

# Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data

PE Blower<sup>1</sup>  
C Yang<sup>1</sup>  
MA Fligner<sup>2</sup>  
JS Verducci<sup>2</sup>  
L Yu<sup>1</sup>  
S Richman<sup>3</sup>  
JN Weinstein<sup>3</sup>

<sup>1</sup>Leadscope Inc, Columbus, OH, USA;  
<sup>2</sup>Department of Statistics, The Ohio State University, Columbus, OH, USA; <sup>3</sup>Laboratory of Molecular Pharmacology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Correspondence: PE Blower, Leadscope Inc, 1275 Kinnear Road, Columbus, OH 43212, USA  
Tel: 001 641 675 376  
E-mail: Pblower@leadscope.com  
<http://www.leadscope.com>  
JN Weinstein, Building 37, Room 5068, NIH, 9000 Rockville Pike, Bethesda, MD 20892, USA  
Tel: 001 301 496 9571  
E-mail: jw4i@nih.gov  
<http://discover.nci.nih.gov>

## ABSTRACT

Genomic studies are producing large databases of molecular information on cancers and other cell and tissue types. Hence, we have the opportunity to link these accumulating data to the drug discovery processes. Our previous efforts at 'information-intensive' molecular pharmacology have focused on the relationship between patterns of gene expression and patterns of drug activity. In the present study, we take the process a step further—relating gene expression patterns, not just to the drugs as entities, but to ~27 000 substructures and other chemical features within the drugs. This coupling of genomic information with structure-based data mining can be used to identify classes of compounds for which detailed experimental structure-activity studies may be fruitful. Using a systematic substructure analysis coupled with statistical correlations of compound activity with differential gene expression, we have identified two subclasses of quinones whose patterns of activity in the National Cancer Institute's 60-cell line screening panel (NCI-60) correlate strongly with the expression patterns of particular genes: (i) The growth inhibitory patterns of an electron-withdrawing subclass of *benzodithiophenedione*-containing compounds over the NCI-60 are highly correlated with the expression patterns of *Rab7* and other melanoma-specific genes; (ii) the inhibitory patterns of *indolonaphthoquinone*-containing compounds are highly correlated with the expression patterns of the hematopoietic lineage-specific gene *HS1* and other leukemia genes. As illustrated by these proof-of-principle examples, we introduce here a set of conceptual tools and fluent computational methods for projecting directly from gene expression patterns to drug substructures and *vice versa*. The analysis is presented in terms of the NCI-60 cell lines and microarray-based gene expression patterns, but the concept and methods are broadly applicable to other large-scale pharmacogenomic database sets as well. The approach (SAT for Structure-Activity-Target) provides a systematic way to mine databases for the design of further structure-activity studies, particularly to aid in target and lead identification. *The Pharmacogenomics Journal* (2002) 2, 259–271. doi:10.1038/sj.tpj.6500116

**Keywords:** molecular class; NCI cell lines; statistical correlation

## INTRODUCTION

Genomic and proteomic technologies will revolutionize drug discovery and development. That much is universally agreed. But, to date, pharmacological and biological databases have not been linked in a way that permits fluent exploration of the relationships that a medicinal chemist or drug designer would most like to see—relationships between the molecular structural features of compounds and the genes or gene products in cells that predict activity of the compound against the cell. The principal aim of this paper is to present a method for doing so. We will use, as an example, the set of compounds tested in the

National Cancer Institute's 60-cell line anti-cancer drug screen and a microarray-generated database of gene expression profiles obtained for the 60 cell lines.<sup>1,2</sup> However, the formalism and methods can be applied more broadly in the context of pharmacogenomic studies.

Since 1990, the National Cancer Institute (NCI) has screened more than 70 000 chemical compounds for their growth inhibitory activity against a panel of 60 human cancer cell lines (the NCI60) in microtiter plate format.<sup>3–5</sup> For each compound and cell line, growth inhibition after 48 h of drug treatment is assessed from changes in total cellular protein using a sulforhodamine B assay. A vector of 60 growth inhibition values, one for each cell line, represents the activity pattern of a compound. These patterns of drug activity across the NCI60 have been found to encode incisive information about mechanisms of drug activity.<sup>3–9</sup> The utility of that information has been enhanced by correlating the activity patterns with molecular structure descriptors of the tested compounds<sup>10,11</sup> and with molecular characteristics of the test cell lines or tissue types.<sup>1,2,6,12–21</sup> The NCI60 panel includes melanomas (8 cell lines), leukemias (6), and cancers of breast (8), prostate (2), lung (9), colon (7), ovary (6), kidney (8) and central nervous system (6) origin.

In a recent study,<sup>1–2</sup> one of the present authors and collaborators used cDNA micro-arrays to generate gene expression profiles for the NCI60 cells. The database generated is being applied to problems in cancer diagnosis, prognosis, prevention and therapy. Most pertinent to the present study, these biological data can be mapped into pharmacological information, as described in Scherf *et al.*<sup>1</sup> However, the correlation itself does not provide the type of insight most useful for drug design and selection of therapy. From the medicinal chemist's or drug designer's perspective, in particular, a method for projecting the genomic information through the pharmacology to molecular descriptors and structural features would be highly desirable. This paper describes such a method. It uses data mining software<sup>22</sup> to identify structural features that are found in compounds whose activity patterns are highly correlated with expression patterns of selected genes. Representative compounds are then used to probe relevant genes and thus gain insight into possible molecular mechanisms of drug action.

The conceptual background of this work<sup>6</sup> involves the databases shown in Figure 1. Database [A] contains the activity patterns of tested compounds, [S] contains molecular structural features of the compounds, and [T] contains gene expression patterns, including those for possible targets or modulators of activity in the cells. The databases [A] and [T] to be analyzed here are publicly available at <http://www.leadscope.com> and <http://discover.nci.nih.gov>. For [S] in this study we used a set containing 27 000 2D structural features.<sup>22</sup>

In practice, the overall [T]-database can include cell properties at the DNA, RNA, protein, functional and pharmacological levels, but in the present analysis we will consider only mRNA transcript expression patterns. More specifically, we will analyze transcript patterns obtained for the 60 cell lines using pin-spotted cDNA microarrays<sup>1,2</sup> pre-

pared by robotically spotting 9706 human cDNAs on glass microscopic slides. The cDNAs represented approximately 8000 different genes.<sup>1</sup> Approximately 3700 of them were previously characterized human proteins, an additional 1900 had homologues in other organisms, and the remaining 4100 were identified only as ESTs. cDNA prepared from a pool of 12 cell lines selected for diversity from the set of 60 was used as an internal standard.<sup>2</sup>

The chemical structure database [S] can, in principle, be encoded in terms of any set of 1-, 2-, or 3-dimensional molecular structural descriptors or physico-chemical properties that are experimentally measured or theoretically calculated. The set of 27 000 2D structural features<sup>22</sup> used here includes the familiar chemical building blocks of a compound: functional groups, aromatics, heterocycles, pharmacophores, etc organized hierarchically from general to specific. Each row in the database [S] corresponds to a structural feature, F. For compound C and feature F, the entry  $S_{FC} = 0$  if F does not occur in C; otherwise,  $S_{FC} = 1/N_F$ , where  $N_F$  is the number of compounds containing F.

To relate the drug activity profiles to the gene expression patterns, activity and expression values were standardized, and the matrix  $[A \bullet T^T]$  of Pearson product-moment correlation coefficients was then obtained by matrix multiplication. Finally, the matrix  $[S \bullet A \bullet T^T]$  (see Figure 1) was generated to associate a structural feature F with a gene. Each element in the  $[S \bullet A \bullet T^T]$  matrix reflects the tendency of a particular substructure to occur in compounds that are active in cell lines that express large amounts of the given gene product. Whereas previous studies of these databases have focused on identification of compounds, the present development (SAT for Structure-Activity-Target) takes this process a step further, identifying structural features that are associated with the observed correlations between gene expression and growth inhibition.

