In Vitro Drug Sensitivity-Gene Expression Correlations Involve a Tissue of Origin Dependency

C. R. Andersson,[†] M. Fryknäs,[‡] L. Rickardson,[§] R. Larsson,[§] A. Isaksson,[§] and M. G. Gustafsson*[,§,||]

The Linnaeus Center for Bioinformatics, Department of Genetics and Pathology, Department of Medical Sciences, and Department of Engineering Sciences, Uppsala University, Uppsala, Sweden

Received March 7, 2006

A major concern of chemogenomics is to associate drug activity with biological variables. Several reports have clustered cell line drug activity profiles as well as drug activity-gene expression correlation profiles and noted that the resulting groupings differ but still reflect mechanism of action. The present paper shows that these discrepancies can be viewed as a weighting of drug−drug distances, the weights depending on which cell lines the two drugs differ in.

1. INTRODUCTION

Chemogenomics is an emerging interdisciplinary field described by Bredel and Jacoby as "the study of the genomic and/or proteomic response of an intact biological system—whether it be single cells or whole organisms—to chemical compounds, or the study of the ability of isolated molecular targets to interact with such compounds".[1] This paper addresses an issue which has been pointed out by several authors yet never fully explored. Correlations between compound activity and biological variables are used as descriptors of the compounds. However, this method produces a different view of the relatedness between compounds than using the activity of the compound by itself. The basis for these differences is intriguing but has been difficult to understand. Here we will first give an idea of the development that led to the description of compounds in terms of correlation coefficients; second we describe the issue in greater detail.

The chemical space has been systematically explored within the context of pharmacology for a long time. High throughput screening of activity in chemical libraries has revealed structure−activity relationships that provide information about moieties important for activity. Compound activity can be gauged in terms of e.g. binding affinity to a protein of interest, but in pharmacological applications it is natural to gauge it in terms of pharmacological effect. Although whole organism models would be the most informative assays, they are not always practical.

In cancer drug research a reasonable trade off between biological complexity of the assay and ethical/economical constraints is achieved by using appropriate cancer cell lines established from patients. Such cell lines include a substantial part of the pathophysiology associated with the disease and may also include clinically relevant pharmacodynamic phenomena such as cellular drug resistance. When cell lines are used for probing activity, the effect of the compound is usually measured by the fraction of cells surviving exposure to the compound. Activities are reported as either the fraction of cells surviving at some fixed concentration or the estimated concentration for which half of the cells die (known as the inhibitory concentration 50, IC_{50}).

The molecular causes and mechanisms of cancer are diverse, so screening for activity in a diverse cancer cell line panel is preferable. A diverse panel allows the discovery of broad spectrum compounds as well as the identification of targeted treatments. This insight motivated the establishment of the NCI-60 cell line panel at the National Cancer Institute, U.S.A. The NCI-60 panel contains 60 cell lines from 9 different cancers, all established from human patients. In an ongoing effort at NCI, chemical libraries are screened for activity in the panel, generating in vitro drug sensitivity profiles for each compound. This effort has shown that drugs with similar sensitivity profiles often have similar mechanims of action,[2] something which was subsequently verified in other cell line panels e.g. by Dahr and co-workers.[3] Thus, sensitivity profiles of novel drugs with unknown mechanism of action can be compared to profiles of known drugs to formulate hypotheses on their mechanisms of action as well as be used to identify important structural characteristics by computing whether drugs with similar activity also have similar structural features.

In addition to classification of mechanisms of activity the cell line panels are also used to uncover the biological variation that accounts for variation in drug response. Ultimately, this may provide target leads for drug optimization by identifying the biological targets of a drug as well as active chemoresistance mechanisms. Furthermore, targeted therapies could be developed by identifying markers for drug activity in particular tissues. In an effort to provide an integrated view of the NCI-60 panel, various biological features of the panel are being characterized. For example, mRNA gene expression microarray technology[4] allows simultaneously assaying the expression level of thousands

* Corresponding author phone: +46-18-4713229; fax: +46-18-555096; e-mail: mg@angstrom.uu.se. Corresponding author address: Dept. Engineering Sciences, Uppsala University, P.O. Box 534, SE-751 21 Uppsala, Sweden.

[†] The Linnaeus Center for Bioinformatics.

[‡] Department of Genetics and Pathology.

[§] Department of Medical Sciences.

[||] Department of Engineering Sciences.

10.1021/ci060073n CCC: $37.00
Published on Web 12/01/2006

of genes. Since the drug sensitivity of a cell line depends on its chemoresponse mechanisms, drug sensitivity can be used to associate genes with chemoresponse by identifiying concordance between gene expression and drug sensitivity across cell lines. Such a combination of chemo- and bioinformatic approaches is one of the core methodologies in chemogenomics.

Weinstein et al.[5] pioneered the study of drug sensitivity and gene expression on a large scale. They noted that clustering the cell lines by mRNA gene expression grouped them by tissue of origin. However, when the cell lines were instead clustered on drug response, some exceptions occurred. This indicates that drug response data contain information about the cell lines that is not related to tissue of origin. Consequently, this indicates that chemoresponse mechanisms are distributed across different tissues in the panel and that it should be possible to link drug activity to gene expesssion.

Chemogenomic data sets typically contain thousands of measurements, many of which are statistically dependent. For example, genes often belong to a coregulated set of genes, and as mentioned drug sensitivity tends to be similar for drugs of the same mechanistic class. When exploring these complex data sets a visual presentation of the information is very useful. Associations between *subsets* of drugs and genes can be visualized by displaying the genome-wide pattern of correlations for all pairs of drug sensitivity and gene expression profiles in an array of colored blocks where the correlation between drugs and genes is indicated by the color of the blocks. Each row corresponds to a drug and each column to a gene. The rows and columns are organized such that drugs in adjacent rows and genes in adjacent columns are similar according to some well defined measure of similarity. This is usually accomplished by having the order of rows and columns correspond to the order of the leaves in dendrograms (family trees) obtained from hierarchical clustering of the drugs and genes, respectively. With this presentation it is possible to visually identify groups of similar genes and drugs with strong correlations between the corresponding drug sensitivity and gene expression profiles. NCI provides a public Web service, CIMminer,[6] for visualizing data in this manner. Using this approach Weinstein et al.[5] found e.g. a set of compounds that were actively transported out of the cells by Pgp/Mdr-1, a well-known system for cellular detoxification. These compounds' activities had negative correlation to activities of Pgp/Mdr-1 and appeared as a coherent region in the map of correlations. Furthermore, a significantly enriched fraction of all compounds known to be Mdr-1 targets was present in the same cluster. This example demonstrates that the chemogenomic approach is able to associate drugs with mechanistically relevant pathways.

As mentioned, the CIMminer approach utilizes hierarchical, also known as agglomorative, clustering for constructing dendrograms. Hierarchical clustering sequentially clusters objects together by choosing the closest pair of objects, where objects may be either individual observations or clusters formed in a previous step. Distance between pairs of observations is determined by the metric in use. The distance between two clusters is determined by another function, the linkage function. For example, the average linkage function calculates the distance between two clusters as the average pairwise distance between observations in one of the clusters to observations in the other cluster. It is well-known that cluster structure (and consequently the arrangement of the correlation map) is greatly affected by the choice of linkage and metric function. There is a large literature available debating the appropriateness of different settings, but, by and large, the choice is arbitrary and left to the investigator.

