Bioorganic & Medicinal Chemistry 20 (2012) 167–176

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Prediction of drug efficacy for cancer treatment based on comparative analysis of chemosensitivity and gene expression data

Peng Wan [a,†], Qiyuan Li [a,‡], Jens Erik Pontoppidan Larsen [a], Aron C. Eklund [a], Alexandr Parlesak [a,§], Olga Rigina [a], Søren Jensby Nielsen [b,¶], Fredrik Björkling [b,c], Svava Ósk Jónsdóttir [a,d,*]

[a] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Bldg. 208, DK-2800 Kgs. Lyngby, Denmark
[b] TopoTarget A/S, Fruebjergvej 3, DK-2100 Copenhagen, Denmark
[c] Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences, Universitetsparken 2, DK-2100 Copenhagen, Denmark
[d] Department of Toxicology and Risk Assessment, National Food Institute, Technical University of Denmark, Mørkhøj Bygade 19, DK-2860 Søborg, Denmark

ARTICLE INFO

Article history:
Received 10 July 2011
Revised 6 November 2011
Accepted 11 November 2011
Available online 22 November 2011

Keywords:
Chemosensitivity
Gene expression
Topoisomerase inhibitors
Molecular structures
NCI60
Cancer

ABSTRACT

The NCI60 database is the largest available collection of compounds with measured anti-cancer activity. The strengths and limitations for using the NCI60 database as a source of new anti-cancer agents are explored and discussed in relation to previous studies. We selected a sub-set of 2333 compounds with reliable experimental half maximum growth inhibitions (GI_{50}) values for 30 cell lines from the NCI60 data set and evaluated their growth inhibitory effect (chemosensitivity) with respect to tissue of origin. This was done by identifying natural clusters in the chemosensitivity data set and in a data set of expression profiles of 1901 genes for the corresponding tumor cell lines. Five clusters were identified based on the gene expression data using self-organizing maps (SOM), comprising leukemia, melanoma, ovarian and prostate, basal breast, and luminal breast cancer cells, respectively. The strong difference in gene expression between basal and luminal breast cancer cells was reflected clearly in the chemosensitivity data. Although most compounds in the data set were of low potency, high efficacy compounds that showed specificity with respect to tissue of origin could be found. Furthermore, eight potential topoisomerase II inhibitors were identified using a structural similarity search. Finally, a set of genes with expression profiles that were significantly correlated with anti-cancer drug activity was identified. Our study demonstrates that the combined data sets, which provide comprehensive information on drug activity and gene expression profiles of tumor cell lines studied, are useful for identifying potential new active compounds.

1. Introduction

As modern cancer therapies seek to identify new anti-cancer agents with increased specificity towards particular cancer types and reduced side effects, there is a great need to identify compounds that act up on selective molecular targets and manipulate only relevant pathways. Identification of such drugs demands fundamental understanding of xenobiotic interaction within the human organism by using combinations of experimental, modeling, and analytical methods.

Much attention is thus being devoted to analyzing and integrating available data, and to examine potential relationships among chemosensitivity (biological response to drug treatment), gene expression data on tissue of origin and biological pathways,[1–3] hereby identifying genetic biomarkers.[4]

The NCI60 screening panel[5–7] is the most extensive compound-activity database of diverse chemicals tested for anti-cancer activity. The panel comprises data on 60 different human cancer cell lines derived from nine organs and has served as the basis for a bulk of studies in the literature.

As similar activity patterns often indicate similar mechanism of action (MOA) and mode of drug resistance,[8,9] a number of different algorithms have been developed in order to utilize the information in the NCI60 database for drug discovery.[6,7] Paull et al.[10] showed by analyzing the NCI screening data that compounds with similar mechanism of cell growth inhibition exhibit a similar activity pattern. This work laid the foundation of the COMPARE algorithm,

Abbreviations: LE, leukemia; BR, breast; ME, melanoma; OV, ovarian; PR, prostate; AA, alkylating Agents; TI, topoisomerase I inhibitors; TII, topoisomerase II inhibitors; R/D, RNA/DNA antimetabolites; DA, DNA antimetabolites; SID, PubChem's substance identifier; RMA, robust multi-chip analysis.

* Corresponding author. Tel.: +45 35886192; fax: +45 35887699.
E-mail addresses: peng.wan@merck.com (P. Wan), sojo@food.dtu.dk (S.Ó. Jónsdóttir).
† Present address: Biostatistics and Research Decision Sciences, Merck Research Laboratory, 2F, B22, Universal Business Park, No. 10, Jiuxianqiao Road, Chaoyang District, Bejing 100015, PR China.
‡ Present address: Dana-Farber Cancer Institute, Harvard Cancer Center, Boston, MA 02215, USA.
§ Present address: Metropolitan University College, Global Nutrition and Health, Pustervig 8, DK-1126 Copenhagen K, Denmark.
¶ Present address: Exiqon A/S, Skelstedet 16, DK-2950 Vedbæk, Denmark.

0968-0896/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2011.11.019

which ranks all the compounds in the NCI database based on their similarity of response in the 60 cell lines with the corresponding response of a seed structure.[10–12] The COMPARE approach has contributed to successful identification of many new anti-cancer agents.[11,13,14]

By using Tanimoto similarity testing Shivakumar and Krauthammer[15] showed that structurally similar compounds had highly correlated anti-cancer activity patterns within the NCI60 drug screen. Chakravarti and Klopman identified several significant structural features (biophores) related to the cytotoxicity of the compounds tested against NCI60 cell lines using the MultiCASE (Multi-Computer Automated Structure Evaluation) method.[16]

The NCI60 cell lines have been subject to extensive profiling based on mRNA,[17–21] microRNA[22] and protein[20,23,24] expression studies. These data were summarized in the recent CellMiner relational database and query tool.[25]

The gene-drug correlations between chemosensitivity profiles of cell lines towards drugs and the corresponding gene expression patterns of untreated cell lines were investigated in several studies. These analysis were used to identify genes where the expression was affected by specific drugs[26,27] and to propose candidate genes as predictive markers for anti-cancer drug sensitivity.[4,28–33] On the other hand, Wallqvist et al.[34] concluded that gene expression levels do not generally correlate with chemical response, but do rather reflect a generic toxic condition.

In recent studies, integrative feature selection schemes were developed to identify chemosensitivity determinants from genome-wide transcriptional profiles and protein expression levels in NCI60 cells.[35,36] It has been demonstrated that integrating chemosensitivity and gene expression data provides valuable help in the search and verification of drug targets.[37] Drug–gene associations were successfully predicted by integrative evaluation of gene expression and drug response data,[38] and in the same study transcriptional responses to drug treatment were predicted and tested on data from the Connectivity map.[39]

In the present work, we extracted a sub-set of compounds with reliable experimental half maximum growth inhibition values (GI_{50}) from the NCI60 anti-cancer drug screen database. A series of analyses was conducted in order to examine the specificity of these compounds in relation to the tissue of origin. First, variations in experimental GI_{50} activity values were compared with the variations in the measured gene expression profiles for the corresponding cell lines. Second, we analyzed the chemosensitivity data set in order to identify compounds with high efficacy and selectivity towards a specific tissue of origin. Third, the structure–activity relationship (SAR) profiles were explored, and compounds exhibiting high structural similarity with standard anti-cancer agents identified. Fourth, the correlation between drug activity patterns and gene expression profiles for known anti-cancer agents was examined.

2. Material and methods

2.1. Data mining from PubChem BioAssay database

The PubChem BioAssay database,[40] was downloaded to a local system. An in-house program was written to parse and load XML documents into MySQL, and cell activity data for human tumor cell line growth inhibition assays were extracted. The cell line panel incorporates a total of 73 different human tumor cell lines of diverse histological origin derived from nine types of tumors, including all the NCI60 cell lines. For each cell line, growth inhibition was tested for between one thousand and forty thousands compounds,[41] for which three dose response parameters GI_{50} (50% growth inhibition), TGI (total growth inhibition) and LC_{50} (50% of cells killed) were measured for each compound. All the analysis in this paper are based on the GI_{50} values, as it is the most commonly used response parameter for estimating growth inhibition.