Other techniques have been used to model the anti-cancer activity of NCI compounds using various molecular property sets. For example, Shi *et al.*<sup>10,11</sup> calculated chemical structural descriptors of ellipticine analogs using molecular modeling software and derived correlations with growth inhibition using genetic function approximation. Fan *et al.*, used similar methods to analyze topoisomerase I inhibitors.<sup>23</sup> Cho *et al.*<sup>24</sup> analyzed structure-activity relationships for the NCI-H23 cell line using a recursive partitioning technique with several types of atom and physico-chemical property class descriptors. The MCASE program, developed by Klopman *et al.*<sup>25,26</sup> was used to dissect the compounds of an NCI60 training set into all possible structural fragments with 2–10 non-hydrogen atoms and statistically correlate each fragment with drug activity. In that study, MCASE identified multi-drug resistance (MDR) reversal agents. Roberts *et al.*<sup>22</sup> described general purpose data mining techniques that combine structural analysis and dynamic property filtering, and illustrated how these techniques could be used to identify compound classes in the NCI cancer screening data with unexpectedly high growth inhibitory activity. Each of these analyses can be thought of as implicitly or explicitly involving an  $[S \bullet A]$  matrix but not an  $[S \bullet A \bullet T^T]$  matrix. In other

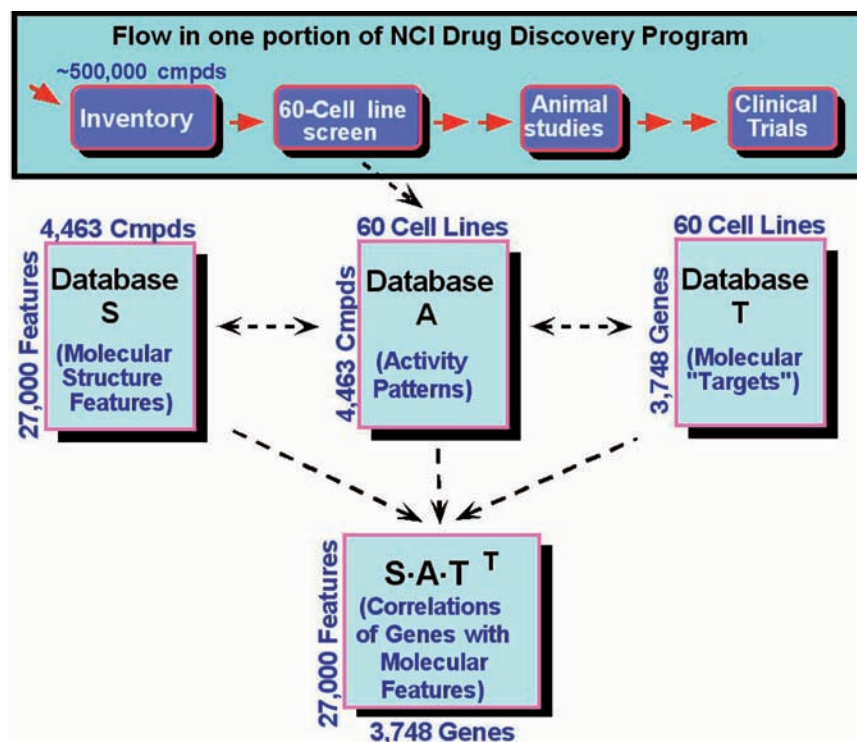


Figure 1 Conceptual framework for statistical analysis relating structural features of compounds to patterns of gene expression in the NCI60 human cancer cell lines. Database [A] contains compound activity patterns, [S] contains molecular structural features of the tested compounds and [T] contains differential gene expression for potential molecular targets in the cells. Modified from Figure 1 of Reference 6.

words, they did not carry the analysis through to the gene expression level as is done here. The present study describes a way to explore the large-scale correlation of molecular structural features with molecular characteristics of cells. It provides a systematic and fluent approach to the mining of genomic, proteomic and other 'omic',<sup>27–30</sup> databases to aid in the identification of new targets and lead compounds.

## RESULTS

### Cell Groupings and Genes Selected for Analysis

When the NCI60 cell lines were clustered on the basis of gene expression pattern, they tended to group by tissue of origin,<sup>1,2</sup> but there were exceptions. Table 1 lists seven relatively robust cell clusters based on the hierarchical cluster tree in Figure 2a of Reference 2. Grouping by tissue of origin is most apparent for the CNS, colon, leukemia, melanoma and renal panels. The melanoma cluster included two lines, MDA-MB-435 and MDA-N, that were derived from a pleural effusion in a patient with breast cancer, but showed the distinctive gene expression patterns of melanotic melanomas. MDA-N is an erb/B2 transfectant of MDA-MB-435. MDA-MB-435 may represent a second primary in the patient, a breast cancer with strong neuroendocrine features, or a contamination prior to derivation of MDA-N. The LOX IMVI cell line, which was not included in the melanoma cluster, is amelanotic and undifferentiated; it lacks the set of genes characteristic of melanin production and melanoma.<sup>1,2</sup> The

breast and non-small cell lung cancers did not form large coherent groups, and there were only two prostate cancer lines. The CNS cell cluster contained one breast cancer line, BT-549. The ovarian cancers were intermediate, with four cell lines in a coherent group and the other two elsewhere in the tree. There was a mixed cluster of 11 cell lines containing three non-small cell lung cancers, two each of breast, ovarian and prostate, and one each of melanoma and renal. Nine additional cell lines were not included in the cell clusters because they belonged to no coherent group with more than three cell lines. Some of these groupings are, to a degree, arbitrary, but the melanotic melanoma and leukemia categories are well defined.

Using the Studentized range test based on the seven cell panels, we selected 476 genes to distinguish 'extreme' panels, ie panels whose average expression level for the gene was significantly higher, or lower than those seen in other panels. Partial results are shown in Table 2, which reports genes with high Studentized range scores. The first column gives the clone identifier, and the last column gives a brief description of the corresponding gene identity. The second column contains the value of the Studentized range statistic, and the columns labeled *Min* and *Max* are the cell clusters that yield the minimum and maximum, respectively, of the means in equation (2). The column labeled *upper 10% mean* gives the average correlation of the highest 10% of the 4463

**Table 1** Cell line clusters

Panel*	Cell lines
Leukemia:	LE:CCRF-CEM, LE:HL-60(TB), LE:K-562, LE:MOLT-4, LE:RPMI-8226, LE:SR
Colon:	CO:COLO 205, CO:HCC-2998, CO:HCT-116, CO:HCT-15, CO:HT29, CO:KM12, CO:SW-620
Ovarian:	OV:IGR-OV1, OV:OVCAR-3, OV:OVCAR-4, OV:SK-OV-3
Melanoma:	BR:MDA-MB-435**, BR:MDA-N**, ME:M14, ME:MALME-3M, ME:SK-MEL-2, ME:SK-MEL-28, ME:SK-MEL-5, ME:UACC-257, ME:UACC-62
Misc:	BR:MDA-MB-231-ATCC, UK:NCI-ADR-RES, LC:HOP-62, LC:HOP-92, LC:NCI-H226, ME:LOX-IMVI, OV:OVCAR-5, OV:OVCAR-8, PR:DU-145, PR:PC-3, RE:SN12C
Renal:	RE:786-0, RE:A498, RE:ACHN, RE:CAKI-1, RE:RXF-393, RE:TK-10, RE:UO-31
CNS:	BR:BT-549, CNS:SF-268, CNS:SF-295, CNS:SF-539, CNS:SNB-19, CNS:SNB-75, CNS:U251
Not included:	BR:HS-578T, BR:MCF7, BR:T-47D, LC:A549-ATCC, LC:EKVX, LC:NCI-H23, LC:NCI-H322M, LC:NCI-H460, LC:NCI-H522

Clusters are based on the cluster tree shown in Figure 2a of Scherf *et al.*<sup>2</sup> The melanoma cluster includes two breast cancer lines, MDA-MB-435 and MDA-N, which show the distinctive expression patterns of melanomas. The LOX IMVI cell line, which was not included in the melanoma cluster, is undifferentiated and amelanotic. It lacks the set of genes characteristic of melanin production and melanoma. The CNS cluster contains one breast cancer line, BT-549.

\*LE: leukemia, CO: colon, OV: ovarian, ME: melanoma, RE: renal, CNS: central nervous system, BR: breast, LC: non-small cell lung, PR: prostate, UK: unknown. Cells were assigned to panels heuristically on the basis of gene expression patterns in Figure 2a of Reference 2.

\*\*MDA-MB-435 and its Erb/B2 transfectant MDA-N were included with the melanotic melanomas for reasons explained in the text.

gene–compound correlations for the gene.

Of the 476 genes in the full version of Table 2, 391 (82%) had melanoma or leukemia as the *Min* or *Max* cell cluster. Guided by values for the upper 10% mean for genes in Table 2, we identified several genes that are selectively over-expressed in melanoma cell lines and well correlated with the activity patterns of specific compounds. These included *Rab7* human small GTP binding protein<sup>31–33</sup> (ID 486233), *ASAH* lysosomal ceramidase<sup>34</sup> (ID 363919), human mRNA for KIAA0110<sup>35</sup> (ID 323730), *MMP14* membrane-type matrix metalloproteinase<sup>36,37</sup> (ID 270505), *hMYH* human mutY<sup>38,39</sup> (ID 268727), *RET* Ret proto-oncogene (ID 485268), and human fetal brain mRNA for vacuolar ATPase (ID 488599). Separately, we found several other genes showing analogous behavior for leukemias. Included were *LCPI* lymphocyte cytosolic protein 1<sup>40,41</sup> (L-plastin, ID 486676), *HS1*<sup>42,43</sup> hematoopoietic lineage cell-specific protein (ID 260052), *SF2* pre-

mRNA splicing factor (ID 357011), *CENPC* centromere auto-antigen C (ID 488194), and *CARS-cyp* human Clk-associated RS cyclophilin<sup>44</sup> (ID 179994).

