Linking the growth inhibition response from the National Cancer Institute’s anticancer screen to gene expression levels and other molecular target data

Anders Wallqvist1,* Alfred A. Rabow1, Robert H. Shoemaker2, Edward A. Sausville2 and David G. Covell2

1Science Applications International Corporation and 2Developmental Therapeutics Program, National Cancer Institute at Frederick, NIH Frederick, MD 21702, USA

Received on December 20, 2002; revised on April 23, 2003; accepted on May 6, 2003

ABSTRACT

Motivation: Data mining tools are proposed to establish mechanistic connections between chemotypes and specific cellular functions. Drawing on a previous study that classified the cellular response patterns of growth inhibition measurements \( \log(\text{GI}_{50}) \) from the National Cancer Institute’s (NCI’s) anticancer screen, we have examined additional data for mRNA expression, sets of known molecular targets and mutational status against these same tumor cell lines to relate chemosensitivity more precisely to biochemical pathways.

Results: Our analysis finds that gene expression levels do not, in general, correlate with \( \log(\text{GI}_{50}) \) measurements, instead they reflect a generic toxic condition. Within the remaining set of non-generic conditions, examples were found where a correlation suggesting a biochemical basis for cellular cytotoxicity could be supported. These included reconfirmation of previously observed associations between mutant and wild-type status of p53, and chemosensitivity to alkylating agents, while extending these results to reveal associations with \( \gamma \)-induced expressions of MDM2, WAF1 and GADD45, signals that were not apparent in measurements of basal mRNA expression levels for any of these genes. Additional examinations revealed that mRNA expression levels directly correlated with paclitaxel chemosensitivity to mitosis, while also identifying additional chemotypes as P-glycoprotein substrates. Our analysis revealed well-known direct associations between \( p16 \) mutant status and chemotypes implicated in cell cycle control, and extended these results to include expression levels for three additional tyrosine kinase proteins (TEK, transgelin and \( h\text{Cdc4} \)). Links were also found that suggested associations between chemosensitivity and the endocrine, paracrine ligand–receptor loops, via expression of the adrenergic receptor, calcium second messenger pathways via expression levels of carbonic anhydrase and cellular communication pathways via fibrillin.

Contact: wallqvist@ncifcrf.gov

1 INTRODUCTION

Recent advances in genomics and bioinformatics offer tremendous potential for the identification of therapeutic targets and the development of small molecule drugs (Reidhaar-Olson et al., 2001). The areas of high-throughput drug screening (HTS) and gene expression monitoring represent two key technologies in establishing a molecular basis for drug activity (Schellens et al., 1996; Krynetski and Evans, 1999). Accompanying these technological achievements has been the transformation of biological studies from small-scale experiments to high-throughput screening processes that generate large data sets, require sophisticated analytical tools and often lead to complicated interpretations. The analysis presented here focuses on integrating large-scale cell-based drug screening data with gene expression monitoring using high-density microarray data.

Publicly available data sources are used for our analysis. All data sets originate from screening and microarray measurement available from the National Cancer Institute’s (NCI’s) anticancer drug discovery program. This screening effort was originally established to test small synthetic molecules in the NCI compound repository for potential antitumor activity (Boyd, 1995; Boyd and Paull, 1995). Historically, the power of these measurements has been derived from pairwise statistical correlations of \( \log(\text{GI}_{50}) \), for drugs tested against various tumor cell lines; the results of which have been used to establish similarities in mechanisms of action, modes of resistance and molecular structure (Paull et al., 1989; O’Connor et al., 1997; Weinstein et al., 1997; Shi et al., 1998a,b). Currently, the screening database (http://www.dtp.nci.nih.gov) contains \( \log(\text{GI}_{50}) \) data for over 100 000 compounds tested in 60–100 tumor cell lines.
Linking growth inhibition response to gene expression levels

and represents the largest publicly available database of this type. A second source of information is the molecular target (MT) data set, which consists of a compilation of various experiments performed across the 60 tumor cell lines in the NCI screen. Recently, these same tumor cells have been monitored for gene expression using cDNA arrays (Ross et al., 2000). Several attempts to integrate the NCI’s HTS data with gene expression profiles have been proposed (Ross et al., 2000; Butte et al., 2000; Scherf et al., 2000; Staunton et al., 2001; Wallqvist et al., 2002). All these approaches require algorithms that enhance detection of features within each separate data set and, then, use these relationships to propose links between chemical response and gene expression.

The approach proposed here focuses on four measured phenotypes (e.g. growth inhibition, mRNA expression, mutational status and MT measurements) across all 60 tumor cell lines in the NCI screen. Statistical means are developed to distinguish global, non-specific cellular response patterns from the subset of patterns that can be supported as unique to this set of data. It is within this set of unique patterns that chemical associations between chemotypes and biochemical pathways, vis-à-vis gene expressions or MTs, are postulated. The analysis assumes that large-scale data collections inherently contain considerable noise and the challenge is to provide effective methods to separate meaningful signals from noise. We identify a subset of associations between cellular phenotypes that reinforce previous notions regarding chemosensitivities and reveal novel associations between drug subsets and cellular functions.