2.2. Data sets

2.2.1. Chemosensitivity data set (data set A)

A cleaned data set of $-\log_{10}(GI_{50}$ [mol/l]) values for 30 NCI60 tumor cell lines was used in our analysis. All available NCI60 cell lines from five different organs were selected, that is, nine melanoma (ME), two prostate (PR), seven ovarian (OV), six breast (BR) and six leukemia (LE) cell lines. The overall data set was cleaned by removing all compounds that had missing data points or were considered to have unreliable experimental $-\log_{10}(GI_{50})$ values for one or more of the 30 cell lines considered. An experimental value was considered unreliable (1) if a compound was only tested once, or (2) if a mean value from multiple experiments for a given compound was listed without a standard deviation. The reproducibility among the experimental values could not be evaluated in the absence of a standard deviation for multiple data points or if only one experimental value was available. Thus we generated a sub-set of compounds with reliable experimental values for all 30 cell lines.

Out of more than 60,000 compounds in the raw data set, the cleaning procedure reduced the data set to 2333 compounds for the final analysis. The mean $-\log_{10}(GI_{50}$ [mol/l]) for all the compounds in data set A is 5.48 with a standard deviation of 0.80 and range from 2.21 to 11.51. By using the same filtering criteria for extracting a comparable cleaned data set for all the NCI60 cell lines, a sub-set containing less than 300 compounds was extracted. As we wanted to ensure appropriate data quality for our analysis, we choice to focus this work on cell lines from a smaller number of target organs.

2.2.2. MOA (mechanism of action) data set (data set B)

$-\log_{10}(GI_{50})$ values for 62 known standard anti-cancer agents were extracted from data set A. These compounds are within the list of 122 compounds with anti-cancer activity and reasonably well known mechanism of action (MOA).[42] These 62 compounds were included in both data set A and B.

2.2.3. Gene expression data set (data set C)

The gene expression data were derived from NCI60 transcript profile data sets based on Affymetrix HG-U133A&B chips. The raw data set containing 60 cell lines each profiled with a total of 44,566 probe sets was downloaded from the Genomics & Bioinformatics Group (GBG) web site.[43] The gene expression data set was pre-processed and filtered in the following way. For better possibility of carrying out comparative analysis between the chemosensitivity and gene expression data, the same 30 cells lines that were covered by data set A were considered. Data from each of the two array platforms (HG-U133A and HG-U133B) of the gene expression data set was normalized individually using robust multichip analysis (RMA).[44] The standard deviation (*SD*) was calculated for each gene as a measure of the variability in the gene expression pattern across the 30 cell lines, and the data set was filtered to generate a sub-set of genes with $SD > 0.8$. In the case of multiple probe sets measuring the same gene, we chose the probe set that yielded maximal *SD*. The remaining genes were removed. This resulted in a sub-set containing the 1901 genes with the largest variability among the 30 cell lines (data set C), which we used for our analysis.

2.3. Chemical structure analysis

Chemical structure information for the compounds contained in data set A were extracted from the PubChem Substance database in 2D SDF (standard data format). Structure information was

available for 1669 of the compounds from data set A and for 44 drugs from data set B. MDL[45] MACCS 2D fingerprints (166 Keys) were calculated for each molecular structure using MOE (Molecular Operating Environment), a modeling software suite from Chemical Computing Group.[46] The MACCS fingerprints were subsequently transformed into strings of binary fingerprints, representing the presence or absence of specific chemical substructures, and the similarity between the compounds in data sets A and B was estimated by calculating Tanimoto coefficients[47] for the binary MACCS fingerprints. Before carrying out this analysis, all the compounds contained in both data sets were removed from data set A.

2.4. Correlation analysis

All correlations presented in this manuscript were Pearson product–moment correlations (PCCs). The PCC (r) between the chemical response vectors of 2333 compounds was calculated for each pair of cell lines. Similarly the PCC between the gene expression profiles of 1901 genes for each cell line pair was calculated. The average PCC was evaluated for each tissue of origin and between different tissues of origin based on chemosensitivity (r_c) and gene expression (r_g), respectively. Then the calculated global degree of similarity between chemosensitivity and gene expression profiles, using the variable 'correlation of correlation' (cc), as introduced by [20]. This correlation is given by

$$cc = \frac{\sum_{i<j} r_{cij} r_{gij} - \frac{1}{N}\sum_{i<j} r_{cij} \sum_{i<j} r_{gij}}{\sqrt{\left(\sum_{i<j} r_{cij}^2 - \frac{1}{N}\left(\sum_{i<j} r_{cij}\right)^2\right)\left(\sum_{i<j} r_{gij}^2 - \frac{1}{N}\left(\sum_{i<j} r_{gij}\right)^2\right)}} \quad (1)$$

where r_{cij} and r_{gij} represent the correlation between all distinct cell lines i and j based on their chemosensitivities and gene expression, respectively. N denotes possible pairwise combinations for any given cell panel. For all 30 cell lines, $N = 30 \times (30 - 1)/2 = 435$ correlations were calculated.

The gene–drug correlations were also evaluated by PCC, but in this case the chemosensitivity values for each compound in the 30 cell lines were correlated with the gene expression data of the same compound in the 30 lines. Similarly, the correlation between chemosensitivities for two selected compounds in the 30 cell lines was explored.

All color charts and color code matrices were generated using the R statistical environment.[48]

2.5. Clustering

Self-organizing maps (SOM)[49] were applied for the clustering of the cell activity data and gene expression data using the SOM-Toolbox for Matlab 5.[50,51] Basically, the SOM tool performs as unsupervised clustering based on similarities between data vectors, in this case chemosensitivity or gene expression patterns for the cell lines studied. Thus cell lines with similar patterns cluster naturally together, and this method gives a visual and easily interpretable overview of multi-dimensional data. The use of SOM is illustrated in greater detail in a metabolite profiling study.[52]

In our work, logistic transformation, which scales the data vectors to be in the range 0 and 1, was applied on the data sets. Linear initialization and batch training algorithm were used. The following parameters were used: Hexangular map lattice, sheet map shape and map grid size of 8×4. Average quantization error and topographic error were used as measures of map quality. The training length was set to 100 epochs based on stabilization of the quantization error. The U-matrix representation was used to visualize the clustering of cell lines on individual neurons (map units) and the distances between neighboring neurons. The distances, which were calculated as the Euclidean distances between the data vectors representing each neuron, were subsequently used to evaluate the overall clustering of the data. Light gray color represents small distance, and all neurons in a light gray area were considered to occupy the same cluster. Cell lines separated by dark color are more dissimilar and the dark ridges in the SOM graph represent cluster boundaries. The labels for the cell lines assigned to each neuron are shown on the corresponding empty grids. Hit histograms were used to indicate how the best matching units of the data sets were distributed on the SOMs. Marker style was set to pie charts. The size of the pie in each unit describes the number of data points (cell lines) clustering on each neuron and the colors represent the different classes.

Hierarchical clustering was used to cluster the compounds in data set A with respect to similarities in the chemosensitivity profiles across the 30 cell lines. The average linkage method and Euclidean distance were used in the R statistical environment.[48] The clustering was done in order to make the graph easier to read.

2.6. Identification of tissue specific compounds

In order to identify compounds with high potency to become effective within a specific group of cell lines, a series of Python scripts were written. Clusters of cell lines with similar gene expression profiles were identified. For each of these clusters, the average potency was calculated, and the data set was searched for compounds with one log unit larger average potency in one cluster compared to the other clusters.

3. Results and discussion

3.1. Clustering analysis of the chemosensitivity and the gene expression data sets

In order to examine the degree of tissue specificity exhibited by the chemosensitivity data, self organizing maps (SOM) clustering and correlation analysis were performed for data set A (GI_{50} data described in the methods section) and data set C (gene expression data), respectively. The overall differences in chemosensitivity were then compared with the differences in gene expression profiles of the corresponding tissues.

Figure 1 shows a SOM representation of the clustering of the different cell lines (Table 1) based on the chemosensitivity data (data set A). It can be seen that the leukemia cell lines and the remaining cell lines are separated (the dark colored belt on the underlying distance map). Thus the cell lines are clearly divided into two separate clusters, with two cell lines, ME1 (LOX-IMVI) and BR1 (MCF7), placed on the cluster boundary. Three main clusters were formed for the corresponding gene expression data (data set C) (Fig. 2), one with the leukemia cells, another containing the BR2 (MDA-N) breast cells and all but one of the melanoma cells, and a third with the remaining cell lines. Two of the breast cell lines, BR1 (MCF7) and BR4 (T-47D), were placed on one of the cluster boundaries. BR2 (MDA-N) clustered with the melanoma cell lines, especially with ME9 (MDA-MB-435), in both data sets A and C. These two cell lines were previously suggested to be breast cancer cell lines, but are now classified as melanoma cells.[17,22]

Clear separation was seen between the basal-like (BR3, BR5, BR6) and the luminal-like (BR1, BR4) breast cancer cells in both data set A and C, reflecting the difference between these two cell types. Furthermore, while all the ovarian and the prostate cell lines clustered closely together with respect to tissue of origin, the distance map revealed that these cells show somewhat larger variability with respect to chemical response (medium gray color, Fig. 1).