For each of these genes, the corresponding column of the [S•A•T<sup>T</sup>] matrix contains the average correlation coefficient between activity and expression level for each of the structural features. In other words, this column identifies structural features of compounds for which the compound activities are well correlated with the expression patterns of the specific genes. For each gene, these correlations were standardized over all features to create a feature *z*-score. Figure 2 shows two compound classes, *benzothiophenedione*<sup>45,46</sup> and *indolonaphthoquinone*,<sup>47,48</sup> that are well correlated with several melanoma genes and several leukemia genes, respectively. That is, they have the highest feature *z*-scores for these genes.

### Anticancer Quinones

Since both compound classes in Figure 2 are heterocyclic quinones, we surveyed other classes of cytotoxic quinone anti-cancer agents.<sup>49–51</sup> There have been a number of recent studies of *indolequinone* anti-tumor agents<sup>52–55</sup> based on *Mitomycin C* and the aziridiny analog *EO9* (see Figure 3). These compounds are substrates for NAD(P)H:quinone oxidoreductases (NQO1 and DT diaphorase), and some show selective activity against DT diaphorase-rich<sup>55</sup> cell lines. A number of topoisomerase II inhibitors<sup>56</sup> such as *doxorubicin* (*adriamycin*), *daunorubicin*, and *mitoxanthrone* (see Figure 3) possess an *anthraquinone* substructure. *Actinomycin D* is a DNA inter-calating agent that contains a *quinoneimine* in a *phenoxazine* ring. Lastly, some *naphthimidazole-4,9-diones* show good activity and selectivity in the NCI60 panel. Based on these anti-cancer agents, we constructed a substructure query for each class. The results shown in Table 3 give information on the average correlations of common quinone classes with several of the genes listed above.

The two substructures in Figure 2 resulted in a class of 23 *dihydrobenzodithiophene-4,8-diones*<sup>45,46</sup> and a class of 20 *indolo-1,4-naphthoquinones*,<sup>47,48</sup> respectively. These two classes had the highest and second highest feature *z*-scores with *Rab7* and *LCPI*, respectively, of any structural feature we studied for these genes. Furthermore, among the 407 quinones in the full set of 4463 compounds, nine of the 22 compounds best correlated with gene *Rab7* were *benzodithiophenediones*, and 13 of the 54 best correlated with *LCPI* were *indolonaphthoquinones*.

Table 3 shows the feature *z*-scores for several classes of compounds using selected genes from Table 2. For example, using compound–gene correlation coefficients for the *Rab7* gene<sup>31–33</sup> from Table 2, the *z*-score for the benzothiophenedione subset is 10.5. Thus, the *benzothiophenedione* class is highly enriched with compounds for which the 60-cell activity patterns are well correlated with the expression pattern of *Rab7*. Similarly, using compound–gene correlation coefficients for the *LCPI* gene<sup>40,41</sup> from Table 2, the *z*-score for the *indolonaphthoquinone* subset is 5.64.

For comparison, the right four columns in Table 3 give average GI<sub>50</sub> values and feature *z*-scores for the compound

**Table 2** Selected genes with high values of studentized range statistic over cell clusters

IMAGE ID	Student range	Min	Max	Upper 10% mean	Description
486844	16.06	col	cns	0.34	GJA1 gap junction alpha-1 protein Chr.6
488479	13.92	leu	ren	0.24	TPM1 tropomyosin alpha chain (skeletal muscle) Chr.15
486471	12.82	leu	cns	0.30	CALD1 caldesmon Chr.7 actin- and myosin-binding protein
343073	11.76	leu	ren	0.19	RGS12 homo sapiens regulator of G-protein signalling 12 (RGS12) mRNA, complete cds Chr.4
417819	11.61	mis	mel	0.29	ASAH N-acylsphingosine amidohydrolase (acid ceramidase) Chr.8
363919	11.44	mis	mel	0.26	ASAH N-acylsphingosine amidohydrolase (acid ceramidase) Chr.8
428625	11.21	ren	leu	0.43	H.sapiens mRNA for orphan nuclear hormone receptor Chr.1
324181	10.95	ova	mel	0.27	SIAT4C H.sapiens mRNA for Gal-beta(1-3/1-4)GlcNAc alpha-2.3-sialyltransferase Chr.11
417918	10.57	ren	leu	0.45	NR113 H.sapiens mRNA for orphan nuclear hormone receptor Chr.1
260052	10.26	mis	leu	0.64	HCLS1 hematopoietic lineage cell specific protein Chr.3
307225	10.26	mel	leu	0.45	LBR lamin B receptor Chr.1
323730	10.18	ova	mel	0.37	Human mRNA for KIAA0110 gene, complete cds Chr.6
489181	10.01	leu	cns	0.27	THBS1 thrombospondin 1 Chr.15
486787	9.46	leu	cns	0.23	CNN3 calponin 3, acidic Chr.1
486676	8.75	mis	leu	0.51	LCP1 lymphocyte cytosolic protein 1 (L-plastin) Chr.13
509943	8.15	ova	leu	0.46	DKC1 dyskeratosis congenita 1 dyskerin,multifunctional protein,mainly expressed in the rapidly dividing cells of the epithelia and the hemopoietic system
270505	8.01	leu	mel	0.26	MMP14 H.sapiens mRNA for membrane-type matrix metalloproteinase 1 Chr.
486233	7.38	leu	mel	0.39	RAB7 human small GTP binding protein mRNA, complete cds Chr.3 protein transport probably involved in vesicular traffic
268727	7.05	leu	mel	0.29	Human mutY homolog (hMYH) gene, complete cds Chr.1
357011	6.99	mel	leu	0.52	SRP46 splicing factor, arginine/serine-rich, 46kD
61531	6.67	ren	leu	0.51	ZNF24 human zinc finger protein mRNA, complete cds Chr.10
161373	6.66	cns	leu	0.47	Human PMS6 mRNA (yeast mismatch repair gene PMS1 homologue), partial cds (C-terminal region) Chr.7
486751	6.43	col	mel	0.28	PIGF phosphatidylinositol glycan, class F Chr.2
127821	6.03	leu	mel	0.28	ACP5 acid phosphatase type 5 Chr.19
488599	5.89	col	mel	0.42	ATP6S14 human fetus brain mRNA for vacuolar ATPase, complete cds Chr.12
488194	5.66	mel	leu	0.53	CENPC centromere autoantigen C Chr.4
179994	5.51	mel	leu	0.55	PPIG peptidyl-prolyl isomerase G (cyclophilin G) human Clk-associated RS cyclophilin CARS-Cyp mRNA, complete cds Chr.2
415434	5.38	ren	col	0.55	H.sapiens mRNA for imogen 38 Chr.13
509602	5.37	cns	leu	0.48	NOP56 homo sapiens mRNA for nucleolar protein Chr.20

The first column (*IMAGE ID*) is the Washington University Clone ID number, and the last column is a brief description from the data source. The second column (*Student range*) is the value of the studentized range statistic in equation (2), and the columns labeled *Min* and *Max* are the cell clusters that give the minimum and maximum of the means. The column labeled *upper 10% mean* is the average correlation of the highest 10% of the 4463 gene-compound correlations for the gene.

classes with the leukemia and melanoma cell line panels. For each compound, we calculated its average  $GI_{50}$  values over the melanoma and leukemia cell line panels, giving AveMEL and AveLEU values for the compound. Then, using these two values, we calculated the mean  $GI_{50}$  values and feature z-scores for each compound class listed in Table 3. In contrast, note that actinomycins and anthraquinones, two potent and widely used classes of cancer chemotherapeutic

agents, have average or negative correlations with the expression patterns of genes listed in Table 3.

The *benzodithiophenediones* and *indolonaphthoquinones* display opposite behavior with respect to the genes in Table 3 in the following sense: The *benzodithiophenediones* have well above average correlation with *Rab7*, *KIAA0110* and *MMP14*, and below average correlation with *LCP1*, *HS1* and *CARS-cyp*, whereas the *indolonaphthoquinone* class shows the

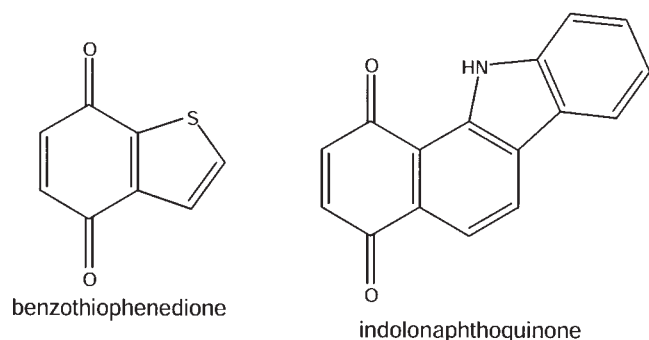


Figure 2 Substructural queries defining the *benzothiophenedione* and *indolonaphthoquinone* classes.