2 DATA TREATMENT

Our analysis begins with the recently published results from the publicly available log(GI50) data in the NCI’s tumor screening database (Rabow et al., 2002). This analysis organized over 20K compounds using a self-organizing map (SOM) procedure (Kohonen, 1995), updated here to include the 36K compounds available in the most recent Developmental Therapeutics Program (DTP) release. The SOM method uses iterative minimization procedure to group multidimensional input data into lower dimensional, characteristic response vectors. The weighted distance between the data vectors and the response vectors are minimized. The method assigns relatively more response vector to regions that are data rich and as a result finely divide the data space of high information content. The resultant response vectors are then projected to a two-dimensional (2D) space for visualization. Data generated in the DTP screen comprises growth inhibition for each screened compound across the NCI’s 60 tumor cells. The multivariate nature of this data (growth inhibition measurements across 60 tumor cells for over 36K compounds) has been the subject of many proposed analyses (Paul et al., 1989; O’Connor et al., 1997; Weinstein et al., 1997; Shi et al., 1998a,b; Butte et al., 2000; Ross et al., 2000; Scherf et al., 2000; Staunton et al., 2001; Wallqvist et al., 2002; van Osdol et al., 1994; Keskin et al., 2000). Among the statistical approaches proposed for analysis, clustering methods have been shown to offer considerable capacity for ‘mining’ this data. Notably, our previous study in this area (Rabow et al., 2002) has revealed novel cellular response categories when using non-hierarchical clustering based on SOMs. These results are in contrast with earlier efforts based largely on methods of hierarchical clustering, principal component analysis, k-means clustering and support vector machines (van Osdol et al., 1994; Keskin et al., 2000; Musumarra et al., 2001a,b), where variable degrees of success have been found for associating chemotypes and mechanism of action. SOM visualizations of these clustering results are made using a 2D hexagonal projection, as shown in Figure 1a, comprising 1066 clusters (also referred to as nodes). This data reduction was shown to aid in the interpretation of cellular sensitivity profiles across the 60 cell lines relevant for identifying relationships between specific groups of compounds and their putative biological activity (Rabow et al., 2002). The cell-based nature of this screen is thought to reflect biological activity; however, the chemical activity of each tested reagent is believed to be largely responsible for the measured cytotoxicity. The details of this analysis are available in Rabow et al. (2002) and at our web site http://spheroid.ncifcrf.gov. For purposes of the study described here, each of these 36K compounds are assigned to one of the 1066 clusters of our SOM. Compounds within each cluster have a unique pattern of growth inhibition, and the node to which these compounds cluster is defined as a target fingerprint. Thus, a fingerprint encodes the growth inhibition across the 60 tumor cell lines resulting from exposure to compound sets contained within each SOM node.

Other measurements that have been made across the 60 tumor cell lines can also be correlated with the log(GI50) data set via their cellular response. Thus, we have used two additional publicly available data sets of measurements across the 60 tumor cell lines. The first set of data contains 6.5K mRNA expression profiles of the constitutive gene expression in the 60 cell lines, as measured using cDNA microarray technology (Ross et al., 2000). The second data set is the so-called MT database which is accessible from the DTP web server at http://www.dtp.nci.nih.gov. This latter data set contains a wealth of different types of measurements, including enzyme activity and mutational status, all measured across the 60 tumor cell lines. There are 195 cellular profiles from the MT data set.

We filtered the gene expression data across the 60 cell lines (Ross et al., 2000) to include measurements that exhibited the strongest intensity in signal as well as variability across tumor cell lines. Data was included if at least 28 cell lines had an intensity measurement above 1500 and had at least a factor of 3 difference in gene expression for a minimum of three cell lines (Dan et al., 2002). This criteria yielded a data set of 998 genes to be included in our analysis. A variety of...
Fig. 1. (a) The SOM of the NCI’s anticancer drug screen based on the log(GI50) responses for 10 cancer cell panels comprising 80 different cell lines (Rabow et al., 2002). Our SOM analysis currently organizes over 36K compounds into 1066 clusters in a 26 × 41 2D hexagonal map. Each cluster is characterized by a response fingerprint, based on cellular chemosensitivity of the agents within a cluster. The color scheme on the map is indicative of the similarity between neighboring nodes; the red color corresponds to high similarity whereas the brown and dark blue colors describe low similarity. Four broad regions of cellular response have been characterized via literature verifications: mitosis in the M region of the map, nucleic acid biosynthesis in the S regions, phosphatase/kinase-mediated cell cycle regulation in the P regions and membrane transport in the N region. The different regions were delineated on the map using a hierarchical clustering scheme of the node fingerprints. The resulting clades define in a more general manner the different regions on the map corresponding to different mechanism of actions. Work is ongoing to delineate further the Q and R regions, which so far have not been characterized. The 1066 SOM clusters define the target fingerprints used in the analysis reported herein. For a fully searchable map and instructions on how to utilize this information the reader is directed to our website http://spheroid.ncifcrf.gov. (b) The bulk of the data analyzed here comprises a generic response according to its correlation map. Each test fingerprint is compared with all target fingerprints and used to establish a similarity score. This score is normalized with the underlying standard deviation $\sigma$ of the random distribution and all maps are expressed in terms of this standard deviation. In the coloring scheme red is indicative of high correlations, whereas blue indicates regions of anti correlations. Seventy percent of the response patterns consist of two anti-correlated maps. Shown in the figure are typical correlation responses for gene expression levels with the entire SOM in the two characteristic responses. The origin of the high $\sigma$ values shown is primarily due to the leukemia cell lines, which show a strong correlated response with the mRNA gene test fingerprints. The strongest correlation patterns are located in the lower left-hand side of the maps and in the upper right-hand corner. The mRNA patterns in these regions exhibit a preponderant concordance with genes involved in cellular growth.
methods have been proposed for filtering raw expression data (Ramassawmy et al., 2001; Staunton et al., 2001; Su et al., 2001; Zembutsu et al., 2002); all of which select gene subsets with the most diverse response signal. Our level of filtering is comparable with many of these published methods.

