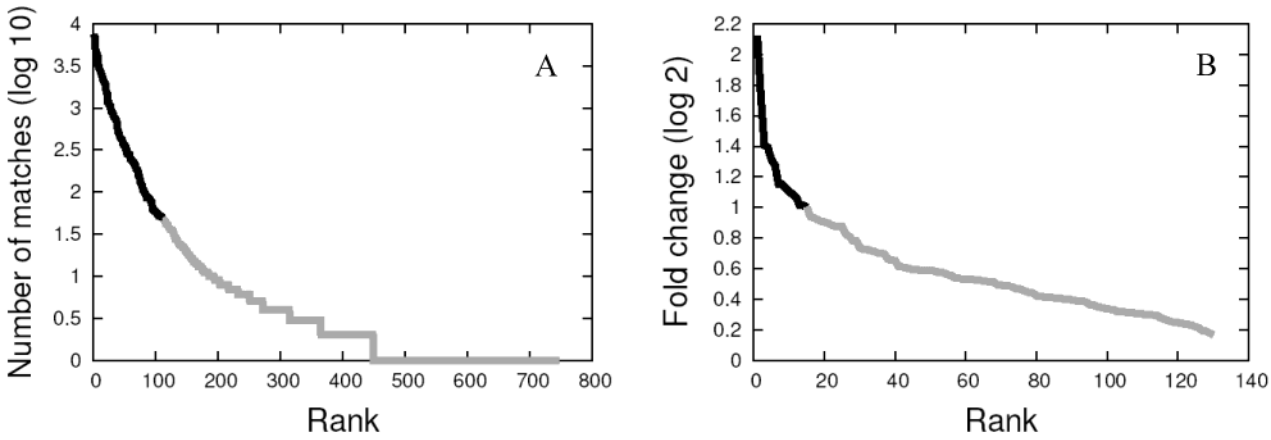


Figure 1. Survey of AACs represented in the NCI anticancer agent database. (A) Plot of the number of compounds matching each of the 747 AACs on the log scale, ranked by decreasing number of matches. The bold line covers 110 fragments contained in at least 30 compounds that are used in the subsequent analysis. (B) Rank plot showing the $\log_2$ fold changes between the average GI_{50} for compounds containing a given AAC relative to the average GI_{50} for compounds lacking the AAC. The fold change that is shown is the largest in magnitude over 59 cell lines. The bold line covers the 15 fragments that have a ≥2-fold change in at least one cell line.

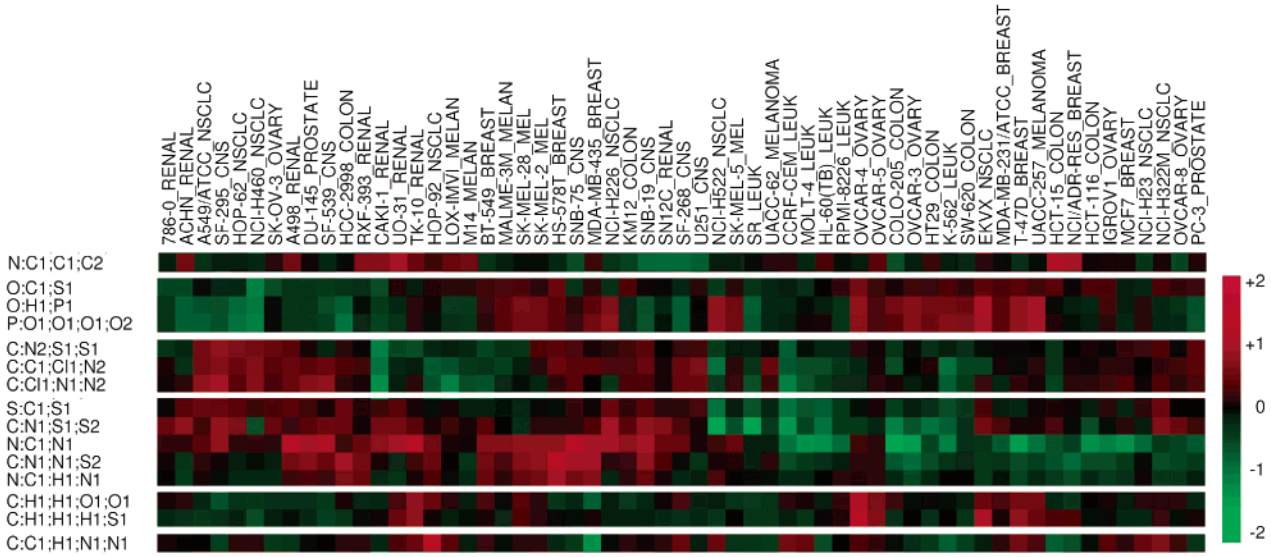


Figure 2. Cluster analysis of cell lines and structure fragments. For each AAC−cell line pair, the $\log_2$ GI_{50} ratio is calculated between compounds containing the AAC and compounds lacking it. Results were clustered with respect to both AACs and cell lines, and visualized as a heat map. Only the 15 AACs with a 2-fold change in at least on cell line are shown.

cell type-selective targeting activities. There are two possible explanations for the clusters. One explanation is that several AACs may inhibit cell growth through a similar mechanism, even though the AACs are contained in distinct sets of compounds. A more trivial explanation is that several AACs may overlap to form a common substructure so that a single class of compounds is represented by several different AACs.

To address this issue, for each distinct pair among the 15 selected AACs, we calculated the number of compounds containing both AACs, divided by the total number of distinct compounds containing either AAC. Table 1 shows the results of this analysis. Two AACs (O:H;P1 and P:O1;O1;O1;O2) exhibit 79% overlap. Evidently, these AACs overlap to form a phosphate group, and therefore occur in approximately the same set of compounds. The N:C1;H1;N1 and C:N1;S1;S2 AACs exhibit 27% overlap, and no other pair of AACs exhibits >12% overlap. Thus, with the exception of the O:H;P1 and P:O1;O1;O1;O2 AACs, and perhaps N:C1;H1;N1 and C:N1;S1;S2, observed associations between different AACs are likely due to similarities in the biological response of the cells to distinct classes of compounds.

Identification of Compounds with AAC-Specific GI_{50} Values. To understand more completely the nature of GI_{50} effects associated with particular AACs, for each of the 45 AAC−cell line pairs, density plots were prepared showing the GI_{50} distribution for compounds containing the AAC and the GI_{50} distribution for compounds lacking it (Figure 3). After visual inspection, many of these distributions had a strong multimodal character or were heavily skewed. This suggests that the observed 2-fold difference in the mean GI_{50}