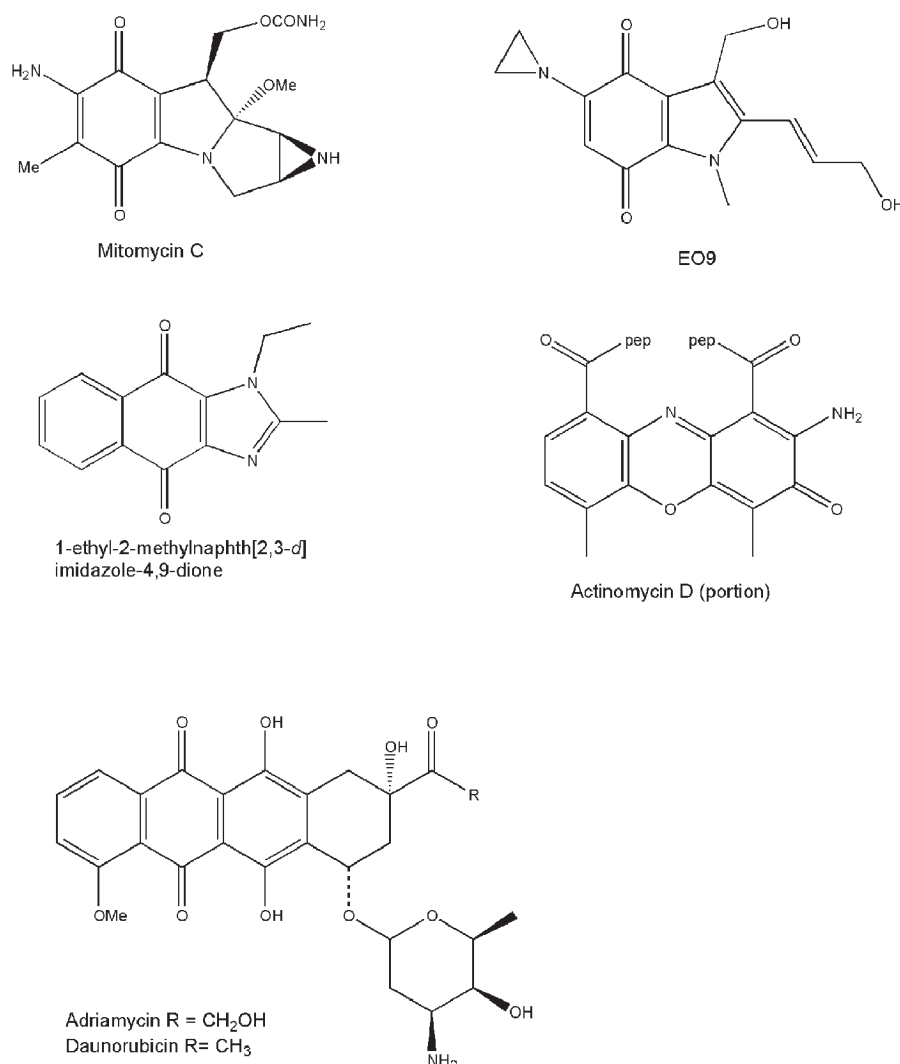


Figure 3 Cytotoxic quinone anti-cancer agents in clinical use.

opposite behavior. This difference is particularly pronounced for the leukemia gene *CARS-cyp*<sup>44</sup> (ID 179994). For this gene, the 23 *benzodithiophenediones* have a feature z-score of  $-7.25$ , whereas the 20 *indolonaphthoquinones* have a feature z-score of  $+5.75$ . Similar differences are seen with the leukemia gene *HS1* (ID 260052) and the melanoma gene *MMP14* (ID 270505). Furthermore, even though the *benzodithiophenediones*, as a class are strongly correlated with over-expression of melanoma genes, members of this class with strong electron withdrawing substituents correlate better with *LCP1*. In a COMPARE analysis<sup>5-9,46</sup> none of the *benzodithiophenediones* had a Pearson correlation coefficient  $>0.6$  against any compound in the NCI 'Standard Agent' database, perhaps indicating that the *benzodithiophenediones* act by a mechanism different from that of any of the standard agents.

**Table 3** Statistics for average compound–gene correlations for several classes of compounds and selected genes

Class	Count	Feature Z-scores for selected genes						Ave. leukemia		Ave. melanoma	
		Rab7	KIA110	MMP14	LCP1	CARS	HS1	Z-score	pG <sub>150</sub>	Z-score	pG <sub>150</sub>
Actinomycin	12	1.69	3.99	1.02	0.67	-1.35	-0.2	5.56	7.57	6.38	7.22
Antraquinone	65	-6.62	-6.61	-3.77	-1.06	2.11	1.05	11.33	7.29	10.23	6.69
Aziridinyloquinone	11	0.44	0.49	2.66	-3.34	-3.76	-3.57	2.25	6.48	4.13	6.63
Benzothiophenedione	23	10.50	9.39	8.48	-1.09	-7.25	-5.57	2.91	6.44	6.84	6.78
Indoleione	7	-0.11	0.14	1.56	-2.18	-1.51	-1.88	1.04	6.18	1.75	6.03
Indolonaphthoquinone	20	-2.03	-1.42	-4.94	5.64	5.75	5.79	0.48	5.87	-1.48	5.01
Naphthimidazoledione	6	-0.36	1.63	0.33	-0.56	-0.77	-1.16	2.29	6.82	2.77	6.54
O-quinone	26	4.45	3.17	2.03	1.19	-1.17	-0.7	0.04	5.76	0.33	5.52
P-quinone, 2-amino	95	8.0	6.96	6.0	0.43	-3.52	-2.76	2.45	6.34	4.78	6.61
P-quinone, 2-halo	65	3.88	3.19	-1.11	6.26	4.64	5.03	2.45	6.51	1.47	5.80
Quinone methide	9	0.11	0.86	0.56	-0.27	0.86	0.94	0.28	5.82	0.12	5.51
Quinoneimine	46	5.47	6.29	3.66	0.49	-2.88	-1.87	2.52	6.18	3.96	5.97
All quinones	407	8.82	6.75	3.93	4.43	0.14	0.6	7.48	6.14	8.56	5.8

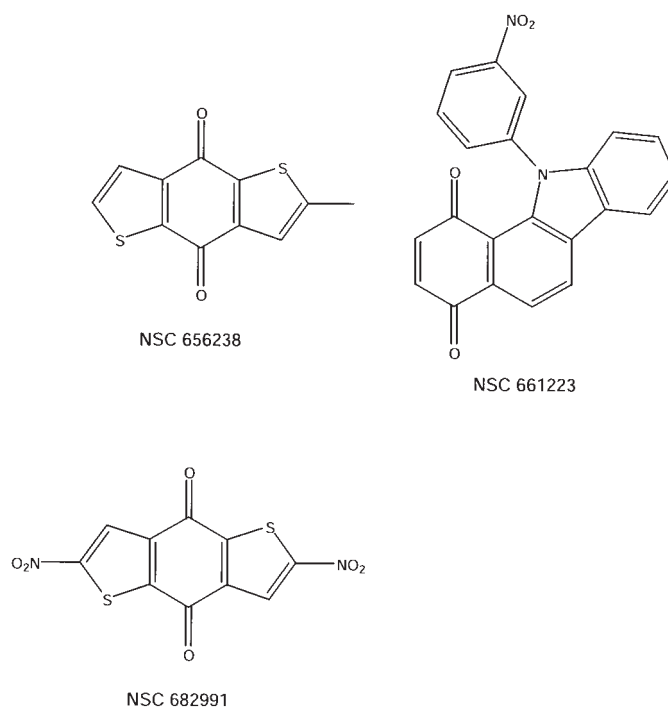
The first column gives the name of the structural class and corresponds to structures in Figures 2 and 3. The second column gives the number of compounds in each class. The third through eighth columns give the feature z-scores for each compound class and gene, defined as:

$$Z = \frac{\text{class mean} - \text{overall mean}}{\text{standard error}}$$

The four righthand columns show average G<sub>150</sub> values and feature z-scores for the compound classes with the leukemia and melanoma cell line panels. For each compound, we calculated the average G<sub>150</sub> value (AveMEL and AveLEU) over the melanoma and leukemia panels. Using these two values, we then calculated the average G<sub>150</sub> value and feature z-score for each compound class.

### Gene Expression Correlations of Representative Compounds

In the last section, we began by identifying genes that differentiated particular cell subsets and projected them through the cells and compounds to correlated substructures. In this section, we do the reverse, starting with substructures of interest and projecting through compounds and then through cells to correlated genes. To understand more fully how compound activity might be related to molecular biology in the NCI60 panel, we used one representative from each of the two quinone classes (NSC 656238 from the *benzodithiophenedione* class and NSC 661223 from the *indolonaphthoquinone* class; see Figure 4) as probes to look for genes whose expression patterns are highly correlated with the compounds' activity profiles. Partial results are shown in Table 4a and b. The full version of Table 4, which gives correlations between 3748 genes and the three compounds of Figure 4 is available at <http://www.leadscope.com> and <http://discover.nci.nih.gov>. In the tables, genes are in rows, and the three columns labeled NSC 656238, NSC 661223 and NSC 682991 contain compound–gene correlation coefficients. Table 4a shows selected genes with high positive correlations with NSC 656238, and Table 4b shows selected genes with high positive correlations with NSC 661223. Note that the correlations are quite high and that for all of the listed genes, the correlation coefficients with these two compounds have opposite signs, in agreement with the trends seen in Table 3. In the full version of Table 4, there are 335 distinct genes that have high correlations ( $\geq +0.5$  or  $\leq -0.5$ ), with at least one of the two probe compounds.