No filtering was performed on the MT data; thus, all this information was included for analysis. Each of these data sets are freely available from http://www.dtp.nci.nih.gov. The log(GI50) values, gene expression data, as well as the MT data were Z-score normalized as described before (Rabow et al., 2002). This normalization serves to extract the variability of response across the cell lines to yield a characteristic cellular sensitivity profile or fingerprint.

Similarity comparisons are made between each target fingerprint and the fingerprints for measurements of mRNA expression, MTs and mutational status, each fingerprint being referred to as the test fingerprint. The similarity is calculated as the Pearson product-moment correlation coefficient \( r_{ij} \) between the \( i \)-th target fingerprint from all SOM nodes and the \( j \)-th gene expression or MT test fingerprint, defined by

\[
 r_{ij} = \frac{\sum_{k=1}^{60} (x_{ik} - \bar{x}_i)(y_{jk} - \bar{y}_j)}{\sqrt{\sum_{k=1}^{60} (x_{ik} - \bar{x}_i)^2} \sum_{k=1}^{60} (y_{jk} - \bar{y}_j)^2} \tag{1}
\]

where the summation runs over the measured properties across the 60 tumor cell lines. Implicit in this analysis is that strongly positively or negatively correlated test and target fingerprints may indicate underlying relationships between chemosensitivity and cellular processes. Alternatively, the occurrence of highly correlated fingerprints may provide potential markers useful for predicting one fingerprint on the basis of another (i.e. chemosensitivity from mRNA expression or MT expression, or gene expression, MT or mutational status are made by generating random distributions of similarity scores for all the \( i \)-th test fingerprints against each node of the SOM. All similarity scores between test and target fingerprints are then normalized according to the calculated random standard deviation, \( \sigma_{\text{random}} \),

\[
 r'_{ij} = r_{ij}/\sigma_{\text{random}} \tag{2}
\]

The \( r'_{ij} \) scores for all the \( i \)-th map positions are then displayed onto the complete SOM for each \( j \)-th gene expression or MT measurement. This information provides a visual reference reflecting the magnitude of positive or negative correlation for each test fingerprint against each node of the SOM. All comparisons between log(GI50) and gene expression, MT measurements, or mutational status are made by generating maps of correlation strength for each measurement. These maps are referred to as \( \sigma \)-relation maps, or simply \( \sigma \)-maps. In these maps, the regions of highest positive and negative correlations are color-coded in red and blue, respectively, and serve to identify associations between chemosensitivity and any of the data sets analyzed herein. The level of significance for identifying unique observations has been arbitrarily established at greater than \( \pm 3\sigma \), which corresponds to a P-value <0.0025.

A further step of our design evolved from our observations that many test fingerprints generated nearly similar to \( \sigma \)-maps or their inverse. Map-to-map correlation statistics were generated as above to detect the occurrence of statistically indistinguishable \( \sigma \)-maps. As our results will document, this additional step revealed the existence of a generic \( \sigma \)-map that included a majority of the test fingerprints analyzed in this study.

\section{3 GENERIC RESPONSE}

\subsection{3.1 Chemotypes}

In contrast to finding the one best matching fingerprint we examine the local correlation across the entire SOM. The projection of the \( \sigma \)-normalized correlation scores onto the complete SOM provides a means for inspecting whether the best similarities also appear as closely clustered neighbors within the SOM. Regions where the best correlations
are clustered together lend confidence to the existence of an actual relationship between test and target fingerprints, while a broad scattering of the best correlations across many clusters may be less likely to indicate a ‘true’ relationship. It is reasonable to assume that a given response across the tumor cell panel may result from many causes. The possibility for many factors underlying a cellular response necessitates highly critical procedures for assigning cause and effect. Map-to-map correlations provide a quantitative mean for distinguishing unique maps from unremarkable test fingerprints and, thus, aid in excluding weakly supported causalities.

