BIOINFORMATICS

Vol. 30 ECCB 2014, pages i497–i504
doi:10.1093/bioinformatics/btu456

Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis

Suleiman A. Khan[1,*], Seppo Virtanen[1], Olli P. Kallioniemi[2], Krister Wennerberg[2], Antti Poso[2,3] and Samuel Kaski[1,4,*]

[1]Department of Information and Computer Science, Helsinki Institute for Information Technology HIIT, Aalto University, 00076 Espoo, [2]Institute for Molecular Medicine Finland FIMM, University of Helsinki, 00014 Helsinki, [3]School of Pharmacy, Faculty of Health Sciences, University of Eastern Finland, 70211 Kuopio and [4]Department of Computer Science, Helsinki Institute for Information Technology HIIT, University Of Helsinki, 00014 Helsinki, Finland

ABSTRACT

Motivation: Analysis of relationships of drug structure to biological response is key to understanding off-target and unexpected drug effects, and for developing hypotheses on how to tailor drug therapies. New methods are required for integrated analyses of a large number of chemical features of drugs against the corresponding genome-wide responses of multiple cell models.

Results: In this article, we present the first comprehensive multi-set analysis on how the chemical structure of drugs impacts on genome-wide gene expression across several cancer cell lines [Connectivity Map (CMap) database]. The task is formulated as searching for drug response components across multiple cancers to reveal shared effects of drugs and the chemical features that may be responsible. The components can be computed with an extension of a recent approach called Group Factor Analysis. We identify 11 components that link the structural descriptors of drugs with specific gene expression responses observed in the three cell lines and identify structural groups that may be responsible for the responses. Our method quantitatively outperforms the limited earlier methods on CMap and identifies both the previously reported associations and several interesting novel findings, by taking into account multiple cell lines and advanced 3D structural descriptors. The novel observations include: previously unknown similarities in the effects induced by 15-delta prostaglandin J2 and HSP90 inhibitors, which are linked to the 3D descriptors of the drugs; and the induction by simvastatin of leukemia-specific response, resembling the effects of corticosteroids.

Availability and implementation: Source Code implementing the method is available at: http://research.ics.aalto.fi/mi/software/GFAsparse

Contact: suleiman.khan@aalto.fi or samuel.kaski@aalto.fi

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Modeling and understanding the diverse spectrum of cellular responses to drugs is one of the biggest challenges in chemical systems biology. Some of the responses can be predicted for targeted drugs, which have been designed to bind to a specific protein that triggers the biological response. The binding of a drug to a target largely depends on the structural correspondence of the drug molecule and the binding cavity of the target molecule, which can be modeled in principle, given ample computational resources. Off-target effects are harder to predict. They are dependent on the cell types, individual genetic characteristics and cellular states making the spectrum of responses overwhelmingly diverse. The less well-known the drug's mechanism of action and the characteristics of the disease, the harder the prediction from first principles becomes. The most feasible way to approach this challenge in an unbiased way, which does not require prior knowledge of all on- and off-target interactions of drugs, is to collect systematic measurements across different drugs, cell types and diseases and search for response patterns correlating with the characteristics of the drugs. The patterns found can be used as evidence for hypotheses on underlying action mechanisms or directly in predicting the responses.

The Connectivity Map (CMap; Lamb *et al.*, 2006) described the basis for a data-driven study of drug–effect relationships at a genome-wide level. CMap hosts the largest collection of high-dimensional gene expression profiles derived from treatment of three different human cancer cell lines with over one thousand drugs. The CMap data have been used in a multitude of studies revealing new biological links between drugs and between drugs and diseases. Genome-wide gene expression responses from the CMap have been used to discover clusters of drugs having similar mechanisms of action, resulting in novel findings, such as effects of heat shock protein (HSP) inhibitors and identification of modulators of autophagy (Iorio *et al.,* 2010). The CMap data have also been successfully used in large-scale integrative studies including the analysis of regulation of drug targets (Iskar *et al.*, 2010), hERG annotations to predict novel inhibitors (Babcock *et al.*, 2013) and drugs' interactions with protein networks (Laenen *et al.*, 2013).

Quantitative structure–activity relationship analysis (QSAR; Cramer *et al.*, 1988) is a widely adopted approach to studying drug responses. Traditionally, univariate biological activities are predicted using a range of methods, including classical regression, support vector machines and Random Forests. The key challenge when moving from traditional QSAR to system-wide analysis of chemical effects is how to relate structural features to genome-wide cellular responses.

Integration of chemical structures with genome-wide responses has become a major research direction in chemical systems biology (Iskar *et al.*, 2012; Xie *et al.*, 2012). Keiser *et al.* (2009) studied structural similarities between ligand sets while Klabunde and Evers (2005) used protein–ligand complexes to predict off-targets. To infer potential indications for drugs, Gottlieb *et al.* (2011) combined similarities from chemical structures, gene expression profiles, protein targets and several

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

other datasets. Atias and Sharan (2011) modeled linkage between structural descriptors of drugs and their side effects using canonical correlation analysis (CCA; Hotelling, 1936). Structures have also been used with genomic datasets to predict toxicity and complex adverse drug reactions (Russom *et al.*, 2013). Recently, Menden *et al.* (2013) combined structures of drugs and mutation information of cell lines to predict drug cytotoxicity in a series of cell lines.

Relationships between structural descriptors of drugs and their gene expression profiles have also been studied. Cheng *et al.* (2010) examined similarities between chemical structures and molecular targets of 37 drugs that were clustered based on their bioactivity profiles. Low *et al.* (2011) classified 127 rat liver samples to toxic versus non-toxic responses, based on combined drug-induced expression profiles and chemical descriptors, and identified chemical substructures and genes that were responsible for liver toxicity. In a broader setting, when the goal is to find dependencies between two data sources (chemical structures and genomic responses), correlation-type approaches match the goal directly, and have the additional advantage that a predefined classification is not required. Khan *et al.* (2012) generalized structure response analysis to multivariate correlations with CCA on the CMap. Because of the limitations of classical CCA, their study was restricted to a limited set of descriptors (76) and genomic summaries (1321 genesets), and did not attempt to take into account the data from the three separate cell lines.