**Figure 4** Representative quinones from the *benzodithiophenedione* and *indolonaphthoquinone* classes used as probes to identify genes for which the expression patterns are highly correlated with the compound's activity.

**Table 4** Selected compound–gene correlations for NSC 656238, NSC 661223 and NSC 682991 (see Figure 4)

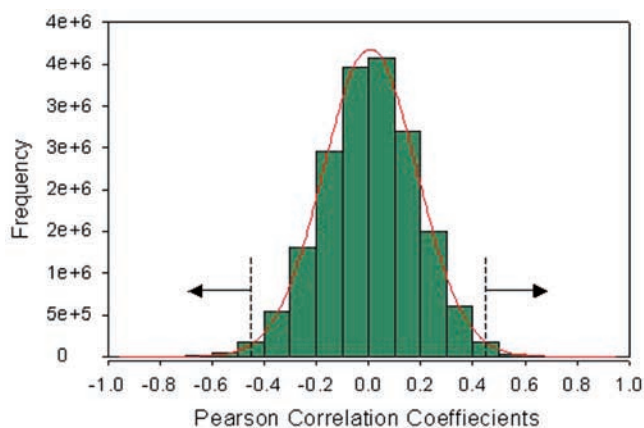
IMAGE ID	NSC656238	NSC661223	NSC682991	Description
<b>a</b>				
486233	0.67	−0.2	0.14	Human small GTP binding protein Rab7 mRNA, complete cds Chr.3
270505	0.65	−0.68	−0.24	H.sapiens mRNA for membrane-type matrix metalloproteinase 1 Chr.
472138	0.62	−0.57	−0.23	Homo sapiens sgk gene Chr.6
268727	0.61	−0.3	−0.06	Human mutY homolog (hMYH) gene, complete cds Chr.1
486086	0.59	−0.39	−0.01	PIGF phosphatidylinositol glycan, class F Chr.2
366379	0.58	−0.59	−0.17	RNASE1 ribonuclease, RNase a family, 1 (pancreatic) Chr.14
486471	0.58	−0.62	−0.31	CALD1 caldesmon Chr.7 (actin and myosin binding protein)
487096	0.58	−0.21	−0.08	Human COP9 homolog (HCOP9) mRNA, complete cds Chr.2
486751	0.55	−0.23	0.08	PIGF phosphatidylinositol glycan, class F Chr.2
429540	0.54	−0.77	−0.32	FLJ10140 hypothetical protein, chr. 22
127821	0.51	−0.31	0.07	ACP5 acid phosphatase type 5 Chr.19
510118	0.51	−0.72	−0.38	ANX5 annexin V (endonexin II) Chr.4
363919	0.5	−0.28	−0.02	ASAH N-acylsphingosine amidohydrolase (acid ceramidase) Chr.8
366489	0.5	−0.79	−0.31	H. sapiens mRNA for RAB13
<b>b</b>				
260052	−0.48	0.84	0.51	HCLS1 hematopoietic lineage cell specific protein Chr.3
488194	−0.37	0.8	0.41	CENPC centromere autoantigen C Chr.4
357011	−0.37	0.75	0.35	SRP46 splicing factor, arginine/serine-rich, 46kD, Chr. 11
179994	−0.46	0.7	0.43	PPIG peptidyl-prolyl isomerase G (cyclophilin G)
469356	−0.49	0.67	0.35	Small nuclear ribonucleoprotein SM D1 Chr.18
43129	−0.2	0.61	0.33	TOP1 DNA topoisomerase I Chr.20
204301	−0.11	0.61	0.53	CDC25A, m-phase inducer phosphatase 1
365824	−0.3	0.6	0.32	Human zinc finger protein (MAZ) mRNA Chr.16
471071	−0.39	0.6	0.38	HNRPDL heterogeneous nuclear ribonucleoprotein D-like protein. Homo sapiens mRNA for A+U-rich element RNA binding factor, complete cds Chr.4
486676	−0.1	0.6	0.44	LCP1 lymphocyte cytosolic protein 1 (L-plastin) Chr.13
234351	−0.16	0.59	0.52	ING3 inhibitor of growth family, member 3
489415	−0.34	0.59	0.32	SFRS3, splicing factor arginine/ser rich-3 pre-mRNA splicing factor SRP20 Chr.11

The first column (*IMAGE ID*) is the Washington University Clone ID number, and the last column is a brief description from the data source. The columns labeled NSC 656238, NSC 661223 and NSC 682991 contain compound–gene correlation coefficients. Table 4a is sorted by correlation with NSC 656238; Table 4b by correlation with NSC 661223. Both tables show genes with compound–gene correlations  $\geq 0.5$  or  $\leq (-0.5)$ . All unidentified ESTs were omitted.

To put these values in context, Figure 5 shows the distribution of Pearson correlation coefficients between the 4463 compound activity patterns and the 3748 gene expression patterns over the NCI60 cell lines. Of the 17 million gene–compound correlations, only 1% are in either tail region, above 0.45 or below  $-0.45$ .

Of the 335 genes mentioned above, 325 or 97% of them have correlation coefficients with opposite signs for compounds NSC 656238 and NSC 661223. This pattern of opposite signs is also present, but to a lesser extent for the full set of 3748 genes. In comparing the correlations of the

activities of NSC 656238 and NSC 661223 with the expression levels of the full set of genes, approximately 70% of the correlation pairs are of opposite sign ( $r = -0.62$ ), and those with the same sign tend to be closer to zero. Finally, although NSC 682991 is in the *benzodithiophenedione* class, it behaves in terms of its correlations with the full set of 3748 genes more like the *indolonaphthoquinone* NSC 661223,  $r = 0.77$ . Its correlations are much less related to those of NSC 656238 ( $r = -0.27$ ), even though both are in the *benzodithiophenedione* class. A possible explanation of this apparent paradox will be offered in the Discussion section.



**Figure 5** Histogram of Pearson correlation coefficients of 3748 genes and 4463 compounds. The correlation coefficients across the NCI60 cell lines are approximately normally distributed around zero.

## DISCUSSION

Genomic studies are producing large databases of molecular information on cancers and other cell and tissue types. As is universally recognized, these databases represent an unparalleled opportunity for pharmaceutical advance. The challenge is to link the data to the drug discovery and development processes. An 'information-intensive' approach<sup>6</sup> formulated several years ago (by one of the present authors and colleagues) provided a blueprint for one productive way to meet that challenge. It provided a way to organize and inter-relate potential therapeutic targets, molecular mechanisms of action of compounds tested and modulators of activity within cancer cell lines. It also suggested a way to project genomic information on the cells used for testing through the activity patterns of compounds to molecular structural characteristics of those compounds.<sup>6</sup> However, that suggestion was not pursued, and it was not converted into a fluent methodology for exploration or into a software package for doing so. Required was a way to couple the genomic (or proteomic) information with structure-based data mining to provide insights fruitful for follow-up in experimental structure-activity studies. Here we have presented such a method, based on the relational database system schematized in Figure 1. Included are gene expression levels for 3748 genes in 60 cell lines (T-matrix), activity values for 4463 compounds in 60 cell lines (A-matrix), and binary indices of occurrence of 27 000 structural features in 4463 compounds (S-matrix). As a proof-of-principle example of the approach, we have used it to identify subclasses of quinones well correlated with genes that are selectively expressed either in melanomas or in leukemias. A brief discussion of these agents and their genomic associations follows.

Of the 4463 compounds in the NCI set used in this analysis, 462 (10.4%) are quinones, quinoneimines or quinone methides. The mechanisms of quinone cytotoxicity<sup>49–51</sup> are complex and varied. However, two principal pathways are well established. First, quinones act as redox-active molecules that can undergo either 1- or 2-electron reductions,

depending on the cellular environment. The mechanism for 1-electron reduction involves redox cycling between quinone and semiquinone radical states, leading to consumption of NADH and formation of hydroperoxy radicals. Depending on the cellular environment, other reactive oxygen species, including superoxides, hydrogen peroxides and hydroxyl radicals, can be formed. These reactive species can, in turn, cause peroxidation of lipids, oxidation and strand breaks in DNA, consumption of reducing equivalents (eg, NAD(P)H or glutathione), and oxidation of other macromolecules. In the second pathway, unhindered quinones act as Michael acceptors, causing cellular damage through alkylation of thiol or amino groups of glutathione, proteins and DNA. Mitomycin C and E09, for example, undergo reductive alkylation<sup>53</sup> by mechanisms that involve opening the aziridinyl ring.