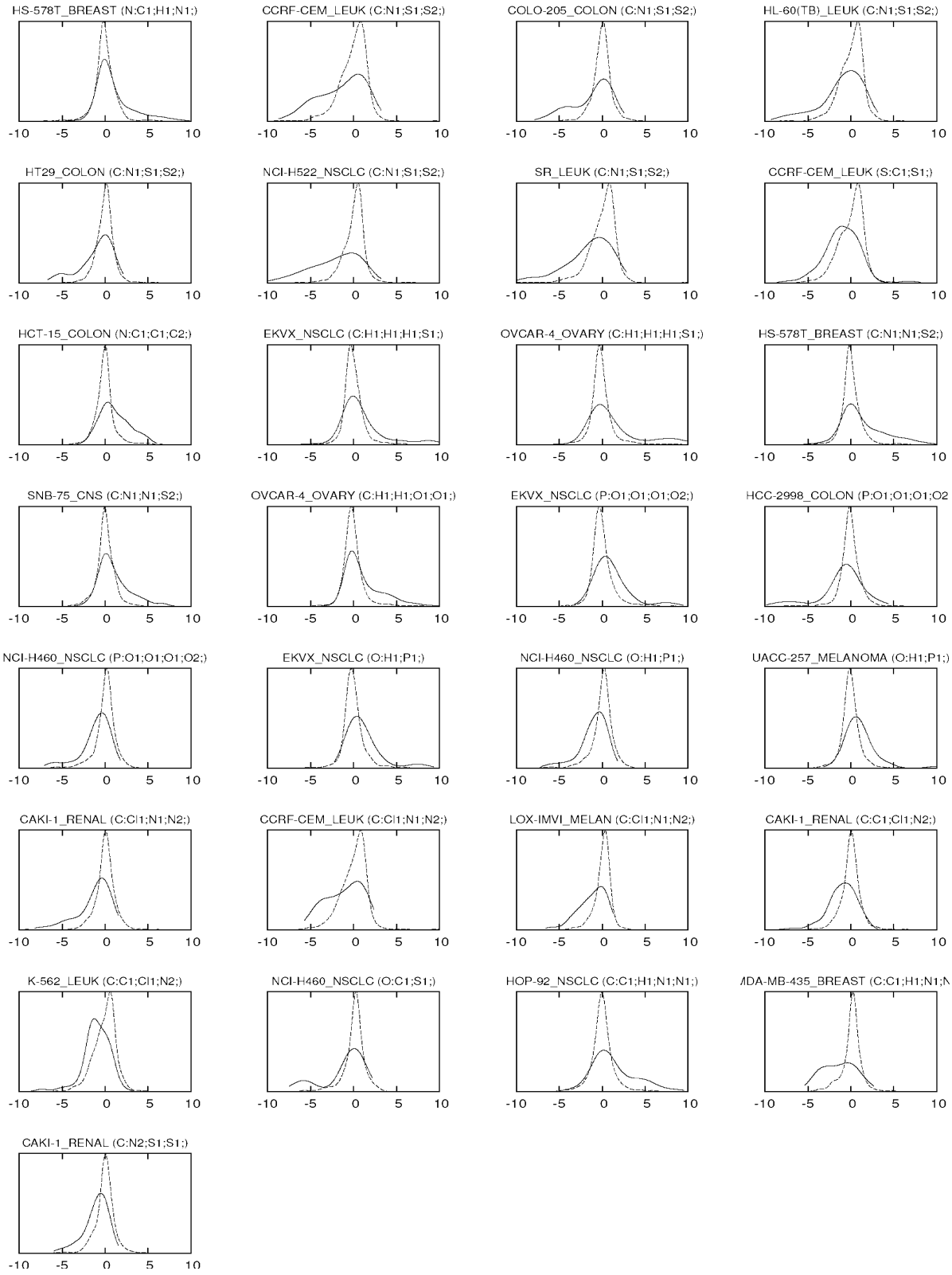


Figure 3. *GI_{50}* densities for each cell line−AAC pair for which a 2-fold difference exists in the *GI_{50}* values between compounds containing the AAC (—) and compounds lacking it (- - -). Note that most density plots are multimodal or highly skewed.

is not uniquely determined by the AAC, since many of the compounds containing the AAC have average *GI_{50}* values. This result is consistent with the fact that the activity of a molecule in a particular cell line is affected in many complex ways by the structure and physicochemical properties of the entire molecule, and not just by any individual AAC. Nevertheless, the skewness and multimodal nature of the *GI_{50}* distribution indicate that a substantial minority of compounds

Table 1. Co-Occurrence of AACs in Individual Molecules, Expressed as a Percentage of Molecules Showing the Two Indicated AACs[a]

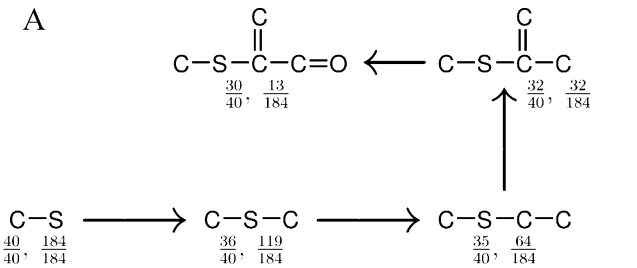
	N:C1;H1;N1	C:N1;S1;S2	S:C1;S1	N:C1;C1;C2	C:H1;H1;H1;S1	C:N1;N1;S2	C:H1;H1;O1;O1	P:O1;O1;O1;O2	O:H1;P1	C:Cl1;N1;N2	C:C1;Cl1;N2	O:C1;S1	C:C1;H1;N1;N1	C:N2;S1;S1
C:N1;S1;S2	1													
S:C1;S1	0	8												
N:C1;C1;C2	1	0	0											
C:H1;H1;H1;S1	1	2	1	4										
C:N1;N1;S2	27	2	2	0	0									
C:H1;H1;O1;O1	3	1	0	2	0	0								
P:O1;O1;O1;O2	0	0	0	1	0	0	1							
O:H1;P1	0	0	0	1	0	0	0	79						
C:Cl1;N1;N2	0	0	0	0	0	0	0	4	4					
C:C1;Cl1;N2	1	0	1	0	2	0	0	0	0	12				
O:C1;S1	0	0	0	2	8	0	0	0	0	0	0			
C:C1;H1;N1;N1	0	0	1	0	0	0	0	0	0	0	0	0		
C:N2;S1;S1	0	0	0	12	0	0	0	0	0	0	0	0	0	
N:C1;N1	0	3	0	0	1	7	0	0	0	1	0	0	0	0

[a] The overlap is greatest between O:H1;P1 and P:O1;O1;O1;O2 (phosphates) AACs, followed by C:N1;N1;S2 and N:C1;H1;N1 AACs (carbothioamides).

containing each AAC exhibits extreme toxicity or resistance against one or more cell lines.

To distinguish the compounds with extreme GI_{50} values from those with average GI_{50} values based on the GI_{50} densities, we partitioned the compounds containing each structure fragment into two subsets. As the extreme GI_{50} group, we selected compounds whose GI_{50} values fall into a region of the GI_{50} distribution where the density for compounds containing the fragment is at least 3 times greater than the density for compounds lacking the fragment. All other compounds containing the fragment are in the average GI_{50} group. This partition can be made separately for each cell line with a 2-fold effect. For AACs with more than one cell line meeting the 2-fold threshold, we found a high degree of similarity in the extreme compounds across the cell lines. Therefore, we pooled all compounds that were extreme in at least one cell line into the extreme group, and the remaining compounds were placed into the average group. This gives a single partition of the compounds containing each AAC.

The fragment expansion algorithm was then used to expand each AAC into two larger fragments: one that was highly specific to the average *GI*$_{50}$ compounds and one that was highly specific to the extreme *GI*$_{50}$ compounds. Illustrating the expansion algorithm result for the C:H1;H1;H1;S1 AAC clearly demonstrates how this statistical technique identifies large molecular fragments that are responsible for differential cytotoxicity, starting from an AAC core (Figure 4A). By the fourth step of the expansion, the substructure matches 30 of 40 of the extreme *GI*$_{50}$ compounds, but only matches 13 of 184 of the average *GI*$_{50}$ compounds. Applied to the 15 AACs selected in the initial screen, the fragment expansion algorithm is able to identify larger substructural motifs that are associated with differential cytotoxicity profiles in several of the compounds (Figure 4B and Table 2 of the Supporting Information).