In this article, we present the first probabilistic approach to the problem of integrated analysis of effects of chemical structures across genome-wide responses in multiple model systems. We extend the earlier work in three major ways: (i) instead of using only two data sources (as in classical CCA), we used the recent Bayesian group factor analysis (GFA) method (Virtanen *et al.*, 2012) that generalizes the analysis to multiple sources, here three cell lines and two sets of chemical descriptors. (ii) Our Bayesian treatment with feature-level priors enabled us to cope better with the uncertainties in the high-dimensional data. (iii) We included a more informative set of 3D chemical descriptors to complement the widely used 2D fingerprints, which are recognized to only explain limited aspects of drugs (Schneider, 2010).

Our goal was to uncover the big picture of relationships between chemical structure parameters and genome-wide responses, in a data-driven fashion (Fig. 1). The data came from CMap, 11 327 gene expression responses in three cell lines (HL60-Blood Cancer/Leukemia, MCF7-Breast Cancer and PC3-Prostate Cancer; Lamb *et al.*, 2006) and from two sets of chemical descriptors: 780 3D Pentacle descriptors of drugs (Duran *et al.*, 2008) and 2769 functionally relevant structural fragments (FCFP4; Glen *et al.*, 2006) as 2D fingerprints of the drugs. These five datasets consist of samples from the 682 drug treatments, coupled by the detailed drug identity. We analyzed the statistical relationships between the datasets by decomposing them into a set of interpretable components. Our method quantitatively outperformed previous studies, thereby validating the approach. We rediscovered findings reported earlier as well as identified novel drug associations and detailed structure response relationships.

2 METHODS

2.1 Gene expression datasets

We used the CMap (Lamb *et al.*, 2006) gene expression data as a measure of the biological response of the three cancer cell lines to drug treatments, forming the gene expression datasets. The CMap hosts over 7100 gene expression profiles including technical replicates treated with 1309 drugs and is the largest available resource of its kind. Responses from a subset of these drugs (682) were measured on all of the three cell lines, namely, HL60 (leukemia), MCF7 (breast cancer) and PC3 (prostate cancer cell line).

We obtained the raw gene expression profiles from the CMap and used the data from the most abundant microarray platform (HT-HG-U133A). The data were preprocessed using the Robust Multiarray Averaging (RMA; Irizarry *et al.*, 2003) and *drug treatment versus control* (log2) differential expression was calculated batchwise (Khan *et al.*, 2012). Technical replicates were merged by taking the mean of each gene. This resulted in gene expression profiles for the 682 drugs having measurements over all three cell lines. To reduce noise, we adapted the approach of Iorio *et al.* (2010) for our setting, by retaining the expression of top 2000 up- and 2000 downregulated genes for each sample, while considering the rest as noise (set to zero). The threshold was large to retain diverse effects and removed small values. These profiles formed three biological response datasets (one for each cell line), each being a differential gene expression matrix of 682 drugs times 11 327 genes.

2.2 Chemical descriptor datasets

The chemical space of drugs was represented using two different types of chemical descriptors, namely, the 2D fingerprints '*FCFP4*' and 3D descriptors '*Pentacle*'. The FCFP4 (functional connectivity fingerprints of radius 4; Glen *et al.*, 2006) are circular topological fingerprints designed specifically for structure–activity modeling and similarity searching. They are rapidly computable and heavily used in a wide variety of applications (Rogers and Hahn, 2010). Each dimension of the fingerprints represents a certain 2D fragment of the compounds, interpretable as presence of certain substructures, typically stereochemical information, and allows easy visual inspection of structures. Therefore, FCFP4 can be used to identify the core 2D substructures that make compounds structurally similar and are responsible for biological activity.

The more complex 3D descriptors Pentacle (Duran *et al.*, 2008) capture the functional properties of the compounds using molecular interaction fields. They are able to group together compounds with dissimilar chemical structures and yet having the same type of molecular field properties. This is especially important in our study where the aim is to find small molecules that share biological functions despite structural dissimilarity. Most of the traditional fingerprints, like MACCS (Molecular Access System) and FCFP4, are superior to recognize 2D structural similarity but unfortunately unable to recognize structurally unrelated and yet biologically similar compounds binding into the same binding pocket. The opposite is true with most (if not all) field-based similarity methods like Pentacle, which find more effective distant similarities; therefore, we decided to combine both approaches. In the earlier work, Khan *et al.* (2012) had used VolSurf descriptors to represent molecular properties. Although VolSurf is an optimal method for physicochemical properties estimation, it is not able to describe pharmacophore features extensively, unlike the Pentacle descriptors, and thus is not an option in our study.