Table 1: Ten Cell Lines Used and Their Corresponding Tissue of Origin

origin	name of cell line
leukemia	CCRF-CEM, CEM-VM1
renal adenomocarcinoma	ACHN
small cell lung cancer	NCI-H69, H69AR
myeloma	RPMI8226-S, 8226/DOX40, 8226/LR5
lymphoma	U937-GTB, U937 VCR

The visualization requires the genes and drugs to be arranged by dendrograms, but the choice of profiles from which gene−gene and drug−drug distances are to be calculated is somewhat arbitrary. The natural choice would be to calculate the distance between measured mRNA gene expression and drug sensitivity profiles, respectively. However, when CIMminer is used, drugs and genes are clustered on their correlation coefficient profiles (CCPs). The CCP for a drug consists of all the Pearson correlation coefficients for its sensitivity profile computed for all gene expression profiles. Similarly, the CCP for a gene consists of the correlation coefficients for its activity profile computed for all drug sensitivity profiles. As pointed out by e.g. Scherf et al.[7] and Dan et al.,[8] drugs with similar mechanisms of action tend to cluster together whether clustered based on CCPs or directly based on drug sensitivity profiles, but the resulting dendrograms are not identical. The quantitative understanding of this difference has remained unclear, and it is important to determine whether the novel organization obtained with the more complex clustering of CCPs provides a more informative picture of how drug and gene activities are related.

In this work we reconsider clustering of drug and gene profiles for visualization of their correlations. We show that clustering of CCPs instead of clustering of the original drug sensitivity and gene expression profiles can be viewed as the result of using a novel dissimilarity measure that depends strongly on the relationships between the cell lines employed.

2. METHODS

The biological data used in the present paper is described in full in Rickardson et al.[9] Briefly, a cell-line panel (Table 1) consisting of the parental cell lines RPMI 8226 (myeloma), CCRF-CEM (leukemia), U937 GTB (lymphoma), and NCI-H69 (small cell lung cancer); the drug-resistant sublines 8226/Dox40 8226/LR5, CEM/VM-1, U937 vcr, and H69AR; and the primary resistant ACHN (renal adenocarcinoma) was assayed. For each cell line 3903 gene activities were measured successfully using our in-house cDNA microrrray gene expression system. The in-house spotted microarrays carried 7458 cDNA clones included in the Human Sequence Verified Set (Research Genetics, Huntsville, AL). A complete list of genes printed on the arrays is available online.[10] Sensitivity profiles (IC_{50}) were recorded for 66 anticancer drugs (Table 2) that were obtained from commercial sources or from NCI, dissolved according to the manufacturer's

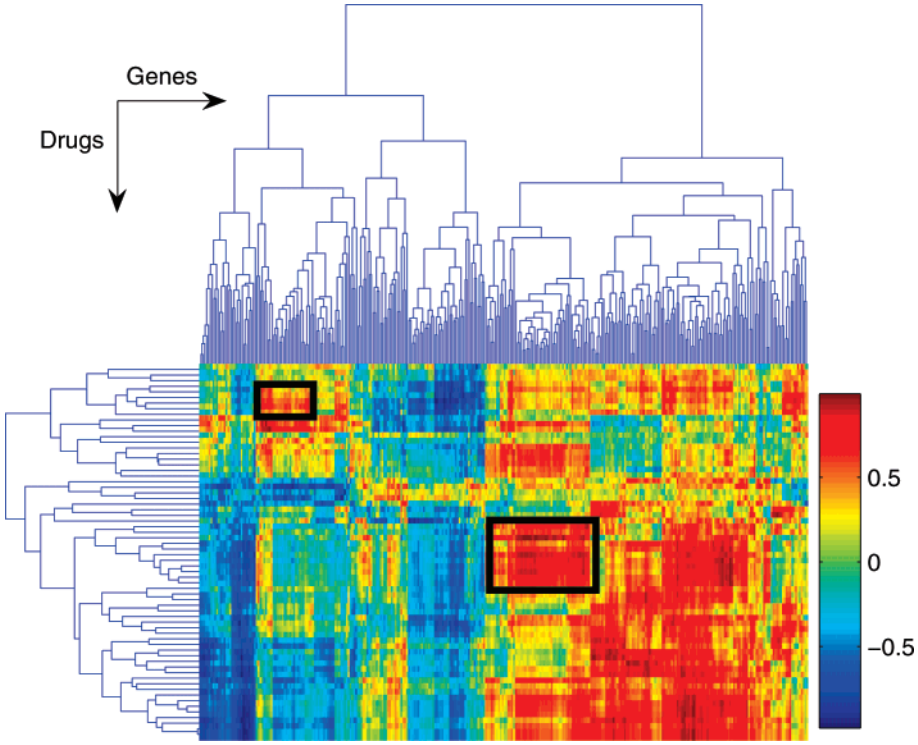

Figure 1. Map of correlation coefficients ordered by clustering of correlation coefficient profiles (CCPs) for both drugs and genes using Matlab and the CIMminer default settings (average linkage, Euclidean metric). Each colored block in the map shows the corresponding Pearson correlation coefficient ρ between a drug sensitivity profile and a gene expression profile, colors ranging from dark blue ($\rho = -1$) to dark red ($\rho = 1$). Indicated by the black borders are groups of genes and drugs that show large positive correlations between their profiles.

Table 2: 66 Drugs Used To Study Drug-Gene Interactions

mechanism of action	name of drug
antimetabolites	acivicin, aminopterin, aphidicolin, 5-azacytidine, L-alanosine, cladribine, cyclocytidine, cytarabine, 3-deazauridine, 2-azacytidine, diglycoaldehyde, fludarabine, 5-fluorouracil, ftorafur, hydroxyurea, 6-mercaptopurine methotrexate, PALA, pentostatin, 6-thioguanine, thymidine
alkylating agents	busulfan, carboplatin, chlorambucil, cisplatin, 4-HC, J1, mechlorethamine, melphalan, mitomycin C, P2, sarcolysine
topoisomerase I-inhibitors	camptothecin, SN-38, topotecan
topoismerase II-inhibitors	amsacrine, bisantrene, daunorubicin, doxorubicin, epirubicin, etoposide, idarubicin, mitoxantrone, teniposide
proteasome inhibitors	bortezomib, lactacystin, MG-132, MG-262
tubulin active agents	colchicine, docetaxel, maytansine, paclitaxel, podophyllotoxin, vinblastine, vincristine, vindesine, vinorelbine, estramustine
others	aclarubicin, anguidine, cycloheximide, flavoneacetate, Hoechst 33342, MBGB, MIBG, spirogermanium

instructions, and tested in five concentrations, obtained by 10-fold serial dilution. The investigational alkylating agents J1 and P2 were kind gifts from Oncopeptides AB (Stockholm, Sweden). The Fluorometric Microculture Cytotoxicity Assay (FMCA) used is based on measurement of fluorescence generated from hydrolysis of fluoroscein diacetate (FDA) to fluorescein by cells with intact plasma membranes (described in detail previously[11]). Hierarchical clustering and all calculations were performed using stock functions in the Matlab (Mathworks Inc.) environment.