B

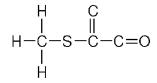
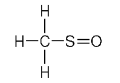
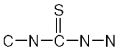
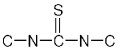
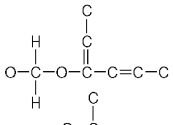
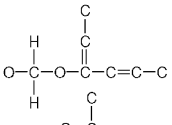
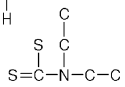
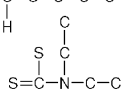
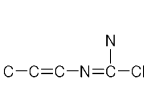
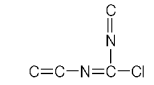
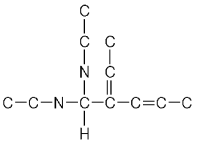
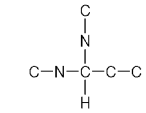
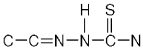
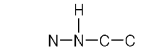
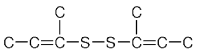
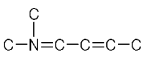
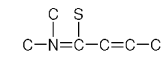
AAC	Extreme Structure	Average Structure
C:H;H;H;S		
C:N1;N1;S2		
C:H;H;O1;O1		
C:N1;S1;S2		
C:Cl1;N1;N2		
C:C1;H;N1;N1		
N:C1;N1		
N:C1;H;N1		
S:C1;S1		
N:C1;C1;C2		
P:O1;O1;O1;O2		
C:C1;Cl1;N2		

Figure 4. Results of the fragment expansion algorithm. (A) Partial results obtained during the stepwise expansion of the S:H1;H1;H1;C1 AAC. Fractions indicate the representation of the fragment in the extreme vs average group of compounds. As the fragment is expanded, greater discrimination between the two groups of compounds is achieved. (B) Expanded structures for the most significant extreme vs average *GI*$_{50}$ AACs, together with the cell lines for which a ≥2-fold change in the *GI*$_{50}$ value is found in association with the AAC.

Structure−Activity Relationship Analysis. After fragment expansion analysis, we proceeded to analyze SARs between the subset of compounds possessing the expanded template incorporating the AAC and the particular observed cytotoxicity profiles. Additional insights relevant to the mechanism of action of the compounds could be obtained by (1) analyzing the scientific literature in terms of the

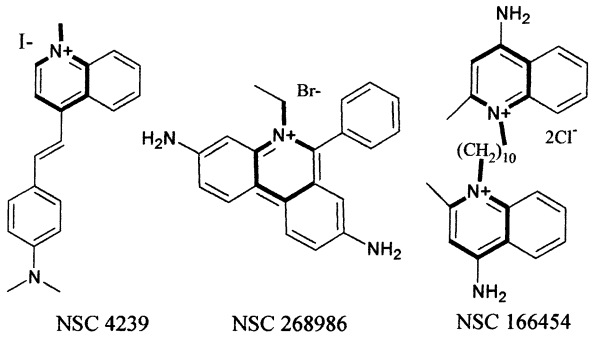


Figure 5. Selected compounds containing the expanded N;C1;C1;C2 AAC (expanded fragments in bold).

reported mechanism of action of compounds possessing the expanded AAC, (2) inspecting the structures of the compounds possessing the expanded AAC for global patterns relating the structure and function of the compounds, and (3) finding genes with a high or low level of expression in the cell lines that are most sensitive or resistant to molecules containing a particular substructure. Because transcriptional profiling data are available and some of the molecules in the library have been previously studied in the scientific literature, candidate mechanisms of action for many of the structure fragments identified by our method can be proposed. In the subsections that follow, each expanded AAC identified in this study is analyzed in terms of how its specific cytotoxicity profile is associated with differential gene expression of the cell line that was tested, and how the chemical structure of the expanded AAC (or the lack of ability to expand the AAC) suggests candidate mechanisms determining the selective activity of anticancer agents (or lack thereof) against specific cancer cell lines.

(1) Delocalized Imminium Cations. The N:C1;C1;C2 (imminium) AAC is generally embedded in a larger aromatic, conjugated structure (Figure 5). Because of the highly conjugated character of these molecules, the positive charge is delocalized. Delocalized lipophilic cations are known to accumulate in mitochondria, as a result of the electrochemical potential across the mitochondrial membrane of actively respiring cells.[22,23] At least three of the compounds sharing the AAC have been previously demonstrated to accumulate in mitochondria (Figure 5): the styryl compound 4M2M[24,25] (NSC 4239), ethidium bromide[26−30] (NSC 268986), and dequalinium[23,31,32] (NSC 166454). Ethidium bromide and dequalinium are toxic to mitochondria as they specifically induce mitochondrial DNA depletion.[23,26,30] F16,[33,34] a close styryl analogue of NSC 4239, also accumulates in mitochondria and induces apoptosis by perturbing respiratory function.

To gain additional insights into the toxicity signature of these compounds, we analyzed the genes that were over- or underexpressed in cell lines that were sensitive and resistant to these compounds. We found that compounds containing this arrangement are relatively more toxic to cell lines SF-268, SN-12C, and SNB-19 and relatively less toxic to cell lines HCT-15, ADR-RES, and SF-268 (Figure 3). Upon examination of gene expression data for all cell lines, the N:C1;C1;C2 GI_{50} profile is more correlated with the levels of the ABCB1 (also known as MDR1 or P-glycoprotein) gene than any other gene (Table 1 of the Supporting Information). For the three copies of ABCB1 in three transcriptional profiling experiments, the nine correlations (r) were as follows: 0.57, 0.62, 0.64, 0.54, 0.58, 0.56, 0.49, 0.51, and 0.56. In addition, the GI_{50} profile is strongly correlated ($r = 0.62$) with independent reports of MDR-1 (ABCB1) protein expression data,[6] and with the rates of rhodamine efflux.[35] These results suggest that, while these

(22) Rosania, G. R. Supertargeted chemistry: identifying relationships between molecular structures and their sub-cellular distribution. *Curr. Top. Med. Chem.* **2003**, *3*, 659−685.

(23) Modica-Napolitano, J. S.; Aprille, J. R. Delocalized lipophilic cations selectively target the mitochondria of carcinoma cells. *Adv. Drug Delivery Rev.* **2001**, *49*, 63−70.

(24) Shedden, K.; Brumer, J.; Chang, Y. T.; Rosania, G. R. Chemoinformatic analysis of a supertargeted combinatorial library of styryl molecules. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2068−2080.

(25) Rosania, G. R.; Lee, J. W.; Ding, L.; Yoon, H. S.; Chang, Y. T. Combinatorial approach to organelle-targeted fluorescent library based on the styryl scaffold. *J. Am. Chem. Soc.* **2003**, *125*, 1130−1131.

(26) Seidel-Rogol, B. L.; Shadel, G. S. Modulation of mitochondrial transcription in response to mtDNA depletion and repletion in HeLa cells. *Nucleic Acids Res.* **2002**, *30*, 1929−1934.

(27) Miller, S. W.; Trimmer, P. A.; Parker, W. D., Jr.; Davis, R. E. Creation and characterization of mitochondrial DNA-depleted cell lines with "neuronal-like" properties. *J. Neurochem.* **1996**, *67*, 1897−1907.

(28) Herzberg, N. H.; Middelkoop, E.; Adorf, M.; Dekker, H. L.; Van Galen, M. J.; Van den Berg, M.; Bolhuis, P. A.; Van den Bogert, C. Mitochondria in cultured human muscle cells depleted of mitochondrial DNA. *Eur. J. Cell Biol.* **1993**, *61*, 400−408.

(29) Hayashi, J.; Takemitsu, M.; Goto, Y.; Nonaka, I. Human mitochondria and mitochondrial genome function as a single dynamic cellular unit. *J. Cell Biol.* **1994**, *125*, 43−50.

(30) Hayakawa, T.; Noda, M.; Yasuda, K.; Yorifuji, H.; Taniguchi, S.; Miwa, I.; Sakura, H.; Terauchi, Y.; Hayashi, J.; Sharp, G. W.; Kanazawa, Y.; Akanuma, Y.; Yazaki, Y.; Kadowaki, T. Ethidium bromide-induced inhibition of mitochondrial gene transcription suppresses glucose-stimulated insulin release in the mouse pancreatic β-cell line βHC9. *J. Biol. Chem.* **1998**, *273*, 20300−20307.