Thus, five different clusters were identified in data set C; leukemia, melanoma, ovarian and prostate, basal-like breast, and luminal-like breast cancer cells. For data set A, the leukemia cells

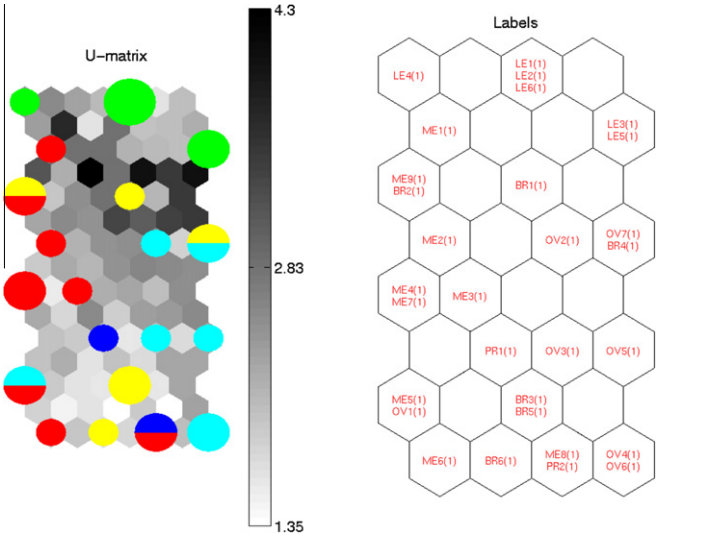

Figure 1. The SOM clustering of the 30 cell lines based on the chemosensitivity data (data set A) in U-matrix (left) and label map (right) representations. The cell lines are colored according to their tissue of origin: green—leukemia (LE), yellow—breast (BR), red— melanoma (ME), cyan—ovarian (OV), blue—prostate (PR). The label map shows the position of each cell line using the codes listed in Table 1. The underlying gray-scaled distance map of the U-matrix is used to identify cluster boundaries within the data, with the main cluster boundary clearly seen as a dark colored belt separating the leukemia cancer cells from the other cell lines.

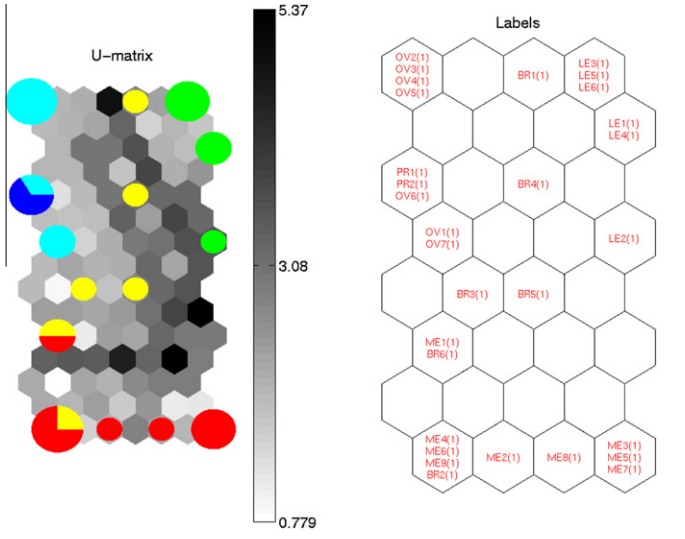

Figure 2. The SOM clustering of the 30 cell lines based on gene expression profiles (data set C) in U-matrix (left) and label map (right) representations. The cell lines are colored according to their tissue of origin: green—leukemia (LE), yellow—breast (BR), red—melanoma (ME), cyan—ovarian (OV), blue—prostate (PR). The label map shows the position of each cell line using the codes listed in Table 1.

Table 1
The names of the 30 NCI60 cell lines included in the current study, listed together with the abbreviations used in this study.

Label	Cell line
ME1	LOX-IMVI
ME2	M14
ME3	MALME-3M
ME4	UACC-62
ME5	UACC-257
ME6	SK-MEL-2
ME7	SK-MEL-5
ME8	SK-MEL-28
ME9	MDA-MB-435
PR1	PC-3
PR2	DU-145
OV1	NCI/ADR-RES
OV2	OVCAR-3
OV3	IGROV1
OV4	SK-OV-3
OV5	OVCAR-4
OV6	OVCAR-5
OV7	OVCAR-8
BR1	MCF7
BR2	MDA-N
BR3	BT-549
BR4	T-47D
BR5	MDA-MB-231/ATCC
BR6	HS-578T
LE1	RPMI-8226
LE2	SR
LE3	CCRF-CEM
LE4	K-562
LE5	MOLT-4
LE6	HL-60(TB)

formed a separate cluster, and a clear separation was seen between basal-like and luminal-like breast cells, however no distinct clusters were found for the remaining tissues.

Previously published hierarchical clustering analyses of gene expression, protein expression and chemosensitivity data for all the 60 cell lines from the NCI60 drug screen add further support to the above classification. As in our study, the gene expression data were generally seen to cluster well with respect to the tissue of origin, and strong consensus can be seen between clustering based on the cDNA set,[18,17] and the Affymetrix HG-U133A data set.[22] The same main clusters were also identified in the mRNA consensus set and the protein expression set.[20] Leukemia and melanoma cells were seen to form separate clusters in all cases. Clear separation between basal-like and luminal-like breast cells was observed in each of these data sets, in which the luminal-like breast cells clustered closely together and the basal-like cells were in most cases a part of a larger cluster rich in CNS cells. The clustering of ovarian and prostate cells was clearer in our work, possibly because we focused our analysis on 30 cells from five types of tissue. Thus, it can be concluded that the division of the cells into five clusters is supported by other, previously performed, comparable analyses.

For the chemosensitivity data a number of clusters related to tissue of origin can be identified both in our analysis and in other studies,[18,53,21] although the clustering is considerably less clear than that for the gene expression and the protein expression data.

The similarities and dissimilarities between the individual cells lines were explored in greater detail by calculating the Pearson correlation coefficient (PCC) between the chemical response vectors of each cell line pair (Fig. 3a), and the corresponding PCC between their gene expression profiles (Fig. 3b). It is seen that the variation in the gene expression profiles for the 1901 genes in data set C is generally considerably larger than the variation in chemical response to the 2333 compounds in data set A. This is not surprising, as the gene expression data contain specific information on the genetic variation from one cell line to another, often not reflected by the chemosensitivity data. Some tissue specificity is, however, seen in the chemosensitivity data as well.

In both cases the melanoma cells showed strong correlation with other melanoma cells, also illustrated by the relatively high mean PCC value (Table 2). Similarly, the breast cancer cells gave the lowest internal mean PCC in both data sets due to the large variation between basal and luminal cells. In parallel, the breast cells exhibited the strongest global degree of similarity (cc_{cg}) between data set A and C, estimated by the correlation of correlation method in Eq. (1) (Table 2). This analysis also revealed strong correlation between the ovarian and the prostate cancer cells in both data sets A and C, which was not as easy to identify in the SOM analysis for data set A.