Using this approach, we have determined that ~70% of the test fingerprints have nearly identical σ projections on our SOM. Thus, a majority of the test fingerprints for mRNA expression and MTs are indistinguishable among themselves and within this set of target fingerprints. This rather unexpected finding indicates that much of this data is consistent with a common response, most likely characteristic of a generalized cell poison. The compounds that are characteristic of a generalized toxic response appear to share the cytotoxic gene response profile and is thus, in our view, indicative of common or similar death pathway. These non-unique responders fall into two classes, each of which is strongly inversely correlated. Their σ-map patterns shown in Figure 1b are dominated by strong similarity scores at the lower left and upper right quadrants of the complete SOM, regions largely populated by agents with cellular sensitivity to only the panel of leukemia cell lines; all other lines having no remarkable patterns of sensitivity for the compounds clustered in these regions. The exception to this observation is a single SOM cluster where all the leukemia lines are quite sensitive to the compounds in that map cluster. This cluster includes bis-anhydrobenzoquinone, 5-methoxysterigmatocysin and hydramycin, and according to our analysis these compounds have a high general toxicity to most of the tumor cell lines. It is noteworthy that map clusters associated with the generic response are mostly distinct from clusters containing a set of 171 standard anti-leukemia agents with the test fingerprint of expression levels for the MDR gene across the 60 tumor cell lines shows a distinct pattern with the log(\(\text{GI}_{50}\)) fingerprints not repeated for any of the other measurements in the DTP’s MTs database (http://www.dtp.nci.nih.gov/mtargets/mt_index.html). The mRNA expression levels were determined by quantitative PCR (Alvarez et al., 1995). This unique pattern, shown in Figure 3a, reveals a significant and well-defined negative correlation in the anti-mitotic portion of the complete SOM, where clusters of taxanes, colchicines and stilbenes are located. This strong negative correlation suggests a testable hypothesis about these compounds regarding their sensitivity to MDR. The actual correlation of data vectors is shown in Figure 4, where we compare the target fingerprint with the test fingerprint of expression levels for the MDR gene. Based on the hypothesis that high MDR mRNA levels are accompanied by low cell sensitivity to the drugs found in

4 NON-GENERIC RESPONSES FOUND IN THE MT DATA SET

The sections below describe cases where test fingerprints have sufficient statistical merit to be considered unique from the generalized response. Our analysis will document the concordance of our analysis with previously reported findings for multidrug resistance, stress-induced effects on p53 and MDM2, and mutant status for these 60 tumor cell lines. Following these sections, we describe the set of unique MT and mRNA expression fingerprints that serve to identify novel relationships between chemosensitivity and cellular phenotypes.

4.1 Multidrug resistance (MDR)

The MDR-1 gene encodes the P-glycoprotein which plays a role in drug resistance for many anticancer drugs. The mechanism of P-glycoprotein action is to function as an efflux pump for reducing intracellular drug concentration. Observed resistance can develop over time as drug therapy progresses via a feedback loop, e.g. for the anti-mitotic drug paclitaxel, drug resistance is a direct cause of paclitaxel binding to the nuclear orphan receptor SXR, which among other functions, activates the expression of the MDR gene and consequently P-glycoprotein (Synold et al., 2001). Paclitaxel also has the ability to activate other signaling pathways, including the MAP kinase regulation of various transcriptional factors needed for cell survival and cell death (Ding et al., 2001).

The mRNA expression for the MDR gene across the 60 tumor cell lines shows a distinct pattern with the log(\(\text{GI}_{50}\)) fingerprints not repeated for any of the other measurements in the DTP’s MTs database (http://www.dtp.nci.nih.gov/mtargets/mt_index.html). The mRNA expression levels were determined by quantitative PCR (Alvarez et al., 1995). This unique pattern, shown in Figure 3a, reveals a significant and well-defined negative correlation in the anti-mitotic portion of the complete SOM, where clusters of taxanes, colchicines and stilbenes are located. This strong negative correlation suggests a testable hypothesis about these compounds regarding their sensitivity to MDR. The actual correlation of data vectors is shown in Figure 4, where we compare the target fingerprint with the test fingerprint of expression levels for the MDR gene. Based on the hypothesis that high MDR mRNA levels are accompanied by low cell sensitivity to the drugs found in
Fig. 3. (a) $\sigma$-map of the MDR gene expression levels with log($GI_{50}$) target fingerprint. In the coloring scheme red is indicative of high correlations, whereas blue indicates regions of anti-correlations. The strongest ($>3\sigma$) anti-correlated response is located in the upper left-hand portion of the map, coinciding with the region of the SOM that clusters chemical agents known to be active against the mitotic process, e.g. taxanes, stilbenes, peltatins and podophyllotoxins. (b) $\sigma$-Maps showing the difference between the stress induced response and the basal non-stressed gene correlation with the cell sensitivity profiles from the log($GI_{50}$) map. These genes are active in p53-mediated pathways. While the basal levels of the mRNA expression levels show no strong regionalized response, the induced levels of MDM2, WAF1 and GADD45 show an enhanced correlation with the DNA synthesis region of the log($GI_{50}$) map.
this map cluster, it follows that the drugs in this anti-correlated region should be sensitive to MDR drug resistance. A selection of especially sensitive agents to MDR efflux, presumably by serving as P-glycoprotein substrates. Agents in this list include bisantrene (Zhang et al., 1994; Lee et al., 1994) and olivomycin, as well as paclitaxel, all consistent with the existence of drug efflux as a therapeutic complication for these agents.