(31) Weiss, M. J.; Wong, J. R.; Ha, C. S.; Bleday, R.; Salem, R. R.; Steele, G. D., Jr.; Chen, L. B. Dequalinium, a topical antimicrobial agent, displays anticarcinoma activity based on selective mitochondrial accumulation. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 5444−5448.

(32) Gamboa-Vujicic, G.; Emma, D. A.; Liao, S. Y.; Fuchtner, C.; Manetta, A. Toxicity of the mitochondrial poison dequalinium chloride in a murine model system. *J. Pharm. Sci.* **1993**, *82*, 231−235.

(33) Fantin, V. R.; Leder, P. F16, a mitochondriotoxic compound, triggers apoptosis or necrosis depending on the genetic background of the target carcinoma cell. *Cancer Res.* **2004**, *64*, 329−336.

(34) Fantin, V. R.; Berardi, M. J.; Scorrano, L.; Korsmeyer, S. J.; Leder, P. A novel mitochondriotoxic small molecule that selectively inhibits tumor cell growth. *Cancer Cell* **2002**, *2*, 29−42.

compounds may be accumulating in the mitochondria of actively respiring cells, the ABCB1 multidrug resistance mechanism may be countering mitochondrial accumulation and toxicity.

That lipophilic cations such as ethidium bromide, dequalinium, and styryl compound 4M2M are substrates for P-glycoprotein is a testable hypothesis. Several molecules possessing delocalized imminium cations have been noted in the literature as good substrates for P-glycoprotein-mediated multidrug resistance, suggesting that ABCB1 may be a general detoxification mechanism for this class of compounds. For example, various rhodamines[35−37] and the cyanine dye JC1[38] are routinely used for mitochondrial labeling of living cells, yet they also are also substrates for P-glycoprotein.

Interestingly, the substructure expansion algorithm did not identify an extension of the N:C1;C1;C2 AAC that was specific to the extreme GI_{50} group of compounds, which indicates that there may be other toxicity mechanisms associated with the toxicity of N:C1;C1;C2, perhaps independent of mitochondrial accumulation. For example, NSC 4238, a closely related isomer of NSC4239, accumulates in other parts of the cell[24,25] and is represented in the average GI_{50} group. However, there was a significant extension of the AAC that was specific to average GI_{50} compounds. We found that the C2 atom in the fragment has a single bond to sulfur in 40−50% of compounds with average GI_{50} values, while for compounds containing extreme GI_{50} values, at most 6% had a C2−S link. This suggests that incorporating a single sulfur atom into compounds containing N:C1;C1;C2 may substantially inhibit their ability to bind P-glycoprotein, potentially reducing the susceptibility of N:C1;C1;C2-containing compounds to P-glycoprotein-mediated drug efflux.

(2) Alkyl-Lysophospholipid Analogues. Two of the AACs (O:C1;S1 and P:O1;O1;O1;O2) correspond to sulfonates (Figure 6A) and phosphates (Figure 6B), respectively, and have relatively similar GI_{50} profiles (Figure 2). In the phosphate subset of compounds, there are at least two distinct structural subclasses: phosphate prodrugs [both nucleotides (NSC 81206; Figure 6B) and non-nucleotides (NSC 610458; Figure 6B)] and alkyl-lysophospholipids (NSC 324368; Figure 6B). These two classes of compounds are expected to have highly distinct cell type-specific targeting mechanisms, as phosphate prodrugs are presumably active only after phosphate hydrolysis,[9] while the alkyl-lysophospholipids are active without hydrolysis.[39−41] The sulfonates (NSC 268965, 72151, and 281817; Figure 6A) are somewhat similar to each other and structurally resemble the alkyl-lysophospholipid compounds because of their polar methanesulfonate headgroup and long hydrophobic tail.

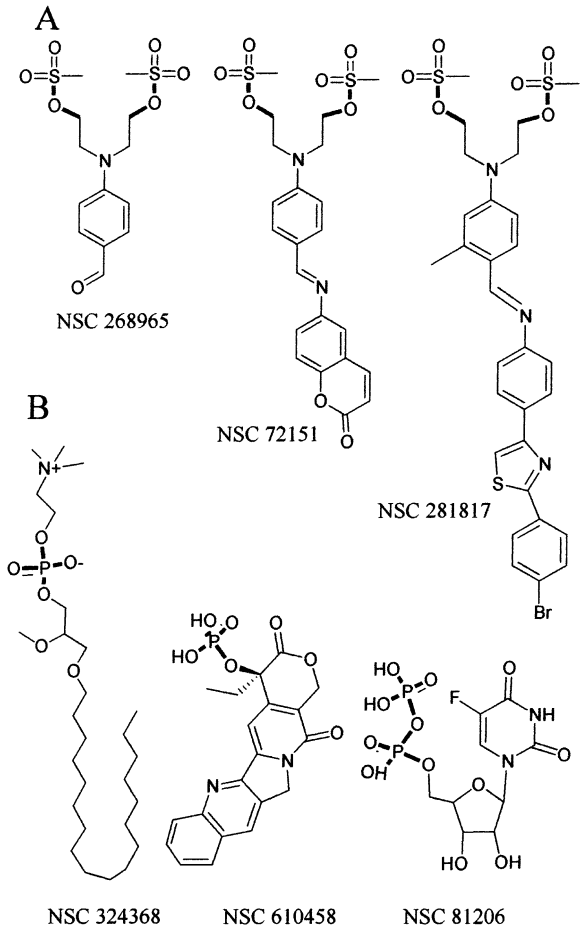


Figure 6. Selected compounds containing the expanded (A) O:C1;S1 and (B) P:O1;O1;O1;O2 and O:H1;P1 AACs (expanded fragments in bold).

Arguably, alkyl-lysophospholipid analogues are the main determinant of the cytotoxicity profile for this entire group of compounds, for three reasons. First, the 13 alkyl-lysophospholipids (or related analogues) exhibit more ex-

(35) Lee, J. S.; Paull, K.; Alvarez, M.; Hose, C.; Monks, A.; Grever, M.; Fojo, A. T.; Bates, S. E. Rhodamine efflux patterns predict P-glycoprotein substrates in the National Cancer Institute drug screen. *Mol. Pharmacol.* **1994**, *46*, 627−638.

(36) Saengkhae, C.; Loetchutinat, C.; Garnier-Suillerot, A. Kinetic analysis of rhodamines efflux mediated by the multidrug resistance protein (MRP1). *Biophys. J.* **2003**, *85*, 2006−2014.

(37) Loetchutinat, C.; Saengkhae, C.; Marbeuf-Gueye, C.; Garnier-Suillerot, A. New insights into the P-glycoprotein-mediated effluxes of rhodamines. *Eur. J. Biochem.* **2003**, *270*, 476−485.

(38) Kuhnel, J. M.; Perrot, J. Y.; Faussat, A. M.; Marie, J. P.; Schwaller, M. A. Functional assay of multidrug resistant cells using JC-1, a carbocyanine fluorescent probe. *Leukemia* **1997**, *11*, 1147−1155.

(39) Van Der Luit, A. H.; Budde, M.; Verheij, M.; Van Blitterswijk, W. J. Different modes of internalization of apoptotic alkyl-lysophospholipid and cell-rescuing lysophosphatidylcholine. *Biochem. J.* **2003**, *374*, 747−753.

(40) van der Luit, A. H.; Budde, M.; Ruurs, P.; Verheij, M.; van Blitterswijk, W. J. Alkyl-lysophospholipid accumulates in lipid rafts and induces apoptosis via raft-dependent endocytosis and inhibition of phosphatidylcholine synthesis. *J. Biol. Chem.* **2002**, *277*, 39541−39547.