3. RESULTS

To illustrate the difference between clustering genes and drugs based on their CCPs and on their original drug sensitivity and gene activity profiles we used drug sensitivity and gene expression data from 10 different human cell lines previously described.[9] For ease of presentation we selected only genes whose expression pattern had a standard deviation greater than one, leaving 324 out of 3903 genes (including all 3903 genes did not change the qualitative results). We choose to study the difference between clustering CCPs and original measurements for average linkage and the Euclidean distance metric. It could be argued that the angular separation similarity measure is more relevant in this setting; however, the main conclusion is the same for angular separation, but a formal investigation requires a little bit more algebra. The corresponding results for angular separation is available as Supporting Information.

In Figure 1 we show a map of correlation coefficients for each pair of original in vitro drug sensitivity profiles of 66 drugs (rows) and the expression profiles of 324 genes (columns) across 10 cell lines. Thus, the image in Figure 1

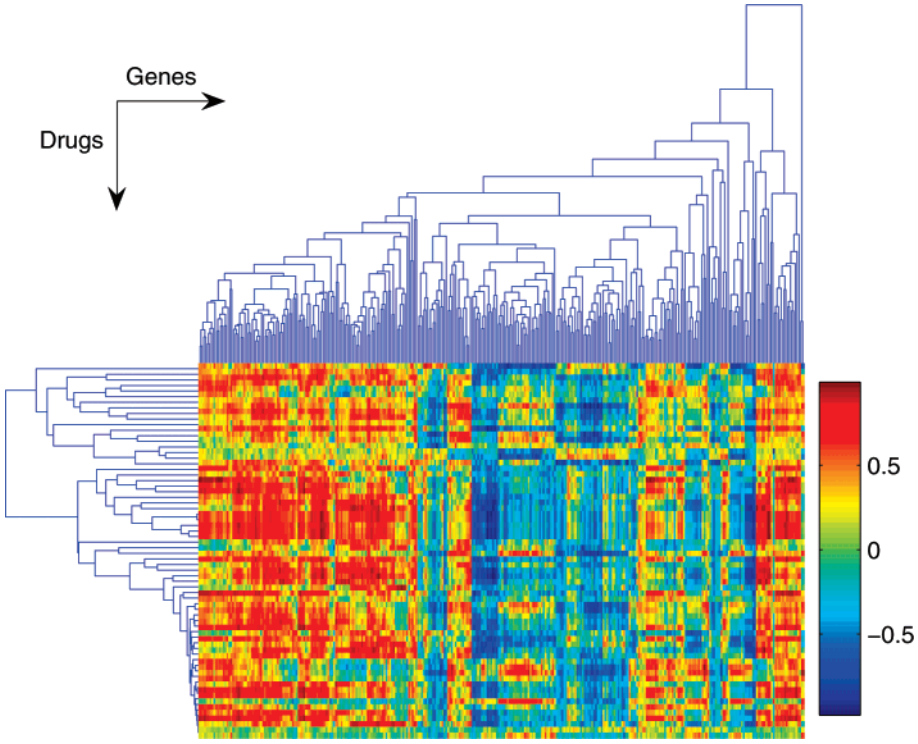

Figure 2. Alternative visualization of correlation coefficients. Here the dendrograms and the associated ordering of genes and drugs are obtained from clustering original drug sensitivity profiles and gene expression profiles (average linkage, Euclidean metric). Apparently this yields different dendrograms and visualization of the correlations from Figure 1. Each block in the correlation map shows the corresponding Pearson correlation coefficient between a drug sensitivity profile and a gene expression profile, colors ranging from dark blue ($\rho = -1$) to dark red ($\rho = 1$). There is much less structure in this correlation map in comparison to the one constructed based on clustering of correlation coefficient profiles as in Figure 1.

consists of 66 × 324 blocks, each of which correspond to a drug-gene pair. The color of each block indicates the magnitude of the Pearson correlation coefficient between the corresponding drug's sensitivity profile and gene's expression profile. Rows and columns are organized by hierarchical clustering using average linkage for merging of profiles/clusters and Euclidean distance as a metric for individual profiles. It is possible to identify locally connected regions consisting of large correlation coefficients such as those marked with a black border. This suggests that there are relatively large groups of genes whose expression patterns all have strong correlation with the sensitivity profiles of a group of drugs. In view of previous studies, such groups of drugs and genes would be highly interesting to pursue further and possibly uncover a mechanistic model explaining the activity of the group of drugs in terms of the biological correlates.

In Figure 1, adjacent genes and drugs are related in the sense that their CCPs are similar. One might also consider ordering rows and columns of the correlation map by clustering drugs and genes separately. We show the resulting arrangement in Figure 2 where the map of correlations is instead ordered by two dendrograms obtained by clustering of the original drug sensitivity and gene expression profiles using the same defaults settings as before (average linkage, Eucidean distance). In this visualization, of the same correlations as in Figure 1, it is difficult to identify any large locally connected regions of gene-drug pairs with large correlation coefficients. Thus, the subsets of genes associated with subsets of drugs found when organizing according to clustering based on the CCPs in Figure 1 find little support in the original data.

To assess the difference in further detail, we focus on the clustering of the drugs. In Figures 3 and 4 we show the dendrograms associated with the drugs for CCP and IC_{50} profiles, respectively. The distances are different since the Euclidean distances between CCPs are bounded (all values lie in the interval $[-1, 1]$). It is obvious from inspection that the compounds are differently grouped for the two data sets. To facilitate a comparison, we argue that an ideal grouping of the compounds would roughly group them by their mechanism of action. We study the grouping by using the dendrogram to divide the compounds into groups, cutting the dendrogram from the root and outward. For instance, cutting the dendrograms at the first bifurcation (from the root) groups about 80% of the alkylating agents in one cluster and 60 of the antimetabolites in the other for the independent clustering. This shows that at least the alkylating agents appear as a somewhat homogeneous group in the dendrogram. With the CCP data, the figures are reversed with about 60% of the alkylating agents and 70% of the antimetabolites grouped together. Continuing, cutting the independent clustering at the next bifurcation splits off compound MBGB, but the clusters do not coalesce into mechanistic groups for the next 5 splits. For the CCPs, dividing the tree into 5 clusters produces a cluster containing all four proteasome inhibitors. However, other mechanistic classes are present in the same cluster, and they are interspersed amoung the proteasome inhibitors; among the proteasome inhibitors only Bortezomib and MG-262 are clustered at the lowest level. We believe this top-down comparison of the grouping does not supply any preference to either of the dendrograms.

