Establishing Connections between Microarray Expression Data and Chemotherapeutic Cancer Pharmacology

Anders Wallqvist,2 Alfred A. Rabow, Robert H. Shoemaker, Edward A. Sausville, and David G. Covell2

Science Applications International Corporation [A. W., A. A. R.], and Developmental Therapeutics Program, National Cancer Institute at Frederick, NIH [R. H. S., E. A. S., D. G. C.], Frederick, Maryland 21702

Abstract
We have investigated three different microarray datasets of ~6 K gene expressions across the National Cancer Institute’s panel of 60 tumor cell lines. Initial assessments of reproducibility for gene expressions within each dataset, as derived from sequence analysis of full-length sequences as well as expressed sequence tags (EST), found statistically significant results for no more than 36% of those cases where at least one replicate of a gene appears on the array. Filtering the data based only on pairwise comparisons among these three datasets creates a list of ~400 significant concordant expression patterns. The expression profiles of these smaller sets of genes were used to locate similar expression profiles of synthetic agents screened against these same 60 tumor cell lines. A correspondence was found between mRNA expression patterns and 50% growth inhibition response patterns of screened agents for 11 cases that were subsequently verifiable from ligand-target crystallographic data. Notable amongst these cases are genes encoding a variety of kinases, which were also found to be targets of small drug-like molecules within the database of protein structures. These 11 cases lend support to the premise that similarities between expression patterns and chemical responses for the National Cancer Institute’s tumor panel can be related to known cases of molecular structure and putative cellular function. The details of the 11 verifiable cases and the concordant gene subsets are provided. Discusisons about the prospects of using this approach as a data mining tool are included.

Introduction
Gene expression profiling from mRNA expression microarrays has become a powerful tool in assessing the cellular response in normal and cancer cells (1–3). The prevailing viewpoint proposes a complex network of genes working together to regulate homeostasis within the cell. The complexity of this cellular network has often been underestimated, because hundreds of genes may be time-dependently up- or down-regulated in response to a single effector (4, 5). Despite the complexity of data interpretation, gene expression profiling has clear clinical applications in its ability to subgroup tumors that cannot otherwise be differentiated (6–9). Recently, Alizadeh et al. (10) were able to identify clinically distinct types of diffuse large B-cell lymphomas using gene expression techniques. Khan et al. (11) used an artificial neural network to successfully classify clinical examples of small, round blue-cell tumors.

In this paper we examine the connection between gene expression data across the NCI’s5 60 tumor cell lines and chemical screening experiments conducted on the same cell lines to establish cytotoxicity, as measured by concentration of agents to GI50 (12–14). Correlations between gene expression patterns and GI50 patterns across this tumor cell panel are postulated, in our analysis, to suggest possible interactions between chemical agents and either gene products or nucleic acids. One should note that no drugs are introduced in this set of gene expression measurements. It is only via their correlated responses that we postulate any connections between putative targets and chemical response.

Previous attempts to identify relationships between molecular targets and chemicals, based on expression patterns observed in the NCI’s anticancer drug screen, have been reported with varied degrees of success (15, 16). Much of this difficulty results from the lack of abundant gene-drug relationships, that can also be experimentally verified, thus also making it difficult here to fully evaluate our general hypothesis. As an alternative it is possible to question our general hypothesis about linkages between chemical responses and putative targets without explicit administration of these agents (4, 17). However, our strategy does not take into account additional concerns related to making conclusions from mRNA microarray data. For example, Tamm et al. (18) critique the use of mRNA as a measure of expression by pointing to the well-known fact that post-transcriptional regulation of expression is most likely of equal importance for the expression of some genes. Others have amplified this viewpoint by suggesting the necessity of measuring actual protein levels, a prevailing feeling among members of the growing proteomics community.

Received 11/26/01; revised 1/29/02; accepted 2/6/02.

1 Supported in whole or in part with federal funds from the National Cancer Institute, NIH, under Contract No. N01-CO-56000.
2 To whom requests for reprints should be addressed, at Science Applications International Corporation, Frederick, MD 21702 (to A. W.) or Developmental Therapeutics Program, National Cancer Institute at Frederick, NIH, Frederick, MD 21702 (to D. G. C.).