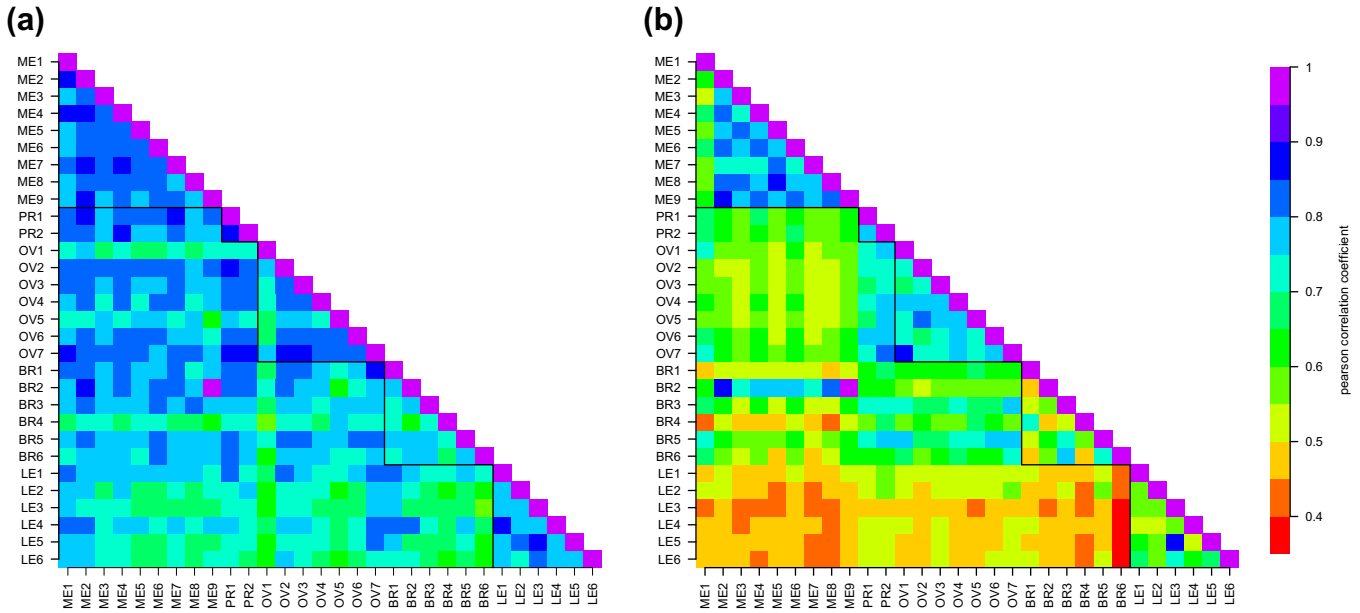

Figure 3. Pairwise correlation matrices for all the 30 cell lines based on chemosensitivity data in data set A (a) and gene expression profiles in data set C (b). Pearson correlation coefficient was calculated for each cell line pair, see the color code on the right hand side. The cell lines are marked with the codes listed in Table 1.

Table 2
The average PCCs between cell line types, based on the chemosensitivity data (data set A), r_c, and the gene expression data (data set C), r_g, respectively, and the correlation of correlation, cc_{cg}, between the chemosensitivity and gene expression data

	ME	PR	OV	BR	LE
r_c					
ME	0.824				
PR	0.812	0.856			
OV	0.779	0.806	0.783		
BR	0.774	0.792	0.751	0.750	
LE	0.733	0.739	0.715	0.704	0.801
r_g					
ME	0.751				
PR	0.612	0.784			
OV	0.576	0.750	0.754		
BR	0.588	0.667	0.647	0.592	
LE	0.466	0.529	0.499	0.465	0.612
cc_{cg}					
ME	0.270				
PR		–			
OV			−0.001		
BR				0.561	
LE					0.471

Thus our results indicate that it is possible to find significant correlations between chemosensitivity and gene expression data in cases where cell types from the same tissue are sufficiently dissimilar, as in the two types of breast cancer cell lines. On the contrary, it might be more difficult to identify such effects in cases where the different cells from the same tissue are highly similar, as in the ovarian cancer cell lines.

3.2. Tissue specificity with respect to chemical response

The strong similarity observed in the chemical response among the tumor cell lines from different tissues of origin render it difficult to identify tissue-specific compounds with global analysis methods. This can be illustrated by the heat map in Figure 4, where the compounds in data set A were clustered according to the PCC similarity of their chemical response vectors in the 30 cell lines. As an exception, the leukemia cells showed significantly higher chemosensitivity than the other cells for most of the chemicals in data set A.

As tissue specificity is important for avoiding toxic lesion in non-malignant tissue, data set A was analyzed with the aim of identifying compounds exhibiting tissue specificity. The cell lines were divided into five groups, each group with similar tissue of origin according to the clusters identified in the gene expression data: (1) All melanoma cancer cells without ME1 (LOX-IMVI), (2) all prostate and ovarian cancer cells, (3) luminal-like breast cancer cells, (4) basal-like breast cancer cells, and (5) all leukemia cancer cells. For each of these groups, the mean pGI_{50} ($-\log(GI_{50}$ [mol/l])) value (labeled mean pGI_{50}) and the corresponding standard deviation were calculated for all compounds listed in data set A.

As already seen in the previous analysis, most compounds in data set A showed similar response throughout the panel of cell lines. Half of the data set (1134 compounds) were considered as weakly active in the different tissues, with mean pGI_{50} values between 5.0 and 7.0. 2150 compounds had mean pGI_{50} lower than 7.0 in all five groups of cells, and thus only 183 compounds exhibited mean activity larger than this limit in one or more of the groups.

We searched specifically for examples of compounds with significantly higher mean potency in one of the groups compared to the others. One compound was found with mean $pGI_{50} \geqslant 6.0$ for the leukemia cells, while the corresponding values for the other cell line groups were lower than 5.0. Similarly, we searched for compounds with mean $pGI_{50} \geqslant 7.0$ in one group and lower than 6.0 in the remaining four groups. Four compounds in the luminal-like breast cell group and two in the leukemia group matched this criteria. By raising both limits by one, two and one compounds were identified for the same groups, respectively (Table 3).

Although we have only identified a few compounds with relatively high efficacy and specificity towards one of the five tissue types defined in our analysis, this analysis demonstrates that such compounds can be found within the data set. Especially it would be advantageous to be able to identify potential drugs with potency in the nano-molar (or even the pico-molar) range, rather than in the micro-molar range, as high potency often provides compounds which are target selective and tentatively less toxic. From this

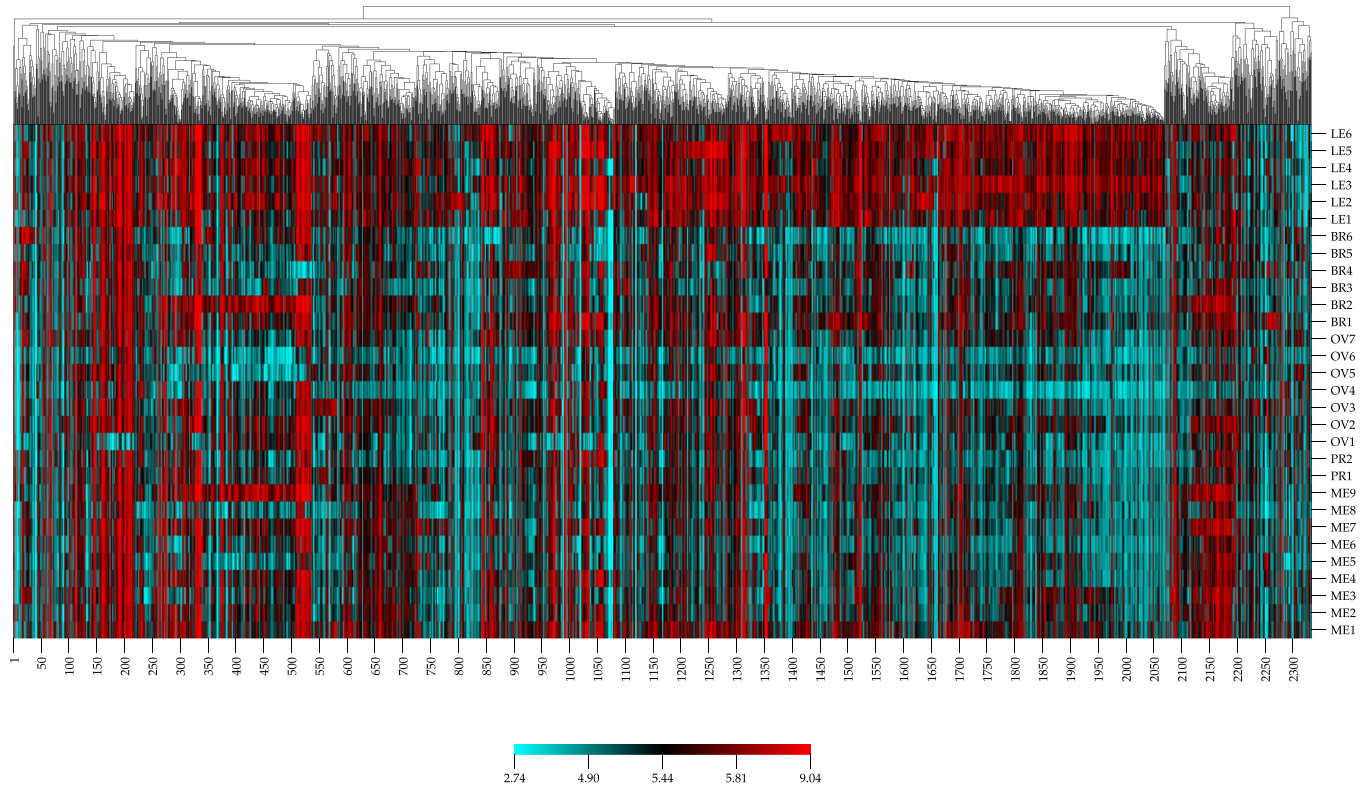

Figure 4. A heat map illustrating variations in chemosensitivity for all the compounds in data set A, the color scale shows the $-\log_{10}(GI_{50})$ values. Compounds with similar chemosensitivity profiles across all 30 cell lines were clustered together.