A bottom-up comparison was also performed, using the number of pairs of cluster leaves that belong to the same

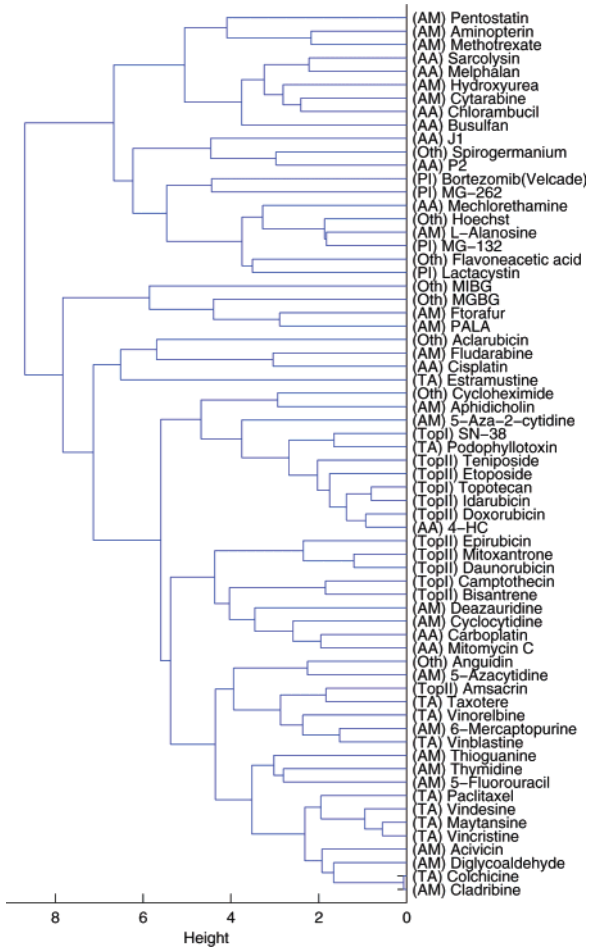

Figure 3. The compound dendrogram corresponding to Figure 1. Mechanistic classes are indicated in parentheses: AA alkylating agent, AM antimetabolite, TA tubulin active, TopI topoisomerase I inhibitor, TopII topoisomerase II inhibitor, PI proteasome inhibitor, Oth other.

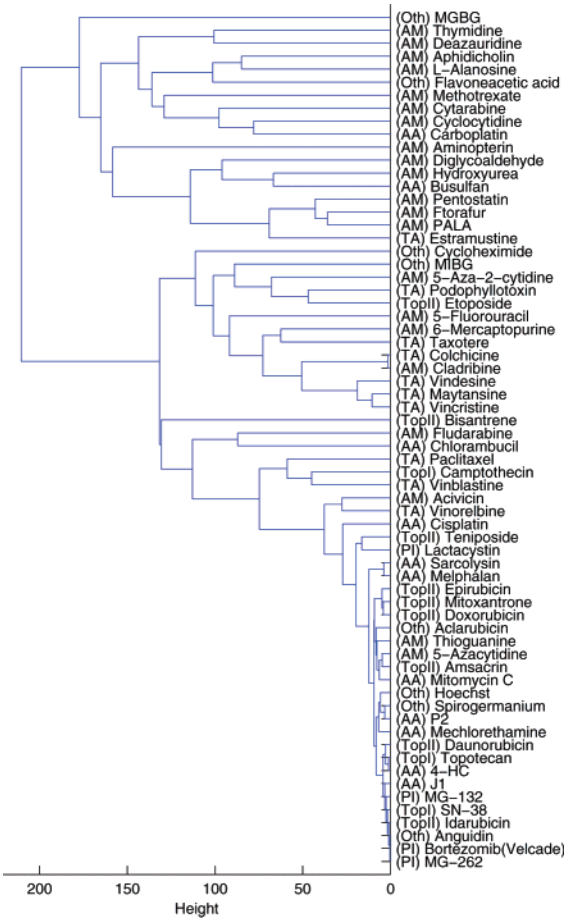

Figure 4. The compound dendrogram corresponding to Figure 2. Mechanistic classes indicated within parentheses as in Figure 3.

mechanistic group as a statistic. Using CCPs, 6 out of 20 (30%) of all cluster leaf pairs are annotated to the same mechanism, in the IC_{50} clustering, 8 out of 15 (53%). Given the small sample sizes, this does not provide any strong evidence of any of the clusterings grouping more of the same together. Overall, there are surprisingly many differences between the two data sets, considering that if two drugs have identical drug sensitivity (IC_{50}) profiles they will also have identical CCPs.

We conclude, as others have before, that clustering the drugs on either IC_{50} profiles or correlation coefficient profiles roughly groups compounds by known mechanisms of action. However, to understand the nature of the differences, the two different approaches must be put on equal footing. Next we show how this is possible by demonstrating that distances between CCPs may be written as functions of the corresponding IC_{50} profiles. Consequently, it is possible to interpret the change of features describing the drugs as a change in metric used for calculating the distances. Throughout the text we use bold lowercase letters ($\mathbf{x}$) to denote column vectors and bold uppercase letters to denote matrices ($\mathbf{X}$). A tilde will be used to indicate mean centered and normalized vectors ($\tilde{\mathbf{x}}$) and matrices ($\tilde{\mathbf{X}}$), where for matrices the mean centering and normalization have been applied to the rows. We start by noting that the Pearson correlation coefficient between observations of two random variables X and Y such as gene expression and drug sensitivity across n_c cell lines can be expressed as $\sum_{n=1}^{n_c} \tilde{x}_n \tilde{y}_n = \tilde{\mathbf{x}}^T \tilde{\mathbf{y}}$, where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ are n_c-element column vectors of mean centered and normalized observations of X and Y. It follows that the mapping of mean centered normalized $n_d \times 1$ drug sensitivity profiles $\tilde{\mathbf{d}}_i$ into $n_g \times 1$ drug-gene correlation profiles $\mathbf{f}_i$ can be expressed as

$$\mathbf{f}_i = \begin{pmatrix} \tilde{\mathbf{g}}_1^T \\ \tilde{\mathbf{g}}_2^T \\ \vdots \\ \tilde{\mathbf{g}}_{n_g}^T \end{pmatrix} \tilde{\mathbf{d}}_i = \tilde{\mathbf{G}} \tilde{\mathbf{d}}_i \qquad (1)$$

where $\tilde{\mathbf{g}}$ are mean centered and normalized gene expression profiles and there are n_g genes under consideration. We assume that $n_c < n_d$ and $n_c < n_g$. This is the case not only for the data set used here ($n_c = 10$, $n_d = 66$, $n_g = 324$) but also in general; generating or obtaining new cell lines or patient samples will continue to be the limiting factor in data acquisition in the forseeable future. Note that the normalized and mean centered drug profiles $\tilde{\mathbf{d}}_i$ can be written as a function of the original profiles $\mathbf{d}_i$ as

$$\tilde{\mathbf{d}}_i = \frac{\left(\mathbf{I} - \frac{1}{n_c}\mathbf{1}\mathbf{1}^T\right)\mathbf{d}_i}{\sqrt{\mathbf{d}_i^T\left(\mathbf{I} - \frac{1}{n_c}\mathbf{1}\mathbf{1}^T\right)\mathbf{d}_i^T}} \qquad (2)$$

where **1** is a column vector with all its elements equal to one. We now investigate how the Euclidean distances between the drugs are altered by transformating them into CCPs. The Euclidean distance between two original drug sensitivity profiles $\mathbf{d}_i$ and $\mathbf{d}_j$ is defined as