3 The abbreviations used are: NCI, National Cancer Institute; GI50, 50% growth inhibition; PDB, Protein Data Bank; STO, staurosporine; cAMP, cyclic AMP; PDGF, platelet-derived growth factor; BDH, 3-hydroxybutyrate dehydrogenase; AND, dehydroepiandrosterone; DHT, dihydrotestosterone; QDE, quercetin; DIA4, diaphorase 4 or menadione oxidoreductase; PTTPRC, protein tyrosine phosphatase receptor type C; MMP, matrix metalloproteinase; AP0D, apolipoprotein D; ADH5, alcohol dehydrogenase 5; CTS3, cathepsin H.
In this analysis we examine three gene expression datasets for coherence between related gene expression patterns. At the outset we acknowledge that much controversy has been raised around the fact that gene expression data can be of highly varying quality and that it is necessary to make repeated measurements to get a clear picture of the true expression profile (19–21). As an alternative, our analysis will treat each of these datasets as a single replicate, and based on coherence of their patterns across the 60 tumor cell lines, extract the “best” set of gene expressions amongst these experiments. In fact, the quality of these data supports only a few hundred significant gene expressions; however, based on our requirement of concordance, this smaller data set can be advanced with higher confidence for additional analysis. It is among these few hundred concordant gene expressions that we hunt for evidence of target-drug relationships based on the additional observations of GI50 values already provided a great deal of information about chemobiomolecular synthesis and cell cycle control (22). This analysis offers only clues about these interactions, validations according to mechanisms of cellular action that include diverse measurements of gene expression and GI50 data. Examinations of these ligand targets for their biochemical function followed by establishing the link between these expressed proteins and mRNA expressions via sequence alignment is used here to make connections between gene expression and chemical activity.

Our analysis finds that in the relatively restricted set of concordant gene expressions, 11 verifiable gene-drug relationships are found. Previous attempts to find similar correlations have yielded, at best, one or two such relationships (15, 16). While our 11 observations represent only a small number when compared with the total space of possible interactions, these results are encouraging by demonstrating that target-drug interactions can be extracted from diverse measurements of gene expression and GI50 data. The tools necessary to connect these measurements are incorporated in the world wide web.4

Data Treatment for Finding Significantly Differentiated Gene Expressions

Currently three publicly available datasets using different microarray technologies exist for the 60 human cancer cell lines used by the NCI Developmental Therapeutics Program to screen anticancer agents. Two datasets are referred to as the Millennium and the Weinstein/Botstein/Stanford data sets, and are freely available from the Developmental Therapeutics Program website (15).5 The Stanford data is derived from cDNA arrays and are referenced to a pool of mRNA from 12 cell lines. The Millennium data set employed chips from Affymetrix where the relative abundance of each gene is calculated from the average difference of intensities between matching and mismatched oligonucleotides. Affymetrix chips were also used in the third dataset, which originates from the Whitehead Institute (16).6 The cell lines investigated using these arrays constitute cells derived from cancers of breast, central nervous system, colorectal, leukemia, lung, melanoma, ovarian, prostate, and renal systems of origin. Given that we have three sets of gene expression measurements across the same number of cell lines it is of interest to compare these experimental results. Our primary aim here is not to address quality control for measurements obtained from different technologies by different laboratories. Rather, as we will demonstrate, despite major reported difficulties regarding microarray data reproducibility, these three datasets can provide novel information regarding target-drug associations. Thus, where there may be a tendency to discount highly noisy datasets, our results reveal that meaningful information can be extracted when carefully analyzed. In order to compare the same genes in each dataset we first selected the set of ESTs within each microarray that have a pairwise common Unigene (24) designation, as derived from the sequence accession number. This yields 4791 putative common genes among the Millennium and the Stanford datasets, 4152 between the Whitehead and the Stanford datasets, and 5759 between the Millennium and the Whitehead datasets. There were 2105 genes common to all three of the datasets. The total number of genes contained in the three datasets was 5390, 9564, and 7006 for the Millennium, Stanford, and Whitehead datasets, respectively. For additional data processing we invoke standard Z-score normalization by setting the average measured value for a particular gene across all of the cell lines to zero and normalize each value by the group’s standard deviation (22). The Z-score normalization was used to provide dimensionless data designation across the microarray and GI50 datasets. This normalization has no effect on the correlation analysis provided herein.