Table 3
Compounds showing selectivity with respect to tissue of origin, measured by the average potency of at least one log unit larger in one of the cell line groups compared to the remaining cell line groups

Limits	Upper limit sub-group	Compound SID	Melanoma	Ovarian-prostate	Breast luminal-like	Breast basal-like	Leukemia
6.0/5.0	LE	491601	4.62 (0.29)	4.50 (0.27)	4.93 (0.52)	4.66 (0.36)	6.27 (0.33)
7.0/6.0	BR-luminal	521002	4.80 (0.73)	5.40 (0.99)	7.16 (0.65)	4.63 (0.15)	4.89 (0.77)
	BR-luminal	530489	5.20 (0.41)	5.07 (0.26)	7.49 (0.24)	4.94 (0.05)	5.87 (0.45)
	BR-luminal	530656	5.06 (0.32)	5.15 (0.39)	7.65 (0.08)	5.05 (0.18)	5.56 (0.09)
	BR-luminal	531399	5.23 (0.20)	5.74 (0.64)	7.33 (0.26)	5.37 (0.05)	5.87 (0.57)
	LE	148445	5.88 (1.14)	5.99 (0.96)	5.77 (0.79)	4.56 (0.04)	7.08 (0.54)
	LE	490387	5.60 (0.26)	5.19 (0.39)	5.91 (0.48)	5.21 (0.35)	7.10 (0.19)
8.0/7.0	BR-luminal	534283	6.20 (1.39)	6.23 (1.30)	8.60 (0.00)	5.38 (0.07)	6.74 (1.43)
	BR-luminal	534284	6.16 (1.42)	6.33 (1.19)	8.59 (0.02)	5.32 (0.07)	6.83 (1.10)
	LE	528801	6.94 (1.17)	6.85 (1.09)	6.87 (1.03)	6.59 (0.79)	8.07 (0.24)

The upper and lower potency limits used in each search are listed, along with the SID (PubChem structure id) numbers of the compounds matching this criteria and the mean ($-\log(GI_{50}$ [mol/l])) values for each cell line group (cluster). The standard deviations are given in the parenthesis.

perspective the three compounds with mean $pGI_{50} \geqslant 8.0$ were considered the most interesting.

3.3. Chemical structure analysis of the chemosensitivity dataset

The similarity in potency across the panel of cell lines was also examined in relation to structural similarity of compounds active in cancer cell growth inhibition.

Data set A was characterized with respect to the known anti-cancer agents in data set B using a simple Tanimoto-based estimate of molecular fingerprints, describing the presence and absence of specific chemical groups. To avoid comparison between the same structures, all known anti-cancer agents were removed from data set A before analysis. Then, all available 2D molecular structures were extracted from the PubChem Substance database, comprising 1669 structures for data set A and 44 for data set B (GI_{50} data for known anti-cancer agents).

As seen in the heat map representation in Figure 5, the compounds in data set A had the closest structural correlation with the topoisomerase inhibitors I and II, and the least resemblance to the alkylating agents in data set B. Only a few compounds showed strong similarity with some of the anti-cancer agents (red color), and a significant number of the compounds showed fair similarity (orange and yellow color). The remaining compounds were considered to be structurally dissimilar from known anti-cancer agents, with a Tanimoto similarity <0.6.

We extracted the compounds with Tanimoto similarity $\geqslant 0.8$ with at least one of the topoisomerase II inhibitors in data set B, 21 hits for eight different topoisomerase II inhibitors in total. Eight of these compounds, which were verified by visual inspection as being highly structurally similar compared to the respective query structures, were selected for further analysis. Table 4 lists the range of growth inhibition values of these eight compounds, along with the Tanimoto similarity and the PCCs between the pGI_{50}

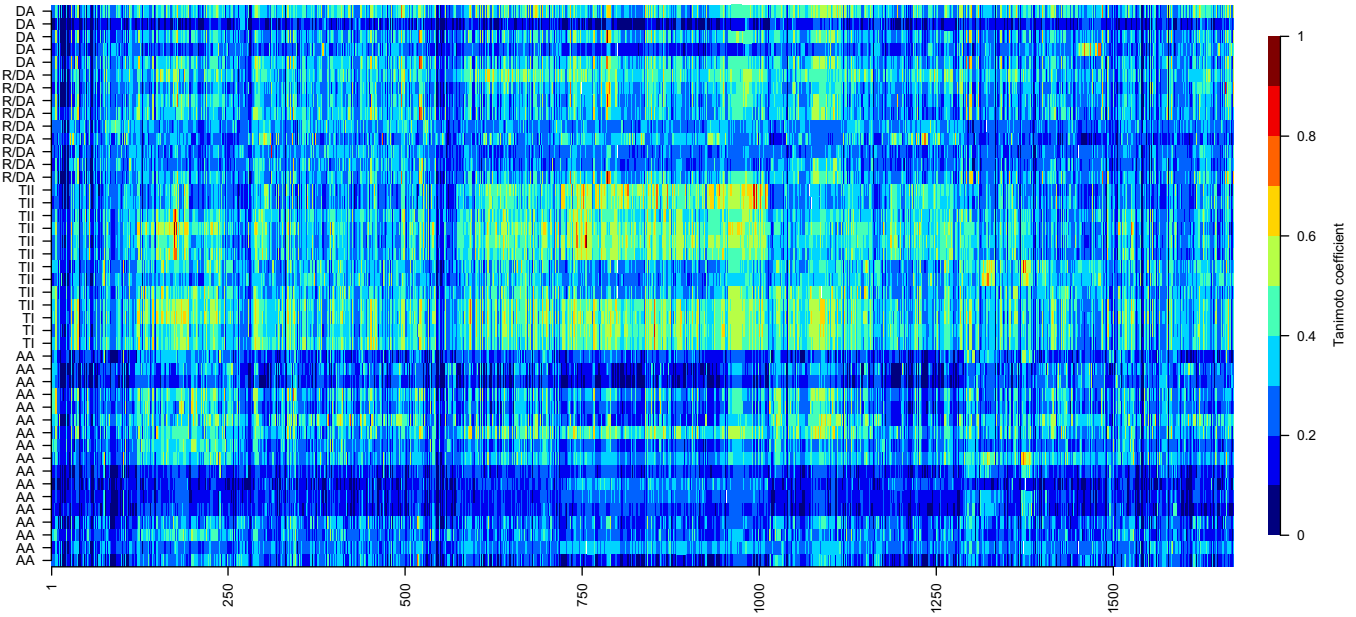

Figure 5. A heat map representing the Tanimoto similarity between sub-sets of compounds in data set A (columns) and data set B (rows) with available structure data in the PubChem Substance database. The Tanimoto coefficient for each compound–drug pair was represented according to the color code shown on the right hand side of the map. It was seen that data set A exhibited the strongest similarity to topoisomerase II (marked TII) inhibitors.