$$d(\mathbf{d}_i,\mathbf{d}_j) = \sqrt{(\mathbf{d}_i - \mathbf{d}_j)^T(\mathbf{d}_i - \mathbf{d}_j)} \tag{3}$$

and it follows that after the transformation in eq 1, the Euclidean distance between two correlation profiles $\mathbf{f}_i$ and $\mathbf{f}_j$ can be written in terms of the mean centered and scaled original profiles as

$$d(\mathbf{f}_i,\mathbf{f}_j) = d(\tilde{\mathbf{G}}\tilde{\mathbf{d}}_i, \tilde{\mathbf{G}}\tilde{\mathbf{d}}_j)$$

$$= \sqrt{(\tilde{\mathbf{G}}\tilde{\mathbf{d}}_i - \tilde{\mathbf{G}}\tilde{\mathbf{d}}_j)^T(\tilde{\mathbf{G}}\tilde{\mathbf{d}}_i - \tilde{\mathbf{G}}\tilde{\mathbf{d}}_j)}$$

$$= \sqrt{(\tilde{\mathbf{d}}_i - \tilde{\mathbf{d}}_j)^T\tilde{\mathbf{G}}^T\tilde{\mathbf{G}}(\tilde{\mathbf{d}}_i - \tilde{\mathbf{d}}_j)} \tag{4}$$

Equation 4 shows that when the mean centered and normalized drug sensitivity profiles $\tilde{\mathbf{d}}_i$ are transformed into CCPs, the Euclidean distances between the resulting CCPs may be interpreted as weighted distances in the space of mean centered and normalized drug sensitivity profiles defined by a matrix (metric tensor) $\mathbf{W}_G = \tilde{\mathbf{G}}^T\tilde{\mathbf{G}}$. Thus, from the distance between correlation profiles $d(\mathbf{f}_i,\mathbf{f}_j)$ we may write down a new measure $d^*(\mathbf{d}_i,\mathbf{d}_j) = d(\mathbf{f}(\mathbf{d}_i),\mathbf{f}(\mathbf{d}_j))$ as

$$d^*(\mathbf{d}_i,\mathbf{d}_j) = \sqrt{\left(\frac{\mathbf{d}_i}{\sqrt{\mathbf{d}_i^T\left(\mathbf{I} - \frac{1}{n_c}\mathbf{1}\mathbf{1}^T\right)\mathbf{d}_i^T}} - \frac{\mathbf{d}_j}{\sqrt{\mathbf{d}_j^T\left(\mathbf{I} - \frac{1}{n_c}\mathbf{1}\mathbf{1}^T\right)\mathbf{d}_j^T}}\right)\mathbf{W}_G\left(\frac{\mathbf{d}_i}{\sqrt{\mathbf{d}_i^T\left(\mathbf{I} - \frac{1}{n_c}\mathbf{1}\mathbf{1}^T\right)\mathbf{d}_i^T}} - \frac{\mathbf{d}_j}{\sqrt{\mathbf{d}_j^T\left(\mathbf{I} - \frac{1}{n_c}\mathbf{1}\mathbf{1}^T\right)\mathbf{d}_j^T}}\right)} \tag{5}$$

The differences between the dendrograms seen in Figures 3 and 4 come from using either d or d^* as the distance measure. This puts the CCPs on equal footing with drug sensitivity profiles as desired and enables us to make more incisive statements about difference in clustering results.

We note that there are two contributions to changes in distances: the normalization and mean centering as well as the weighting by the matrix $\mathbf{W}_G$. The effect of the normalization and mean centering operation represented by eq 2 can be interpreted as changing to the angular separation similarity measure which is well described in the literature. Here we therefore focus instead on the influence of the weighting matrix $\mathbf{W}_G$ by considering the distance between two drug profiles which already have zero mean and unit variance, i.e., $\mathbf{d}_i = \tilde{\mathbf{d}}_i$ as well as $\mathbf{d}_j = \tilde{\mathbf{d}}_j$. For such profiles, $d^*(\mathbf{d}_i,\mathbf{d}_j) = \sqrt{\delta\mathbf{d}^T\mathbf{W}_G\delta\mathbf{d}}$ where $\delta\mathbf{d} = \mathbf{d}_i - \mathbf{d}_j$.

To gain some insights about this result we express $\delta\mathbf{d}$ in a coordinate system in which the eigenvectors $\mathbf{w}_i$ of $\mathbf{W}_G$ are the basis vectors. This yields $\delta\mathbf{d} = \sum_i\alpha_i w_i$ and $d^*(\mathbf{d}_i,\mathbf{d}_j) = \sqrt{\sum_i\alpha_i^2\lambda_i}$ where λ_i is the eigenvalue of $\mathbf{W}_G$ that corresponds to eigenvector $\mathbf{w}_i$, i.e., $\mathbf{W}_G\mathbf{w}_i = \lambda_i\mathbf{w}_i$. Thus the weighted distance between the drugs depends on the direction of the difference $\delta\mathbf{d}$, whereas a conventional Euclidean distance

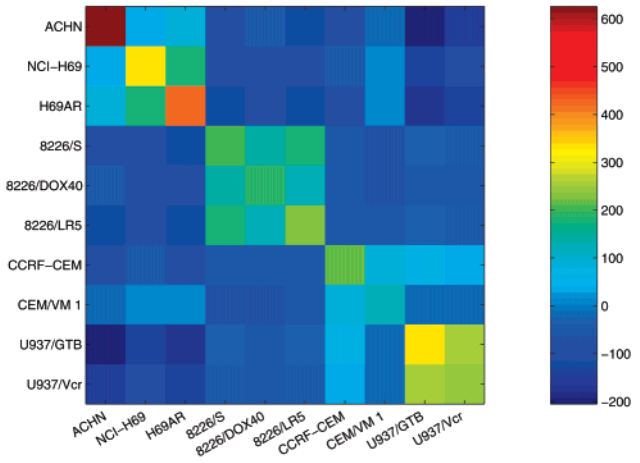

Figure 5. Visualization of the elements in the matrix $\mathbf{W}_G$ which defines the weighted distance measure (metric) for drugs. The numerical values of $\mathbf{W}_G$ are color coded according to the colorbar to the right. Apparently this matrix contains information about the origins of the cell lines.

($\mathbf{W}_G = I$) does not. More specifically, the largest distance is obtained for vectors where the difference is parallel with the eigenvector of the largest eigenvalue. Analogously, the smallest nonzero distance is obtained for vectors with a difference that is parallel to the eigenvector with the smallest eigenvalue. In looser terms, this means that calculating distances from the CCPs embeds a directionality into the drug sensitivity profile space that is determined by the gene expression profiles. Also, as may be concluded from above, a change $\delta\mathbf{d}$ that only affects mean value and scale will not affect the distance between the CCPs since all such changes will be lost in the normalization and mean centering.