Pentacle field distance descriptors were computed using Pentacle v 1.0.4 (http://www.moldiscovery.com/soft_pentacle.php), by Molecular Discovery. The descriptors were calculated for all the available 10 probe sets, namely, D^2, O^2, N^2, T^2, DO, DN, DT, ON, OT, NT, where D is dry probe to represent hydrophobic interactions, O is carbonyl oxygen probe to represent hydrogen bond donor feature of the molecules and N flat probe of Nitrogen is the hydrogen bond acceptor, while T is TIP probe representing shape of the molecule, in terms of steric hot spots.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

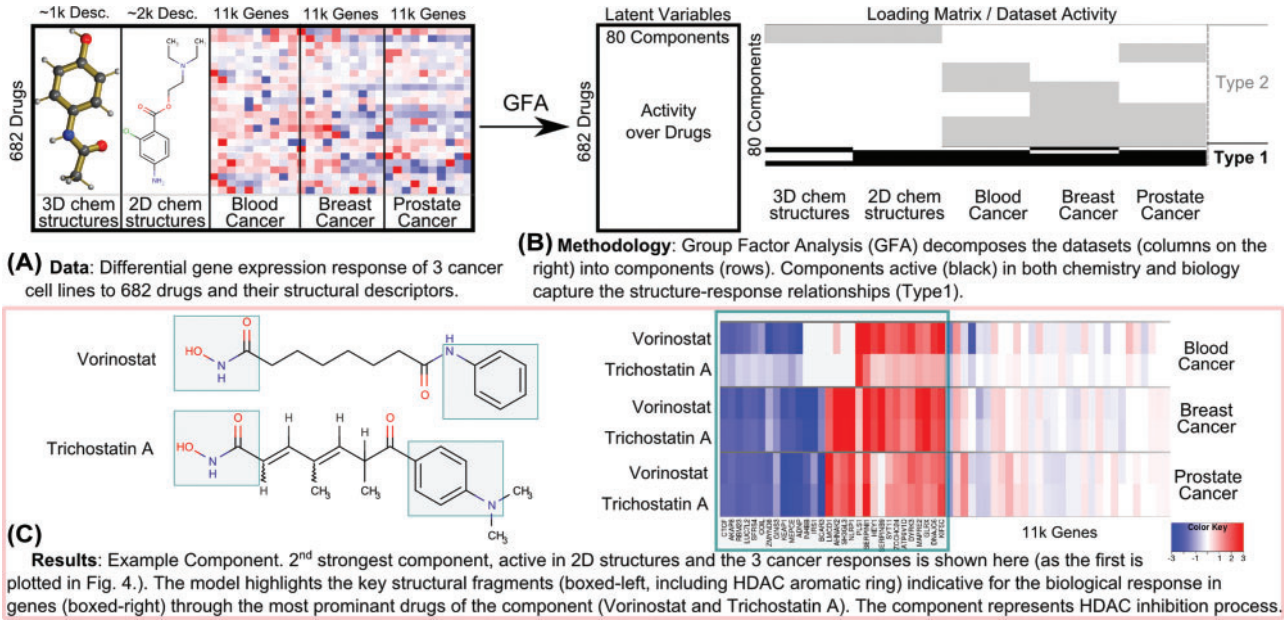


Fig. 1. Overview of the symmetric multi-structure to multi-response decomposition. (**A**) The five datasets spanning the common 682 drugs are (**B**) decomposed into components by GFA. Components of Type 1 represent shared patterns in both chemistry and biology, whereas Type 2 describes biology-only or chemistry-only variation (not as useful in our case). (**C**) Each shared component identifies key structures and genes of an underlying biological process

For each probe set, 78 descriptors were obtained, representing the interaction potentials of probes at different distances, resulting in 780 descriptors in total. Distances in the Pentacle descriptors are true distances between putative interaction sites (hot spots) and are thus connected to the size of the compound and distances between potential pharmacophoric features. This results in a 682 × 780 data matrix, with each row being a drug and the 780 columns representing the Pentacle descriptors. This forms the first chemical dataset in our study.

The 2D FCFP4 represent the chemicals as structural fragments. In FCFP, the fragments are not predefined, rather computed dynamically and thus can represent variation in novel structures. The FCFP4 fingerprints were computed using Pipeline Pilot Student Edition software (http://accelrys.com/products/pipeline-pilot/), by Accelrys. A total of 2769 unique structural fragments are found, and the fingerprints are represented as a matrix of 682 compounds × 2769 fragment descriptors. This forms the second chemical dataset in our study.

2.3 Model: GFA

We search for relationships between chemical descriptors and biological responses, as clues to the key underlying biological processes. GFA is a model designed to capture such relationships (statistical dependencies) by explaining a collection of datasets ('views') by a set of factors or components, which form a combined low-dimensional representation (Virtanen *et al*., 2012). In the *multi-view* setting, each component is active in a subset of the datasets and is a simplified model of an underlying process visible in those sets. The task solved by GFA is to separate the shared components that capture the structure–biology relationships from the rest of the data: the former are visible in all or a subset of the datasets, whereas components active in a single view describe variation specific to that particular view or noise.

Given a collection of M datasets $X^{(1)} \in R^{N\times D_1} \ldots X^{(m)} \in R^{N\times D_M}$, consisting of N co-occurring samples $\boldsymbol{x}_n^{(m)}$, GFA finds a set of latent components (with upper limit K, see below). Each dataset is assumed to have been generated as a linear combination of latent components $Z \in R^{N\times K}$, with weights of the combination given by a loadings matrix $W^{(m)} \in R^{D_m\times K}$: Assuming normal distributions for simplicity, the model is

$$\begin{aligned} \boldsymbol{x}_n^{(m)} &\sim Normal\left(\boldsymbol{W}^{(m)}\boldsymbol{z}_n, \Sigma^{(m)}\right), \\ \boldsymbol{z}_n &\sim Normal(0, I), \end{aligned} \tag{1}$$

where z_n is the n^{th} row of Z, and $\Sigma^{(m)}$ is a diagonal noise covariance matrix. GFA is special in that the projections W are required to be group-wise sparse, *i.e*. all the elements $\boldsymbol{W}_{:,k}^{(m)}$ are set to zero for the components k that are not active in the m^{th} dataset. The components with non-zero projections between two or more views capture dependencies between the views.