Table 4
The 10 compounds with high structural similarity with known topoisomerase II inhibitors and their corresponding query structures (marked with *)

SID	Tanimoto coefficient	PCC	Lowest $-\log(GI_{50}$ [mol/l])	Highest $-\log(GI_{50}$ [mol/l])
140790*	–	–	OV1: 4.998	LE5: 7.624
403734	0.80	0.52	BR3: 5.935	LE5: 7.392
412389	0.85	0.23	OV1: 4.734	LE4: 7.489
139837	0.98	−0.03	OV1: 5.655	BR4: 10.540
125750	0.86	0.36	ME8: 7.354	BR3: 7.648
570114*	–	–	OV4: 4.469	LE2: 6.025
403734	0.90	0.80	BR3: 5.935	LE5: 7.392
576309	0.84	0.09	BR4: 7.241	BR1: 8.602
572071	0.81	0.76	ME3: 5.713	LE5: 6.723
125750	0.81	0.42	ME8: 7.354	BR3: 7.648
561989*	–	–	OV1: 4.789	LE2: 7.702
301203	0.89	0.86	OV1: 4.921	LE2: 7.211
137181*	–	–	ME9: 5.343	LE5: 7.479
529810	0.84	0.62	OV3: 4.431	LE6: 6.060

The compounds are listed with their SIDs, their Tanimoto similarity with their respective query structure and the PCCs between their activity profile and the correspond activity profile of the query structure. The lowest and highest measured growth inhibition values ($-\log(GI_{50}$ values) for the 30 cell lines are also listed for each compound.

values of the compounds and the corresponding values of the query structures for the 30 cell lines. Four hits were identified for each of the compounds menogaril (140790) and *N*,*N*-dibenzyl dunomycin (570114), respectively. Two of the structures identified based on these two compounds were the same. One hit was found for rubidazone (561989) and another one for *m*-AMSA (137181) (Table 4 and Fig. 6). All the compounds in Figure 6 are seen to contain three inter-connected aromatic rings, which is a common substructure in topoisomerase inhibitors II that intercalate into DNA.

Menogaril and rubidazone are anthracycline compounds that possess significant anti-tumor activity due to the inhibition of macromolecular synthesis [54,55], and rubidazone has been used extensively for the treatment of acute leukemias and sarcomas.

Two of the hits found for *N*,*N*-dibenzyl dunomycin had an activity profile similar to the drug, with a slightly higher potency range, and thus they might have a similar binding mode. The two remaining hits had a higher and more narrow potency range, possibly because they have polar substituents at a position where *N*,*N*-dibenzyl dunomycin has non-polar groups attached. The hits found for menogaril had more polar substituents compared to those found within the drug (query structure), all these molecules had similar or higher activity compared to menogaril. For *m*-AMSA, a structurally similar but somewhat less potent hit was identified. The hit found for rubidazone had strong similarity in both structure and activity pattern. Hence, these examples and other similar studies[15] illustrate that structurally similar compounds often have similar activity (potency) patterns. Compounds can on the other hand have similar activity patterns without being structurally similar.

The current method uses a simple similarity measure to identify hits based on the overall structural similarity with the query structure, but it does not identify hits with specific structural motives like pharmacophore type screens and three-dimensional fingerprints Table 5.

As demonstrated in the above analysis, identification of potential hits with a very simple similarity measure is a viable option. The current or other methods can thus be used to screen through large databases in order to select new compounds for subsequent experimental testing. Whereas a method like COMPARE has been proven to be extremely useful for identifying potential new drugs by searching for compounds with similar activity profiles, the use of structural similarity can provide new candidate compounds for measuring new activity profiles.

3.4. Correlation analysis between gene expression and drug sensitivity

Despite the difficulty in filtering out gene–drug correlations relevant to the drug action from the huge amount of genes coding for biological processes involved in cell proliferation and differentiation, we attempted to identify genes that may be associated with drug sensitivity. Correlation analysis was done for the panel of the 30 cell lines in the attempt to identify gene expression values which correlated well with the drug activity.

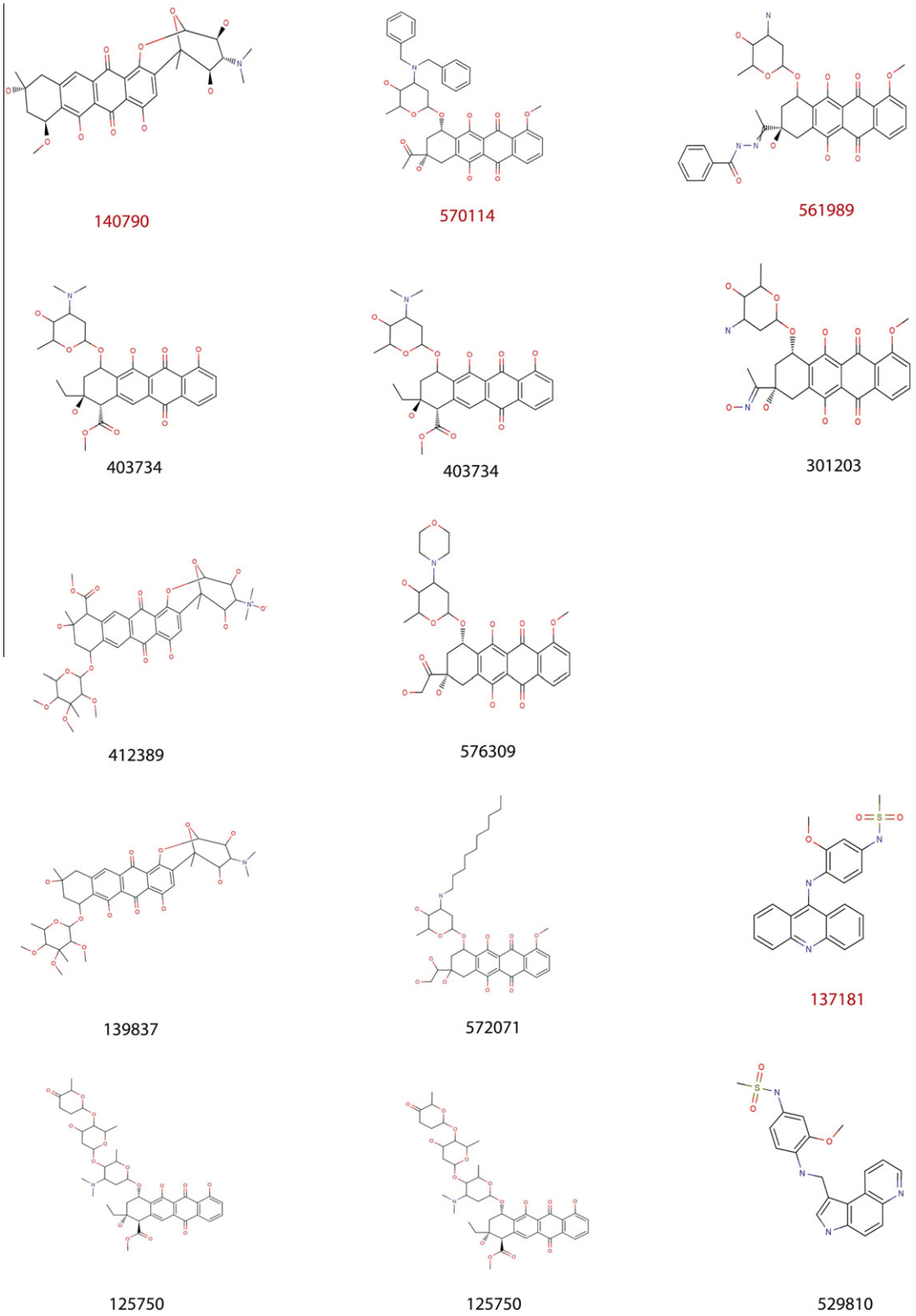

Figure 6. Chemical structures and SIDs of eight compounds from data set A with the strongest structural similarity to at least one of the topoisomerase II inhibitor in data set B. The corresponding four query compounds (topoisomerase II inhibitors with SIDs 140790, 570114, 561989 and 137181, all marked in red).