(41) Bergmann, J.; Junghahn, I.; Brachwitz, H.; Langen, P. Multiple effects of antitumor alkyl-lysophospholipid analogs on the cytosolic free Ca^{2+} concentration in a normal and a breast cancer cell line. *Anticancer Res.* **1994**, *14*, 1549−1556.

treme GI_{50} values, having a median absolute value (across compounds and cell lines) of 1.28 compared to 0.62 for the other 39 compounds possessing the phosphate AAC. Second, if we calculate the GI_{50} profile based on the alkyl-lysophospholipid analogues and compare it to the profile obtained using all 52 compounds containing the AAC, the resulting profiles have a strong correlation of 0.87 even though the alkyl-lysophospholipids are a minority of the 52 compounds. Third, if we consider the correlations between individual non-alkyl-lysophospholipids and the alkyl-lysophospholipid-derived GI_{50} profile, the mean absolute correlation is only 0.17, and only 8 of 39 compounds have a correlation greater than 0.3. Taken together, this indicates that the phosphate and/or sulfonate GI_{50} profile is primarily determined by a cell type-specific cytotoxic activity of alkyl-lysophospholipid-like molecules and not by the cell type-specific targeting of phosphate and/or sulfonate prodrugs.

Interestingly, among the compounds that were clearly different from alkyl-lysophospholipid, several displayed a relatively high correlation with the mean alkyl-lysophospholipid GI_{50} levels. These included nucleoside phosphates such as NSC 670654 (not shown; $r = 0.72$) and phosphate prodrugs such as NSC 610458 (Figure 6B; camptothecin phosphate, $r = 0.57$). Camptothecin was present in the database in both phosphorylated and dephosphorylated forms. Unexpectedly, the GI_{50} values for camptothecin phosphate showed higher correlation with the alkyl-lysophospholipid signature ($r = 0.57$) than with dephosphorylated camptothecin ($r = 0.32$, 0.41, and 0.47 in three replications). This suggests that the cell type-specific targeting mechanism of camptothecin phosphate may be related to its similarity to alkyl-lysophospholipids, rather than to its similarity to camptothecin. Thus, camptothecin phosphate behaves as an alkyl-lysophospholipid analogue.

The potential mechanistic similarity between phosphate prodrugs and alkyl-lysophospholipids is important, because alkyl-lysophospholipids behave as surfactants[42] and affect the cell surface receptor-mediated intracellular signaling mechanism.[41,43] Interestingly, TNFR (tumor necrosis factor receptor) and PKIA (inhibitor subunit of protein kinase A) were among the 10 genes having strongest correlation with the P:O1;O1;O1;O2 GI_{50} profile ($r = -0.58$, -0.40, and -0.54 for TNFR and $r = -0.46$, -0.44, and -0.47 for PKIA, in triplicate experiments). Both TNFR and PKIA are components of important mediators of intracellular signal transduction pathways.

(3) Nucleobase and Nucleoside Analogues. Three of the AACs, C:C1;Cl1;N2, C:Cl1;N1;N2, and C:N2;S1;S1, are embedded in a chloropyrimidine [NSC 373854 and 699704 (Figure 7A) and NSC 58573 (Figure 7B)], chloropurine [NSC 614491 (Figure 7A) and NSC 685829 (Figure 7B)], or thiazolo/thiazolium [NSC 657718, 699666, and 699668 (Figure 7C)] substructures. These compounds exhibit similar patterns of resistance and sensitivity across the cell lines (Figure 2), and are particularly active against the CAKI-1 renal cell line (Figure 3), a cell line that tends to be relatively resistant to many other anticancer agents. While chlorinated pyrimidines and purines are clearly related to each other, their relationship with thiazolo/thiazolium compounds is less obvious.

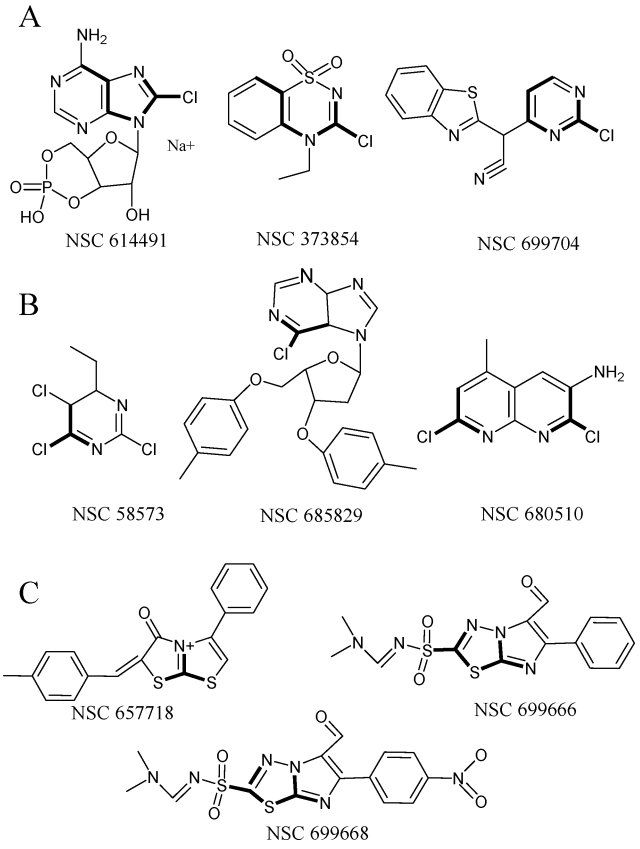


Figure 7. Selected compounds containing the expanded (A) C:C1;Cl1;N2, (B) C:Cl1;N1;N2, and (C) C:N2;S1;S1 AACs (expanded fragments in bold).

Analysis of gene expression data in relation to the cytotoxicity profiles of these compounds reveals that cell type-specific uptake and transport of nucleobase or nucleoside analogues[44−46] may constitute their primary targeting mechanism. There are two transporter genes that are particularly suggestive in this regard: the SLC29A1 gene overexpressed in sensitive cell lines ($r = -0.43$, -0.41, and

(42) Stafford, R. E.; Fanni, T.; Dennis, E. A. Interfacial properties and critical micelle concentration of lysophospholipids. *Biochemistry* **1989**, *28*, 5113−5120.

(43) Yan, J. J.; Jung, J. S.; Lee, J. E.; Lee, J.; Huh, S. O.; Kim, H. S.; Jung, K. C.; Cho, J. Y.; Nam, J. S.; Suh, H. W.; Kim, Y. H.; Song, D. K. Therapeutic effects of lysophosphatidylcholine in experimental sepsis. *Nat. Med.* **2004**, *10*, 161−167.

(44) Mangravite, L. M.; Badagnani, I.; Giacomini, K. M. Nucleoside transporters in the disposition and targeting of nucleoside analogs in the kidney. *Eur. J. Pharmacol.* **2003**, *479*, 269−281.

(45) Lu, X.; Gong, S.; Monks, A.; Zaharevitz, D.; Moscow, J. A. Correlation of nucleoside and nucleobase transporter gene expression with antimetabolite drug cytotoxicity. *J. Exp. Ther. Oncol.* **2002**, *2*, 200−212.

(46) Baldwin, S. A.; Beal, P. R.; Yao, S. Y.; King, A. E.; Cass, C. E.; Young, J. D. The equilibrative nucleoside transporter family, SLC29. *Pfluegers Arch.* **2004**, *447*, 735−743.

−0.51 for correlations with C:Cl1;N1;N2 in triplicate experiments; $r = -0.43$, −0.41, and −0.51 for correlations with C:C1;Cl1;N2 in triplicate experiments) and the ABCB6 gene overexpressed in resistant cell lines ($r = 0.55$, 0.43, and 0.50 for correlations with C:C1;Cl1,N2 in triplicate experiments; $r = 0.48$, 0.39, and 0.47 for correlations with C:Cl1;N1;N2 in triplicate experiments). For thiazoles, the SLC29A2 gene, a close relative of SLC29A1, is associated with sensitivity (Table 1 of the Supporting Information).