To increase the interpretability of the model, we extend GFA by introducing *element-wise* sparsity in addition to the *group sparsity* for the projection matrices, matching the biological prior assumption that each process typically activates only a subset of genes. We introduce element-wise automatic relevance determination (ARD; Neal, 1995) prior for the projection weight matrices, pushing irrelevant weight values $W_{d,k}^{(m)}$ toward zero and making each component element-wise sparse. For the group sparsity, we apply the group spike and slab prior where the binary variable $\boldsymbol{H}_k^{(m)}$ controls the activity of the k^{th} component in the group m. The prior is

$$\begin{aligned} \boldsymbol{W}_{d,k}^{(m)} &\sim \boldsymbol{H}_k^{(m)} Normal\left(0, (\alpha_{d,k}^{(m)})^{-1}\right) + \left(1 - \boldsymbol{H}_k^{(m)}\right)\delta_0, \\ \boldsymbol{H}_k^{(m)} &\sim Bernoulli(\pi_k), \\ \pi_k &\sim Beta(a^\pi, b^\pi), \\ \alpha_{d,k}^{(m)} &\sim Gamma(a^\alpha, b^\alpha). \end{aligned} \tag{2}$$

If $\boldsymbol{H}_k^{(m)}$ becomes zero, all values in $\boldsymbol{W}_{:,k}^{(m)}$ will be set to zero. To complete the model description, we set an uninformative before the diagonal elements of the precision matrix $(\Sigma^{(m)})^{-1}$. Here we made two assumptions, (i)

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

normal distributions for simplicity and (ii) sparsity. Sparsity was implemented by combining the previously (Klami *et al.*, 2013) separately used beta-Bernoulli formulation and the element-wise normal-gamma ARD.

We represent our (M = 5) datasets as matrices of drugs versus features. The rows represent the samples (drugs), and the columns are the features (genes or chemical descriptors). Drugs pair all the views, *i.e.* a row in all matrices corresponds to the same drug. A total of $N = 682$ drugs were used in the study. The features of the chemical descriptors, Pentacle (m = 1) and FCFP4 (m = 2) are $D_1 = 780$ Pentacle probe fields and $D_2 = 2769$ fragments, respectively. The biological responses of the three cell lines (m = 3,4,5) are represented by differential expression of $D_m = 11\,327$ genes each.

The hyperparameters are set to $a^\alpha, b^\alpha = 10^{-3}$ and $a^\pi, b^\pi = 1$, to obtain uninformative priors. We initialize the model by sampling the latent variables from the prior. The model parameters ($\boldsymbol{W}^{(m)}_{:,k}$, $\boldsymbol{H}^{(m)}_{k}$, $\Sigma^{(m)}$, $\alpha^{(m)}_{d,k}$, π_k, Z) are then learned from the data using Gibbs sampling. The number of components is optimally learned from data by initializing K to be large enough, such that sparsity assumptions push some to be inactive. Here for computational reasons, we set $K = 80$, a value significantly larger than the actual number of shared components, and let the noise model represent the rest of the data. For sampling, we ran 10 chains and selected for further analysis the one having its likelihood closest to the mean of non-outlier chains. The first 5000 samples were discarded as the burn-in, and the chain was run for 1000 more iterations, with a thinning factor of 5. The mean value of the samples was used as a representation of the model. As a sanity check, we verified that our shared components had over 70% similarity in top genes and descriptors with the second (non-used) chain. The model's complexity is $O(NDK^2 + K^3)$ where $D = \text{sum}(D_{1:M})$. The current implementation ran for 5 days on a standard desktop computer consuming 6 GB memory.

For interpretation, we represent each component by listing the high-valued latent scores Z and projection values W. For the latent scores, we performed a permutation test to detect the most significantly (q-value < 0.05) activated drugs, while for the projections we inspected the top 30 elements.

3 RESULTS

Figure 2 gives an overview of the types of components discovered by the model. For studying structure–activity relationships, the most important are the components *shared* by one or more chemical view and one or more of the cancer subtypes. The components active in only the expression datasets represent drug responses not captured by the used chemical descriptors, and components only active in the chemical datasets represent biologically irrelevant structural variance. Additionally, components active in only a single dataset may represent dataset-specific noise. We found 11 shared components, which will be discussed below. The detailed structure–response relationships discovered from all the shared components are visualized in Supplementary Figure S1 and tabulated in a usable format in Supplementary Table S2.

3.1 Validation via chemical biology ontology

We started by quantitatively evaluating how closely related the drugs in the shared components are in terms of known chemical–biology relationships and compared our data with those of two previous studies (Iorio *et al.*, 2010; Khan *et al.*, 2012) that investigated drug actions using the same CMap database version.

The established chemical–biology relationships were obtained from the ontology Chemical Entities of Biological Interest (ChEBI; Degtyarenko *et al.*, 2008), which is the largest such ontology of small compounds. ChEBI links compounds with respect to chemical structure, biological roles they are known to play and their applications. Examples of classifications are antibiotic, coenzyme and agonist (biological); donor, ligand, inhibitor (chemical); and pesticide, antiasthmatic (applications). ChEBI was downloaded as a graph and contained paths between 328 (of 682) of our compounds via 611 ontology terms (http://www.ebi.ac.uk/chebi).

The average similarity (inverse path distance) of drugs within the shared GFA components was consistently higher than the corresponding similarities of Khan *et al.*, (2012) and Iorio *et al.*, (2010) and random sets of compounds (Fig. 3). The largest path length (16) in ChEBI linked all drugs, whereas the smallest (2) linked only the most similar. Interestingly, the difference in GFA and others on small path lengths was higher than that on larger ones, indicating that drugs closely connected in ChEBI were even better found by GFA.