We calculated the PCC between the activity patterns of each of the 62 anti-cancer agents (data set B) and the expression profiles of each of 1901 genes (data set C) in all 30 tumor cell lines, giving 117862 PCC values in total. To focus our analysis on genes that showed the strongest correlation with chemosensitivity or chemoresistance, we imposed a threshold value ($|PCC| > 0.7$). Thus, 128 genes with the strongest positive ($PCC > 0.7$) or negative ($PCC < -0.7$) correlation with at least one of the anti-cancer agents were identified. Whereas the alkylating agents showed many non-specific gene-drug correlations as expected, the gene–drug correlations identified for topoisomerase inhibitors I and II and R/DA and DA antimetabolites were generally more

Table 5
The 62 MOA drugs in data set B are listed with their common names and SIDs, as well as the short labels assigned in this work

Compound class	Names	SID
AA	Asaley	562499
AA	BCNU	479311
AA	Carboxyphthalatoplatinum	301233
AA	CBDCA	568180
AA	CHIP	569367
AA	Chlorambucil	69624
AA	Chlorozotocin	446503
AA	Clomesone	460978
AA	Cyclodisone	462884
AA	Dianhydrogalactitol	301193
AA	Hepsulfam	459000
AA	Hycanthone	427813
AA	Melphalan	301128
AA	Mitomycin C	87663
AA	Mitozolamide	463980
AA	Nitrogen mustard	538248
AA	Piperazine	575677
AA	Pipobroman	86412
AA	Spirohydantoin mustard	443958
AA	Teroxirone	147076
AA	Tetraplatin	576794
AA	Thio-tepa	72371
AA	Triethylenemelamine	75085
AA	Uracil nitrogen mustard	92079
AA	Yoshi-864	301162
TI	Camptothecin derivative	501764
TI	camptothecin derivative	577550
TI	Camptothecin derivative	484408
TI	Camptothecin derivative	484522
TI	Camptothecin derivative	484520
TI	Morpholinodoxorubicin	576309
TII	Amonafide	454647
TII	*m*-AMSA	137181
TII	Anthrapyrazole derivative	576375
TII	Pyrazoloacridine	466692
TII	Bisantrene HCl	575181
TII	Deoxydoxorubicin	301228
TII	Mitoxantrone	572869
TII	Menogaril	140790
TII	*N*,*N*-Dibenzyl daunomycin	570114
TII	Oxanthrazole	575923
TII	Rubidazone	561989
TII	VM-26 (teniposide)	416461
TII	VP-16 (etoposide phosphate)	427063
R/DA	5-Azacytidine	405114
R/DA	5-Fluorouracil	82653
R/DA	Acivicin	301209
R/DA	Dichlorallyl lawsone	418677
R/DA	Brequinar	577126
R/DA	Ftorafur (pro-drug)	430704
R/DA	5,6-Dihydro-5-azacytidine	301226
R/DA	Methotrexate	67627
R/DA	*N*-(Phosphonoacetyl)-L-aspartate (PALA)	567041
R/DA	Pyrazofurin	427861
R/DA	Trimetrexate	463773
DA	2′-Deoxy-5-fluorouridine	88034
DA	5-HP	301166
DA	5-Aza-2′-deoxycytidine	419300
DA	Hydroxyurea	90752
DA	Inosine glycodialdehyde	301177
DA	Thioguanine	538243
DA	Thiopurine	538245

The labels include the MOA class of each compound, AA for alkylating agents, TI and TII for topoisomerase inhibitors I and II, and R/DA and DA for RNA/DNA and DNA antimetabolites.

specific. Thus most of the anti-cancer agents were associated only with one or a few genes.

Protein annotations encoded by the identified genes in the Universal Protein Resource (UniProt),[56] revealed only a few genes as potentially relevant for drug action. The expression profile of the v-myb myeloblastosis viral oncogene homologue (MYB) gene was positively correlated with menogaril, *N*,*N*-dibenzyl dunomycin, hydroxyurea and dichlorallyl lawsone. This gene is known to be involved in DNA replication and repair. In previous investigations, the related c-Myb has been identified as the candidate factor for regulating topoisomerase IIa gene expression in proliferating hematopoietic cells. Topoisomerase IIa acts as a logical downstream effector for the growth-stimulatory effects of the Myb transcription factor family.[57]

A few other genes of importance to cell proliferation and apoptosis were identified in our analysis. For example, myeloid leukemia factor 1 (MLF1) correlated negatively with menogaril and VM-26, and transcription factor AP-2 alpha (TFAP2A) was negatively correlated with *N*,*N*-dibenzyl dunomycin.

4. Conclusions

In this study, we mined and investigated drug activity patterns and gene expression profiles of 30 cancer cell lines from five different organs from the NCI60 cell line panel. Five groups of cells with similar tissue of origin were identified by SOM unsupervised clustering and correlation analysis, i.e. leukemia, melanoma, prostate and ovarian, basal-like breast, and luminal-like breast cancer cells. Thus we used similarities in the gene expression profiles rather than organ types for classifying the cancer cell lines with respect to tissue of origin. This was done in order to investigate specificity in drug response with respect to tissue of origin. High efficacy compounds that show selectivity towards one of the five groups of cells lines were identified.

The breast cancer cells had the largest genetic variation and showed the most variable response to the different chemical agents. This effect was determined by cellular subtypes of breast cancers, basal and luminal. In this case, a strong correlation between gene expression and chemosensitivity data was observed. By treating the two breast cell lines as separate groups, high efficacy compounds selective to the BR luminal cancer cell lines were found.

We also performed a structural similarity analysis to further facilitate the potential use of the chemosensitivity data set. Eight potential topoisomerase II inhibitors were identified using this analysis. This analysis suggested that it is possible to find promising compounds effective in cancer cell growth inhibition with a simple similarity measure. Such methods can be used to screen through large databases in order to select compounds from large data sets for subsequent experimental testing.

Authors' contributions

P.W. carried out the largest part of the data mining and data analysis work and wrote parts of this manuscript, S.O.J. supervised P.W.'s work, made the analysis on tissue specificity and wrote the main body of this manuscript, Q.L. and J.E.P.L. contributed to the data mining and data analysis work, O.R. wrote the mysql program used for parsing and searching the data, S.J.N. helped with evaluating genes identified by gene–drug correlation, A.C.E., Q.L., A.P. and F.B. helped supervising the project and contributed with ideas and input to the project. All authors read and contributed to this manuscript.

Acknowledgements

Peng Wan, Jens Erik Pontoppidan Larsen and Svava Ósk Jónsdóttir acknowledge financial support by the Danish Research Council for Technology and Production Sciences (FTP) and the Program Commission on Nanoscience, Biotechnology and IT (NABIIT) under the Danish Strategic Research Council.