In order to capture similar differential gene expressions among these datasets we look to similarities between the total expression profile across the 60 cell lines. In Fig. 1 we display the normalized distribution of correlation coefficients of the data vectors for each gene that has two or more members within the same Unigene cluster but restricted to

5 Internet address: http://dtp.nci.nih.gov.
6 Internet address: http://www.genome.wi.mit.edu/MPR.
the same microarray. There were 987 such genes in the Millennium data set, 869 in the Stanford set, and 520 in the Whitehead dataset. The correlation coefficient of two expression profiles $\hat{A}$ and $\hat{B}$ across the 60 cell lines was calculated from:

$$r_{\hat{A},\hat{B}} = \frac{\hat{A} \cdot \hat{B}}{|\hat{A}||\hat{B}|}$$  \hspace{1cm} (1)$$

These three distributions of the set of pairwise correlation coefficients between respective microarrays are skewed towards a slightly positive correlation. At the 5% significance level, as derived from the randomized distribution, we find from numerical integration that only 11% of the measurements are correlated in the Stanford datasets compared with 19% and 36% in the Whitehead and Millennium datasets, respectively. Thus, while the bulk of the measurements within each microarray experiment appears to be noise, significant information appears to be retained above background (i.e. random) levels. A comparison of the gene expression profiles across the 60 tumor cell lines between the same genes (as defined by Unigene) across these datasets reveals a similar picture. Thus, in Fig. 2 we show the normalized distribution of all of the pairwise correlation coefficients of data vectors for each dataset compared with data vectors derived from a random distribution. At a 5% significance level about 26–37% of the data vectors are correlated with each other.

Our next step is to order each gene expression pattern according to the existing set of response profiles available from chemical screens across the 60 cell-line panel. The basic premise for correlating gene expression and chemical response data assumes an underlying relationship between chemical activity (GI$_{50}$) and gene expression, that when challenged with chemicals evokes a cytotoxic response. The exact nature of this response can be quite complex, most likely involving multiple targets of biochemical pathways. Regardless of mechanism, our premise is that correlations between gene expression patterns and GI$_{50}$ patterns indicate, albeit crudely, linkages between chemical response and gene expression.

Our strategy for making this linkage lies in our previous analysis of the NCI’s screening data. As noted earlier, we have clustered GI$_{50}$ values for $\sim$36 K screened compounds (22) to organize this data into $\sim$1 K clusters that represent different types of cellular activity using a self organizing map (SOM). Using this organization, the gene expression profiles are matched to similar GI$_{50}$ profiles from screened data. The matching projection is done by calculating the Euclidian distance between the data vector and all of the node vectors of the GI$_{50}$ map, and selecting the location with the minimal distance. These projections of gene expression data onto chemical response data are a means of relating each measurement according to its activity pattern across the 60 tumor cells. However, our projections are not based on the complete set of measured genes for each microarray, rather the analysis is conducted only on the concordant subset of gene expressions. Thus, the gene expression data is first filtered according to concordance, then projected onto chemical response space.

An appropriate question about gene projections onto chemical clusters is the reliability of placement. The similarity measure for map projections is their Euclidean distance. The data vector from gene expressions measured across the 60 tumor cells is placed in the cluster having the smallest Euclidean distance. We estimate the a priori probability of a chance occurrence of having two data vectors coprojecting to the same location on the GI$_{50}$ map by calculating the ratio of all of the vectors that coproject to those that do not. This yields a $P$ of $3.8 \times 10^{-3}$ for the coprojection procedure for finding significantly differentiated gene expressions in the datasets.

A total of 106 genes are identified based on pairwise filtering of the Millennium and the Stanford datasets, 175
from the Stanford and the Whitehead datasets, and 154 from the Millennium and Whitehead datasets. This gives a total of 376 unique genes that survive this filtering technique. A listing of these genes is available together with their Unigene identifier. The 376 selected genes represent many types of cellular functions, although this set is dominated by genes involved in signal transduction, with the remainder falling in broadly defined categories: adhesion/extracellular matrix, tumor suppressor, ribosomal/transcription, immune response, melanoma, and proliferation clusters. Only 29 of them have been characterized as housekeeping genes (25).

The recent paper by Staunton et al. (26) attempts to identify drug-gene relationships in ways similar to ours but using a more extensive prefiltering of the datasets into cells of extreme drug (in)sensitivity. While our analysis cannot be directly compared with theirs, we find a 34% overlap between their sets of reported genes and those found by us to convey the most information in our analysis.

As examples we provide here a brief description of genes that appear to have strongly concordant gene expression profiles across these three datasets. We emphasize strongly that where these genes have similar response patterns across the 60 tumor cell lines, conclusions about these observations in regard to cell function are not addressed.

**EDNRB** (Hs.82002, endothelin receptor type B, cluster 16.8) or endothelin receptor B is expressed in all of the human melanoma cell lines, though metastatic melanoma expresses this receptor relatively less (27, 28). Inspection of the response pattern for the tumor cell panel reflects the high expression of EDNRB within the melanoma panel (data not shown). A similar strong pattern is observed for a smaller set of breast cancer cell lines.