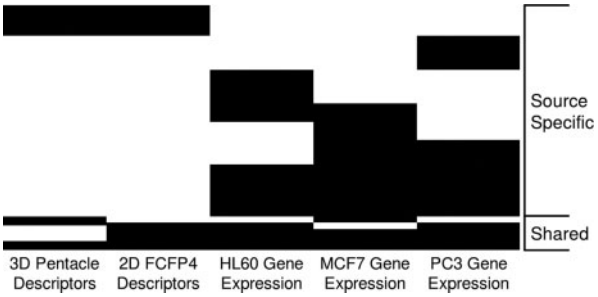


Fig. 2. Summary of the GFA components. The plot demonstrates activity (black is active) of each component (y-axis) over the five input datasets (x-axis). Each component is active in some or all of the datasets. Components shared by (active in) both chemical descriptor and expression datasets capture structure–response relationships

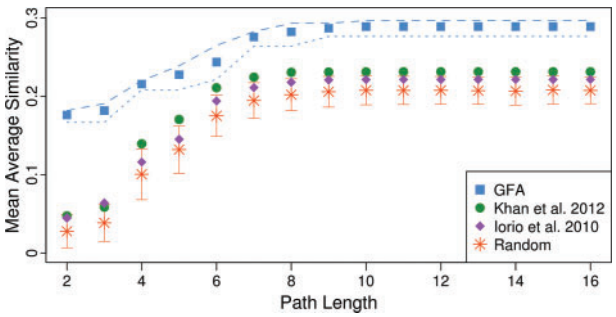


Fig. 3. Quantitative validation of chemical biology similarity of drugs in shared GFA components. Drugs in the same GFA component (blue squares) had a consistently higher mean average similarity (y-axis) in ChEBI than either of the earlier studies, and random sets of compounds, over the entire range of ChEBI path lengths (x-axis). To access the relative contribution of the 3D descriptors we additionally plotted results with components containing them (dashed line) and components containing only 2D descriptors (dotted line), demonstrating that both descriptors are valuable. Error bars (red) are one stdandard over 1000 randomly generated sets

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

Table 1. Shared components having cell line-specific response

	Drug description	Biological interpretation	Structural P.
SP1	Antimetabolite (8-Azaguanine) used for antineoplastic activity and anisomycin a protein synthesis inhibitor. 8-azaguanine has been used in leukemia (Colsky *et al*., 1955).	Protein synthesis inhibition in HL60 and PC3 cells only. It could be interesting to explore 8-azaguanine as an anti-prostate cancer drug. In a recent study, Wen *et al*. (2013) also indicated 8-azaguanine for potential therapeutic efficacy in prostate cancer.	2D ring structures of 8-azaguanine
SP4	Antiestrogen drugs	Response visible in MCF7 (estrogen receptor) cell line only.	Pentacle ON/ OT fields.

Note: The components (rows) are summarized by their top drugs (Column 1), biological response (Column 2) and the structural properties (Column 3).

3.2 Component interpretations

We next analyzed the shared components in detail. Each component connects a set of structural drug properties and gene expression changes, forming a hypothesis of a structure–activity relationship. A component can be characterized by the set of drugs that activate it the most, and by the set of genes that are expressed differentially when the component is active.

We first compared the findings with the two other studies that have investigated drug actions using the same CMap database (Iorio *et al*., 2010; Khan *et al*., 2012). Of the 11 shared GFA components, the majority of the drugs in seven components were similar to the clusters found by Iorio *et al*. (2010), while three components captured structurally driven cell-specific responses they had missed. Compared with the other earlier study (Khan *et al*., 2012), the majority of the drugs in 6 of the 11 GFA components matched a corresponding structure-response subcomponent of Khan *et al*., (2012), again indicating conformance to known results. Our components also revealed several novel drug actions because of cell-type specificity and advanced 3D descriptors that were missed by both of these earlier studies, and are presented below.

Detailed interpretation of all the 11 shared components is presented in Supplementary Table S1. The components are numbered in the order of the amount of variation they captured; the cell line-specific components identified by the model are separately ordered with the prefix SP. One component (SP3) captured outlier response of a single drug and was omitted from further analysis.

The majority of the components captured effects shared among all the three cell lines, whereas five components had responses that were cell line-specific (Components SP1, SP2, SP4), dominant in a specific cell line (Component 7) or revealed some cell line specificity indications for an interesting drug (Component 1). The 2D structural features were active in most components, identifying similarities in structurally analogous drugs. The pentacle descriptors captured similarities in five components, four of which indicated *novel responses* of drugs that have not been reported before. We discuss these four novel components in detail below. One of them had cell line-specific effects (SP2), whereas the remaining cell line-specific components (SP1 and SP4) are summarized in Table 1.

Component 1 was characterized by cardenolides. The top seven drugs of the component, lanatoside C, digitoxigenin, digoxin, digoxigenin, ouabain, helveticoside and strophantidin belonged to this class. The primary activity of the other drugs anisomycin, lycorine and cicloheximide is protein synthesis inhibition, and bisacodyl is used as a laxative through stimulation of secretion in the colon. Cardenolides act on Na^+/K^+ pumps and are known for ion flux alterations. Interestingly, the other compounds of Component 1 also appeared to affect membrane potassium ion flux. Bisacodyl and anisomycin activate K^+ flux, lycorine is known to reduce membrane potential (indicative of potassium efflux) and, indicative of affecting K^+, emetine needs to be administered with potassium to reduce cardiotoxicity. Interestingly, bisacodyl exhibited the response in MCF7 and PC3 cells only, suggesting that its target may be expressed selectively.

On the structural side, the top four FCFP4 fragments collectively represented the correct core 2D response triggering substructure in all the seven cardenolides, as detailed in Figure 4. The other two key drugs, bisacodyl and anisomycin, were different from cardiac glycosides in terms of 2D structures, but the Pentacle descriptors indicated potential field similarities on ON, OT and NT probes. These probes referred to existence of common structural pharmacophoric features: hydrogen bonding and shape-related features. The 3D descriptors may therefore indicate that these drugs bind the same ion channels as the cardenolides.