In the present study, we found that several genes selectively over-expressed in melanomas have expression patterns that are well correlated with the activity patterns of a subclass of *benzodithiophenedione* compounds. This class shows a distinctive substituent effect: *Benzo*dithiophenediones with strong electron-withdrawing substituents (eg, NSC 682991; see Figure 5) show low or negative correlation with many of the genes that are over-expressed in melanomas (see Table 4a), whereas members with electron-donating substituents (eg, NSC 656238) show high positive correlations with those genes. For example, NSC 656238 is 10 times more potent against the melanoma cell lines than is NSC 682991. Electron-withdrawing substituents such as nitro groups raise the reduction potential of the quinone moiety, making it a better oxidant than it is in compounds with electron-donating groups. A plausible hypothesis for the cytotoxicity of a *benzodithiophenedione* is that it may disrupt an essential cellular redox process. This hypothesis is consistent with the roles of genes over-expressed in melanomas. In particular, *Rab7*<sup>31–33</sup> is the gene most strongly correlated with the electron-donating *benzodithiophenediones*. For example, the correlation coefficient with NSC 656238 is 0.67. Genes in the *Rab* family are small GTP binding proteins that ensure specificity of the docking of transport vesicles. In particular, *Rab7* has recently been identified as a key regulatory protein for aggregation and fusion of late endocytic lysosomes. Cells expressing a dominant-negative *Rab7* mutant have been reported not to form lysosomal aggregates.<sup>31</sup> The dispersed lysosomes exhibit sharply higher pH, presumably due to disruption of the vacuolar proton pump. Interestingly, in this context, another gene highly correlated with NSC 656238 is *ACP5* ( $r=0.51$ ). *ACP5* (Clone ID 127821) is a unique lysosomal membrane ATPase responsible for maintaining the pH. Several other lysosomal proteins are also well correlated with the electron-donating *benzodithiophenediones*. Two other ATPases, ATP6B2 (Clone ID 380399) and ATP6E (Clone ID 417475), have correlation coefficients of 0.40 and 0.46, respectively, with NSC 656238. Both of these ATPases are reported to be lysosomal H<sup>+</sup> transporters. Other lysosomal genes, *ASAH* (Clone ID 417819) and *LAMP2* (357407), also show high correlations (0.50 and 0.40, respectively) with this quinone. Thus, genes well-cor-

related with this particular quinone class seem to be enriched in lysosomal proteins that are involved in vacuolar proton pump activity.

This substituent effect suggests a possible link between the oxidation potential of quinones, the proton pump, and the electron transport chain. A plausible hypothesis is that NSC656238 may act as a surrogate oxidizing agent in the electron transport chain. Ubiquinone-10 is the electron acceptor for mitochondrial oxidative phosphorylation. Menadione (2-methylnaphthoquinone), a compound known to compete with ubiquinone in the oxidative phosphorylation chain, also shows a reasonable correlation with *Rab7* ( $r = 0.40$ ). The reduction potentials of menadione and ubiquinone are known,<sup>57,58</sup> but the reduction potential of NSC656238 has not been reported. We speculate that its oxidizing potential allows it to compete successfully with ubiquinone in the electron transport chain, as does menadione. The oxidizing potential of the quinone moiety would be a key factor in such a mechanism. Although compound NSC 682991 is a better oxidant, it may be reduced by cellular protective agents such as glutathione. Thus, at low concentrations, it may not be available to compete with ubiquinone and therefore may be effective only at higher concentrations.

We have illustrated a way to couple information on differential gene expression with structure-based data mining. The approach provides insights that may allow selective targeting of cellular mechanisms preferentially operating in specific tissues. The *benzodithiophenedione* series that emerged from this study is a clear example. This is a well-defined and structurally homogeneous series of quinones, which are well correlated with the expression patterns of *Rab7* and other melanoma genes. The substituent effect seen in this series suggests a relationship between the oxidation potential of a compound and its correlation with the expression patterns of specific genes. This relationship prompts new questions that can be pursued experimentally: Is there a quantitative relationship between the oxidation potential of the *benzodithiophenedione* series and melanoma cytotoxicity? If so, is there a direct relationship between the selective cytotoxicity of NSC 656238 and *Rab7* or the *ATPases* that are over-expressed in melanomas? The data currently available do not permit answers to these questions, but the analyses described here do provide indirect evidence of connections that can be tested in experimental structure-activity studies.

In this article, we have described a general analytical method, designated SAT, for discovering relationships between compound classes and potential molecular targets. The method uses statistical techniques to select genes with characteristic expression patterns, then applies structure-based data mining software to identify compound substructural classes that are well-correlated with the expression patterns of those genes. Selected members of the class identified can then be used as molecular probes to identify additional compound-gene associations and thereby refine hypotheses or focus further experiments. This semi-empirical method projects genomic information from cells through compound

activity patterns to molecular structural features of drugs or potential drugs. It can also do the reverse, identifying genes whose expression levels (or other characteristics) correlate strongly with structural features of a particular drug, or drug candidate. The SAT approach to pharmacogenomic analysis can shed light on molecular mechanisms and has the potential to accelerate the process of drug discovery in several ways: (i) it can be used to prioritize genes for follow-up studies as potential therapeutic targets; (ii) because the analysis projects genomic information to molecular substructure through the [S] matrix, it allows extraction of a preliminary structure-activity relationship (SAR) directly from the SAT correlations; (iii) the preliminary SAR can, in turn, be used for early pharmacophore development or to select new, untested drug candidates from an actual or virtual library of compounds; and (iv) it can be used to prioritize candidate compounds for detailed gene expression analysis or other biological studies.

## METHODS

### Databases

For the target matrix [T] in this study, we used a 3748-gene subset of the 9704-cDNA database. The subset was selected to include only genes whose identities had been sequence-verified<sup>1</sup> and which had <10% missing data values over the 60 cell lines. For the activity matrix [A], we selected a set of 4463 compounds that had been tested in the NCI Developmental Therapeutics Programs sulforhodamine B assay two or more times and for which we had structure records. The compound activity values used were based on  $GI_{50}$ , the concentration needed for 50% growth inhibition. More specifically, activity was parameterized as  $-\log(GI_{50})$ . These databases are available at <http://www.leadscope.com> and <http://discover.nci.nih.gov>.

### Correlation Matrix

The activity matrix [A] ( $4463 \times 60$ ) contains compound activity data for the 60 cell lines, and the target matrix [T] ( $3748 \times 60$ ) contains gene expression patterns over these same cell lines.<sup>6</sup> For each compound, the activity level can be considered as a variable defined over the 60 cell lines; likewise, for each target gene the expression level can be considered as a variable over these same 60 cell lines. If there were no missing data, the Pearson correlations between the compound activities and gene expressions could be computed by first standardizing the rows of each matrix into Z-scores,

$$Z = \frac{\text{variable} - \text{mean}}{\text{standard deviation}}$$

then forming the matrix product  $[A \bullet T^T]$ , and finally dividing each entry by  $n - 1 = 59$  to obtain the correlation matrix.<sup>6</sup> However, approximately 7% of the compound activity data and about 2% of the gene expression data were missing, so the algorithm for obtaining all pair-wise correlations had to be modified. The correlation coefficient between the activity of the *ith* compound and the expression level of the *jth* target was actually computed as

$$r_{A_i T_j} = \frac{S_{A_i T_j}}{S_{A_i} S_{T_j}} \quad (1)$$

where  $s_{A_i}$  and  $s_{T_j}$  are the standard deviations of the activity of the *i*th compound and the expression of the *j*th target, respectively, and  $s_{A_i T_j}$  is the covariance between these variables. In formula (1) all  $N_i$  activity values available for the *i*th compound and all  $N_j$  expression values available for the *j*th target are used to compute the activity and expression means and standard deviations. To account for the pattern of missing data, the denominator used in computing the covariance  $s_{A_i T_j}$  is

$$N_{ij}^* = N_{ij} - 1 + \left(1 - \frac{N_{ij}}{N_i}\right) \left(1 - \frac{N_{ij}}{N_j}\right),$$

where  $N_{ij}$  is the number of cell lines for which *both* activity and expression were measured. This divisor  $N_{ij}^*$  has been shown<sup>59</sup> to give an unbiased estimate of the correlation if data points are missing at random. Missing data in [A] and [T] are not entirely random in distribution, but the effect of the non-randomness is expected to be second-order.