### 4.2 Induced stress response

In response to cellular stresses that cause DNA damage, the transcription factor p53 becomes active and initiates a cascade of signals that ultimately induce cellular apoptosis. The \( \gamma \)-ray induced mRNA expression levels of MDM2, WAF1 and GADD45 have been measured across the 60 tumor cell lines (O’Connor et al., 1997). These genes encode three proteins that interact with p53: MDM2 inhibits p53-mediated transactivation in the nucleus (Chen et al., 1994) and targets p53 degradation in the cytoplasm via the ubiquitination pathway; WAF1 regulates the cyclin-dependent kinase CDK2 and is in turn promoted by p53 to give a direct mechanism relating tumor suppressor genes and cell cycle control; and GADD45 is a growth arrest- and DNA damage-inducible gene and functioning as one component of the p53 pathway. Comparing the map of basal expression levels with the target fingerprints of the log(GI50) shows no correlation with the map created with the induced levels (Amundson et al., 2000). In fact, their slightly negative correlation of \( -1.2\sigma \) indicates that basal mRNA expressions do not reflect the effects of \( \gamma \)-ray induced cellular stress. Both the basal and induced expressions are, however, correlated among these three genes, MDM2, WAF1 and GADD45, with an average correlation coefficient of 3.6\( \sigma \) and 4.5\( \sigma \), respectively. The \( \sigma \)-maps for both the basal and the induced levels are given in Figure 3b. The induced MDM2, WAF1 and GADD45 maps show strong positive and unique correlations in the S1–S5 region of the log(GI50) map, a region which comprises DNA damaging agents and inhibitors of DNA repair, indicating that the \( \gamma \)-ray induced response is similar to the log(GI50) target fingerprints observed for these agents. Similarly, data relating to the G1 arrest after X-ray exposure (Amundson et al., 2000) shows a strong positive correlation with these agents (4.4\( \sigma \)). The DNA mismatch repair pathway protein MLH1 was also measured (Taverna et al., 2000) and found to retain a high correlation with the log(GI50) response in the S1–S5 region of the SOM. Taken together, these results establish a clear connection between compounds involved directly in DNA damage and additional processes known to affect the DNA damage and repair systems of the cell. The correlation signals seen in DNA damaged systems are a consequence of the signaling cascade propagated either through the p53 pathway or other unknown DNA damage repair pathways, to initiate repair. This strong correlation establishes a link between gene response and a broad spectrum of chemical agents involved in DNA damage and DNA synthesis inhibition. Perhaps a more important conclusion of this result is that basal mRNA expression levels do not, a priori, reflect chemosensitivity. Only following \( \gamma \) radiation are the mRNA expression levels sufficiently different from their basal gene expressions to relate their test fingerprints to a specific chemosensitivity.

### 4.3 Mutational status of selected genes

We have examined the correlation of test fingerprints based on the mutational status of selected genes with the target fingerprints from the log(GI50) SOM. Wild-type genes were arbitrarily assigned a value of +1, mutant genes −1 and 0 for an unknown status. The value of these assignments does not affect the reported results; any indicator of group distinctions would be sufficient.

The p53 status of the NCI 60 cancer cell lines has been experimentally characterized (O’Connor et al., 1997). The correlation of p53 status and log(GI50) SOM is shown in Figure 5a. A strong similarity is found between this \( \sigma \)-map and the previously described \( \sigma \)-maps for the stress-induced MDM2, WAF1 and GADD45 mRNA expressions, with an average map-to-map correlation coefficient of 3.3\( \sigma \). Regions of greatest \( \sigma \) scores implicate agents that cause DNA damage or interfere with DNA repair machinery. Consistent with the previously described absence of information provided by basal mRNA expression, the corresponding data for the p53 mRNA expression levels were not correlated with p53 mutational status (0.7\( \sigma \)).