Analogous to the conclusion above about how the drug clustering is affected by the change from a Euclidean metric to a generalized metric, essentially the same conclusions can be made regarding the gene clustering. In the gene clustering, the corresponding matrix of interest is the $n_c \times n_c$ dimensional matrix $\tilde{\mathbf{W}}_D = \tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$ constructed from the $n_d \times n_c$ matrix

$$\tilde{\mathbf{D}} = \begin{pmatrix} \tilde{\mathbf{d}}_1^T \\ \tilde{\mathbf{d}}_2^T \\ \vdots \\ \tilde{\mathbf{d}}_{n_d}^T \end{pmatrix}$$

In summary, for the Euclidean distance measure used for clustering, the matrix $\mathbf{W}_G = \tilde{\mathbf{G}}^T\tilde{\mathbf{G}}$ for the drugs and the matrix $\mathbf{W}_D = \tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$ for the genes must be considered when changing from the original profiles to the CCPs. In the Supporting Information, details about the corresponding changes in the angular separation measure of similarity are considered.

Based on the gene expression from the 10 cell lines and the drug sensitivity profiles for the 66 drugs, in Figure 5, the elements of the matrix $\mathbf{W}_G$ are displayed. From Figure 5 we find that the matrix has a distinct diagonal block structure. One example of a diagonal block consists of the relatively large positive coefficients corresponding to cell lines 8226/S, 8226/DOX40, and 8226/LR5 forming a 3×3 block matrix. Interestingly, cell line 8226/S is a myeloma parental cell line, and cell lines 8226/DOX40 and 8226/LR5 are two

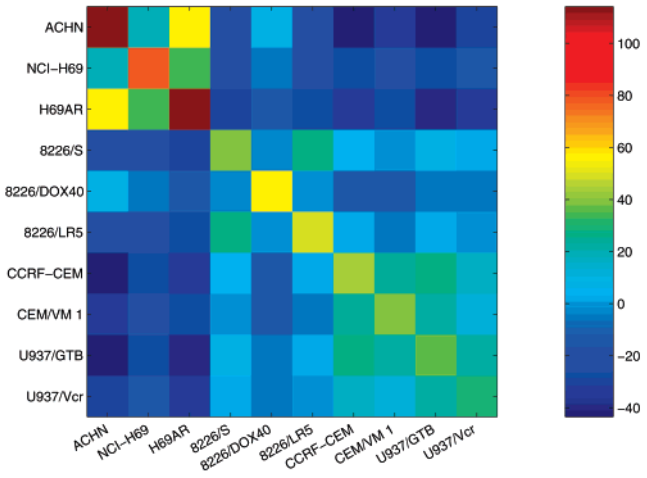

Figure 6. Visualization of the elements of the matrix $\mathbf{W}_D$ which defines the weighted distance measure (metric) for genes. The numerical values of the elements are indicated by a color scaling represented by the colorbar to the right. It is noticeable that this matrix strongly reflects the relationships between the cell lines.

drug-resistant sublines, selected for resistance against topoisomerase II inhibitor doxorubicin and the alkylating agent melphalan, respectively. Another example of a diagonal block matrix are the coefficients corresponding to the cell lines U937/GTB and U937/Vcr which also involve a distinct group of coefficents. These two cell lines are a pair of parental and drug-resistant lymphoma cells, and U937/Vcr is selected for vincristine resistance. We also see that the small cell lung cancer cell line H69 and cell line H69AR, selected for doxorubucin resistance, form a block with large coefficients. Parental and selected cell lines CCRF-CEM and CEM/VM1 also form a block. Finally, cell line ACHN is not part of a diagonal block matrix, and this is consistent with the fact that the corresponding cell line ACHN has no closely related cell lines in the panel. Thus, the matrix $\mathbf{W}_G$ reflects the different origins of the cell lines studied and shows that the new distance measure implicitly introduced via the use of CCPs is a consequence of these relationships.

We also computed the matrix $\mathbf{W}_D$ associated with distances between CCPs for the genes, see Figure 6. A slightly different pattern appears here. Notably, the 3×3 block of the 8226 cell lines now show higher coefficients for (8226/S, 8226/LR5) than for (8226/S, 8226/DOX40). Another difference is that ACHN is now included in a 3×3 block with the H69 cell lines. Nevertheless, the predominant pattern is that of similar cross terms between closely related cell lines.

To see the effect of these tissue of origin effects we eliminated the effects of the rescaling by clustering normalized mean centered drug profiles using average linkage and the Euclidean distance metric. Any difference between such a clustering and clustering the CCPs comes from the directionality. Next we studied the cluster leaf pairs formed in this clustering as we did earlier before rescaling. The majority of pairs were present in both clusterings, such as e.g. the leaf pair consisting of the isomeres melphalan and sarcolysin. However, e.g. the leaf pair podophyllotoxin and 4-HC, with roughly the same distance as melphalan and sarcolysin, were grouped with SN-38 and doxorubicin, respectively, in the CCP clustering. With the reasoning outlined above, this could be explained by different directions of the differences between their original sensitivity profiles. In Figure 7 a so-called delta graph shows the differences in IC_{50} in the respective cell lines. Apparently the main difference between sarcolysin and melphalan is a large difference in the 8226/LR5 cell line. The difference between

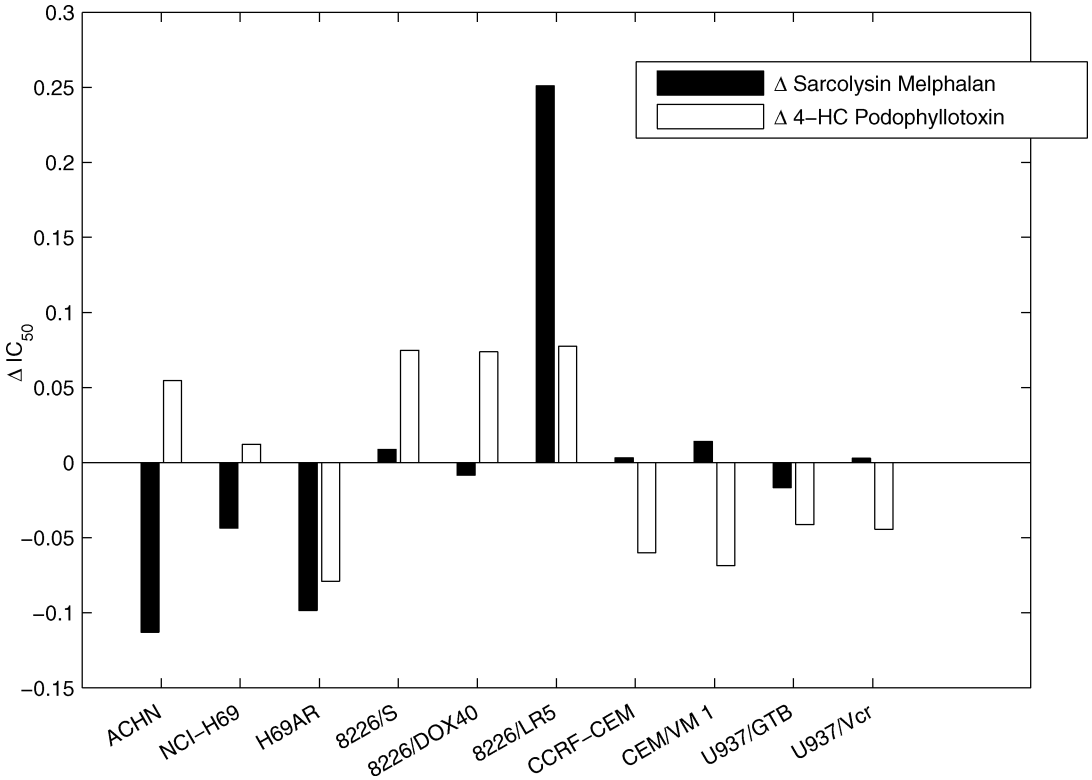

Figure 7. Difference in IC_{50} across the cell lines for between sarcolysin and melphalan as well as podophyllotoxin and 4-HC. Although the distance between these pairs is roughly the same after normalization, podphyllotoxin and 4-HC are broken apart when CCPs are clustered. This is due to the directionality of the difference coming into play with CCPs.

podophyllotoxin and 4-HC, on the other hand, can be seen to follow suite in the tissues, yielding a large projection to the eigenvectors corresponding to the largest eigenvalues. For details of the eigenvector analysis see the Supporting Information.