**FN1** (Hs.287820, Fibronectin 1 or LETS, cluster 7.22) is a fibronectin, which is an important class of extracellular multi-adhesive matrix proteins. As such, fibronectins are ligands to the integrin family of cell adhesion molecules and partake in the regulation of cytoskeletal organization. The strong signal for fibronectin expression has also been corroborated by previous measurements of cancer expression profiles using a variety of alternative methods (1, 29, 30). The strong fibronectin signal within the renal panel lines is quite evident and coincides with the observations that fibronectin may be a critical factor in the regulatory role of extracellular matrix proteins in metastatic invasion of renal cancer cells (31).

**LCP1** (Hs.76506, Lymphocyte cytosolic protein 1 (L-plastin), cluster 23.9) is an actin regulating protein. Structural proteins like actin, may be involved in the development and progression of cancer (32). Regulation of these genes is accomplished by a number of genes, L-plastin among them. L-plastin is an actin binding protein that has tissue-specific expression patterns. L-plastin is specifically expressed in hematopoietic cells but has also been found to be highly expressed in cell lines derived from mammary solid tumors. Dysregulation of actin-binding proteins during carcinogenesis may, thus, be the direct link between the observed up-regulation of L-plastin in the cancer cell lines, although the exact role or L-plastin in the tumor process remains unknown (33). Upregulation of L-plastin has been linked to testosterone in breast and prostate cancer cells (34). This observation might suggest a corresponding subpanel sensitivity to testosterone. The expression profile of L-plastin is strongest within the leukemia and breast cancer panels, near a region on our anticancer map demonstrated to have sensitivity to selected steroid molecules, NSCs 624018 and 633664.

**MCAM** (Hs.211579, melanoma adhesion molecule, MUC18, cluster 15.7) is a transmembrane glycoprotein and is a member of the immunoglobulin superfamily. The protein is closely related to a number of cell adhesion molecules. Tumor progression and metastasis in human malignant melanoma is associated with MCAM. Consistent with this expression pattern we observe enhanced expression activities in the melanoma panel.

**S100P** (Hs.2962, S100 calcium-binding protein P, cluster 8.10) is a low molecular weight calcium-binding protein, which is associated with the regulation of cellular processes such as cell cycle progression and differentiation. Overexpression of S100P has been postulated to play an important role in the immortalization of human epithelial cell in vitro and in tumor progression in vivo (35). Other S100 calcium binding proteins are also found to be correlated among these three expression datasets. These include the S100A4 gene \( r = 0.70, P < 0.01 \), whereas the S100B gene expression is only weakly correlated \( r = 0.27, P < 0.01 \). S100P is down-regulated after androgen deprivation in an androgen-responsive prostate cancer cell line (36). As in the L-plastin case described above, the gene expression profile of S100P is most similar to a region on our anticancer map that is sensitive to steroid molecules, NSC 689621 and 652123.

It is important to note that previous analysis of portions of these datasets have also identified L-plastin and S100P as important genes. The methods used in these reports were considerably more complicated that the simple filtering method proposed here.

**Identifying Molecules That Affect Expression Levels**

Relating expression patterns from cell lines to a chemical response represents an important validation step, usually involving considerable biochemical effort. As an alternative we seek verifications based on surveys of the PDB structural library of ligand complexes. The steps taken in this verification are outlined in Fig. 3. The procedure for relating chemical response and gene expression levels is to first identify the set of proteins within the PDB that are homologous to the genes that are coprojected on the GIs map. The entire projected gene datasets that have a homologous PDB sequence spans the entire map. The number of unique genes on the microarray that have at least one ligand-bound PDB homolog is 1231, 1197, and 1523 for the Millennium, Stanford, and Whitehead datasets, respectively. The average coverage of the complete SOM map for these expressed genes is 71%. The imposed concordance criteria narrows the selection to the most likely gene/drug associations. Methods of protein sequence alignments were performed using FASTA, version 3.3, with standard gap-parameters and the BLOSUM50 similarity matrix (37). The next step is to determine whether there
is a structural match between the ligand bound in the PDB structure and screened NSC compounds. The complete set of ligands within the PDB database were extracted by scanning for heteroatom records, and any fully or partially present ligands were collated as a PDB ligand. This includes ligands that are associated with only DNA records as well as modified residues that are covalently attached to other residues. This collation provides a sample of possible ligand/protein and ligand/DNA interactions. Small ions and unsuitable metal ligands were deselected from this list, leaving a total of 1919 PDB small molecule ligands suitable for structural comparisons to the screened NSC compounds.