The p16 gene encodes the cyclin-dependent kinase inhibitor p16\( ^{NK4A} \) which is active against the CDK4: cyclin D and CDK6: cyclin D complexes (Serrano et al., 1993). These complexes are crucial for phosphorylation of retinoblastoma protein and cell cycle progression. Mutations in p16 have been
Linking growth inhibition response to gene expression levels

**Fig. 5.** (a) The mutational status of the p53, p16 and hCdc4 genes correlated with the node vectors of the log(GI50) map. In the coloring scheme red is indicative of high correlations, whereas blue indicates regions of anti correlations. These response patterns exhibit a strong and unique character in this data set. The p53 map is correlated with the DNA synthesis region in the lower right-hand portion of the map, while the p16 and hCdc4 mutational statuses exhibit both positive and negative correlation in the upper right-hand corner, which encompasses the kinase region of the SOM. (b) σ-Maps of the constitutive mRNA expression levels with the the node vectors from the SOM in Figure 1a. Data are shown for the ErbB3, fibrillin and adrenergic receptor gene expression levels as well as for the mRNA status of carbonic anhydrase. These maps can all be classified as belonging to well-delineated groups of genes.
noted in a wide variety of human cancers (Sellers and Kaelin Jr, 1997; Liggett Jr and Sidransky, 1998). Since p16 is known to have tumor suppression activity, agents having a similar log(GI50) target fingerprint might also have a similar antitumor effect. The mutational status of the p16 gene has been assessed in the NCI’s 60 cancer cell lines by Kubo et al. (1999). The corresponding σ-maps between p16 mutational status with the target fingerprints in the log(GI50) map are given in Figure 5a and show a clear positively correlated region in the upper right-hand portion of the SOM. Chemical agents in this region have been implicated in regulation of cell cycle progression (Rabow et al., 2002). A nearly identical σ-map (3.2σ) to the p16 mutational map is found for the mRNA levels of TEK, an endothelial tyrosine kinase also known as the human angiopoietin 1 receptor, which functions in vascular development (Sato et al., 1995). Modifications of p16 status have been shown to lead to downregulation of vascular endothelial growth factor expression (VEGF) (Miki et al., 2000). Additional highly correlated σ-maps include those for mRNA expression of transgelin (2.8σ) (Shields et al., 2002) and fibroblast growth factor receptor 2 (FGFR2) (2.8σ) (Forootan et al., 2000). These similar response patterns suggest the possibility of shared roles for these proteins with p16 in cell cycle control of growth and tumor progression.

Another regulator of cell cycle control is the human F-box protein hCdc4 whose mutational status has recently been determined (Moberg et al., 2001; Strohmaler et al., 2001). The protein hCdc4 is thought to regulate cyclin E by directly binding and targeting cyclin E for ubiquitin-mediated degradation. The σ-map for the mutational status of hCdc4 across the NCI 60 cell lines, shown in Figure 5a, reveals a strong negative correlation in the cell cycle control region of the SOM. This σ-map is unique to this data set in that its σ pattern is not duplicated by any of the other measures examined here (though it shows a small overall negative correlation with the p16 σ-map); instead it shows a large negative correlation in the cyclin-dependent kinase region of our SOM.

5 NON-GENERIC RESPONSES FOUND IN THE GENE EXPRESSION DATA SET

Thirty percent of the σ-maps appear to belong to a set of well-defined test fingerprints that are highly correlated with target fingerprints. In this section, we examine the publicly available data from the gene expression levels across the NCI’s 60 tumor cell lines (Ross et al., 2000). These measurements are based on tumor cells that have not been exposed to drugs and, thus, are assumed to represent constitutive gene expressions under in vitro conditions of growth for these cell lines. While our earlier analysis has shown that constitutive expression levels of p53, for example, are not predictive of chemosensitivity to DNA damaging agents, the following sections describe the set of under and overexpressed genes that may reflect cellular sensitivity.

ErbB3: The mRNA levels of ErbB3 have a distinct correlation pattern with the SOM log(GI50) fingerprints. ErbB3 is a transmembrane receptor tyrosine kinase that is implicated in growth regulation. The natural ligand for ErbB3 is neuregulin, a polypeptide growth factor, which causes heterodimerization with ErbB2 and subsequent activation of downstream signal transduction. The σ-map in Figure 5b shows a singular high correlation in cluster 9.16 of the SOM. Among the compounds clustered at this position is tetrahydro-brefeldin1,13-dione, giving a theoretical indication of the activity of brefeldin A which is a known regulator of neuregulin (Loeb et al., 1998) and consequently a possible effector of ErbB3 activity. A highly similar (5.2σ) σ-map is also present for EphA6, another receptor tyrosine kinase that is thought to be involved in, among other things, vascular development. Neuregulin is also a regulator of Eph receptor expression (Lai et al., 2001), again pointing to the possibility that brefeldin could also be an effector of Eph receptor activity. Two other receptor tyrosine kinases involved in angiogenesis, TIE and TEK (Jones et al., 2001), have a moderately similar (3.4σ) correlation with the EphA6 map.