The mitochondrial localization and transport function of both SLC29 and ABCB6 genes also suggest a mitochondrial toxicity mechanism. The SLC29 genes are a family of nucleoside and nucleobase transporters, with a high level of expression in the kidney.[46] SLC29A1, also known as hENT1, is an equilibrative, nucleoside membrane transporter responsible for the disposition of nucleoside analogues in the urine.[44,46] SLC29A2 is a closely related nucleobase transporter.[46] Recently, localization of hENT1 to mitochondria suggests that it may also be a mitochondrial nucleotide exchanger involved in the transport of nucleosides across the mitochondrial membrane.[47] In this regard, hENT1 expression has been associated with the toxicity of a variety of antiviral and anticancer nucleosides.[44,45,47] Less is known about the function of the ABCB6 gene, although it is a mitochondrial membrane protein homologous to other transmembrane transporters,[48,49] and its mutation has been implicated in disorders of mitochondrial iron homeostasis.

(4) Organosulfur Compounds and Organometallic Complexes. The expanded template structures of four different AACs belong to a family of metal chelators and organometallic complexes. These AACs are C:N1;N1;S2 [thiosemicarbazones and hydrazinecarbothioamides (Figure 8A,B)], N:C1;N1 (Figure 8B,C), N:C1;H1;N1 (Figure 8A), and C:N1;S1;S2 [carbodithioates (Figure 8D)]. These metal complexes possess above average growth inhibitory activity against all leukemic cell lines, as well as against several solid tumor cell lines (Figures 2 and 3). Deducing drug sensitivity or resistance mechanisms from gene expression values is challenging in this case. Because of tissue-specific differences in gene expression, the toxicity and resistance of any agent that is active against a particular tissue type (e.g., leukemic cells) will correlate with every gene that is relatively over- or underexpressed in that tissue type. Since leukemic cells generally grow in suspension, they do not form an extracellular matrix and fail to express cytoskeletal genes involved in cell spreading and adhesion. For this reason, numerous genes such as collagens and other cytoskeletal constituents are associated with the toxicity profile of this class of compounds (Table 1 of the Supporting Information). Thus, while these associations may be mechanistically significant to some extent, it is likely that many of them are indirectly associated with toxicity. Consequently, they are considered uninformative with respect to the mechanism of action of the drug.

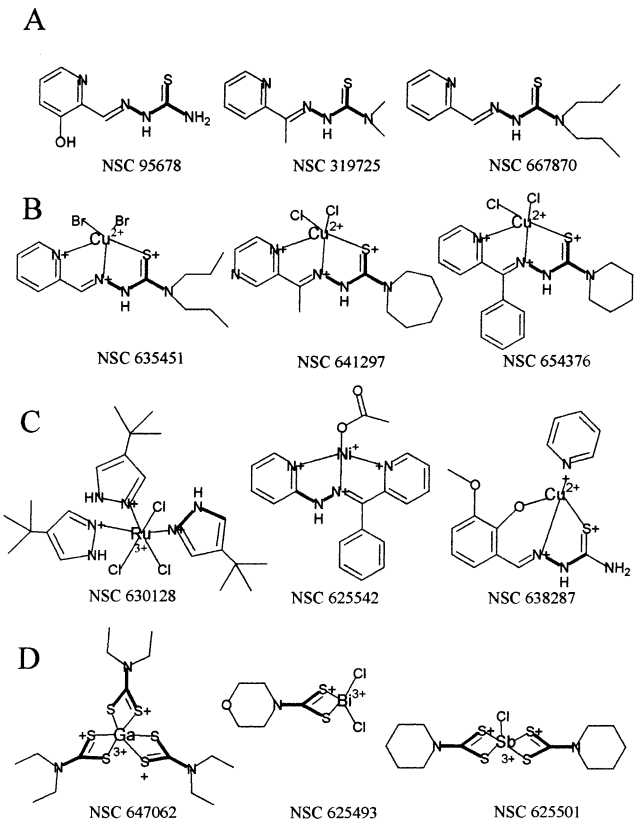


Figure 8. Selected compounds containing the expanded (A) N:C1;H1;N1 and C:N1;N1;S2, (B) N:C1;N1 and C:N1;N1;S2, (C) N:C1;N1, and (D) C:N1;S1;S2 AACs (expanded fragments in bold).

Therefore, instead of relying on gene expression data to analyze the targeting mechanism of the cluster, we could gain insights directly from structure−activity relationships observed in this subset of compounds. Expanded templates of the C:N1;N1;S2 AAC overlap with N:C1:N1 and N:C1;H1;N AACs (Table 1), forming the tridentate N*−N*−S* metal binding carbothioamide motif.[50−54] The uncomplexed N*−N*−S* thiosemicarbazones [NSC 95678, 319725, and 667870 (Figure 8A)] tend to be 1 or 2 orders

(47) Lai, Y.; Tse, C. M.; Unadkat, J. D. Mitochondrial expression of the human equilibrative nucleoside transporter 1 (hENT1) results in enhanced mitochondrial toxicity of antiviral drugs. *J. Biol. Chem.* **2004**, *279*, 4490−4497.

(48) Mitsuhashi, N.; Miki, T.; Senbongi, H.; Yokoi, N.; Yano, H.; Miyazaki, M.; Nakajima, N.; Iwanaga, T.; Yokoyama, Y.; Shibata, T.; Seino, S. MTABC3, a novel mitochondrial ATP-binding cassette protein involved in iron homeostasis. *J. Biol. Chem.* **2000**, *275*, 17536−17540.

(49) Lill, R.; Kispal, G. Mitochondrial ABC transporters. *Res. Microbiol.* **2001**, *152*, 331−340.

(50) Nocentini, G.; Federici, F.; Armellini, R.; Franchetti, P.; Barzi, A. Isolation of two cellular lines resistant to ribonucleotide reductase inhibitors to investigate the inhibitory activity of 2,2′-bipyridyl-6-carbothioamide. *Anticancer Drugs* **1990**, *1*, 171−177.

(51) Nocentini, G.; Barzi, A. The 2,2′-bipyridyl-6-carbothioamide copper(II) complex differs from the iron(II) complex in its biochemical effects in tumor cells, suggesting possible differences in the mechanism leading to cytotoxicity. *Biochem. Pharmacol.* **1996**, *52*, 65−71.

of magnitude less toxic than the complexed N*−N*−S* carbothioamides [NSC 635451, 641297, and 654376 (Figure 8B and data not shown)], indicating that the toxicity of these compounds is most likely mediated through the bound metal. Nevertheless, metal binding itself is not the primary determinant of the cell type-specific targeting mechanism, as uncomplexed carbothioamides (Figure 8A) possess cytotoxicity profiles similar to those of metal complexes (Figure 8B and data not shown).

Since N:C1;N1 AAC represents the most structurally diverse of all organometallic complexes and is often part of molecules that do not contain sulfur atoms [NSC 630128 and 625542 (Figure 8C)], it appears that the sulfur atom may not be an essential component for the targeting mechanism of this cluster of compounds. The N:C1;N1 AAC forms a complex with a variety of metals [for example, Ru (NSC 630128), Ni (NSC 625542), and Cu (NSC 638287); Figure 8C], in a variety of different binding modes. Some of the N:C1;N1 AAC-containing compounds are the tridentate N*−N*−S* carbothioamides in metal-bound form [NSC 635451, 641297, and 654376 (Figure 8B)], overlapping with the C:N1;N1;S2 AAC. Others are in a class of their own [for example, NSC 630128 and 625542 (Figure 8C)] and bear little resemblance to the other metal chelators that have been identified. Thus, the targeted activity of many organosulfur compounds appears to be associated with a general activation or resistance response to organometallic complexes, and not very specific to a particular type of complex.