In order to describe the chemical similarity of the PDB ligand to the compounds deposited in the NCI database we use a bit-vector assignment to describe each molecule. This is an electronically convenient way of describing a molecule in order to catch the flavor of its possible interactions. In such a description the molecule is dissected for properties that are coded in an on/off bit, e.g., presence or absence of aromatic fragments, carboxyl groups, hydrogen-bond donor, and so forth. We have used the properties defined by the regular E-screen bit-vectors (38), which encodes 431 bits. We then use the Tanimoto coefficient as measure to identify compounds containing similar chemical elements or fragments via a bit-vector similarity, defined as the number of bits in common divided by the total number of bits. This is calculated for bit-vectors $\vec{A}$ and $\vec{B}$ as:

$$T(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}| + |\vec{B}| - \vec{A} \cdot \vec{B}}$$

(2)

The Tanimoto coefficient is a measure of the number of common substructures shared by two molecules as described by this bit-vector mask. In this work the E-screen bit-vector mask was generated and used as a similarity measure between NSC compounds and PDB ligands. Although a high similarity does not guarantee that two compounds will behave the same in a biological screen, high structural similarity can be used to identify structural binding motifs of similar compounds bound to a protein target (39, 40).

Similarities between molecules are thus measured via the Tanimoto coefficient of a discrete bit-vector of length 431 for each compound. This Tanimoto coefficient identifies common molecular fragments between two compared molecules and ranges from 0 to 1. In this case we have used a cutoff of 0.75 as being of significant similarity (41). Thus, if we find a similar ligand in the PDB we query the parent structure for its function, and if its function is similar to that of the original query gene we consider evidence for verification of a target-dug association. Because the number of ligands in the PDB is rather modest, we cannot expect to verify each individually selected significant gene; instead we use this process to verify the basic premise of similarity between gene expression and drug response.

In order to estimate the joint occurrence of a coprojection and the chance occurrence that a PDB ligand has a Tanimoto score $>0.75$ with a NSC compound we calculate the ratio of all of the PDB ligand: NSC compound pairs that have such a Tanimoto coefficient to those that do not. This yields a $P$ of $8 \times 10^{-4}$. The a priori probability of a joint occurrence of these two events is then the product of these two probabilities and yields a final $P$ of $3 \times 10^{-6}$ for the procedure.

We have used a pairwise comparison strategy to extract information from the three gene expression datasets. Our analysis finds evidence for 11 putative chemical-gene relationships. While this number represents a low percentage of the total number of concordant genes, the remaining not-yet-verifiable genes represent the subject of future investigations into their potential chemical-gene relationships.

**Genes and Chemicals Connected via Cellular Profiles**

The connection between gene expression and GI50 values is investigated for each pairwise concordant set of microarray experiments, Millennium-Whitehead, Millennium-Stanford, and Whitehead-Stanford. Genes identified by our procedure as having a corresponding PDB ligand molecule are listed in Table 1, and their cross-reference to PDB and NSC ligands is listed in Table 2. Fig. 4 provides structural representations of the NSC compounds and their analogous PDB ligand structure. Note that these structural conformations are arbitrary and might not represent the actual bound conformation. The following sections briefly discuss the 11 verifiable cases.
The gene expression profile of CAMK1 (calcium/calmodulin-dependent protein kinase I) is most similar to the G150 profiles observed for a set of chemicals that includes NSC compound 618487, which is identical to the PDB ligand STO shown in Fig. 4A. Thus, a direct connection is established between the gene expression of CAMK1 and the chemical response of STO against its PDB ligand target. In this instance both the tested ligand and the crystallized ligand are identical, to leave little doubt about this connection. The phosphorylation and inhibition of CAMK1 by other calcium-dependent kinases also makes it a likely candidate to be involved in modulating the balance between cAMP- and Ca\(^2+\)-dependent signal transduction pathways (42). CAMK1 is homologous to several other kinase proteins in the PDB database: 1STC, 1BXG, and 1CKP, which contain either the catalytic subunit of the cAMP-dependent protein kinase \(\alpha\) or a human cyclin-dependent kinase 2. Our analysis permits only speculations about the potential binding of STO to these other kinase molecules. Examination of the cellular profiles finds the renal panel to be most sensitive to the STO. Surveys for PDB proteins homologous to CAMK1 find an \(\alpha\)-catalytic subunit of a cAMP-dependent protein kinase (1STC). This observation is significant, because 1STC also shares homology with the MAP2K4 sequence. Mitogen-activated protein kinase pathways are signal transduction cascades with distinct functions in mammals. MAP2K4 is a potent physiologic activator of the stress-activated protein kinases. 1STC is bound by a ligand having structural similarity to NSC compound 645327 shown in Fig. 4A. Although this compound and STO are chemically quite different, both compounds display some structural similarity in their fused ring systems that might suggest a common pharmacophore and cellular activity.