Fibrillin: The σ-map for the mRNA levels of fibrillin exhibits its strong positive correlations in the upper right-hand corner of the SOM, as shown in Figure 5b. Fibrillin is a major component of extracellular microfibrils. Closely related correlation maps are found for the mRNA test fingerprints of vegfc, dynein and peroxidasin, which are also involved in cellular growth. A functional analysis of the genes in this region of the map shows an enhanced preference for proteins involved in connective tissue/muscle, growth factors and cytoskeletal functions. The response pattern here is distinct from the earlier described generic pattern exhibited by a large class of ECM genes. The ability to localize the fibrillin gene to a specific region of chemosensitivity suggests that selective fibrillin targeting may provide opportunities for enhanced sensitivity within highly expressing cellular populations.

Adrenergic β2 receptor: A distinct correlation pattern is formed by another group of genes represented by the adrenergic β2 receptor, as shown in Figure 5b. This group of σ-maps exhibits two strong regions of anti-correlation in the mitotic and sub-mitotic regions of the SOM in Figure 1a. This anti-correlation reflects the possibility that drug molecules in these regions are antagonists to proteins or to biological pathways associated with these genes. The most highly similar (>5.0σ) gene expression patterns to the adrenergic receptor are those of the Apobec-1 complementation factor, thymosin β4, an expressed sequence tag (GenBank accession number W69176) whose amino acid sequence is 69% similar to the Ca2+/calmodulin-dependent protein kinase kinase and C-jun kinase 2. The membrane-bound adrenergic β2 receptor is activated by epinephrine and other catecholamines, and acts via signal transducing G-proteins to activate adenylate cyclase, thus regulating cAMP concentration in the
cell (Lodish et al., 1995). This receptor has been linked to the relaxation of smooth muscle tissue, apoptosis, adipogenesis (Amer et al., 1993; Shen et al., 1998) and actin depolymerization in cytoskeletal rearrangements (Kwon et al., 1996). The adrenergic receptor has also been shown to lead to C-jun N-terminal kinase activation via the second messenger cAMP (Yamauchi et al., 2001), to provide a direct linkage between gene expression levels of adrenergic receptor and C-jun kinase. The apoptotic and adipogenesis activities are not mediated via adenyate cyclase but through an alternate transduction pathway (Wang and Malbon, 1996; Hirshman et al., 2001). Since the pattern in Figure 5b for adrenergic β2 receptor is common to many genes, the cellular response encoded by the log(GI50) measurements does not reflect a single response to any one of these genes but instead appears to be most related to the class of cAMP controlled pathways. The Apobec-1 complementation factor is necessary for function of the catalytic subunit of apolipoprotein B mRNA editing enzyme. As apolipoprotein B is the main component of very-low-density lipoprotein (VLDL) particles, and VLDL metabolism is regulated by cAMP and cAMP-dependent kinases (Bjornsson et al., 1994), this high correlation of the Apobec-1 complementation factor σ-map with other cAMP-dependent processes can be understood. Likewise, fragments of the thymosin β4 protein, an actin polymerization inhibitor, have been shown to use cAMP in the immuno-neuroendocrine system (Galoyan et al., 1994; Voelter et al., 1995). The role of the Ca2+/calmodulin-dependent protein kinase is not clear, but is ample evidence that the cAMP and Ca2+ signal transduction pathways interact with each other through β-adrenergic receptor stimulation (Bartel et al., 2000). Even though the gene expression patterns correlate with each other and point to the interplay between regulation of these gene products, the exact physiological connection of these proteins in the cell cannot be determined. We note, however, that the anti-correlated portions of the adrenergic receptor σ-map correspond to agents that affect mitosis and ion channels. This correspondence is completely consistent with the known cytoskeletal functions of the adrenergic receptor and thymosin β4 protein. Compounds associated with these anti-correlated response patterns are known inhibitors of the mitotic process, including stilbenes, peltatins and podophyllotoxins. The interplay between cAMP and Ca2+ signaling pathways, combined with an anti-correlation response fingerprint in the ion channel region of the SOM, is also consistent with the activity of the known calcium channel blockers, verapamil and tetrandrine.