### Selection of Genes

We examined several methods for identifying genes that are strongly correlated with compound activities. Included were: (1) selection of genes differentially expressed in particular tissues of origin; and (2) selection of genes highly correlated with the activity levels of large numbers of compounds. Here, we focus on the first method. First, the cluster analyses in Ross *et al*<sup>1</sup> and Scherf *et al*<sup>2</sup> were used to group the cell lines into seven subsets, which corresponded roughly to tissue of origin. To find genes whose average level of expression distinguished among the seven panels, we used a version of the Studentized range procedure. Because some cell subsets (eg the leukemia panel) had expression levels that were more tightly clustered (less variable) than did the other panels, we used an unpooled estimate of variance. The unpooled Studentized range statistic is given by the expression:

$$\text{Studentized range} = \frac{(\bar{X}_1 - \bar{X}_0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}} \quad (2)$$

where  $\bar{X}_1$  is the mean expression level for the panel with the highest average expression for the given gene, and  $s_1^2$  is the corresponding variance;  $\bar{X}_0$  and  $s_0^2$  are analogous values for the cell panel with the lowest mean expression for the gene. We calculated the statistic for all 3748 genes and selected those for which the Studentized range value was greater than 5.08. Based on a Bonferroni adjustment for all 21 possible pair-wise comparisons between subsets, we would expect the values of fewer than 1% of the genes to exceed 5.08.

The rows of [T] corresponding to genes with high values of the Studentized range statistic (see Table 2) formed a submatrix [T<sub>S</sub>]. For each of the selected genes in [T<sub>S</sub>] we examined the distribution of the 4463 gene–compound Pearson

correlation coefficients (columns of the correlation matrix [A•T<sub>S</sub><sup>T</sup>]) to identify genes well-correlated with compound activity. This distribution of correlation coefficients was summarized for each gene by computing the average correlation of the highest 10% of the 4463 gene–compound correlations. These values are reported in Table 2. Large values indicate a gene that is well-correlated with compound activity.

The examples of SAT analysis presented here focus on one particular basis for selection of genes—differential expression between tissues of origin. However, many other bases for gene selection could equally be the starting point. For example, one could choose to focus on genes with the highest variance in expression level over the NCI60 cell lines, or on genes that simply happen to be the subject of one's research. The following steps are independent of the basis on which the gene is selected. Analogously, if one were starting from a structural feature, or features, and finding related genes, any basis for selection of the feature could serve as a starting point.

### Mining of Structural Features

Once a set of genes was selected, the final step was to identify structural features well correlated with the expression levels of those genes. Any structural feature F (eg, 1,4-benzoquinone) is either present or absent in each compound. Let  $N_k$  denote the number of compounds with feature  $F_k$ , where  $k$  is an index over the structural features. The structural feature matrix [S] has potentially 27 000 rows corresponding to the full set of 2D structures considered and 4463 columns corresponding to compounds in the NCI database subset analyzed. For feature  $F_k$ , the corresponding row in [S] has entry  $1/N_k$  for compounds in which the feature is present or 0 for compounds for which the feature is absent. The structural information was incorporated by forming the matrix product<sup>6</sup> [S•A•T<sup>T</sup>] in the LeadScope software (LeadScope, Inc, Columbus, OH, USA). Each row corresponds to a structural feature  $F_k$ , each column corresponds to a gene, and each element is the mean of the correlation coefficient between the gene and all compounds containing feature  $F_k$ . Although the size of matrix [S•A•T<sup>T</sup>] would be 27 000 × 3748 if all structural features were represented in at least one compound and all genes were used, in practice we always selected smaller subsets of features and genes for analysis.

The *j*th column of the matrix [S•A•T<sup>T</sup>] can be analyzed to identify structural features most highly associated with expression of the *j*th gene. To identify structural features that are enriched with compounds for which the 60-cell activity patterns are highly associated with expression patterns of a gene, for each structural feature we calculated a *feature z-score*;

$$Z = \frac{\text{feature mean} - \text{overall mean}}{\text{standard error}}$$

In this equation, the feature mean is the average correlation with the *j*th gene of all compounds containing the feature, the overall mean is the average correlation with the *j*th gene

of all compounds, and the standard error is the standard deviation of the correlation with the *j*th gene of all compounds divided by the square root of the number of compounds with the feature. Further details are given in Reference 22.

After selection of a gene column in  $[S \bullet A \bullet T^T]$ , the structural classes with the highest *feature z-scores* (ie those features that tend to have the highest average correlation) were identified. For example, using the melanoma gene *Rab7* (ID 486233 in Table 2),<sup>31–33</sup> we found that the *2-arylcarbonylthiophene* class had the highest *feature z-score*. For the leukemia gene *LCPI* (ID 486676 in Table 2),<sup>40,41</sup> the *7-carbonylindole* class had the highest *feature z-score*. By sorting the structural classes in order of decreasing *feature z-score* and examining the compounds in the high-scoring structural classes, we could usually postulate a structural class that defined the membership more precisely than did the highest scoring feature in the hierarchy. We then formulated a substructure query to define the postulated class. In the two cases described here, we defined the substructure queries labeled *benzothiophenedione* and *indolonaphthoquinone* shown in Figure 2. Note that these substructures are more precise extensions of the highest scoring features in the hierarchy; viz, *2-arylcarbonylthiophene* and *7-carbonylindole*.

## DATA

The [A] and [T] databases analyzed here are publicly available at <http://www.leadscope.com> and <http://discover.nci.nih.gov>. The [S] matrix is available from LeadScope, Inc on request. These sites also provide the full version of Table 4, which gives correlations between 3748 genes and the three compounds of Figure 4.

## DUALITY OF INTEREST

Authors PE Blower, C Yang and L Yu are employees of LeadScope, Inc, Columbus, OH, USA, which produces the software used in this study.

## REFERENCES

- Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P *et al*. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000; **24**: 227–235.
- Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L *et al*. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 2000; **24**: 236–244.
- Boyd MR, Paull KD. Some practical consideration and applications of the National Cancer Institute *in vitro* anti-cancer drug discovery screen. *Drug Dev Des* 1995; **34**: 91–109.
- Monks AP, Scudiero DA, Johnson GS, Paull KD, Sausville EA. The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. *Anti-Cancer Drug Des* 1997; **12**: 533–541.
- Paull KD, Shoemaker RH, Hodes L, Monks A, Scudiero DA, Rubinstein L *et al*. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J Natl Cancer Inst* 1989; **81**: 1088–1092.
- Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ, Kohn KW *et al*. An information intensive approach to the molecular pharmacology of cancer. *Science* 1997; **275**: 343–349.
- Weinstein JN, Kohn KW, Grever MR, Viswanadhan VN, Rubinstein LV, Monks AP *et al*. Neural computing in cancer drug development: predicting mechanism of action. *Science* 1992; **258**: 447–451.
- Paull KD, Hamel E, Malspeis L. Prediction of biochemical mechanism of action from the *in vitro* antitumor screen of the National Cancer Institute. In: Foye WE (ed). *Cancer Chemotherapeutic Agents*. American Chemical Soc Books, 1993, pp 1574–1581.
- Weinstein JN, Myers TG, Buolamwini JK, Raghavan K, van Osdol W, Licht J *et al*. Predictive statistics and artificial intelligence in the US National Cancer Institutes drug discovery program for cancer and AIDS. *Stem Cells* 1994; **12**: 13–22.
- Shi LM, Myers TG, Fan Y, O'Connor PM, Paull KD, Friend SH *et al*. Mining the National Cancer Institute anticancer drug discovery database: cluster analysis of ellipticine analogs with p53-inverse and central nervous system-selective patterns of activity. *Mol Pharmacol* 1998; **53**: 241–251.
- Shi LM, Fan Y, Myers TG, O'Connor PM, Paull KD, Friend SH *et al*. 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.
- Wu L *et al*. Multidrug-resistant phenotype of disease-oriented panels of human tumor cell lines used for anticancer drug screening. *Cancer Res* 1992; **52**: 3029–3034.
- Lee J-S *et al*. Rhodamine efflux patterns predict P-glycoprotein substrates in the National Cancer Institute drug screen. *Mol Pharmacol* 1994; **46**: 627–638.
- Alvarez M *et al*. Generation of a drug resistance profile by quantitation of MDR-1/P-glycoprotein expression in the cell lines of the NCI anticancer drug screen. *J Clin Invest* 1995; **95**: 2205–2214.
- Bates SE *et al*. Molecular targets in the National Cancer Institute drug screen. *J Cancer Res Clin Oncol* 1995; **121**: 495–500.
- Izquierdo MA *et al*. Overlapping phenotypes of multidrug resistance among panels of human cancer-cell lines. *Int J Cancer* 1996; **65**: 230–237.
- Koo H-M *et al*. Enhanced sensitivity to 1-beta-D-arabinofuranosylcytosine and topoisomerase II inhibitors in tumor cell lines harboring activated ras oncogenes. *J Natl Cancer Inst* 1996; **56**: 5211–5216.
- O'Connor PM *et al*. Characterization of the p53-tumor suppressor pathway in cells of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer Res* 1997; **57**: 4285–4300.
- Freije JM *et al*. Identification of compounds with preferential inhibitory activity against low-Nm23-expressing human breast carcinoma and melanoma cell lines. *Nat Med* 1997; **3**: 395–401.
- Wosikowski K *et al*. Identification of epidermal growth factor receptor and c-erbB2 pathway inhibitors by correlation with gene expression patterns. *J Natl Cancer Inst* 1997; **89**: 1505–1513.
- Staunton JE, Slonim DK, Collier HA, Tamayo P, Angelo MJ, Park J *et al*. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci USA* 2001; **98**: 10787–10792.
- Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE. LeadScope: software for exploring large sets of screening data. *J Chem Inf Comput Sci* 2000; **40**: 1302–1314.
- Fan Y, Weinstein JN, Kohn KW, Shi LM, Pommier Y. Molecular modeling studies of the DNA-topoisomerase I ternary cleavable complex with camptothecin. *J Med Chem* 1998; **41**: 2216–2226.
- Cho SJ, Shen CF, Hermsmeier MA. Binary formal inference-based recursive modeling using multiple atom and physicochemical property class pair and torsion descriptors as decision criteria. *J Chem Inf Comput Sci* 2000; **40**: 668–680.
- Klopman G, Shi LM, Ramu A. Quantitative structure-activity relationship of multi-drug resistance reversal agents. *Mol Pharmacol* 1997; **52**: 323–334.
- Klopman G, Tu M. Diversity analysis of 14 156 molecules tested by the National Cancer Institute for anti-HIV activity using the quantitative structure-activity relational expert system MCASE. *J Med Chem* 1999; **42**: 992–998.
- Weinstein JN. Fishing Expeditions. *Science* 1998; **282**: 627–628.
- Weinstein JN. Pharmacogenomics: teaching old drugs new tricks. *N Eng J Med* 2000; **343**: 1408–1409.
- Weinstein JN, Buolamwini JK. Molecular targets in cancer drug discovery: cell-based profiling. *Curr Pharm Des* 2000; **6**: 473–483.
- Weinstein JN. Searching for pharmacogenomic markers: the synergy between omic and hypothesis-driven research. *Disease Markers* 2001; **17**: 77–88.
- Bucci C, Thomsen P, Nicoziani P, McCarthy J, van Deurs B. Rab7: a key to lysosome biogenesis. *Mol Biol Cell* 2000; **11**: 467–480.
- Meresse S, Steele-Mortimer O, Finlay BB, Gorvel JP. The rab7 GTPase