PDGFRA (platelet-derived growth factor receptor, \(\alpha\) polypeptide, PDGFR2) is a membrane-spanning growth factor receptor with tyrosine kinase activity. Overexpression of the PDGFRA subcomponent in the PDGF signaling system has been implicated in the development and malignant progression of diffuse gliomas (43). From their similarities in cellular response profiles, we identify NSC compound 672971 as a candidate ligand based on its structural similarity to the PDB ligand ANP 5'-adenyly-imido-triphosphate, DHT, and 3-beta-hydroxy-5-androst-17-one, QUE 3,5,7,3'-4'-pentahydroxyflavone (quercetin), VK3 menadione, OBA 2-(oxalyl-amino)-benzoic acid, PLH methylamino-phenylalanly-leucyl-hydroxamate, AZE axorphenthine, IUS iso-uroxyecholic acid and ALD carboxybencyleucinyleucynyleucinyleucinal.

### Table 1 Gene designations

<table>
<thead>
<tr>
<th>Gene</th>
<th>UniGene</th>
<th>Millennium</th>
<th>Stanford</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAMK1</td>
<td>Hs.184402</td>
<td>GC22306</td>
<td>GC18090</td>
<td>10.23</td>
</tr>
<tr>
<td>MAP2K4</td>
<td>Hs.75217</td>
<td>GC25430</td>
<td>GC14360</td>
<td>12.15</td>
</tr>
<tr>
<td>PDGFRA</td>
<td>Hs.74615</td>
<td>GC25705</td>
<td>GC10098</td>
<td>19.1</td>
</tr>
<tr>
<td>BDH</td>
<td>Hs.76893</td>
<td>GC20746</td>
<td>GC15388</td>
<td>27.15</td>
</tr>
<tr>
<td>PRKCB1</td>
<td>Hs.77202</td>
<td>GC23444</td>
<td>GC14583</td>
<td>27.6</td>
</tr>
<tr>
<td>PRKCB1</td>
<td>Hs.77202</td>
<td>GC23445</td>
<td>GC14583</td>
<td>27.6</td>
</tr>
<tr>
<td>DIA4</td>
<td>Hs.80706</td>
<td>GC21197</td>
<td>J03934</td>
<td>15.10</td>
</tr>
<tr>
<td>PTPRC</td>
<td>Hs.17012</td>
<td>GC25403</td>
<td>Y00062</td>
<td>25.6</td>
</tr>
<tr>
<td>MMP1</td>
<td>Hs.83169</td>
<td>GC26040</td>
<td>X54925</td>
<td>7.17</td>
</tr>
<tr>
<td>APOD</td>
<td>Hs.75736</td>
<td>GC19985</td>
<td>J02611</td>
<td>8.17</td>
</tr>
<tr>
<td>ADH5</td>
<td>Hs.78989</td>
<td>GC18296</td>
<td>M81116</td>
<td>10.14</td>
</tr>
<tr>
<td>CTHS</td>
<td>Hs.76476</td>
<td>GC10087</td>
<td>X16832</td>
<td>13.6</td>
</tr>
</tbody>
</table>

| Gene/PDB linkage

Cross-references of gene data with the PDB and the underlying NSC compounds, and its associated PDB ligand. The expectation value of the alignment of the GenBank EST from the Unigene clustering of ESTs with the PDB sequence is given in the fourth column. The Tanimoto coefficient between the NSC compound and the PDB ligand is given in the last column. The PDB ligands associated with the protein structures are STO, ANP 5'-adenyly-imido-triphosphate, DHT, and 3-beta-hydroxy-5-androst-17-one, OUE 3,5,7,3'-4'-pentahydroxyflavone (quercetin), VK3 menadione, OBA 2-(oxalyl-amino)-benzoic acid, PLH methylamino-phenylalanly-leucyl-hydroxamate, AZE axorphenthine, IUS iso-uroxyecholic acid and ALD carboxybencyleucinyleucynyleucinyleucinal.