Component 3 captured protein synthesis inhibition. All drugs in the component are known to inhibit protein synthesis but each in a different way. The only exception, alexidine, is a derivative of clorhexidine, which is used as an antibacterial mouth wash. Interestingly, it has been described to have anticancer cell activity through an unknown target (Yip *et al*., 2006). The model identified pentacle probe fields of D2, DO and DT (shape and lipophilicity-related probes) that relate alexidine's protein synthesis inhibition response with the known protein synthesis inhibitors.

Component 5 was HSP90 inhibition response. The component contains the three similar drugs geldanamycin, tanespimycin, alvespimycin, and on the 2D structure level dissimilar 15-delta prostaglandin J2 (PGJ2) and puromycin. Geldanamycin and its two analogs tanespimycin and alvespimycin are HSP90 inhibitors, and the latter two have been explored in the clinic as anticancer drugs. PGJ2 has also been described as having anticancer activity through an unknown mechanism, causing inhibition of several cancer survival signals. Puromycin is reported

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

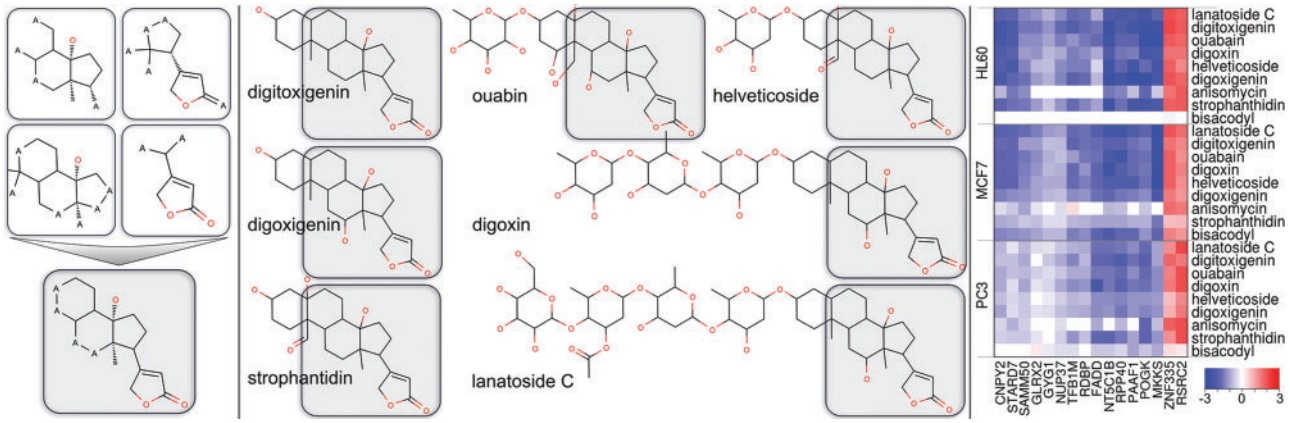


Fig. 4. Structure identification in Component 1. Left: the top four FCFP4 structural fragments identified by the model as strongly relating to the response of the drugs (right). When combined, these fragments represent the core response triggering structure *steroid backbone* (shaded gray) in all the cardenolides

as an aminonucleoside antibiotic with a primary function of terminating ribosomal protein translation. At the response level, this component appeared to be strongly inducing a heat shock response with many HSP and related genes being upregulated (see Fig. 5, left). The expression profile strongly indicated that PGJ2 and puromycin are also inhibiting HSP90. PubChem drug-target data demonstrate that HSP90 targets have been reported as active in geldanamycin and its derivatives, while untested/unspecified for both puromycin and prostaglandin.

On the structural side, the 2D descriptors confirmed that puromycin and prostaglandin are dissimilar to the three geldanamycin analogs. However, the Pentacle descriptors clearly indicated that N2, DN and NT fields shared a strong pattern across all the five drugs. The patterns were only visible in features of smaller distances of these large molecules, indicating that only a small region of these compounds (polar atoms of all compounds) created the activity, whereas the rest of the structure is just needed to maintain the shape. This fitted well with the observation that the drugs are overall structurally dissimilar. At the smaller distances, the structure responsible for biological response was characterized by N2: ligands hydrogen bonding capacity, DN: hydrogen bonding and lipophilicity and NT: hydrogen bonding/shape-based descriptors. In geldanamycin and prostaglandin, this distance (see Fig. 5 where N2 descriptor is plotted) was connected to polar ring atoms and more precisely corresponding hydrogen bonding positions. These same positions, although in a different conformational arrangement (but with almost identical distance), are critical in the binding of geldanamycin to HSP90. Hence, while the expression data strongly argue for PGJ2 inhibiting HSP90 activity at some level, the structural information suggests that this effect could be through a direct binding to HSP90 enzymes.

Component SP2 was characterized by responses to a set of corticoids, other steroids such as etynodiol, and surprisingly different drugs simvastatin and repaglinide. There appears to be a dual response: an HL60-specific metabolic regulatory response and an HL60 and PC3-selective anti-inflammatory response (Fig. 6) with the MCF7 not exhibiting these responses at all, indicating that the relevant target or signal may be selectively expressed in HL60 and PC3 cells. Both simvastatin (a cholesterol-lowering HMG-CoA reductase inhibitor drug) and repaglinide (a diabetes drug) are highly dissimilar at the 2D level when compared with the corticosteriods, but both interestingly have been reported to have anti-inflammatory activities, likely because of targets other than the primary target(s). Once again, Pentacle descriptors capture the underlying similarities between these drugs through NT and N2 fields, suggesting that the common gene expression patterns induced by the different drugs (corticosteroids, simvastatin and repaglinide) is a result of binding the same targets.