- controls the maturation of *Salmonella typhimurium*-containing vacuoles in HeLa cells. *EMBO J* 1999; **18**: 4394–4403.
- 33 Press B, Feng Y, Hoflack B, Wandinger-Ness A, Mutant. Rab7 causes the accumulation of cathepsin D and cation-independent mannose 6-phosphate receptor in an early endocytic compartment. *J Cell Biol* 1998; **140**: 1075–1089.
- 34 Hong SB, Li CM, Rhee HJ, Park JH, He X, Levy B *et al*. Molecular cloning and characterization of a human cDNA and gene encoding a novel acid ceramidase-like protein. *Genomics* 1999; **62**: 232–241.
- 35 Nagase T, Miyajima N, Tanaka A, Sazuka T, Seki N, Sato S *et al*. Prediction of the coding sequences of unidentified human genes III. The coding sequences of 40 new genes (KIAA0081–KIAA0120) deduced by analysis of cDNA clones from human cell line KG-1. *DNA Res* 1995; **2**: 37–43.
- 36 Holmbeck K, Bianco P, Caterina J, Yamada S, Kromer M, Kuznetsov SA *et al*. MT1-MMP-deficient mice develop dwarfism, osteopenia, arthritis and connective tissue disease due to inadequate collagen turnover. *Cell* 1999; **99**: 81–92.
- 37 Apte SS, Fukai N, Beier DR, Olsen BR. The matrix metalloproteinase-14 (MMP-14) gene is structurally distinct from other MMP genes and is co-expressed with the TIMP-2 gene during mouse embryogenesis. *J Biol Chem* 1997; **272**: 25511–25517.
- 38 Shinmura K, Yamaguchi S, Saitoh T, Takeuchi-Sasaki M, Kim SR, Nohmi T *et al*. Adenine excisional repair function of MYH protein on the adenine:8-hydroxyguanine base pair in double-stranded DNA. *Nucleic Acids Res* 2000; **28**: 4912–4918.
- 39 Ohtsubo T, Nishioka K, Imaiso Y, Iwai S, Shimokawa H, Oda H *et al*. Identification of human MutY homolog (hMYH) as a repair enzyme for 2-hydroxyadenine in DNA and detection of multiple forms of hMYH located in nuclei and mitochondria. *Nucl Acids Res* 2000; **28**: 1355–1364.
- 40 Wang J, Brown EJ. Immune complex-induced integrin activation and L-plastin phosphorylation require protein kinase A. *J Biol Chem* 1999; **274**: 24349–24356.
- 41 Jones SL, Wang J, Turck CW, Brown EJ. A role for the actin-bundling protein L-plastin in the regulation of leukocyte integrin function. *Proc Natl Acad Sci* 1998; **95**: 9331–9336.
- 42 Ingley E, Sarna MK, Beaumont JG, Tilbrook PA, Tsai S, Takemoto Y *et al*. HS1 interacts with Lyn and is critical for erythropoietin-induced differentiation of erythroid cells. *J Biol Chem* 2000; **275**: 7887–7893.
- 43 Brunati AM, Donella-Deana A, James P, Quadroni M, Contri A, Marin O *et al*. Molecular features underlying the sequential phosphorylation of HS1 protein and its association with c-Fgr protein-tyrosine kinase. *J Biol Chem* 1999; **274**: 7557–7564.
- 44 Nestel FP, Colwill K, Harper S, Pawson T, Anderson SK. RS cyclophilins: identification of an NK-TR1-related cyclophilin. *Gene* 1996; **180**: 151–155.
- 45 Chao YH, Kuo SC, Ku K, Chiu, I, Wu CH, Mauger A *et al*. Synthesis and cytotoxicity of Methyl-4,8-dihydrobenzo[1,2-b:5,4-b']dithiophene-4,8-dione derivatives. *Bioorg Med Chem* 1999; **7**: 1025–1031.
- 46 Chao YH, Kuo SC, Wu CH, Lee CY, Mauger A, Sun IC *et al*. Synthesis and cytotoxicity of 2-acetyl-4,8-dihydrobenzodithiophene-4,8-dione derivatives. *J Med Chem* 1998; **41**: 4658–4661.
- 47 Kundel MW, Kirkpatrick DL, Johnson JI, Powis G. Cell line-directed screening assay for inhibitors of thioredoxin reductase signaling as potential anti-cancer drugs. *Anti Canc Drug Des* 1997; **12**: 659–670.
- 48 Rogge M, Fischer G, Pindur U, Schollmeyer D.  $\alpha$ -Anellated carbazoles with anti-tumor activity: synthesis and cytotoxicity. *Monatsh Chem* 1996; **127**: 97–102.
- 49 Monks TJ, Hanzlik RP, Cohen GM, Ross D, Graham DG. Quinone chemistry and toxicity. *Toxicol Appl Pharmacol* 1992; **112**: 2–16.
- 50 O'Brien PJ. Molecular mechanisms of quinone cytotoxicity. *Chem Biol Interact* 1991; **80**: 1–41.
- 51 Bolton JL, Trush MA, Penning TM, Dryhurst G, Monks TJ. Role of quinones in toxicology. *Chem Res Toxicol* 2000; **13**: 135–160.
- 52 Phillips RM, Naylor MA, Jaffar M, Doughty SW, Everett SA, Breen AG *et al*. Bioreductive activation of a series of indolequinones by human DT-diaphorase: structure–activity relationships. *J Med Chem* 1999; **42**: 4071–4080.
- 53 Xing C, Wu P, Skibo EB, Dorr RT. Design of cancer-specific antitumor agents based on aziridinylcyclopent[b]indoloquinones. *J Med Chem* 2000; **43**: 457–466.
- 54 Beall HD, Hudnott AR, Winski S, Siegel D, Swann E, Ross D *et al*. Indolequinone antitumor agents: relationship between quinone structure and rate of metabolism by recombinant human NQO1. *Bioorg Med Chem Lett* 1998; **8**: 545–548.
- 55 Fitzsimmons SA, Workman P, Grever M, Paull K, Camalier R, Lewis AD. Reductase enzyme expression across the National Cancer Institute tumor cell line panel: correlation with sensitivity to mitomycin C and EO9. *J Natl Canc Inst* 1996; **88**: 259–269.
- 56 Sengupta SK. Inhibitors of DNA-transcribing enzymes, In Foye WE (ed). *Cancer Chemotherapeutic Agents*. American Chemical Society: Washington, DC, 1993, pp 205–260.
- 57 Mitchell J, Marrian DH. Radiosensitization of cells by a derivative of 2-methyl-1, 4-naphthoquinone, In Morton RA (ed). *Biochemistry of Quinones*. Academic Press: New York, 1965, pp 503–541.
- 58 Nesta P. Radiation chemistry of quinonoid compounds, In Patai S, Rappoport S (eds). *The Chemistry of the Quinonoid Compounds*. John Wiley & Sons: New York, 1988, pp 879–898.
- 59 Schaffer J. *The Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall, 1996.