### Table 2 Gene/PDB linkage

<table>
<thead>
<tr>
<th>Gene</th>
<th>EST</th>
<th>PDB</th>
<th>E-value</th>
<th>NSC</th>
<th>Lig</th>
<th>Tan</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAMK1*</td>
<td>L14186</td>
<td>1STC</td>
<td>1.1e-27</td>
<td>618487</td>
<td>STO</td>
<td>1.00</td>
</tr>
<tr>
<td>MAP2K4*</td>
<td>L36870</td>
<td>1STC</td>
<td>4.4e-09</td>
<td>645327</td>
<td>STO</td>
<td>0.75</td>
</tr>
<tr>
<td>PDGFR4*</td>
<td>M21574</td>
<td>2SRC</td>
<td>1.4e-25</td>
<td>672971</td>
<td>ANP</td>
<td>0.93</td>
</tr>
<tr>
<td>BDHa</td>
<td>M93107</td>
<td>1DHT</td>
<td>1.3e-13</td>
<td>92227</td>
<td>DHT</td>
<td>0.86</td>
</tr>
<tr>
<td>BDHa</td>
<td>M93107</td>
<td>3DHE</td>
<td>1.3e-13</td>
<td>92227</td>
<td>AND</td>
<td>0.83</td>
</tr>
<tr>
<td>PRKCB1*</td>
<td>X07109</td>
<td>2HCK</td>
<td>8.2e-07</td>
<td>169517</td>
<td>QUE</td>
<td>0.82</td>
</tr>
<tr>
<td>DIA4b</td>
<td>J03934</td>
<td>2QR2</td>
<td>1.0e-44</td>
<td>11897</td>
<td>VK3</td>
<td>0.89</td>
</tr>
<tr>
<td>DIA4p</td>
<td>J03934</td>
<td>2QR2</td>
<td>1.0e-44</td>
<td>651207</td>
<td>VK3</td>
<td>0.76</td>
</tr>
<tr>
<td>PTPRCa</td>
<td>Y00062</td>
<td>1C8S</td>
<td>8.1e-35</td>
<td>635528</td>
<td>OBA</td>
<td>0.79</td>
</tr>
<tr>
<td>MMP1b</td>
<td>X05231</td>
<td>1MNc</td>
<td>1.4e-41</td>
<td>672675</td>
<td>PLH</td>
<td>0.76</td>
</tr>
<tr>
<td>APOD</td>
<td>J02611</td>
<td>1FEN</td>
<td>1.6e-06</td>
<td>122759</td>
<td>AZE</td>
<td>0.75</td>
</tr>
<tr>
<td>ADH5a</td>
<td>M30471</td>
<td>1DDA</td>
<td>2.9e-100</td>
<td>49452</td>
<td>IUS</td>
<td>0.75</td>
</tr>
<tr>
<td>CTHSf</td>
<td>X16832</td>
<td>1BP4</td>
<td>1.5e-24</td>
<td>679678</td>
<td>ALD</td>
<td>0.84</td>
</tr>
</tbody>
</table>

* Millennium and Stanford dataset.

* Millennium and Whitehead dataset.

* Stanford and Whitehead dataset.

* Millennium and Stanford dataset.
src family kinase, hck. The similarities in cellular profiles of the gene expression of PRKCB1 and the GI50 response pattern to the compound QUE are consistent with the sequence similarities of its target protein, and structural similarities between NSC compound 169517 and the kinase hck-bound ligand QUE.

DIA4 is part of the detoxification process of quinones derived from the oxidation of benzene metabolites. Diaphorase can also activate bioreductive anticancer drugs. Down-regulation of diaphorase has been shown to induce gastric cancer in certain cell lines (48). Menadione is present in the PDB as a ligand to the human quinone reductase type 2, and two analogous NSC compounds 11897 and 651207 are found to have similar gene expression profiles. Their structural similarity is given in Fig. 4B.

MMP1 is a matrix metalloproteinase that helps to break down interstitial collagen. Overexpression of MMP1 in tumor cells is indicative of the invasive nature of cancer. In the PDB there exists a structure of the catalytic domain of the metalloprotease neutrophil collagenase. The inhibitor bound to this enzyme is PLH, which shares common structural elements with the NSC compound 672675 in Fig. 4B. The gene expression profile projects to the nearest neighbor of the cluster containing this compound, providing a tentative link between chemical agent and gene.

APOD is a member of the a (2 µ)-microglobulin superfamily of carrier proteins termed lipocalins. It shares a high degree of homology to retinol-binding protein. This homology allows us to assign the PDB structure 1FEN as possessing similarities with the APOD gene. The axerophthene ligand 635526. This molecule is analogous to the PDB ligand OBA in PDB deposition 1CB5, which contains the structure of a protein tyrosine phosphatase 1B. The similarity between the ligand and the NSC compound shown in Fig. 4B, coupled with the closely matched gene/protein function, clearly establishes their gene/chemical relationship.
is closely analogous to the NSC compound 122759 and shares its cellular profile with the gene expression for AP0D. The strong similarity between the ligand and the NSC compound in Fig. 4B is evidence for a tentative gene/drug relationship between retinol-like molecules and lipocalins.