4 CONCLUSIONS AND DISCUSSION

We extended the drug response analysis paradigm from standard QSAR, of relating drug properties and univariate responses, to finding relationships between specific structural descriptors of drugs with the *genome-wide responses they elicit in multiple cell lines*. The task was formalized as discovering dependencies between multiple datasets and addressed using the state-of-the-art method GFA. The approach identified structure–genomic response relationships as underlying components of the data and can be used as a tool for exploring such relationships from large-scale measurement datasets.

We quantitatively validated our structure–response components over the established chemical–biology relationships of ChEBI and found them to be better than earlier studies (Iorio *et al.*, 2010; Khan *et al.*, 2012) that did not account for separate cell lines and advanced 3D chemical descriptors. Moreover, several drug groups we identified were consistent with earlier studies, while several revealed interesting novel findings earlier studies had missed, demonstrating that our approach is viable for explorative multi-set structure–activity analysis. These novel findings were clearly attributed to separate cell lines and advanced 3D descriptors in our formulation. In a different setting, Yera *et al.*, (2011) found 3D similarity to be more important

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

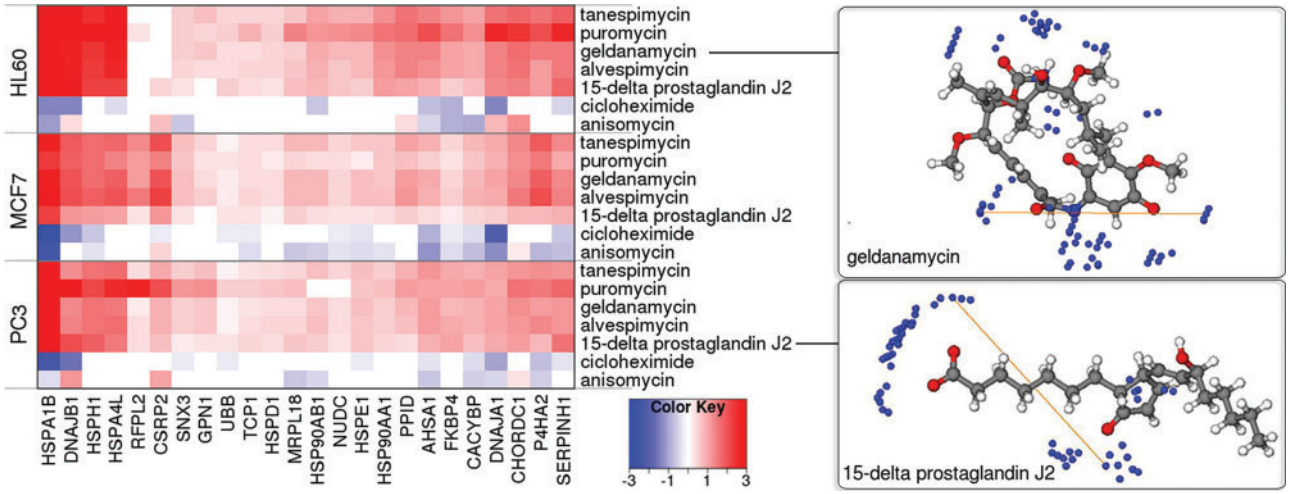


Fig. 5. Component 5 identified a novel HSP90 response of prostaglandin. Left: gene expression response of the top seven drugs in the three cell lines (*y*-axis), over the top genes (*x*-axis) of the component, demonstrates HSP genes being strongly upregulated by the HSP90 inhibitors and by the strikingly different puromycin and prostaglandin. Right: N2 descriptor in geldanamycin and prostaglandin connected to several polar ring atoms (red and blue). The Pentacle feature (N2 distance range) found by GFA as related with HSP gene expression is represented with the yellow line

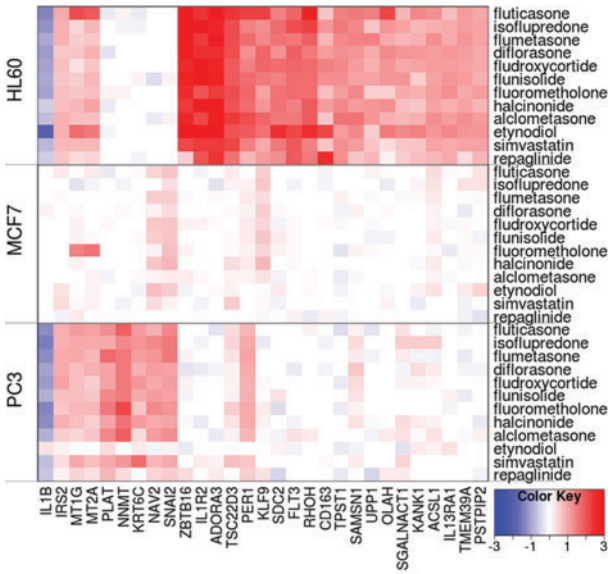


Fig. 6. SP2: corticosteroids showing response specific to HL60 cells, while only minor regulation in PC3 and not at all in MCF7

for off-target identification, and this was partially supported by our study as well.

The discovered components revealed interesting new findings of potential importance for revealing novel action mechanisms of drugs. The 2D fingerprints highlighted important core structural groups primarily responsible for activity of similar drugs, such as the identification of the steroid backbone in cardiac glycosides and aromatic ring in HDAC (Histone deacetylases) inhibitors. The joint analysis of data from multiple cell lines with advanced 3D Pentacle descriptors allowed us to identify relationships between drugs that were not known earlier. If validated, this suggests an approach that could significantly help in medicinal chemistry and drug design. For example, our data led to the identification of a previously unknown and novel shared mechanism of 15-delta prostaglandin J2 (PGJ2) and HSP90 inhibitors. Interestingly, PGJ2 and related prostaglandin analogs have repeatedly been described in the literature for having anticancer activities, but their mechanism of action has not been clarified before (Fionda *et al.*, 2007; Hegde *et al.*, 2011; Zimmer *et al.*, 2010). Furthermore, our analysis revealed that simvastatin, a cholesterol-lowering drug, has a leukemia-specific response similar to a range of corticosteroids. This appears to be a significant finding as lovastatin, a close structural analog of simvastatin, was recently shown to selectively inhibit leukemic stem cells together with several steroids (Hartwell *et al.,* 2013).