Consistent with this observation, the C:N1;S1;S2 AAC [carbodithioates (Figure 8D)] identifies a related type of metal-chelating agent with a unique metal binding mode [NSC 647062, 625493, and 625501 (Figure 8D)]. While some of the C:N1;S1;S2 compounds bind to metals as monomers [NSC 625493 (Figure 8D)], others do so as dimers and trimers [NSC 625501 and 647062 (Figure 8D)]. Most importantly, unlike the case with other organometallic complexes, fragment expansion analysis of the C:N1;S1;S2 AAC reveals that antimony (Sb) complexes are almost exclusively associated with extreme toxicity signatures for this group of compounds (data not shown). Whether Sb increases the targeted cytotoxicity of the thiosemicarbazones and hydrazinecarbothioamides is a question that would need to be addressed experimentally.

(5) Thiocolchicines and Podophyllotoxins. The C:H1;H1;H1;S1 [methylthioether (Figure 9A)] and C:H1;H1;O1;O1 AACs [1,3-dioxo (Figure 9B)] are embedded in chemical structures revealing a remarkable degree of structural relatedness. In the case of the C:H1;H1;H1;S1 compounds [for example, NSC 186301, 111702, and 650773 (Figure 9A)], the C:H1;H1;H1;S1 AAC is often attached to a rigid, planar scaffold. A di- or trimethoxy- or hydroxyphenyl group is generally found at the opposite end of this structure. Similarly, in the case of 1,3-dioxo compounds, a rigid, heterocyclic aromatic system bridges the C:H1;H1;O1;O1 AAC to a mono-, di-, or trimethoxyphenyl group [for example, NSC 211488, 403148, and 350102 (Figure 9B)]. Compounds containing either of these AACs are active against a broad variety of cell lines from different tissues (Figure 2), yet tend to be less active against TK-10 (renal), OVCAR-4 and OVCAR-5 (ovarian), EKVX (non-small cell lung cancer), T-47D (breast), and UACC257 (melanoma) cells (Figures 2 and 3).

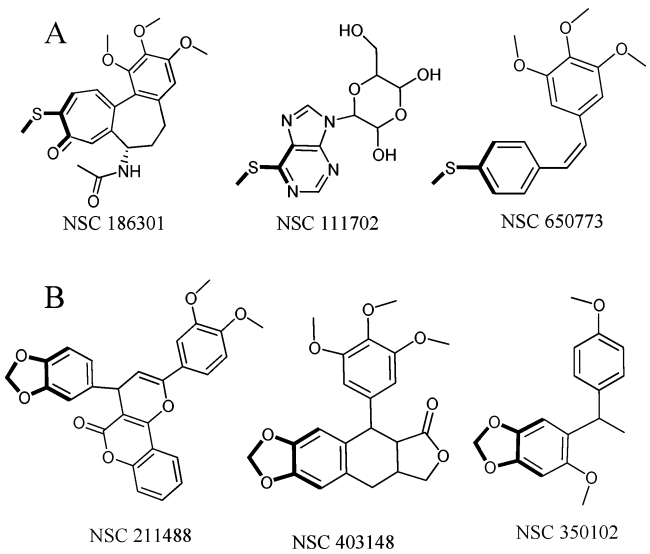


Figure 9. Selected compounds containing the expanded (A) C:H1;H1;H1;S1 and (B) C:H1;H1;O1;O1 AACs (expanded fragments in bold).

Analysis of the scientific literature revealed that the similarities between these compounds are most likely related to their activity as microtubule inhibitors, and the observed SAR is related to their tubulin binding mode.[55−59] The C:H1;H1;H1:S1 AAC is characteristic of the thiocolchicines [for example, NSC 186301 (Figure 9A)], a family of well-characterized microtubule-depolymerizing compounds, while the C:H1;H1;O1;O1 AAC is characteristic of podophyllotoxins [NSC 211488 and 403148 (Figure 9B)], a different

(52) Nocentini, G.; Barzi, A. Antitumor activity of 2,2′-bipyridyl-6-carbothioamide: a ribonucleotide reductase inhibitor. *Gen. Pharmacol.* **1997**, *29*, 701−706.

(53) Antonini, I.; Cristalli, G.; Franchetti, P.; Grifantini, M.; Martelli, S.; Filippeschi, S. 2,2′-Bipyridyl-6-carbothioamide derivatives as potential antitumor agents. *Farmaco* **1986**, *41*, 346−354.

(54) Antonini, I.; Claudi, F.; Cristalli, G.; Franchetti, P.; Grifantini, M.; Martelli, S. N*-N*-S* tridentate ligand system as potential antitumor agents. *J. Med. Chem.* **1981**, *24*, 1181−1184.

(55) Wolff, J.; Knipling, L.; Cahnmann, H. J.; Palumbo, G. Direct photoaffinity labeling of tubulin with colchicine. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 2820−2824.

(56) Sackett, D. L. Podophyllotoxin, steganacin and combretastatin: natural products that bind at the colchicine site of tubulin. *Pharmacol. Ther.* **1993**, *59*, 163−228.

(57) Luduena, R. F.; Roach, M. C. Tubulin sulfhydryl groups as probes and targets for antimitotic and antimicrotubule agents. *Pharmacol. Ther.* **1991**, *49*, 133−152.

(58) Correia, J. J. Effects of antimitotic agents on tubulin-nucleotide interactions. *Pharmacol. Ther.* **1991**, *52*, 127−147.

(59) Burns, R. G. Analysis of the colchicine-binding site of β-tubulin. *FEBS Lett.* **1992**, *297*, 205−208.

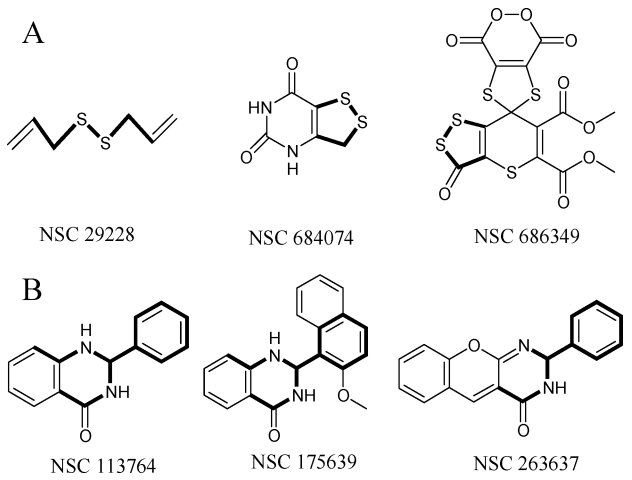


Figure 10. Selected compounds containing the expanded (A) S1:C1;S1 and (B) C:C1;H1;N1;N1 AACs (expanded fragments in bold).

class of well-characterized microtubule inhibitors that are known to interact with the colchicine-binding site of tubulin. While several extreme compounds in this cluster are neither thiocholchicines nor podophyllotoxins [for example, purine compound NSC 111702 and styryl compound NSC 650773 (Figure 9A)], key structural features necessary for tubulin binding are readily apparent.

(6) Disulfides and Quinazolinones. Lastly, less information could be gleaned by analyzing the structure of the remaining two classes of compounds, illustrating the limitations of the approach. The S:C1;S1 AAC characterizes compounds containing a disulfide bond (Figure 10A) and could not be expanded to a larger substructure template (Figure 4). Visual inspection of these compounds revealed a remarkably degree of structural diversity (Figure 10A), represented by the diallyl sulfide (NSC 29228), dithiolenouracil (NSC 684074), and other thiopyran-containing compounds (NSC 686349). Like the organometallic complexes, disulfide-containing compounds exhibit above average toxicity against all leukemic cells. This makes it difficult to identify genes that are specifically associated with the molecular mechanism of action of these compounds.