4. DISCUSSION

It could be argued that clustering of drugs and genes based on their original activity profiles arranges the map of all drug-gene correlation coefficients in a way that is easier to interpret than when the ordering of the coefficients is based on clustering of the corresponding CCPs. Ordering based on clustering of the original profiles using an appropriate combination of linkage function and dissimilarity measure results in maps with locally connected regions of large correlation coefficients, each region consisting of a particular subset of similar drug sensitivity profiles that are similar to a corresponding subset of similar gene expression profiles. This is not necessarily the case for a locally connected region with large correlation coefficients that have been found in a correlation map organized based on CCPs. In this case, all one can expect is that each such a region consists of a subset of drugs with similar CCPs and a subset of genes with similar CCPs. This important difference was noted by Weinstein et al.[5] where correlation coefficients were organized and visualized after clustering the drugs on their original profiles, but the genes were nevertheless clustered based on their CCPs, not on their original expression profiles. Moreover, the current version of CIMminer tool[6] at NCI does not offer clustering of original drug and gene profiles as an option for organization and visualization of the correlation coefficients.

Comparison of drugs via clustering of CCPs instead of original drug sensitivity profiles emphasizes and suppresses similarities based on the evolutionary relationships between the cell lines. In particular, the cross sensitivity of a drug in related cell lines is emphasized. This effect of $\mathbf{W}_D$ and $\mathbf{W}_G$ is not very intuitive, and we offer a technical analysis in the Supporting Information as well as the following example to aid understanding. Suppose that only three cell lines were available and that

$$\mathbf{W}_G = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

This is similar to the patterns observed in $\mathbf{W}_G$ calculated for our real data, i.e., cell lines 1 and 2 in this example are considered to be closely related and thus have large cross terms. Cell line 3 does not have any nonzero cross terms. If we denote the difference in sensitivity between two drugs in cell line i by δ_i we may write out the Euclidean distance between the CCPs of the drugs as

$$\sqrt{{\delta_1}^2 + 2\delta_1\delta_2 + {\delta_2}^2 + {\delta_3}^2}$$

which should be compared to the distance between the original measurements

$$\sqrt{{\delta_1}^2 + {\delta_2}^2 + {\delta_3}^2}$$

Thus, in this example the Euclidean distance between CCPs will differ from the distance between original measurements when $\delta_1\delta_2 \neq 0$. We note that if $\delta_1\delta_2 < 0$, the distance calculated between CCPs will be smaller than the distance between the original measurements. In effect this means that two drugs are considered more similar if drug sensitivity differences in related cell lines go in opposite directions. We also note that differences which are approximately equal, i.e., $\delta_1 \approx \delta_2$, will be given much greater weight. In this case, the differences in a group of closely related cell lines will be counted multiple times, whereas differences for a group of distantly related cell lines are counted only once. The effects identified in this toy example are also present in the more complex matrices $\mathbf{W}_G$ and $\mathbf{W}_D$ for the data set used in the present work. However, since the matrix elements are no longer binary but instead real numbers with different signs and magnitudes, the effects are more difficult to identify. We believe that the identified effects are undesirable since it is a priori expected for drugs to display cross sensitivity in related cell lines. Therefore, two drugs should be considered more similar if the differences in sensitivity are due to cross sensitivity in related samples than if the drugs differ in a set of samples that are not closely related. As indicated here, the opposite may happen when using CCPs.

We expect these phenomena to be present in all databases of mRNA expression and drug sensitivity based on cell line panels containing related samples. In addition to the results presented in this work based on our own cell lines and drug libraries, we examined the NCI 60 Cancer Cell Line database[6] on gene expression and drug sensitivity and found similar effects there. For instance, using CCPs for clustering drugs involves an emphasis on cross sensitivity in the leukemia samples, followed by an emphasis on cross sensitivity in the melanoma samples (data not shown). See the Supporting Information for some of the results obtained.

The fundamental problem here is that drug sensitivities in closely related cell lines are highly correlated. In Figure 8a a fictional set of four drug sensitivity profiles (1−4) accross two closely related cell lines are indicated. Each axis in the graph represents the drug sensitivity in one of the two cell line considered. Indicated by the solid lines are the contours of the underlying frequency distribution of drug activities upon repeated sampling in the chemical space of the drugs (i.e. we are not considering technical variation). Thus, drug profiles 1−4 all have equal probability of being observed. Apparently the Euclidean distance between profiles 2 and 4 is larger than between profiles 1 and 3. The effects of comparing these particular profiles in terms of their corresponding CCPs (eq 1) are shown in Figure 6b; the distance between profiles 1 and 3 is decreased, whereas that between drugs 2 and 4 is increased. This we believe to be undesirable because the effects of using the CCPs will suppress the fact that drugs 2 and 4 seem to be subject to the same chemoresponse mechanisms and that drugs 1 and 3 seem to be subject to different mechanisms. Clearly, drugs 2 and 4 differ in sensitivity by the same amount in both cell lines, whereas drug 1 has a higher sensitivity than drug 3 in one cell line, lower in the other. Thus drugs 1 and 3 appear to be more different than drugs 2 and 4, but in Figure 6b drugs 1 and 3 are instead moved closer and 2 and 4 farther away.