Such systematic explorations raise the possibility for targeted interventions and will become a growing trend in the future as more large-scale datasets like the CMap will become available. For drug designers, it opens up the opportunity to tailor drug molecules to match a desired gene expression fingerprint. For medicinal chemists, it could help to increase understanding of action mechanisms of existing drugs and revealing potential on-label and off-label applications for use in precision medicine.

ACKNOWLEDGEMENTS

We thank Pekka Tiikkainen for generating the FCFP4 descriptors.

Funding: This work was supported by the Academy of Finland [140057 and Finnish Centre of Excellence in Computational Inference Research COIN 251170]; the Jane and Aatos Erkko Foundation; and the FICS doctoral program.

Conflict of Interest: none declared.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014

REFERENCES

Atias,N. and Sharan,R. (2011) An algorithmic framework for predicting side-effects of drugs. *J. Comput. Biol.*, **18**, 207–218.

Babcock,J.J. *et al.* (2013) Integrated analysis of drug-induced gene expression profiles predicts novel hERG inhibitors. *PLoS One*, **8**, e69513.

Cheng,T. *et al.* (2010) Investigating the correlations among the chemical structures, bioactivity profiles and molecular targets of small molecules. *Bioinformatics*, **26**, 2881–2888.

Colsky,J. *et al.* (1955) Response of patients with leukemia to 8-azaguanine. *Blood*, **10**, 482–492.

Cramer,R.D. III *et al.* (1988) Comparative molecular field analysis (CoMFA), effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, **110**, 18.

Degtyarenko,K. *et al.* (2008) ChEBI: a database and ontology for Chemical Entities of Biological Interest. *Nucleic Acids Res.*, **36**, 344–350.

Duran,A. *et al.* (2008) Development and validation of AMANDA, a new algorithm for selecting highly relevant regions in molecular interaction fields. *J. Chem. Inf. Model*, **48**, 1813–1823.

Fionda,C. *et al.* (2007) Inhibition of trail gene expression by cyclopentenonic prostaglandin 15-deoxy-delta12,14-prostaglandin J2 in T lymphocytes. *Mol. Pharmacol.*, **72**, 1246–1257.

Glen,R.C. *et al.* (2006) Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs*, **9**, 199–204.

Gottlieb,A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, **7**, 496.

Hartwell,K.A. *et al.* (2013) Niche-based screening identifies small-molecule inhibitors of leukemia stem cells. *Nat. Chem. Biol.*, **9**, 840–848.

Hegde,S. *et al.* (2011) Δ12-prostaglandin J3, an omega-3 fatty acid-derived metabolite, selectively ablates leukemia stem cells in mice. *Blood*, **118**, 6909–6919.

Hotelling,H. (1936) Relations between two sets of variants. *Biometrika*, **28**, 321–327.

Iorio,F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl Acad. Sci. USA*, **107**, 14621–14626.

Irizarry,R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, 4.

Iskar,M. *et al.* (2010) Drug-induced regulation of target expression. *PLoS Comput. Biol.*, **6**, 9.

Iskar,M. *et al.* (2012) Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Curr. Opin. Biotechnol.*, **23**, 609–616.

Keiser,M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–191.

Khan,S.A. *et al.* (2012) Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs. *BMC Bioinformatics*, **13**, 112.

Klabunde,T. and Evers,A. (2005) GPCR antitarget modeling: pharmacophore models for biogenic amine binding GPCRs to avoid GPCR-mediated side effects. *ChemBioChem*, **6**, 876–889.

Klami,A. *et al.* (2013) Bayesian canonical correlation analysis. *J. Mach. Learn. Res.*, **14**, 965–1003.

Laenen,G. *et al.* (2013) Finding the targets of a drug by integration of gene expression data with a protein interaction network. *Mol. Biosyst.*, **9**, 1676–1685.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Low,Y. *et al.* (2011) Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chem. Res. Toxicol.*, **24**, 1251–1262.

Menden,M.P. *et al.* (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One*, **8**, e61318.

Neal,R.M. (1995) Bayesian learning for neural networks. PhD Thesis, University of Toronto, Canada.

Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model*, **50**, 742–754.

Russom,C.L. *et al.* (2013) Predicting modes of toxic action from chemical structure. *Environ. Toxicol. Chem.*, **32**, 1441–1442.

Schneider,G. (2010) Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.*, **9**, 273–276.

Virtanen,S. *et al.* (2012) Bayesian group factor analysis. In Proceedings of AISTATS. *J. Mach. Learn. Res. W&CP*, **22**, 1269–1277.

Wen,D.Y. *et al.* (2013) A computational bioinformatics analysis of gene expression identifies candidate agent for prostate cancer. *Andrologia*, **46**, 625–632.

Xie,L. *et al.* (2012) Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annu. Rev. Pharmacol.*, **52**, 361–379.

Yera,E.R. *et al.* (2011) Chemical structural novelty: on-targets and off-targets. *J. Med. Chem.*, **54**, 6771–6785.

Yip,K.W. *et al.* (2006) Potential use of alexidine dihydrochloride as an apoptosis-promoting anticancer agent. *Mol. Cancer Ther.*, **5**, 2234–2240.

Zimmer,M. *et al.* (2010) The Connectivity Map links iron regulatory protein-1-mediated inhibition of hypoxia-inducible factor-2a translation to the anti-inflammatory 15-deoxy-delta12,14-prostaglandin J2. *Cancer Res.*, **70**, 3071–3079.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INRIA Paris on September 7, 2014