**ADHS** (class III), \( \chi \) polypeptide is a protein of which the specific function in humans is largely unknown. There exists a highly homologous protein model in the 1DDA PDB deposition, which is an ADH complexed with isoursodeoxycholic acid, a steroid. An analogous NSC steroid compound shown in Fig. 4C, 49452, is found to have a strongly similar gene expression profile, indicating a tentative relationship between these two data profiles.

**CTSH** belongs to a class of cystein-dependent intracellular proteases. The cathepsins have an important function in regulating intracellular protein degradation. The up-regulation of cathepsin gene transcription appears to be characteristic for invasive tumor cells (49). In the structural deposition of 1BP4, papain has been used as a model to test cathepsin inhibitors. The PDB ligand carbenzyloxyleucinyl-leucinyl-leucinal shares structural similarity to the NS compound 679678 as shown in Fig. 4C, providing a link between the protease functions and the activity of structurally similar ligands that may bind cathepsin.

**Conclusion**

Computational tools aimed at improving our understanding of chemotherapeutic cancer pharmacology can also aid the drug discovery process. In this paper we present an analysis that links chemical response space via GI\(_{50}\) measurements to a set of expression profiles of specific gene targets. Because the cellular environment in which these drugs act is very complex it is advantageous to derive an understanding about what genes are affected by which compounds. Our computational tools are not specific enough to pinpoint all of the possible chemical/gene interactions, but they do serve to provide initial hints about which processes or pathways might be affected, either directly or indirectly. Information of this type provides valuable insight into cause and effect, examinations that serve as the basis for additional biochemical studies.

Using a methodology that seeks similarities in cellular response patterns derived from gene expression measurements and chemical screens, connections between gene and chemical space can be made. Our procedure is grounded in the premise that these similarities in cellular response represent associations between gene products and chemical activity. We additionally verify this association by identifying small structurally similar compounds that imply a putative connection to chemotherapeutic cancer pharmacology. These latter relationships are verified here for 11 test cases. Although not emphasized in this work these measures also allow us to differentiate gene/chemical responses based on different cell lines and, thus, also on clinically different cancer types. This may aid the identification of drugs that are specific for certain types of cancers and provide a tool for focusing efforts in the drug discovery process.

Different methods for identifying gene-chemical associations have been proposed by Butte *et al.* (16) and by Scherf *et al.* (15), who also describe the paucity of verifiable connections possible from this same dataset; the former case revealing 1 and the latter case another of the 11 associations reported here. The difference between our approach and theirs is the use of multiple datasets as surrogate replicate measurements of the same data, then filtering these data based on concordant response patterns and finally verifying our gene-chemical relationships by seeking actual structural cases. We find, with reasonably high confidence, assignments of gene-drug relationship for 11 verifiable cases, comprising drug binding to a variety of targets. Known kinase effector molecules taken from the PDB were positively correlated with their corresponding genes and NSC compounds based on similarities in their cellular response profiles. Likewise the BDH gene was found to be projected to a cluster on our WEB-accessible anticancer map with known steroid activity, which could be verified by the corresponding hydroxysteroid dehydrogenase ligand and structure in the PDB archive. None of these 11 associations appears to be spurious, although this cannot be ruled out without additional biochemical investigations of each specific system.

The wealth of data accompanying the post-genomic era offers high promise for understanding cellular processes and deriving strategies to affect these systems. Harvesting this information will not be simple. As our investigation reveals, this data can be quite noisy, but when confronted with data of poor quality, additional computational efforts can be utilized that lead to the extraction of meaningful information. These results are not unanticipated, given that these analyses involve quite large amounts of data that are collected from extremely complex biological systems. Additional complications related to this system are that these measurements are made on somewhat artificial cell lines and not real tumors (1, 8, 50), the GI\(_{50}\) experiments are single valued measurements of a highly complex system, and that only a subset of all the genes in the cell are represented on the microarray chip. Strategies to overcome these criticisms will be devised. Our approach offers one solution by exploring chemical and genetic links that in most cases cannot be easily verified by other means than the route taken here. This strategy does offer hope, by revealing a small set of gene/drug linkages that can be additionally exploited as possible novel data in the search for new chemotherapeutic strategies.

**References**