Carbonic anhydrase: We have also analyzed public data sets where observations have been reported on fewer than the complete set of 60 tumor cell lines. One recent example involves mRNA measurements of carbonic anhydrases (Ivanov et al., 2001). Carbonic anhydrases are zinc metalloproteins that hydrate carbon dioxide. As was done for mutational status, we have assigned a value of +1 to the cell lines where carbonic anhydrase is expressed and −1 to those cell lines which do not express this gene and 0 to those cell lines for which the gene status is unknown. The measurements of carbonic anhydrase 12 (CA-XII) were available for 42 cell lines, whereas data for 37 cell lines were available for carbonic anhydrase 9 (CA-IX).† The correlation of mRNA status for CA-IX with the target SOM fingerprints is given in Figure 5b. The corresponding map for the CA-XII gene expression is highly similar (4.6σ) (data not shown). Both CA-IX and CA-XII are transmembrane glycoproteins that have been recognized as cellular markers for renal cancer (McKiernan et al., 1997; Tureci et al., 1998). More recently, these markers have been observed in a range of different cancers, and are believed to regulate extracellular pH and influence ion transport (Ivanov et al., 2001, 1998; Wykowski et al., 2001). The maintenance of a high pH confers a selective advantage of cancerous versus normal cells as well as enhanced resistance to radiation and chemotherapy. The ion transport effect of carbonic anhydrases is reflected in the positive correlation exhibited by these expression levels with cellular chemoresistance to agents known to be ion channel blockers. This positive correlation is consistent with an upregulation of carbonic anhydrase genes due to the effect of reduced ion transport by these channel blocking agents. A region of large negative correlation is visible in adjacent map nodes as well, and lends additional support to the conclusion that cells highly sensitive to putative ion channel agents are correlated with low expression levels of the carbonic anhydrase gene. Although no known inhibitor of carbonic anhydrase is present in the NCI’s set of screened agents, a structural motif from the well-known inhibitor acetazolamide (Parkkila et al., 2000), a thidiazole ring system, is present in the form of 5-hydroxy-7-amin(1,2,3)thidiazole[5,4-d]pyrimidine (NSC 675278). This compound is clustered in the region of the SOM characterizing channel active agents and shows an appreciable negative correlation (−2.0σ) with the gene expressions of both carbonic anhydrases. This indicates that potential inhibitors of carbonic anhydrase function may also be located in this portion of the map.

6 SUMMARY

Efforts to identify relationships between chemicals and cellular sensitivity will require novel computational, statistical and biochemical approaches. The analysis reported here builds on our previous study where we have classified cellular chemoresistance using the growth inhibition measurements log(GI50) of the NCI’s anticancer screen. That approach used

† Studies designed to examine the effect of using fewer than 60 cell lines on our analysis estimate that 80% of the significant correlations found with data based on all 60 cell lines are also found when using about 40 cell lines. Regarding this as an acceptable degree of consistency when using fewer data points, we proceeded with the analysis of the carbonic anhydrase data. By visual inspection, we can estimate that a minimum of about 20 cell lines will give an adequate representation of the σ-maps.
SOM clustering tools to group the 36K screened compounds into 1066 clusters, which were further organized into regions of cellular response space, grossly divided according to agents affecting mitosis (M), nucleic acid biosynthesis (S), cellular regulation (P) and membrane integrity (N). Using the clustered set of 1066 response fingerprints, the current analysis explores whether additional data from microarray, MT and mutant status measurements, all obtained from the same set of tumor cells, could be projected onto our log(GI50) SOM, such that meaningful associations might be made with our four putative classifications of cellular response (M, N, P, S) and underlying biochemical processes.

The data analyzed herein could be separated into two large populations comprising generic and non-generic responses. The generic response (a) populated regions largely outside the four response classifications described above, (b) were considered as generalized poisons, (c) comprised over 70% of the data examined here and (d) were characterized by mRNA expression patterns dominated by ECM proteins. The remaining 30% of measurements were further examined to reveal linkages between chemosensitivities and biochemical processes. Within this set of non-generic responses, examples were found where a more precise biochemical basis for cellular cytotoxicity could be supported. Proof-of-concept examinations reinforced previously observed associations between mutant and wild status of p53 and chemosensitivity to alkylating agents, while extending these results to reveal associations to γ-induced expressions of MDM2, WAF1 and GADD45, signals that were not apparent in measurements of basal mRNA expression levels for any of these genes. Additional proof-of-concept examinations revealed that mRNA expression levels were directly correlated with paclitaxel chemosensitivity to mitosis, while also identifying additional chemotypes as P-glycoprotein substrates. Our analysis reconfirmed the well-known direct association between p16 mutant status and chemotypes implicated in cell cycle control, and extended these results to include expression levels for three additional tyrosine kinase proteins (TEK, transgelin and hCdc4). Links were also found that suggested associations between chemosensitivity and the endocrine, paracrine ligand–receptor loops, via expression of the adrenergic receptor, calcium second messenger pathways, via expression levels of carbonic anhydrase and cellular communication pathways, via fibrillin.

These results establish mechanistic connections between chemotypes and specific cellular functions that span the classifications of mitosis, cell cycle control, nucleic acid synthesis and membrane integrity. Only a subset of the measurements processed in our analysis conveyed information relevant to establishing linkages between chemotype and biochemical function. While rather stringent levels were imposed to separate this population of data from its background noise, it is clear that such efforts yield dividends with respect to clearly establishing such links. In our instance, all data sets described herein as non-generic could be related to a biochemical function via previously published sources. Within this set of high quality ‘hits’, further studies can be proposed to validate the associations proposed here. In conclusion, while considerable optimism accompanies efforts to relate gene expression levels and MT screens to therapeutic chemotypes potentially, unraveling nature’s web will require carefully designed experimental and computational procedures.

ACKNOWLEDGMENTS

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract no. NO1-CO-56000. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

REFERENCES


Linking growth inhibition response to gene expression levels