Moreover, we know that the elliptic contours in Figure 6a reflect the fact that the related cell lines have overlapping

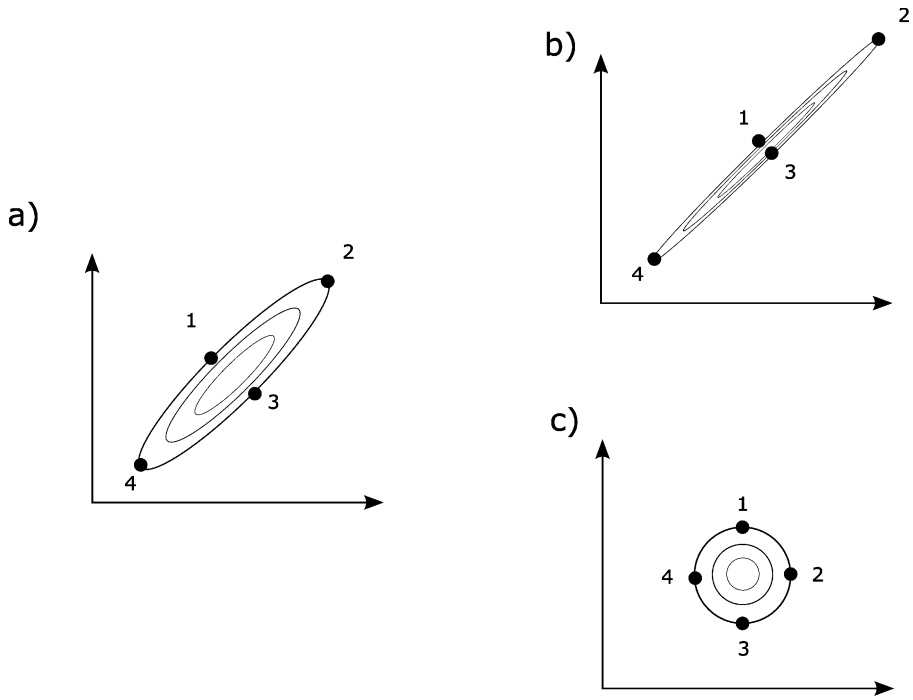

Figure 8. Effects of using CCPs as well as the Mahalanobis distance, a potential alternative. (a) Four original drug sensitivity profiles (1−4) across two cell lines. Each axis corresponds to the sensitivity of one of the cell lines. (b) Original drug profiles represented as CCPs results in a magnification of correlated differences. (c) Relationship between the profiles according to the classical Mahalanobis distance that de-emphasizes correlations. See text for details.

chemoresponse mechanisms. According to Figure 6a, if one of the cell lines is sensitive (insensitive) to a particular drug, it is quite likely that the other cell line also is sensitive (insensitive) to the same drug. This is the case because some of the drug mechanisms are overlapping; if the two cell lines would have been completely different, the contours would have been circular as in Figure 6c. Thus, if one would be interested in the natural goal to compare how different drugs are in terms of to which extent they activate completely different chemoresponse mechanisms, a transformation that results in de-emphasized correlations and relative distances as in Figure 6c is more much more appropriate than what is obtained in Figure 6b. The situation in Figure 6c is obtained by employing the classical Mahalanobis distance measure[13] which compensates for variances (contours) being different in different directions. Thus this use of the Mahalanobis distance may be interpreted as creating an ideal nonredundant cell-line panel which consists only of two cell lines with statistically uncorrelated chemoresponse mechanisms. With such a panel differential drug sensitivity could be read off directly as differential efficacy of chemoresponse mechanisms, and the Euclidean distance would be natural. However, in real data there will always be a technical variation in addition to the variation due to chemoresponse mechanisms, and further research is needed before such alternative measures as the Mahalanobis distance can be recommended.

It is perhaps somewhat surprising that also the dissimilarity between *gene* profiles calculated from the CCPs is strongly influenced by the origin of the cell lines in which they differ since such an influence comes from the drug sensitivity data. It has been noted in other studies that clustering of cell lines based on drug responses does not group cell lines as strongly on tissue of origin as when clustering on gene expression.[7,8] However, we note that the cell lines selected for resistance to a specific drug will behave in exactly the same manner as the parental cell line vis-à-vis all drugs which are not effected by the evolved resistance. This explains why biological variation due to origin seem to dominate the variation in drug sensitivity profiles as well (see also the eigenvector analysis in the Supporting Information). The strength of these effects ultimately depends on the drug library under consideration.

The biased distance measure which is used implicitly when clustering CCPs instead of the original profiles does not seem to optimize any well defined performance criterion, and it does not seem to reflect any valuable prior knowledge. However, if well-founded and meaningful prior information exists, it could be used to define a weighted distance measure that could be used to emphasize relative importance between the differences observed across the cell lines employed. For the time being we would recommend organization of the correlation coefficients based on clustering of the original drug and gene activity profiles. Identification of locally connected regions of high magnitude correlation coefficents in the resulting map provides the investigator with an easily interpretable pattern. We also note that this demonstrates that one should be careful in general when using correlation coefficents as feature vectors.

5. CONCLUSION

This work shows that the common practice to display drug-gene correlation coefficients based on hierarchical clustering of in vitro drug sensitivity and gene expression profiles using the corresponding CCPs does not offer any obvious advantages in comparison with clustering directly on the original drug sensitivity and gene expression profiles.

ACKNOWLEDGMENT

This work was supported by the Swedish Cancer Society, the Swedish Research Council, the Lions Cancer Research

Fund, Beijer Foundation, Wallenberg Consortium North, Marcus Borgström Foundation, Swedish Society for Medical Research, Göran Gustafsson Foundation, Carl Tryggers Foundation, Stockholm Cancer Society, and Swedish Knowledge Foundation. The authors would also like to thank Hanna Göransson at the Department of Medical Sciences, Uppsala University, as well as two anonymous reviewers for input that greatly improved the presentation.

Supporting Information Available: Use of the angular separation metric and an eigenvector analysis of sample dependencies in our data set and the NCI-60 data set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bredel, M.; Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nature Rev. Genet.* **2004**, *4*(5), 262−275

(2) Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadhan, V. N.; Rubinstein, L. V.; Monks, A. P. et al. Neural computing in cancer drug development: predicting mechanism of action. *Science* **1992**, *258*(5081), 447−451.

(3) Dhar, S.; Nygren, P.; Csoka, K.; Botling, J.; Nilsson, K.; Larsson, R. Anti-cancer drug characterisation using a human cell line panel representing defined types of drug resistance. *Br. J. Cancer* **1996**, *74*, 888−896.

(4) Sievertzon, M.; Nilsson, P.; Lundeberg, J. Improving reliability and performance of DNA microarrays. *Expert Rev. Mol. Diagn.* **2006**, *6*(3), 481−492

(5) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; Kohn, K. W. et al. An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science* **1997**, *275*, 343−349.

(6) http://discover.nci.nih.gov (accessed June 6, 2006).

(7) Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L. et al. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **2000**, *24*, 236−244.

(8) Dan, S.; Tsunoda, T.; Kitahara, O.; Yanagawa, R.; Zembutsu, H.; Katagiri, T. et al. An integrated database of chemosensitivity to 55 anticancer drugs and gene expression profiles of 39 human cancer cell lines. *Cancer Res.* **2002**, *62*, 1139−1147.

(9) Rickardson, L.; Fryknäs, M.; Dhar, S.; Lövborg, H.; Gullbo, J.; Rydåker, M. et al. Identification of molecular mechanisms for cellular drug resistance by combining drug activity and gene expression profiles *Br. J. Cancer* **2005**, *93*, 483−492.

(10) http://www.medsci.uu.se/klinfarm/arrayplatform/cDNA_array.htm (accessed Nov 6, 2006).

(11) Larsson, R.; Nygren, P.; Ekberg, M.; Slater, L. Chemotherapeutic drug sensitivity testing of human leukemia cells in vitro using a semiautomated fluorometric assay. *Leukemia* **1990**, *4*, 567−571.

(12) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*(25), 14863−14868.

(13) Webb, A. *Statistical Pattern Recognition*, 2nd ed.; Chichester, U.K.; Wiley: 2002.

CI060073N