The expanded C:C1;H1;N1;N1 AAC template identifies a family of closely related quinazolinones [NSC 113764, 175639, and 263637 (Figure 10B)]. These quinazolinones are most active against the multidrug resistant NCI-ADR and HCT15 cell lines (Figures 2 and 3), which is suggestive of their ability to act as P-glycoprotein inhibitors as reported in the scientific literature.[60] However, gene expression data reveal that the activity of these compounds is correlated with genes whose mechanistic significance is unclear. These genes include histone deacetylase 9 ($r = 0.39$, 0.44, and 0.46, in triplicate experiments) and thymosin $\beta 4$ ($r = 0.45$, 0.42, and 0.46, in triplicate experiments), and are negatively correlated with the expression of the TSSC3 gene ($r = -0.44$, -0.43, and -0.45, in triplicate experiments).

(60) Wang, S.; Ryder, H.; Pretswell, I.; Depledge, P.; Milton, J.; Hancox, T. C.; Dale, I.; Dangerfield, W.; Charlton, P.; Faint, R.; Dodd, R.; Hassan, S. Studies on quinazolinones as dual inhibitors of Pgp and MRP1 in multidrug resistance. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 571−574.

Discussion

We have developed a chemoinformatic analysis strategy and applied it to the NCI's anticancer agent database to identify candidate molecular mechanisms conferring cell type-selective anticancer drug activity. For analysis, each compound in a database is computationally broken down into fragments comprised of an individual atom bonded to its immediate neighbors (AAC). AACs are then used to identify associated toxicity signatures using a rigorous statistical comparison between compounds containing and lacking the AAC. Once a toxicity signature is identified, the AAC is expanded, yielding a larger substructure associated with toxicity. Gene expression measurements on the 59 cell lines are then used to infer biological mechanisms underlying the associations between chemical structures and toxicity profiles. This strategy should also be generally applicable to the exploration of complex chemical structure−activity relationships in large data sets.

At the outset, we hypothesized that individual AACs could serve as "markers" for larger structural motifs associated with specific bioactivity profiles, just as conserved amino acids in the active site of proteins serve to identify classes of proteins catalyzing a specific type of chemical reaction. Individual AACs could be expanded to familiar substructures associated with known drug toxicity and resistance mechanisms, confirming our hypothesis. In addition, analysis of gene expression patterns allows us to infer several hypothetical associations between specific chemical structures and genes over- or underexpressed in sensitive and resistant cell lines. One of these associations involves a mitochondrial/P-glycoprotein detoxification axis, associated with lipophillic cations harboring the N:C1;C1;C2 AAC. Another one of these associations involves a relationship among certain classes of sulfonate compounds, phosphate prodrugs, and the mechanism of action of alkyl-lysophospholipid analogues. Yet another one involves an association among a variety of chloropyrimidines, chloropurine, and thiazole anticancer agents and membrane transport mechanisms previously associated with mitochondrial toxicity. Lastly, we have identified a large group of organosulfur compounds whose cell type-selective activity appears to be more closely associated with the bound metal than with the chemical structure of the chelator.

What are the advantages and disadvantages of using AACs as a starting point for SAR studies in cell-based assays? Previously, it had been shown that the cell type-selective toxicity of small molecules against multiple cancer cell lines is dependent on substructural motifs.[5,7,12,16,17] This is consistent with the fact that the cytotoxic or growth inhibitory activity of molecules is thought to depend on interaction with protein targets. Because most of these interactions generally occur at pockets or clefts in the three-dimensional structure

of proteins, most bioactive structures are associated with molecules whose size and shape can be accommodated by these pockets or clefts. Yet, the usefulness of large substructural motifs for data mining efforts is limited by how substructural motifs are defined in chemical space, and by the number and diversity of compounds harboring those motifs. As an alternative to "fingerprinting" approaches for examining many molecular descriptors in parallel to assess global similarities,[19] AACs offer a simple, economical, all-or-none measure of local similarity. Our results establish that AACs contain relevant information about a compound's cytotoxicity profile.

In the case of AACs, because the minimum number of compounds representing each AAC is large (>30), the contribution of a particular AAC to the cell targeting mechanism is predetermined to be statistically significant from the outset. In this sense, AACs provide a highly reliable marker with which to discover associations between chemical structures and cytotoxicity profiles. Compared to other substructure-based chemoinformatic analysis,[5] our results demonstrate that using a relatively small number of AACs for the initial screen can aid in elucidation of larger structural motifs involved in the molecule's mechanism of action, once the AACs are expanded. Although the starting number of AACs (747) is far smaller than the number of substructures employed in other analyses, fewer false positives are expected with the AACs. However, it can be argued that the AACs will miss some important substructures, and therefore that the method is less sensitive than other methods. To surmount this limitation, linked AACs (pairs, triplets, or *n*-plets) could be used as starting points to identify structure−activity relationships. For simplicity, the current analysis was limited to individual AACs as starting points.

As a caveat, while the results demonstrate the usefulness of this approach for studying the NCI anticancer database, there are also limitations to the approach that could limit its application to other databases; identifying associations between AACs and bioactivity profiles is highly dependent on the size and chemical diversity represented in the collection of compounds being analyzed. First, the ability to find an association depends on the degree of structural diversity associated with the library, as determined by the total number of AACs represented in the library. If diversity is too limited, then most of the molecules in the library will share a small number of AACs. If diversity is too broad, then each AAC will be represented by too few molecules to arrive at a statistically significant association. In addition, identification of AACs determining the functional properties of compounds also depends on the size of the library and the number of AACs represented in each molecule. Finally, the ability to identify association of AACs and bioactivity profiles are dependent on the statistical variance observed in the bioactivity profiles.

Nevertheless, used as a starting point for exploratory analysis of complex structure−GI profile relationships in the NCI database, AACs clearly offer an alternative approach to more complex, substructure similarity-based searches. That AACs serve as markers to identify larger structural motifs associated with specific bioactivity profiles can be explained for the same reasons that conserved amino acids in the active sites of proteins are able to yield useful information about the function of an entire protein. While the AAC alone may not be sufficient for determining a specific type of activity, if the AAC is embedded in a molecule and surrounded by other favorable structural features, the AAC may become the primary determinant of the functional feature of the entire molecule. Consistently, the results demonstrate that if AACs are found to be associated with specific toxicity profiles, it is possible to expand the AAC into a larger substructural template that is important for the toxicity property of the AAC.

In conclusion, the ability to use AACs as a starting point for elucidating structure−activity relationships in the NCI data set should open another window for elucidating the molecular mechanisms targeting the activity of anticancer agents to certain types of cancer cells. On the basis of similarities between the structure of the compounds and the correlation between the activities of the compounds on different cell lines with gene expression differences in those cell lines, it has been possible to infer hypothetical relationships between the structures of compounds and their cell growth inhibitory activity. While experiments are already underway to test these hypotheses, the ability to derive novel, meaningful mechanistic hypotheses from these relationships indicates that mining large, complex data sets should become increasingly relevant for anticancer drug targeting efforts.

Acknowledgment. We thank G. Crippen for critical reading of the manuscript. This work was funded by a pilot grant from the Bioinformatics Program at the University of Michigan, Pfizer, Inc., and the Howard Hughes Medical Institute to G.R.R. and K.S., and an Upjohn-Vahlteich Award to G.R.R.

Supporting Information Available: Correlation of gene expression and AAC GI_{50} measurements across all cell lines or excluding leukemic cell lines and a list of extreme and average subsets of compounds containing each AAC. This material is available free of charge via the Internet at http://pubs.acs.org.

MP049953K