References and notes

1. Huang, R.; Wallqvist, A.; Thanki, N.; Covell, D. G. *Pharmacogenomics J.* **2005**, *5*, 381.
2. Rhodes, D. R.; Kalyana-Sundaram, S.; Mahavisno, V.; Varambally, R.; Yu, J.; Briggs, B. B.; Barrette, T. R.; Anstet, M. J.; Kincead-Beal, C.; Kulkarni, P.; Varambally, S.; Ghoshy, D.; Chinnaiyan, A. M. *Neoplasia* **2007**, *9*, 166.
3. Covell, D. G. *Trends Pharmacol. Sci.* **2008**, *29*, 1.
4. Ooyama, A.; Takechi, T.; Toda, E.; Nagase, H.; Okayama, Y.; Kitazato, K.; Sugimoto, Y.; Oka, T.; Fukushima, M. *Cancer Sci.* **2006**, *97*, 510.
5. Alley, M. C.; Scudiero, D. A.; Monks, P. A.; Hursey, M. L.; Czerwinski, M. J.; Fine, D. L.; Abbott, B. J.; Mayo, J. G.; Shoemaker, R. H.; Boyd, M. R. *Cancer Res.* **1988**, *48*, 589.
6. Shoemaker, R. H. *Nat. Rev. Cancer* **2006**. http://dtp.nci.nih.gov/.
7. Wang, H.; Klinginsmith, J.; Dong, X.; Lee, A. C.; Guha, R.; Wu, Y.; Crippen, G. M.; Wild, D. J. *J. Chem. Inf. Model.* **2007**, *47*, 2063.
8. Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. *Science* **1995**, *275*, 343.
9. Holbeck, S. L.; Collins, J. M.; Doroshow, J. H. *Mol. Cancer Ther.* **2010**, *9*, 1451.
10. Paull, K. D.; Shoemaker, R. H.; Hodes, L.; Monks, A.; Scudiero, D. A.; Rubinstein, L.; Plowman, J.; Boyd, M. R. *J. Natl. Cancer Inst.* **1989**, *81*, 1088.
11. Zaharevitz, D. W.; Holbeck, S. L.; Bowerman, C.; Svetlik, P. A. *J. Mol. Graph. Model.* **2002**, *20*, 297.
12. Fang, X.; Shao, L.; Wang, S. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 249.
13. Fojo, T.; Farrell, N.; Ortuzar, W.; Tanimura, H.; Weinstein, J.; Myers, T. G. *Crit. Rev. Oncol. Hematol.* **2005**, *53*, 25.
14. Evans, A.; Bates, V.; Troy, H.; Hewitt, S.; Holbeck, S.; Chung, Y. L.; Phillips, R.; Stubbs, M.; Griffiths, J.; Airley, R. *Cancer Chemother. Pharmacol.* **2008**, *61*, 377.
15. Shivakumar, P.; Michael Krauthammer, M. *BMC Bioinform.* **2009**, *10*, S17.
16. Chakravarti, S. K.; Klopman, G. *Bioorg. Med. Chem.* **2008**, *16*, 4052.
17. Ross, D. T.; Scherf, U.; Eisen, M. B.; Perou, C. M.; Rees, C.; Spellman, P.; Iyer, V.; Jeffrey, S. S.; Van, R. M.; Waltham, M.; Pergamenschikov, A.; Lee, J.; Lashkari, D.; Shalon, D.; Myers, T.; Weinstein, J. N.; Botstein, D.; Brown, P. O. *Nat. Genet.* **2000**, *24*, 227.
18. Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L.; Kohn, K. W.; Reinhold, W. C.; Myers, T. G.; Andrews, D. T.; Scudiero, D. A.; Eisen, M. B.; Sausville, E. A.; Pommier, Y.; Botstein, D.; Brown, P. O.; Weinstein, J. N. *Nat. Genet.* **2000**, *24*, 236.
19. Staunton, J. E.; Slonim, D. K.; Coller, H. A.; Tamayo, P.; Angelo, M. J.; Park, J.; Scherf, U.; Lee, J. K.; Reinhold, W. O.; Weinstein, J. N.; Mesirov, J. P.; Lander, E. S.; Golub, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10787.
20. Shankavaram, U. T.; Reinhold, W. C.; Nishizuka, S.; Major, S.; Morita, D.; Chary, K. K.; Reimers, M. A.; Scherf, U.; Kahn, A.; Dolginow, D.; Cossman, J.; Kaldjian, E. P.; Scudiero, D. A.; Petricoin, E.; Liotta, L.; Lee, J. K.; Weinstein, J. N. *Mol. Cancer Ther.* **2007**, *6*, 820.
21. Ring, B. Z.; Chang, S.; Ring, L. W.; Seitz, R. S.; Ross, D. T. *BMC Genomics* **2008**, *9*, 74.
22. Blower, P. E.; Verducci, J. S.; Lin, S.; Zhou, J.; Chung, J.; Zunyan Dai, Z.; Liu, C.; Reinhold, W.; Lorenzi, P. L.; Kaldjian, E. P.; Croce, C. M.; Weinstein, J. N.; Sadee, W. *Mol. Cancer Ther.* **2007**, *6*, 1483.
23. Myers, T. G.; Anderson, N. L.; Waltham, M.; Li, G.; Buolamwin, J. K.; Scudiero, D. A.; Pad, K. D.; Sausville, E. A.; Weinstein, J. N. *Electrophoresis* **1997**, *18*, 647.
24. Ma, Y.; Ding, Z.; Qian, Y.; Shi, X.; Castranova, V.; Harner, E. J.; Guo, L. *Clin. Cancer Res.* **2006**, *12*, 4583.
25. Shankavaram, U. T.; Varma, S.; Kane, D.; Sunshine, M.; Chary, K. K.; Reinhold, W. C.; Pmmier, Y.; Weinstein, J. N. *BMC Genomics* **2009**, *10*, 277. http://discover.nci.nih.gov/cellminer/.
26. Nakatsu, N.; Yoshida, Y.; Yamazaki, K.; Nakamura, T.; Dan, S.; Fukui, Y.; Yamori, T. *Mol. Cancer Ther.* **2005**, *4*, 399.
27. Dai, Z.; Barbacioru, C.; Huang, Y.; Sade, W. *Pharm. Res.* **2006**, *23*, 336.
28. Amundson, S. A.; Myers, T. G.; Scudiero, D.; Kitada, S.; Reed, J. C.; Fornace, A. J., Jr. *Cancer Res.* **2000**, *60*, 6101.
29. Dan, S.; Shirakawa, M.; Mukai, Y.; Yoshida, Y.; Yamazaki, K.; Kawaguchi, T.; Matsuura, M.; Nakamura, Y.; Yamori, T. *Cancer Sci.* **2003**, *94*, 1074.
30. Shedden, K.; Townsend, L. B.; Drach, J. C.; Rosania, G. R. *Pharm. Res.* **2003**, *20*, 843.
31. Okabe, M.; Szakács, G.; Reimers, M. A.; Suzuki, T.; Hall, M. D.; Abe, T.; Weinstein, J. N.; Gottesman, M. M. *Mol. Cancer Ther.* **2008**, *7*, 3081.
32. Musumarra, G.; Barresi, V.; Condorelli, D. F.; Sciré, S. *Biol. Chem.* **2003**, *384*, 321327.
33. Barresi, V.; Fortuna, C. G.; Garozzo, R.; Musumarra, G.; Sciré, S.; Condorelli, D. F. *Mol. BioSyst.* **2006**, *2*, 231239.
34. Wallqvist, A.; Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. *Bioinformatics* **2003**, *19*, 2212.
35. Yi, S.; Park, T.; Lee, J. K. *BMC Bioinform.* **2008**, *9*, 76.
36. Ma, Y.; Ding, Z.; Qian, Y.; Wan, Y.; Tosun, K.; Shi, X.; Castranova, V.; Harner, E. J.; Guo, N. L. *Int. J. Oncol.* **2009**, *34*, 107.
37. Covell, D. G.; Wallqvist, A.; Huang, R.; Thanki, N.; Rabow, A. A.; Lu, H. *Proteins: Structure, Function, Bioinformatics* **2005**, *59*, 403.
38. Kutalik, Z.; Beckmann, J. S.; Bergmann, S. *Nat. Biotechnol.* **2008**, *26*, 531.
39. Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.; Subramanian, A.; Ross, K. N.; Reich, M.; Hieronymus, H.; Wei, G.; Armstrong, S. A.; Haggarty, S. J.; Clemons, P. A.; Wei, R.; Carr, S. A.; Lander, E. S.; Golub, T. R. *Science* **2006**, *313*, 1929–1935.
40. Pubchem bioassay database: http://pubchem.ncbi.nlm.nih.gov.
41. Boyd, M. R. In *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials and Approval*; Teicher, B. A., Ed.; Humana Press: Totowa, NJ, 1995; pp 23–41.
42. Boyd, M. R. *Biostatistics* **2003**, *4*, 249.
43. Genomics & Bioinformatics Group. http://discover.nci.nih.gov/index.jsp.
44. Irizarry, R. A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y. D.; Antonellis, K. J.; Scherf, U.; Speed, T. P. *Biostatistics* **2003**, *4*, 249.
45. Molecular Design Limited. http://www.mdli.com.
46. Chemical Computing Group. http://www.chemcomp.com.
47. Jónsdóttir, S. Ó.; Jørgensen, F. S.; Brunak, S. *Bioinformatics* **2005**, *21*, 2145.
48. R: Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. http://www.R-project.org, ISBN 3-900051-07-0.
49. Kohonen, T. *Self-Organizing Maps*; Springer Series in Information Sciences: Berlin, 1995.
50. Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. The Matlab DSP conference: 16–17 November, Espoo, Finland, pp. 35–40.
51. Laboratory of computer and information science. http://www.cis.hut.fi/projects/somtoolbox/.
52. Kouskoumvekaki, I.; Yang, Z.; Jónsdóttir, S. Ó.; Olsson, L.; Panagiotou, G. *BMC Bioinform.* **2008**, *9*, 59.
53. Dan, S.; Tsunoda, T.; Kitahara, O.; Yanagawa, R.; Zembutsu, H.; Katagiri, T.; Yamazaki, K.; Nakamura, Y.; Yamori, T. *Cancer Res.* **2002**, *62*, 1139.
54. Wierzba, K.; Sugimoto, Y.; Matsuo, K.; Toko, T.; Takeda, S.; Yamada, Y.; Tsukagoshi, S. *Jpn. J. Cancer Res.* **1990**, *81*, 842.
55. Sartiano, G. P.; Lynch, W. E.; Bullington, W. D. *J. Antibiot. Tokyo* **1979**, *32*, 1038.
56. Uniprot (Universal protein resource). http://www.uniprot.org.
57. Brandt, T. L.; Fraser, D. J.; Leal, S.; Halandras, P. M.; Kroll, A. R.; Kroll, D. J. *J. Biol. Chem.* **1997**, *272*, 6278.